BUSCA DE MARCADORES MOLECULARES TECIDO-ASSOCIADOS EM REGIÕES TRANSCRITAS NÃO CARACTERIZADAS DO GENOMA HUMANO

BARBARA PEREIRA DE MELLO

Tese apresentada à Fundação Antônio Prudente para obtenção do título de Doutor em Ciências

Área de Concentração: Oncologia

Orientador: Dra. Helena Paula Brentani
Co-Orientador: Dr. Eduardo Abrantes

São Paulo
2009

FICHA CATALOGRÁFICA

Preparada pela Biblioteca da Fundação Antônio Prudente

Mello, Barbara Pereira de

Busca de marcadores moleculares tecido-associados em regiões transcritas não caracterizadas do genoma humano / Barbara Pereira de Mello – São Paulo, 2009.

86p.

Tese (Doutorado)-Fundação Antônio Prudente.

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

Orientador: Helena Paula Brentani

Descritores: 1. ANÁLISE DA EXPRESSÃO DE TRANSCRIÇÃO 2.GENOMA HUMANO. 3.NEOPLASIAS DA PRÓSTATA. 4.ALINHAMENTO DE SEQUENCIA 5. ANÁLISE DE SEQUENCIA. 6. MARCADORES DE TUMOR. 7. BIOLOGIA MOLECULAR COMPUTACIONAL.

Supercalifragilisticexpialidocious

(Sherman Brothers, 1964)

DEDICATÓRIA

Aos meus pais,
por acreditarem no meu sonho,
por me incentivarem a buscá-lo,
pelo interesse no meu mundo,
pelo amor durante toda a jornada,
por estarem felizes junto comigo.

AGRADECIMENTOS

À minha orientadora, Helena Brentani, por ter acolhido a mim e ao meu trabalho, por acreditar comigo que ainda tínhamos (e temos) muito o que fazer pela frente e nos frutos que esse trabalho, iniciado no meu mestrado, ainda pode gerar. Pela orientação, ensinamentos, confiança, cuidado e amizade, pelo companheirismo na alegria e na tristeza.

Ao meu, co-orientador, Eduardo Abrantes, por ter acompanhado de perto e participado de todas as etapas desse trabalho, por todas as discussões, por tudo que me ensinou. Pela confiança, amizade e brigadeiros.

Ao Prof. Ricardo Brentani, Dra. Dirce Carraro e Dr. Luiz Fernando Lima Reis, pelas reuniões, discussões, sugestões e pelo acompanhamento desse trabalho.

À todos os amigos do LBHC: Cecília, Aderbal, Renato, Fábio, Marcelo, Mário, Diogo, Xurrus, Luiz Paulo e César, novamente Edu e Helena, e aos "agregados", Ana e Hugo, pela ajuda direta e indireta no trabalho, pelo cuidado, pela diversão, pela filosofia da sexta-feira e pelos espetinhos e celsadas.

À todos do LGEA, LABRI/LEA e do laboratório da Dra. Sílvia Rogatto, por terem me acolhido em seus laboratórios para realizar a parte experimental do meu trabalho, pela ajuda e diversão. Um obrigada especial para Bianca, Gustavo, Mari e Elisa (LGEA), Alex, Grá, Letícia e Nair, (LABRI/LEA) e Lívia e Miriam (Dra. Sílvia).

Ao Deptº de Anatomia Patológica, ao pessoal do Banco de Tumores e do SAME, especialmente Dr. Fernando Soares, Dr. Hugo, Severino, Carlinhos, Dra. Dirce, Louise, Eloísa e Vera por todo o auxílio com prontuários, amostras e análises.

À Dra. Anamaria Camargo e ao Instituto Ludwig de Pesquisa sobre o Câncer, por ceder as linhagens celulares de próstata utilizadas nesse trabalho e a Profa. Maria Mitzi Brentani e a Dra. Rosimeire Roela da Universidade de São Paulo, pela cultura das células.

Ao MD Anderson Cancer Center e ao pessoal do Arap/Pasqualini Laboratory pela oportunidade, hospitalidade e colaboração durante meu estágio. Em especial ao Dr. Wadih Arap, Renata Pasqualini e Dr. Emmanuel Dias-Neto. À Fabíola do laboratório da Dra. Sílvia Rogatto, pela apoio, companhia e amizade no meu primeiro mês em Houston.

À Comissão de Ética no Uso de Animais e a Dra. Adriana Abalen Dias, pela oportunidade de participar do grupo e aprender novos temas.

Ao Comitê de Ética em Pesquisa. Às colaboradoras da biblioteca, especialmente à Su, Fran e Rosi, pela ajuda com artigos, aulas e tese. À Pós-Graduação do Hospital, especialmente à Ana Maria e Luciana, por toda a ajuda, até de longe.

Ao Prof. Dr. Brentani pela direção exemplar do Hospital A. C. Camargo e pela oportunidade. À Fundação Antônio Prudente e ao Hospital A. C. Camargo, por possibilitarem a realização desse trabalho.

Aos integrantes da Banca de Qualificação, pelo acompanhamento do trabalho, críticas e sugestões.

À FAPESP, à CAPES e ao CNPq pela credibilidade e suporte financeiro.

Aos amigos do Instituto Ludwig de Pesquisa sobre o Câncer, pela ajuda, interesse, torcida e amizade, em especial à Anna Chris, Lara, Lepique e Enrique.

Aos amigos do LABRI/LGEA, novos e velhos: Ana, Bianca, Ju, Letícia, Nair, Camila, Marina, Vlad, Grá, por continuarmos juntos, mesmo em laboratórios diferentes.

À Aline, pela amizade, carinho, cuidado, pela ajuda em tudo, mesmo de longe.

À Lelê e à Nair, por tantas coisas... pela amizade, pelas conversas, desabafos, pela diversão, pela ajuda com experimentos, relatórios, apresentações,... por tudo.

Aos meus amigos: Kiki, Lud, Marta, Flá, Van, Marcy, Juliane, Ju, Felipe, Maite e muitos outros, pelo incentivo, torcida e interesse no meu trabalho.

À Dag, tio Luiz, tia Zeni, tio Zé, tia Ivana, tia Helena, tio Pedro, Mazinha, Tadeu, Cé, Lu, Co, Marcos, Cris, Isa, Dodô, Íris e Mari por saberem o valor do meu trabalho, por acreditarem, torcerem e se orgulharem de mim.

Aos meus avós, Mathilde, Tereza e Ditinho, pelo cuidado, interesse e preocupação com o meu presente e com o meu futuro.

À minha tia D'Euacy, meu tio Renato e minhas primas Nana e Nina, pela torcida, cuidado e carinho.

Ao meu pai, minha mãe e meu irmão, pela confiança, interesse, cuidado, carinho. Por estarem sempre presentes (até em Houston), por se interessarem e valorizarem tanto o meu trabalho. Por acreditarem em mim, pelo orgulho e amor que sentem por mim.

RESUMO

Mello BP. **Busca de marcadores moleculares tecido-associados em regiões transcritas não caracterizadas do genoma humano**. São Paulo; 2009. [Tese de Doutorado-Fundação Antônio Prudente].

Com foco na atividade transcricional do genoma humano, foi desenvolvido um trabalho de mestrado, em que construímos um microarranjo de cDNAs composto por sequências ORESTES resultantes do Projeto Genoma do Câncer Humano (FAPESP/LICR - HCGP) que não se alinharam com sequências de cDNA geradas por outros projetos. Este arranjo foi hibridizado contra cDNAs derivados de diferentes tecidos humanos, normais ou tumorais, resultando na identificação de 3.421 regiões transcritas do genoma humano (83,3% da plataforma) não descritas por outros projetos de sequênciamento como transcritos. Acreditando que parte dessas sequências pudessem representar RNAs não codificadores, variantes de *splicing* ou transcritos antisenso naturais fizemos uma reanálise computacional das sequências avaliadas no trabalho de mestrado e também uma análise de expressão diferencial das mesmas, buscando variações de expressão de transcritos tumor- e/ou tecido-associadas. Identificamos como possíveis ncRNAs 28% das sequências analisadas. Mil e sete sequências foram identificadas como diferencialmente expressas, sendo que 291 representam potenciais ncRNAs. Além disso, três potenciais marcadores tumorais de próstata foram validados por PCR em tempo real. Estudos adicionais de um desses marcadores, PCA3, revelaram um possível papel desse ncRNA na regulação do gene PRUNE2 e a existência de uma retenção intrônica não descrita na sequência de PCA3, aparentemente mais frequente em amostras normais de próstata. Também pudemos contribuir com análises iniciais de um potencial novo marcador de câncer de próstata a ser explorado para a complementação de marcadores já existentes, mas falhos em alguns aspectos. Um artigo referente a parte desse trabalho foi publicado no periódico *Nucleic Acids Research* (MELLO et al. 2009).

SUMMARY

Mello BP. **[Search for tissue-associated molecular markers in non-characterized transcribed regions of the human genome]**. São Paulo; 2009. [Tese de Doutorado-Fundação Antônio Prudente].

With focus on transcriptional activity of the human genome, we developed a master's work, in which we built a cDNA microarray composed of ORESTES sequences resulting from the Human Cancer Genome Project (FAPESP / LICR - HCGP) that did not align with cDNA sequences generated by other projects. This array was hybridized against cDNAs derived from different normal or tumor tissues, resulting in the identification of 3,421 transcribed regions of the human genome (83.3% of the slide) not described by other sequencing projects as transcripts. Believing that some of these sequences may represent non-coding RNAs, and splicing variants of natural antisense transcripts we did a new computational analysis of the sequences found in the master's work and also an analysis of differential expression of these sequences, seeking changes in expression of tumor- and/or tissue-associated transcripts. We identified as possible ncRNAs 28% of the sequences analyzed. One thousand and seven sequences were identified as differentially expressed, and from these 291 represent potential ncRNAs. In addition, three potential tumor markers for prostate cancer were validated by real-time PCR. Further studies of one of these markers, PCA3, revealed a possible role of this ncRNA in the regulation of PRUNE2 gene and the existence of an intron retention not described within PCA3 sequence, apparently more common in normal samples from prostate. We could also contribute with initial analysis of a potential new marker for prostate cancer to be explored in complementing currently markers not very acurated. An article containing part of this work was published in the journal Nucleic Acids Research (MELLO et al. 2009).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

μg	micrograma
μL	microlitro
ATCC	*American Type Culture Collection*
BLAST	*Basic Local Alignment Search Tool*
BLAT	*BLAST-like Alignment Tool*
C. elegans	*Caenorhabditis elegans*
cDNA	*complementary DNA*, DNA complementar
CEP	Comitê de Ética em Pesquisa
CPC	*Coding Potential Calculator*
CT	*cycle threshold*
DNA	*deoxyribonucleic acid*, ácido desoxirribonucléico
dNTP	*deoxynucleotides triphosphate,* desoxinucleotídeos trifosfato
EST	*Expressed Sequence Tags*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FW	*forward*, sentido senso
HCGP	*Human Cancer Genome Project*
HGP	*Human Genome Project*
IGF	*insulin-like growth factor*
LB	Luria-Bertani
LBHC	Laboratório de Biotecnologia do Hospital do Câncer
LGEA	Laboratório de Genômica e Biologia Molecular
LICR	*Ludwig Institute for Cancer Reserach*
log	logarítmo
mg	miligrama
$MgCl_2$	cloreto de magnésio
miRNAs	micro RNAs
mL	mililitro
mM	milimolar
mRNA	*messenger RNA*, RNA mensageiro

NATs	*natural antisense transcripts*, transcritos antisenso naturais
ncRNAs	*non-coding RNA*, RNA não codificador
NIP	neoplasia intraepitelial prostática
nM	nanomolar
ORESTES	*Open Reading Frame EST Sequences*
ORF	*Open Reading Frame*
pb	pares de bases
PCR	*polymerase chain reaction*, reação em cadeia da polimerase
pd(N)15	iniciador pentadecâmero randômico
p	probabilidade
PSA	*prostate-specific antigen*, antígeno prostático especifico
RIN	*RNA integrity*, integridade do RNA
RNA	*ribonucleic acid*, ácido ribonucléico
RNase	ribonuclease
RT-PCR	*reverse transcription polymerase chain reaction*, PCR a partir da transcrição reversa
RV	*reverse*, sentido antisenso
siRNAs	*small interfering RNAs*
SNP	*single nucleotide polymorphsm*
snRNAs	*small nuclear RNAs*
snoRNAs	*small nucleolar RNAs*
UTs	unidades transcricionais

ÍNDICE

1 INTRODUÇÃO

1.1 TRANSCRIÇÃO DO GENOMA HUMANO

Para o entendimento das bases genéticas do desenvolvimento humano e dos mecanismos envolvidos na fisiopatologia das doenças, é importante que se conheça o nosso genoma e seu funcionamento. Esse conhecimento sofreu um significativo avanço com o surgimento da tecnologia de sequênciamento e a conclusão dos projetos genoma (MAXAM e GILBERT 1977; SANGER et al. 1977; LANDER et al. 2001; VENTER et al. 2001). O Projeto Genoma Humano (*Human Genome Project* – HGP), lançado em 1990, revelou um número final de genes no genoma humano de aproximadamente 30 a 40 mil, um número muito menor do que o esperado, de cerca de 50 a 140 mil (LANDER et al. 2001; VENTER et al. 2001). Desde então, esse número vem sofrendo constantes modificações, com o surgimento de novas tecnologias e abordagens genômicas, tais como: a predição genica, com uma acurácia de cerca de 70% para genes humanos, realizada pela área de bioinformática (DE SOUZA et al. 2000; PENN et al. 2000; STERKY e LUNDEBERG 2000; KAN et al. 2001; ROGIC et al. 2001; SHOEMAKER et al. 2001; CAMARGO et al. 2002); a utilização de ESTs (*Expressed Sequence Tags*), que representam mais da metade dos genes humanos (DERISI et al. 1996; HILLIER et al. 1996; SCHULER et al. 1996; MAO et al. 1998); e a sua posterior complementação com o desenvolvimento da metodologia ORESTES (*Open Reading Frame EST Sequences*), permitindo a cobertura das regiões centrais dos transcritos

(DE SOUZA et al. 2000; DIAS-NETO et al. 2000; CAMARGO et al. 2001; CAMARGO et al. 2002; BRENTANI et al. 2003); e apresentando uma maior sensibilidade para capturar transcritos de baixa expressão, dado ao seu efeito normalizador (DE SOUZA et al. 2000; DIAS-NETO et al. 2000; CAMARGO et al. 2001). Apesar do avanço gerado por essas novas tecnologias, o número de genes no genoma humano ainda é muito menor do que o esperado: a estimativa mais recente (11/2009) é de 45.303 genes humanos, um genoma pouco menos de duas vezes maior do que o genoma de *C. elegans* (21.188 genes – 04/2009) (www.ncbi.nlm.nih.gov/projects/Gene/gentrez_stats.cgi), refletindo o fato de que a maior complexidade do genoma humano não está em seu tamanho.

Hoje sabemos que o dogma central da biologia molecular (DNA-RNA-proteína), inspirada em trabalhos clássicos com organismos procariontes, reflete somente parte da complexidade genética dos eucariotos e que existem inúmeros mecanismos de regulação envolvidos nessa relação (MENDES SOARES e VALCARCEL 2006; BIRNEY et al. 2007). Sequências codificadoras compreendem somente 3% do genoma, sendo interrompidas por grandes regiões intergênicas e compostas por pequenos éxons, normalmente intercalados por numerosos e longos íntrons (STERKY e LUNDEBERG 2000; ROGIC et al. 2001; BRINKMAN 2004; SCHWERK e SCHULZE-OSTHOFF 2005; SREBROW e KORNBLIHTT 2006; PAJARES et al. 2007) e a maior complexidade parece não ser refletida por essa estrutura, mas sim por seu funcionamento e regulação (MATTICK 2004). A maior parte do genoma, que até recentemente era considerada “DNA lixo”, é amplamente transcrita, incluindo transcritos que se sobrepõem uns aos outros e, principalmente, transcritos não codificadores. Estes tipos de transcritos podem exercer funções na

regulação genica (MENDES SOARES e VALCARCEL 2006; BIRNEY et al. 2007), uma vez que muitos transcritos não codificadores têm sido identificados sobrepostos a *loci* codificadores de proteínas e em regiões que pensávamos ser transcricionalmente silenciadas. Diversos estudos vêm confirmando a presença de transcrição em regiões não mapeadas em éxons descritos ou mesmo em regiões intergênica (BERTONE et al. 2004; CARNINCI et al. 2005; CHENG et al. 2005; BIRNEY et al. 2007), levando a conclusão de que a simples visão do genoma contendo um conjunto definido de *loci* isolados transcritos independentemente não parece estar completamente correta e de que somente os eventos de *splicing* e *splicing* alternativos, apesar de ainda não serem conhecidos em sua totalidade, não são suficientes para explicar a complexidade do genoma humano. Estes estudos têm resultado na descoberta de novas sequências transcritas, criando uma nova perspectiva sobre o número e extensão dos transcritos humanos. Além disso, hoje é aceito o fato de que apenas uma pequena fração das sequências geradas por métodos de ESTs representam transcritos mitocondriais, transcritos reversos de RNAs ribossômicos, contaminantes bacteriais ou moléculas de mRNAs imaturos (DIAS-NETO et al. 2000; CAMARGO et al. 2001). Os resultados apresentados por RAVASI *et al*. (2006) (RAVASI et al. 2006) mostraram que a maior parte das sequências não codificadoras provenientes da coleção de cDNAs do RIKEN Research Institute é expressa e não derivada de contaminação genômica ou de mRNAs prematuros.

Nesses últimos anos, o uso de ferramentas de bioinformática aliados a estudos experimentais, particularmente envolvendo o genoma todo, tem se tornado uma abordagem comum e promissora para predizer e filtrar novos RNAs não

codificadores (ncRNAs) e RNAs antisenso (BABAK et al. 2005; GUSTINCICH et al. 2006; PANG et al. 2007; WASHIETL et al. 2007; WEILE et al. 2007). Apesar da conservação entre espécies de muitas regiões transcritas de ncRNAs ser fraca, os promotores desse tipo de transcritos são geralmente muito mais conservados evolutivamente e a extensão das regiões conservadas é muito maior do que a de promotores de RNAs codificadores de proteínas (5kb *versus* 500bp, respectivamente) (CARNINCI et al. 2005; GUSTINCICH et al. 2006; SUN et al. 2006). Esse fato aliado à evidências de estrutura secundária, em análises de bioinformática do genoma todo, é de grande valor pois sugere que tais regiões são transcritas e devem exercer algum papel biológico (WASHIETL et al. 2005; PEDERSEN et al. 2006; GRIFFITHS-JONES 2007; WEILE et al. 2007). RNAs não codificadores são moléculas de RNAs que não são traduzidas em proteínas e, portanto, não contém uma ORF (*Open Reading Frame*) (CALIN e CROCE 2006; GARZON et al. 2006; HAMMOND 2006). As principais fontes de ncRNAs são os íntrons e as regiões intergênicas (MATTICK e MAKUNIN 2006). Os íntrons constituem mais de 30% do genoma humano e cerca de 95% da sequência dos genes (VENTER et al. 2001). Estas regiões intrônicas podem ser degradadas ou transcritas para exercer um papel regulatório. Existem evidências de que os RNAs intrônicos possam ser processados em pequenos RNAs, com vidas-médias significativas e localizações sub-celulares específicas (CLEMENT et al. 1999; CLEMENT et al. 2001). Os ncRNAs vem sendo implicados como principais atuantes no controle transcricional e traducional, representando um novo nível de complexidade genômica (GOODRICH e KUGEL 2006; NAKAYA et al. 2007). Além disso, ncRNAs parecem apresentar expressão celular- ou condição-restrita e em menores níveis, quando comparados a genes

codificadores bem caracterizados (NUMATA et al. 2003; KAMPA et al. 2004; GUSTINCICH et al. 2006). Dados mostram que a proporção ncRNAs/RNAs codificadores aumenta de procariotos para mamíferos (MATTICK 2004; FRITH et al. 2005).

Apesar do progresso dos programas de anotação, do sequênciamento de mais transcritos humanos e do início de estudos sistemáticos destes RNAs que não possuem potencial codificador de proteínas e seu reconhecimento como possíveis reguladores em diversos processos na biologia celular (MATTICK 2003; SCHOLZOVA et al. 2007), a atribuição de funções a essas regiões intergênicas transcritas e a regiões intrônicas não codificadoras do genoma humano continua representando um desafio (COLLINS et al. 2004; NAKAYA et al. 2007). Além disso, a maior parte dos trabalhos dos últimos anos que tratam de identificação de novos transcritos, dão pouca atenção as ESTs sem evidência de *splicing* que mapeiam parcialmente ou inteiramente em regiões intrônicas de genes codificadores de proteína (NAKAYA et al. 2007). O *splicing* alternativo e suas variantes, eventos pós-transcricionais, a existência de uma variedade de famílias de RNAs não codificadores (snRNAs, snoRNAs, miRNAs, siRNAs e ncRNAs longos), a detecção de transcritos híbridos contendo pedaços de unidades transcricionais independentes, a presença de transcritos antisenso à mRNAs processados, o silenciamento transcricional e a edição de RNAs são alguns exemplos que atuam na regulação gênica em diversos níveis, levando a consequências biológicas impactantes e decisivas (BRINKMAN 2004; VENABLES 2004; SCHWERK e SCHULZE-OSTHOFF 2005; CALIN e CROCE 2006; FLOREA 2006; GARZON et al. 2006;

HAMMOND 2006; MENDES SOARES e VALCARCEL 2006; PAJARES et al. 2007).

1.2 MARCADORES MOLECULARES DE CÂNCER

A definição de marcador molecular dada pelo BIOMARKERS DEFINITIONS WORKING GROUP (2001) é: "uma característica que objetivamente é medida e avaliada como um indicador de processos biológicos normais, processos patogênicos, ou respostas farmacológicas a intervenções terapêuticas", e deve mostrar alta especificidade ao tecido ou condição em que está sendo medido. A aplicação de marcadores moleculares na detecção e monitoramento do câncer compreende: uso como ferramenta diagnóstica para a identificação daqueles pacientes com a doença ou alguma alteração do seu estado saudável, uso como ferramenta para o estadiamento ou classificação do câncer, uso como indicador de prognóstico e para a predição e monitoramento da resposta clínica à uma intervenção, entre outros (BIOMARKERS DEFINITIONS WORKING GROUP 2001). O câncer é uma doença de base genética; alterações complexas em nível de DNA e de expressão gênica ocorrem em vários estágios da doença, levando a alterações nos níveis de proteína e funções e portanto, alterando o comportamento celular com relação a por exemplo, descontrole da proliferação e potencial de invasão e metástase (BICKERS e AUKIM-HASTIE 2009). Essas alterações em nível de DNA, RNA e proteína podem ser exploradas como marcadores moleculares para o câncer.

Alterações nos mecanismos de regulação gênica podem resultar na

modificação da taxa de produção de mRNAs, na produção de mRNAs previamente inexistentes e/ou em alterações na proporção entre isoformas de mRNAs esperados e suas isoformas, podendo levar à consequências em nível protéico relacionadas a tumorigênese (BRINKMAN 2004; SCHWERK e SCHULZE-OSTHOFF 2005; PAJARES et al. 2007). Em teoria, uma isoforma específica, presente exclusivamente em células tumorais de um tecido específico mas não nas demais células saudáveis, ou em quantidades claramente distintas nesses dois tipos de células, seria um candidato ideal para marcador molecular (PAJARES et al. 2007). Por isso, o estudo sistemático de isoformas alteradas em câncer e de seus reguladores é de extrema importância na busca por marcadores moleculares para diagnóstico, prognóstico, predição de recidiva, terapia gênica e farmacogenômica (BRINKMAN 2004; VENABLES 2004; SCHWERK e SCHULZE-OSTHOFF 2005; FLOREA 2006; PAJARES et al. 2007). O perfil de expressão de ncRNAs pode ser usado como ferramenta para diagnóstico, classificação, prognóstico, predição de recidiva e alvo terapêutico tumoral (CALIN et al. 2005; LU et al. 2005; GARZON et al. 2006; HAMMOND 2006; HUPPI et al. 2007; LIU et al. 2007; OSADA e TAKAHASHI 2007; SCHOLZOVA et al. 2007). Dados crescentes mostram uma relação entre alterações nos níveis de expressão de ncRNAs e o câncer (REIS et al. 2005; NAKAYA et al. 2007; PANZITT et al. 2007; BRITO et al. 2008), evidenciando uma provável função dos ncRNAs na tumorigênese e o potencial desse tipo de transcrito como marcador molecular tumoral (PANZITT et al. 2007). Em câncer de mama, por exemplo, o ncRNA BCYRN1 foi recentemente identificado como marcador molecular de mal prognostico quando super-expresso (IACOANGELI et al. 2004). Em câncer de pulmão, a expressão aumentada do gene MALAT1 indica uma pior

resposta clínica (JI et al. 2003); e em carcinoma hepatocelular, o ncRNA HULC é um dos genes mais super-expressos (PANZITT et al. 2007). Em câncer de próstata, a super-expressão de PCGEM1 (SRIKANTAN et al. 2000; PETROVICS et al. 2004) e PCA3 (BUSSEMAKERS et al. 1999) está implicadas na tumorigênese. Esses dados refletem um forte argumento para a inclusão de transcritos não codificadores nas ferramentas utilizadas para o diagnóstico molecular, as quais, até o momento, são compostas predominantemente por transcritos codificadores de proteínas (REIS et al. 2004).

1.3 MARCADORES MOLECULARES DE CÂNCER DE PRÓSTATA

Com relação ao câncer de próstata, o mais prevalente em homens (Ministério da Saúde 2008), até o momento, o único marcador molecular amplamente aceito é o PSA (*prostate-specific antigen*), um antígeno específico de próstata, porém não de câncer de próstata. O uso do teste de PSA, em mais de uma década, aumentou a detecção do câncer de próstata em estadiamento precoce (ILYIN et al. 2004) e reduziu o numero de pacientes com doença metastática no diagnóstico (SHARIAT et al. 2008). Níveis de PSA não apresentam uma correlação direta com o aumento da graduação e estadiamento do câncer de próstata (SHARIAT et al. 2004) e maiores níveis de PSA são encontrados em estadios mais avançados da doença, levando a utilização da quantificação do PSA como ferramenta prognóstica e para o monitorameno da doença após tratamento, especialmente após prostatectomia radical (SHARIAT et al. 2004). Entretanto, altos níveis e PSA na hiperplasia benigna

prostática, prostatite e até mesmo no tecido normal, resultam em um numero significativo de casos falso-positivos da doença (STAMEY et al. 1987; CHARRIER et al. 2001). Além disso, um numero significativo de tumores de próstata são negativos para PSA, impossibilitando a detecção da doença através desse teste (PRYOR e SCHELLHAMMER 2002). O valor preditivo positivo de PSA (> 4ng/mL) é falho na detecção de um número significativo de tumores de próstata (THOMPSON et al. 2004; THOMPSON et al. 2005), sendo real apenas em 25% dos casos, segundo uma meta-análise de estudos de PSA realizada por (MISTRY e CABLE 2003). Um modelo combinado da quantificação de PSA, proPSA (proteína precursora de PSA) e PSA livre, mostrou uma sensibilidade de 95% e uma especificidade de 37%, maior do que o uso somente de PSA (15%) ou só de PSA livre (27%) (KHAN et al. (2003); revisado por (WRIGHT e LANGE 2007; BICKERS e AUKIM-HASTIE 2009).

Portanto, é evidente a necessidade de novos marcadores para a diferenciação de lesões benignas e malignas e de tumores agressivos e indolentes, por exemplo, levando a um benefício na determinação da estratégia de tratamento e na diminuição de pacientes desnecessariamente submetidos a terapias agressivas. Prostatectomia radical está associada a 20% de risco de incontinência urinária, 70% de risco de impotência e 3,5% de risco de danos intestinais, infecção, trombose e hemorragia (GOPALKRISHNAN et al. 2001) e benefícios significativos na sobrevida não foram demonstrados com a prostatectomia radical precoce em comparação a observação e seguimento de pacientes com doença indolente (BILL-AXELSON et al. 2008). Entretanto, a maior parte dos novos marcadores moleculares de câncer de próstata não se confirmam, em relação aos valores de sensibilidade e especificidade, em

diferentes estudos e falham em determinar a resposta da doença a tratamentos. Por isso, a busca por novos marcadores que possam ser associados àqueles já estabelecidos, como o PSA, é importante (WRIGHT e LANGE 2007; SHARIAT et al. 2008; BICKERS e AUKIM-HASTIE 2009).

Outras kalikreínas pertencentes ao cromossomo 19, assim como o PSA (hK3), também tem sido avaliadas como potenciais marcadores moleculares de cancer de próstata. A hK2 por exemplo, apresenta sua maior expressão em próstata (YOUSEF e DIAMANDIS 2001) e está super-expressa em cancer de próstata (DARSON et al. 1997), e sua acurácia na predição desse tipo de cancer é de 59,7%, entretanto esse valor não supera o valor preditivo do PSA (KUREK et al. 2004; LILJA et al. 2007). IGFs (*insulin-like growth factor*) também vem sendo sugeridas como potenciais marcadores moleculares de câncer de próstata. Altos níveis circulantes de IGF-1 vem sendo associado a um aumento do risco de desenvolvimento de cancer de próstata (CHAN et al. 1998), mas novamente a predição por IGF-1 não acrescenta ao teste de PSA (HARMAN et al. 2000). No caso da proteína ligante IGFBP-2, altos níveis circulantes tem sido observados em pacientes com cancer de próstata, com uma correlação inversa com a agressividade e estadios tardios da doença, entretanto ainda em níveis de expressão mais altos em pacientes com o cancer (SHARIAT et al. 2002). Também, baixos níveis de IGFB-3 foram associados a um risco aumentado de desenvolvimento de cancer de próstata e inversamente correlacionados com o desenvolvimento de metástase óssea (KANETY et al. 1993; CHAN et al. 1998; SHARIAT et al. 2002). Expressão local aumentada de TGF-β1 vem sendo associada com alto grau e estadio tumoral, invasão e progressão metastatica em pacientes com cancer de próstata (STEINER e BARRACK 1992;

TRUONG et al. 1993; SHARIAT et al. 2004) entretanto, não é capaz de diferenciar pacientes saudáveis de pacientes com cancer de próstata. Níveis circulantes elevados de IL-6 e e seu receptor tem sido associados com a doença com características agressivas, estadios avançados, presença de metástase a distancia e sobrevida diminuida (TWILLIE et al. 1995; ADLER et al. 1999; NAKASHIMA et al. 2000; SHARIAT et al. 2001) entretanto, assim como TGF-β1, não são capazes de diagnosticar o cancer de próstata. Expressão aumentada de uPA e seu receptor também vem sendo observada em pacientes com cancer de próstata (MCCABE et al. 2000) e relacionada com invasão tumoral (GAYLIS et al. 1989; KEER et al. 1991), progressão e metástase óssea (HIENERT et al. 1988; MIYAKE et al. 1999; SHARIAT et al. 2007) e além disso, uPA parece ser um forte preditor de recidiva bioquímica pós prostatéctomia (SHARIAT et al. 2008). Outro potencial marcador molecular de cancer de próstata é o EPCA, que apresenta alta sensibilidade e especificidade para o diagnóstico de cancer de próstata (PAUL et al. 2005). Aparentemente, até o momento, o mais promissor marcador molecular de câncer de próstata é o PCA3. PCA3 é descrito como altamente super-expresso em câncer de próstata quando comparado ao tecido não tumoral adjacente e sua expressão é restrita a próstata (BUSSEMAKERS et al. 1999). A super-expressão de PCA3 é confirmada em mais de 95% dos tumores primários e metastáticos de próstata, sendo mais de 60 vezes no tecido tumoral em relação ao normal (BUSSEMAKERS et al. 1999; DE KOK et al. 2002; HESSELS et al. 2003; DERAS et al. 2008). A quantificação de PCA3 em sedimento urinário foi sugerida como teste diagnóstico em 2002 e hoje já existem diversos estudos de sucesso nessa área (HESSELS et al. 2003; FRADET et al. 2004; TINZL et al. 2004; GROSKOPF et al. 2006; MARKS et al. 2007; VAN

GILS et al. 2007), inclusive já sendo utilizados na prática clínica (GEN-PROBE, (GROSKOPF et al. 2006).

Dentro das alterações em nível de DNA exploradas como marcadores moleculares de câncer de próstata, o rearranjo gênico mais comum envolve os fatores de transcrição ETS, ERG (21q22.2) e ETV1 (7p21.2) e o gene regulador de andrógeno TMPRSS2 (21q.22.3) (TOMLINS et al. 2005). Essa fusão gênica é observada em 40% a 80% dos pacientes com câncer de próstata, em cerca de 20% das neoplasias intraepiteliais prostáticas (NIPs) e raramente em lesões benignas (LAXMAN et al. 2006) e parece ser mais comumente associada com *Gleason Score* maior que 7 e com pior prognóstico e desenvolvimento de metástase (DEMICHELIS et al. 2007). Outro tipo de alteração em nível de DNA frequentemente presente em câncer de próstata é a perda de heterozigosidade, descrita em múltiplos sítios cromossômicos (revisado por (WRIGHT e LANGE 2007). Alterações epigenéticas, como a hipermetilação de ilhas de CpG em regiões promotoras de genes supressores de tumor, são um importante evento na tumorigênese, considerado um passo inicial no desenvolvimento do câncer de próstata (revisado por (WRIGHT e LANGE 2007). O principal gene encontrado hipermetilado em câncer de próstata é o GSTP1, utilizado na distinção entre tecidos de próstata benignos e malignos (GOESSL et al. 2000; JERONIMO et al. 2001; GONZALGO et al. 2004). Outros genes como p16, ARF, MGMT (pelo menos um deles, (HOQUE et al. 2005), APC, RASSF1a e RARβ2 (ROUPRET et al. 2007) também foram reportados como hipermetilados em pacientes com câncer de próstata comparado ao tecido normal, sendo que os genes identificados por (ROUPRET et al. 2007) tiveram uma sensibilidade de detecção de 86% e uma acurácia diagnóstica de 89% para câncer de próstata. Alterações em nível

de sequência também vêm sendo relatadas com a identificação de SNPs (*single nucleotide polymorphisms*), principalmente em cinco regiões cromossômicas distintas, relacionados ao risco de desenvolvimento de cancer de prostata. Essas alterações foram observadas em três regiões no 8q23, em uma região no 17q12 e em uma região no 7q.24.3. Hoje já existe comercialmente um teste que analiza a presença desses e de outros SNPs (deCODE Diagnostics Laboratory, US) (revisado por FRADET 2009).

A tecnologia de microarranjos de cDNA vem proporcionando avanços na identificação de marcadores moleculares para o câncer. Por exemplo, em relação ao câncer de próstata, o primeiro gene identificado por microarranjos de cDNA, com possibilidade de uso na prática clínica para um melhor diagnóstico, foi o P504S (AMACR) (XU et al. 2000). A proteína traduzida desse gene já é utilizada clinicamente, auxiliando na distinção do câncer de próstata de doenças benignas (COOPER et al. 2007) e na discriminação de diferentes graus e tipos de câncer de próstata (JIANG et al. 2004). Outros exemplos de marcadores moleculares para o câncer de próstata, identificados por microarranjos de cDNA, são: o gene EZH2, que pode ser usado como marcador de progressão e metástase, indicando uma pior sobrevida (LATULIPPE et al. 2002; VARAMBALLY et al. 2002; RHODES et al. 2003), e a fusão gênica TMPRSS2-ERG (TOMLINS et al. 2005).

2 JUSTIFICATIVA

A partir da exploração dos dados gerados no trabalho de mestrado (MELLO 2007), que teve como foco a atividade transcricional do genoma humano, resultando na identificação de 3.421 regiões transcritas não descritas por nenhum outro projeto, propomos a busca de marcadores moleculares de câncer. Estes possíveis novos transcritos identificados no trabalho de mestrado podem representar variantes de *splicing*, ncRNAs, RNAs primários ou transcritos antisenso naturais (NATs) e, devido ao sabido envolvimento desses tipos de sequências na regulação do genoma humano e portanto, apresentando um potencial como marcadores moleculares de câncer, propomos uma reanálise das sequências que compõem a lâmina utilizada no trabalho de mestrado (MELLO 2007).

Neste trabalho construímos uma lâmina de microarranjos de cDNA, composta por sequências resultantes do Projeto Genoma do Câncer Humano (FAPESP/LICR - HCGP) que não se alinharam com sequências de cDNA (ESTs) geradas por outros projetos e depositadas em bancos de dados públicos (cerca de 30% das sequências obtidas no projeto) (CAMARGO et al. 2001; FONSECA et al. 2006) e com probabilidade de representarem novos transcritos, segundo o algoritmo utilizado no trabalho de mestrado de FONSECA (2005). Utilizamos estratégias de bioinformática para identificar novos transcritos a partir dos dados gerados com a hibridização da lâmina contra cDNAs provenientes de 12 tecidos diferentes e validamos parte dos achados através de RT-PCR.

Ênfase foi dada ao tecido de próstata pelo fato de que, além de apresentar o câncer com maior prevalência nos homens e ser o sexto tipo de câncer mais comum no mundo (Ministério da Saúde 2008), existe uma necessidade de novos marcadores moleculares de diagnóstico, prognóstico e tratamento da doença, uma vez que a maior parte dos novos marcadores moleculares de câncer de próstata não apresentam acurácia, sensibilidade e especificidade satisfatórias para serem traduzidos para a rotina clínica.

3 OBJETIVOS

3.1 GERAL

Identificar marcadores moleculares de câncer a partir das regiões transcritas não caracterizadas identificadas no trabalho de mestrado (MELLO 2007).

3.2 ESPECÍFICOS

- Identificar dentro das 3.421 regiões transcritas identificados durante o mestrado (MELLO 2007), regiões transcritas não mapeadas em éxons anotados.

- Associar as regiões transcritas não mapeadas em éxons anotados com os dados de expressão gerados por microarranjos de cDNA, obtidos durante o mestrado, buscando identificar transcritos tumor-associados e tumor/tecido-associados.

- Validar e explorar os transcritos que se mostrarem interessantes na associação com os dados de microarranjos de cDNA, quanto a tendência a expressão diferencial em câncer de prostata, em amostras de tecidos normais ou tumorais, através de PCR em tempo real.

4 MATERIAIS E MÉTODOS

4.1 IDENTIFICAÇÃO DE REGIÕES TRANSCRITAS NÃO MAPEADAS EM ÉXONS ANOTADOS

4.1.1 Mapeamento genômico das sequências imobilizadas na lâmina

Com o intuito de identificar as sequências ORESTES imobilizadas na lâmina de microarranjos de cDNA, gerada durante o trabalho de mestrado (MELLO 2007), que não correspondem a éxons anotados, realizamos o mapeamento genômico destas sequências utilizando as bases de dados disponíveis na UCSC Genome Bioinformatics (genome.ucsc.edu). Os arquivos contendo as coordenadas genômicas de ESTs depositadas em bancos de dados públicos e os arquivos contendo coordenadas de anotação de transcritos provenientes do NCBI (RefSeq), Ensembl e UCSC Genome Bioinformatics (KnownGene) (disponíveis em hgdownload.cse.ucsc.edu/goldenPath/hg18/database) foram carregados em um banco de dados local (MySQL). Inicialmente, foram excluídas as ESTs com mapeamento em múltiplas localizações genômicas. Em seguida, as sequências ORESTES presentes na plataforma de microarrajos de cDNA, foram mapeadas em relação as coordenadas referentes aos bancos de transcritos utilizados. Sequências ORESTES não mapeadas em unidades transcricionais (UTs) anotadas foram definidas como intergênicas, aquelas mapeadas em UTs anotadas mas sem sobreposição com éxons conhecidos foram classificadas como intrônicas e, finalmente, aquelas sobrepostas a éxons conhecidos (sobreposição de pelo menos uma base) foram classificadas como

exônicas. Estes mapeamentos foram realizados utilizando programas desenvolvidos no nosso laboratório (Laboratório de Biotecnologia do Hospital do Câncer, LBHC), utilizando linguagem de programação PERL.

4.1.2 Predição de ncRNAs

Com o intuito de identificar candidatos a ncRNAs, sequências genômicas correspondentes as ORESTES imobilizadas na lâmina foram analisadas utilizando uma abordagem para identificar somente ncRNAs com evidência de conservação de sequência e estrutura secundária em nível de RNA. Primeiro essas seqüências foram divididas em três grupos, de acordo com seus mapeamentos utilizando as bases de dados RefSeq e KnownGene (hgdownload.cse.ucsc.edu/goldenPath/hg18/database): (1) totalmente exônicas, (2) parcialmente exônicas (sobreposição de pelo menos uma base) e (3) não-exônicas. Para cada grupo, foram realizadas buscas por três características: (1) ORF putativa, (2) potencial codificador e (3) conservação de sequência e estrutura secundária. Para verificar se uma sequência é inteiramente uma ORF, foi utilizado o programa getorf (www.ebi.ac.uk/Tools/emboss), o qual analisa se as três fases de leitura de ambas as fitas da sequência poderiam gerar uma sequência codificadora; também foi verificada se a ORF mais longa identificada por este programa refletia a sequência completa ou com exceção de duas bases nas extremidades. Para refinar essa predição de ORFs, também foi utilizado o programa Coding Potential Calculator (CPC) (KONG et al. 2007) com os parâmetros originais, o qual classifica sequências em codificadoras e não codificadoras (fracamente codificadoras, codificadoras, fracamente não codificadoras ou não codificadoras). Esse programa leva em consideração seis características, sendo três delas baseadas

na extensão, qualidade e integridade de ORFs preditas. As outras três são derivadas da análise dos alinhamentos obtidos utilizando a ferramenta BLASTX contra o banco de dados UniRef90 (blast.ncbi.nlm.nih.gov/Blast.cgi), sendo o número, a qualidade e a estrutura dos alinhamentos. Nós agrupamos as sequências classificadas como não codificadoras ou fracamente não codificadoras e as sequências classificadas como codificadoras ou fracamente codificadoras. Para identificar conservação de sequência e estrutura secundaria, foram realizadas buscas por alinhamentos múltiplos inter-espécies (16 genomas de vertebrados contra o genoma humano, hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way) que sobrepunham regiões das sequências ORESTES. Estes alinhamentos foram analisados usando o programa RNAz (WASHIETL et al. 2005) com os parâmetros originais, a fim de detectar evidências de conservação de estrutura secundária, como substituições compensatórias de bases. Foram consideradas ncRNAs putativos sequências que apresentaram as seguintes características: mapeamento parcialmente exônico ou não-exônico, predição do programa CPC de potencial não codificador e evidência de conservação de estrutura secundária de acordo com o programa RNAz.

4.2 ANÁLISES DE EXPRESSÃO DOS DADOS DE MICROARRANJOS DE CDNA

Fizemos uma reanálise dos dados de expressão gênica obtidos durante o mestrado (MELLO 2007) buscando diferenças de expressão em relação as amostras normais e tumorais. Nesta etapa, três (placenta, pulmão e testículo) de 12 tecidos utilizados nos experimentos de microarranjos de cDNA (MELLO 2007) foram

descartados, uma vez que nesses experimentos só foram utilizadas amostras normais desses tecidos, impossibilitando a análise de expressão diferencial em busca de marcadores moleculares para cânceres desses tecidos. Foram determinadas, para cada sequência, a diferença de expressão em *fold* e a mediana de intensidade de sinal em duas comparações: tecido tumoral *versus* tecido normal para cada tecido estudado e próstata tumoral *versus* todos os tecidos normais. Inicialmente foi escolhido o tecido de próstata para ser trabalhado uma vez que este representa o câncer com maior prevalência nos homens (Ministério da Saúde 2008). Então, usando gráficos MA, nós selecionamos regiões transcritas não mapeadas em éxons anotados, expressas pelo menos 4 ou -4 vezes em tumor de próstata em relação a próstata normal e pelo menos 2 ou -2 vezes em tumor de próstata quando comparado a todos os tecidos normais estudados (valores convertidos em $\log_2$).

4.3 VALIDAÇÃO E DETERMINAÇÃO DA EXPRESSÃO DIFERENCIAL DAS REGIÕES TRANSCRITAS NÃO MAPEADAS EM ÉXONS ANOTADOS POR PCR EM TEMPO REAL

4.3.1 Iniciadores e gene normalizador

Iniciadores foram desenhados para as sequências diferencialmente expressas entre próstata tumoral e normal usando os programas Primer3 (frodo.wi.mit.edu) e Primer Express (Applied Biosystems, US). A escolha do gene normalizador HPRT, utilizado nos experimentos de PCR em tempo real relativos a próstata, foi baseada na literatura a partir de trabalhos que estudaram perfis de expressão gênica em amostras

de tecidos normais ou tumorais de diferentes regiões (VANDESOMPELE et al. 2002; DE KOK et al. 2005; RUBIE et al. 2005).

4.3.2 RNAs e síntese de cDNAs

A utilização de amostras de pacientes foi devidamente autorizada, sob aprovação deste trabalho pelo Comitê de Ética em Pesquisa do Hospital A. C. Camargo (CEP 970/07). Foram utilizadas amostras de próstata normal ou tumoral, provenientes dos bancos de RNAs e de Tumores do Hospital A. C. Camargo. Estes RNAs foram extraídos pela equipe do biobanco, utilizando o método de TRIzol® (Invitrogen, US), de acordo com as instruções do fabricante, passando previamente pelas etapas de confirmação do diagnóstico histológico e semi-microdissecção manual, realizados em colaboração com o Departamento de Anatomia Patológica do Hospital A. C. Camargo.

Da mesma forma, RNAs provenientes de três linhagens celulares de tumor de próstata (PC-3, DU 145 e LNCaP), cedidas pelo Instituto Ludwig de Pesquisa sobre o Câncer de São Paulo e cultivadas pelo Laboratório de Investigação Médica/24 da Faculdade de Medicina da Universidade de São Paulo, foram extraídos (TRIzol®, Invitrogen, US). Essas linhagens foram cultivadas com base nas instruções da ATCC (www.atcc.org) e o RNA total foi extraído quando atingida a confluência de 80-90%.

Todas as amostras de RNAs foram submetidas a tratamento com DNase (RQ1 RNase-Free DNase, Promega, US) de acordo com as instruções do fabricante, previamente a síntese de cDNA para a remoção de DNA genômico contaminante. A concentração dos RNAs foi estimada através de leitura em NanoDrop® (ND-1000

Spectrophotometer, US) e a integridade foi verificada utilizando o 2100 Bioanalyzer (Agilent Technologies, US).

A síntese da primeira fita de cDNA foi realizada para amostras de tecidos normais ou tumorais, com seus respectivos controles negativos, na ausência da enzima transcriptase reversa. As reações partiram de 2µg de RNA total, usando 0,5µg oligo(dT)$_{12-18}$, 0,5mM dNTP mix, 1x first-strand buffer, 10mM DTT, 200U SuperScript™ II Reverse Transcriptase (Invitrogen, US) e 40U RNasin® (Promega, US), em volume final de reação de 20µl, utilizando as condições térmicas sugeridas pelo fabricante, mantidas no Gene Amp PCR System 9700 (PE Applied Biosystems, US).

Para avaliar a eficiência de síntese de cDNA, o produto foi amplificado com 0,15mM dos iniciadores desenhados em diferentes éxons do gene p53 (iniciador senso - 5' GGAGGAGCCGCAGTCAGA 3' e iniciador antisenso - 5' CAAGAAGCCCAGACGGAAAC 3'; produtos amplificados com 340pb), a fim de poder diferenciar a amplificação de DNA genômico (produtos amplificados com 548pb). A reação partiu de 0,5µL de cada cDNA, utilizando 1x PCR buffer, 1,6mM $MgCl_2$, 0,2mM dNTP mix e 1U Platinum Taq DNA Polymerase (Invitrogen, US), em um volume final de reação de 20µL, sob a ciclagem: 95ºC por 3 minutos; 30 ciclos de 95ºC por 45 segundos, 58ºC por 45 segundos e 72ºC por 1 minuto; 72ºC por 5 minutos, finalizando o processo a 4ºC, no mesmo sistema, Gene Amp PCR System 9700 (PE Applied Biosystems, US). Os produtos amplificados foram analisados em gel de agarose 1%, corado com brometo de etídeo.

4.3.3 PCR em tempo real

As reações de PCR em tempo real foram realizadas em duplicata para a monitoração contínua da fluorescência do SYBR Green®, utilizando o 7900HT Fast Real-Time PCR System (Applied Biosystems, US). As reações foram padronizadas usando um conjunto de RNAs provenientes de 3 linhagens celulares de tumor de próstata. As concentrações de cDNA e de iniciadores utilizadas nos experimentos de PCR em tempo real foram padronizadas de modo a se obter o menor CT (*cycle threshold*) possível, em um volume final de 20µl de reação (5µl cDNA, 5µl iniciadores e 10µl master mix, SYBR® Green PCR Master Mix, Applied Biosystems, US), por 10 minutos a 95ºC e 40 ciclos de 15 segundos a 95ºC e 1 minuto a 60ºC, seguido do protocolo de desnaturação térmica. As curvas de eficiência, utilizadas nos cálculos de expressão relativa, foram obtidas a partir da diluição seriada (1:2, 10 concentrações) de DNA genômico de (extraído da linhagem de mama Hb4a), em experimentos realizados em triplicata, sendo aceitos somente os valores de eficiência entre 1,7 e 2,3 (GINZINGER 2002). Os dados foram extraídos usando o programa 7900HT Sequence Detection System, versão 2.3 (Applied Biosystems, US), usando os parâmetros padrões e os produtos amplificados foram fracionados por eletroforese em gel de poliacrilamida 8%, corado com prata, para verificação do tamanho do produto esperado ou a presença de produtos inespecíficos. O cálculo de diferença de expressão foi realizado de acordo com a formula de (PFAFFL 2001) e $\log_2$ foi aplicado a todos os valores de expressão.

Adicionalmente aos experimentos de PCR em tempo real para as nossas sequênciass-alvo, também foram realizados esses experimentos para um marcador molecular de câncer de próstata previamente descrito (AMACR) (XU et al. 2000),

como controle positivo das nossa reações (iniciador senso - 5’ AGAAATTGAGTCTGTGGGAAGCA 3’ e iniciador antisenso - 5’ AGCCATGAATTCCCCATCTG 3’; produtos amplificados com 101pb).

4.3.4 Confirmação da correspondência das ORESTES imobilizadas na lâmina de microarranjos de cDNA aos transcritos validados

A fim de certificar que os fragmentos imobilizados na lâmina verdadeiramente representam as sequências esperadas, validadas por PCR em tempo real, realizamos experimentos de sequênciamento. Para isso, os clones bacterianos contendo as sequências ORESTES originais foram crescidos a 37ºC em meio Luria-Bertani (LB) contendo ampicilina (250mg/mL). Seguindo as instruções dos fabricantes, o DNA plasmidial foi obtido através do método à vácuo do kit Wizard® Plus Minipreps DNA Purification System (Promega, US) e, em seguida, o inserto de cDNA foi amplificado por PCR usando iniciadores M13/pUC (Fermentas, CA) e a enzima Phoneutria Taq DNA Polymerase (Phoneutria, BR), em um volume final de 20μL. Uma fração do produto de PCR (25%) foi submetido a eletroforese em gel de agarose 2%, corado com brometo de etídeo, para confirmação da presença de um fragmento com o tamanho esperado e o restante foi purificado através de tratameto com ExoSAP-IT (USB, US) para ser submetido ao sequênciamento (BigDye® Terminator v3.1 Cycle Sequencing Kit; 3130 Genetic Analyzer, Applied Biosystems, US) no LGEA (Laboratório de Genômica e Biologia Molecular) do Hospital A. C. Camargo. Os resultados gerados foram analisados através da ferramenta BLAST (www.ncbi.nlm.nih.gov/BLAST).

5 RESULTADOS E DISCUSSÃO

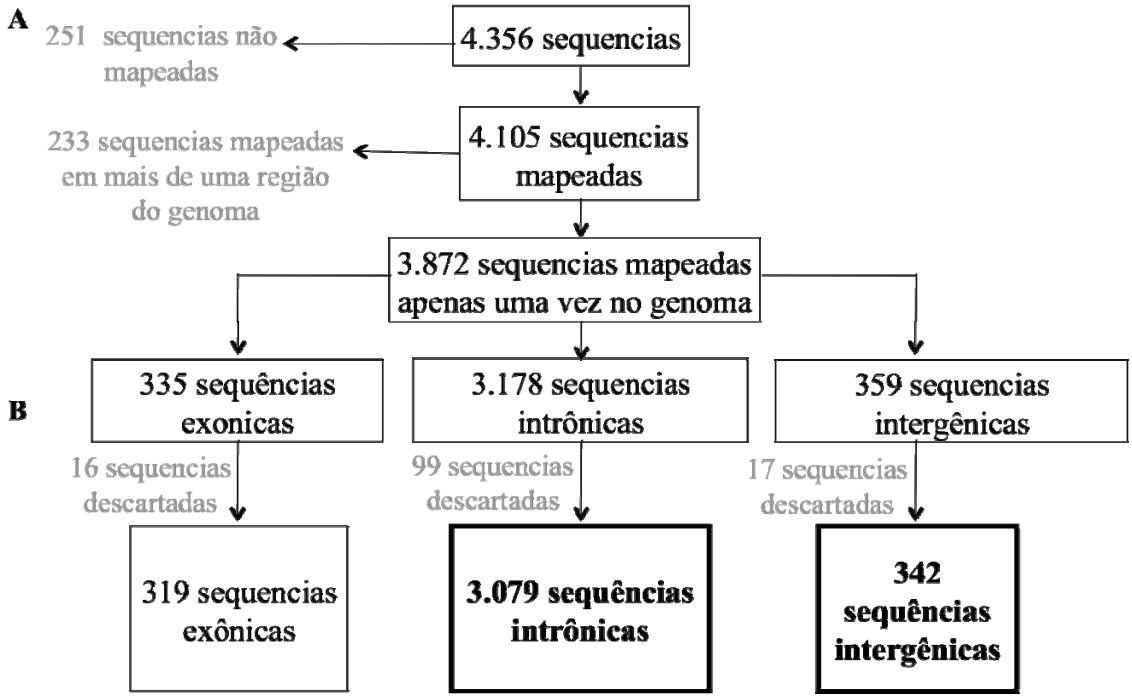
5.1 IDENTIFICAÇÃO DE REGIÕES TRANSCRITAS NÃO MAPEADAS EM ÉXONS ANOTADOS

5.1.1 Mapeamento genômico das sequências imobilizadas na lâmina

O alinhamento das sequências imobilizadas na lâmina (MELLO 2007) com o genoma humano mostrou que essas ORESTES estão preferencialmente mapeadas em regiões transcritas do genoma humano, em sua maioria em regiões intrônicas (Figura 1), indicando que essas ORESTES devem representar novos transcritos, uma vez que estão mapeadas em regiões transcritas do genoma humano mas não sobrepostas a éxons de genes codificadores conhecidos. Apenas uma pequena fração das sequências imobilizadas na lâmina estão sobrepostas a éxons de genes codificadores ou mapeadas em regiões intergênicas (Figura 1).

Esse alinhamento resultou em 3.872 sequências, mapeadas apenas uma vez no genoma, a serem consideradas em nossas análises (sequências não mapeadas ou sequências mapeadas em mais do que uma região do genoma humano foram descartadas, 484 sequências). Um grande número dessas ORESTES (3.767) não possuem evidência de *splicing*. As 3.872 sequências foram divididas em: exônicas (335), intrônicas (3.178) e intergênicas (359), representando 8,6%, 82,1% e 9,3% das sequências mapeadas apenas uma vez no genoma, respectivamente (Figura 1*a*). Nós verificamos, através de um teste de Wilcoxon, que essas sequências imobilizadas na lâmina não apresentaram uma tendência sistemática associada a sua classificação,

uma vez que não houveram diferenças estatisticamente significativas na comparação das medias de intensidade de sinal dos três tipos de sequências (exônicas, intrônicas e intergênicas), corroborando com o potencial das sequências mapeadas em regiões não exônicas representarem sequências transcritas. Após a aplicação de critérios restritivos estabelecidos no trabalho de mestrado (MELLO 2007), baseados em valores de intensidade de sinal, para considerar transcritos válidos, o número final de regiões transcritas não associadas com éxons anotados válidas foi de: 3.079 sequências intrônicas e 342 sequências intergênicas (Figura 1*b*). Rapidamente, para demonstrar o número potencial de novos transcritos humanos foi identificado, em cada lâmina, o elemento com menor valor de intensidade (após correção da fluorescência de fundo) e, dentre os 112 elementos identificados, adotamos como valor de corte o maior valor de intensidade encontrado para cada canal. Em todas as lâminas, os elementos com valor de intensidade menor que este valor de corte foram descartados. Para definir o número de prováveis novos transcritos, foi calculada a mediana dos valores de intensidade de cada elemento em todas as lâminas e somente aqueles que apresentaram mediana do valor de intensidade maior que os valores de corte estabelecidos para cada canal foram considerados contendo uma sequência de cDNA correspondente a um potencial novo transcrito. Também obedeceram a esse critério 319 sequências exônicas (Figura 1*b*), corroborando com o potencial da nossa abordagem em identificar novas regiões transcritas, uma vez que estas sequências foram depositadas em bancos de dados públicos por outros autores, ao longo da realização do trabalho de mestrado (MELLO 2007).



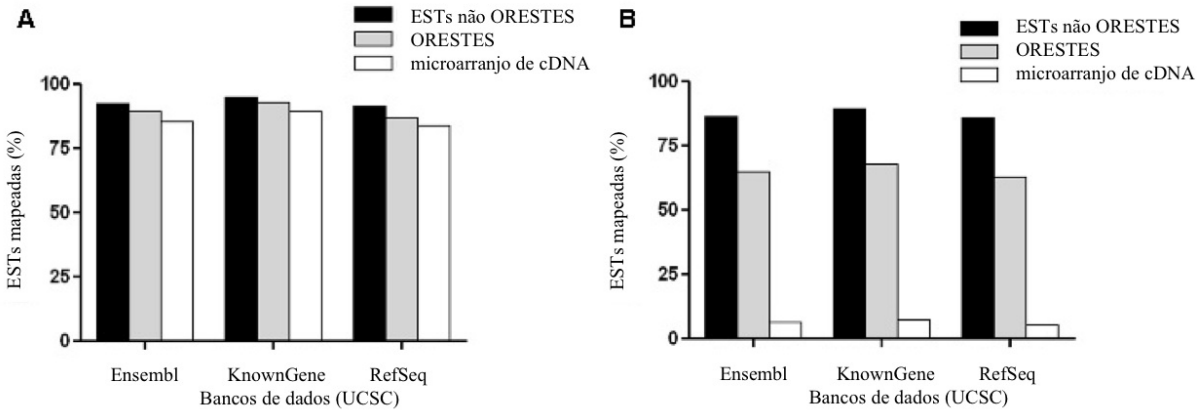
Legenda: *a*- Das 4.356 sequências imobilizadas na lâmina, 251 e 233 sequências foram descartadas porque não mapearam ou mapearam em mais de uma localização do genoma humano, respectivamente. As 3.872 sequências restantes, mapeadas apenas uma vez no genoma humano, foram divididas em sequências exônicas, intrônicas e intergênicas. *b*- Após a aplicação dos critérios de intensidade adotados no mestrado (MELLO 2007), que resultou na eliminação de 16, 99 e 17 sequências dos grupos de sequências exônicas, intrônicas e intergênicas, respectivamente, os números de transcritos válidos na lâmina foram de 319, 3.079 e 342, para sequências exônicas, intrônicas e intergênicas, respectivamente.

Figura 1 - Mapeamento genômico das 4.356 sequências que compõem a lâmina de microareanjos de cDNA.

5.1.2 Predição de ncRNAs

De acordo com o programa getorf (www.ebi.ac.uk/Tools/emboss) que detecta ORFs em uma sequências, apenas 9% das sequências mapeadas apenas uma vez no genoma poderiam representar genes codificadores de proteínas, indicando um grande potencial nos nossos dados para a busca de sequências não codificadoras. Além disso, apesar da metodologia ORESTES apresentar uma tendência de cobertura de regiões centrais de transcritos, com grande probabilidade de representar ORFs,

nossos dados mostraram que apenas 7,6% das sequências não exônicas apresentaram ORFs putativas. Em contraste, 47,3% das sequências totalmente exônicas apresentaram uma ORF putativa (Figura 2).



Legenda: *a*- ESTs mapeadas em regiões transcritas. *b*- ESTs mapeadas em regiões exônicas. *Barras pretas*- ESTs; *barras cinza*- ORESTES; *barras brancas*- ORESTES que compõem o microarranjo de cDNA.

Figura 2 - Mapeamento genômico de ESTs, de acordo com três bancos de dados distintos.

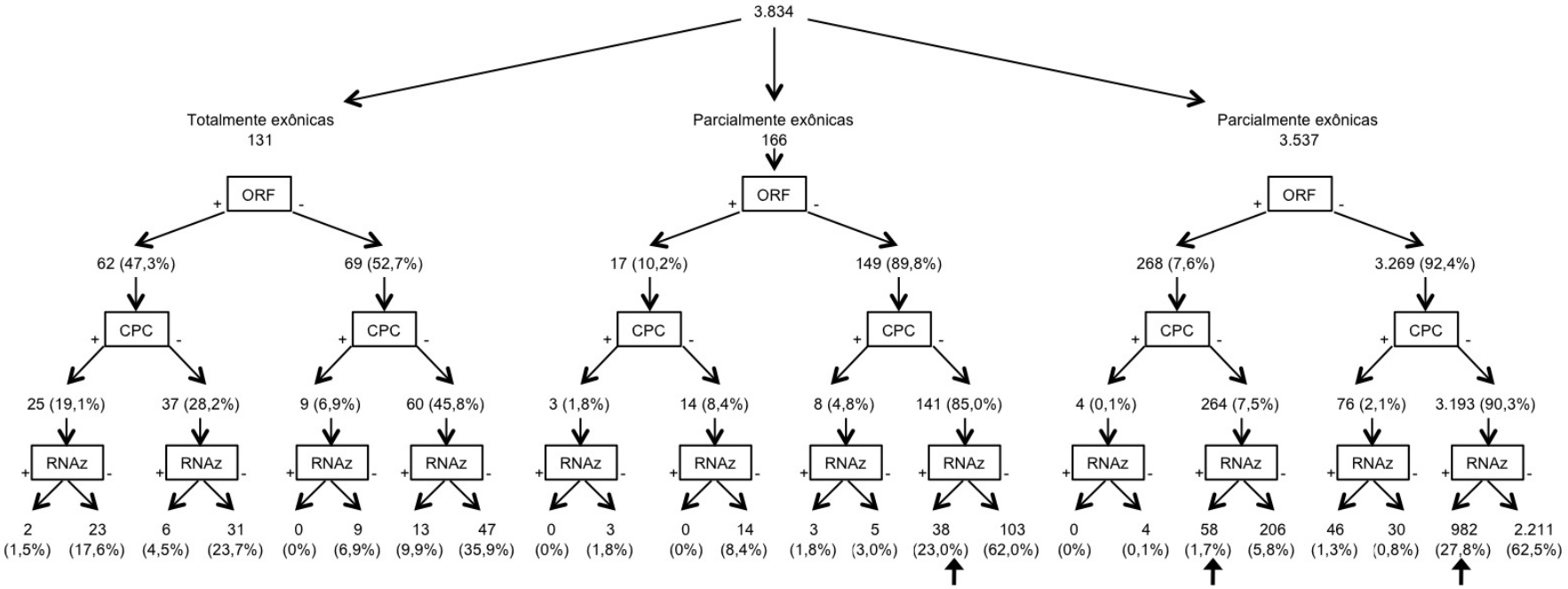
Para estas análises nós agrupamos as sequências que não mapearam em éxons anotados (intrônicas e intergênicas), compondo um grupo de 3.537 sequências não-exônicas. As sequências exônicas foram divididas em totalmente exônicas (131) e parcialmente exônicas (166). Uma vez que para essas análises nós consideramos somente as bases de dados KnownGene e RefSeq para classificar as sequências, foram descartadas 38 sequências previamente classificadas como exônicas de acordo com o mapeamento inicial, feito a partir dos três bases de dados: RefSeq, Ensembl e KnownGene (hgdownload.cse.ucsc.edu/goldenPath/hg18/database). Foram consideradas ncRNAs putativos sequências que apresentaram as seguintes características: mapeamento parcialmente exônico ou não-exônico, predição do

programa CPC de potencial não codificador e evidência de conservação de estrutura secundária de acordo com o programa RNAz.

ncRNAs frequentemente apresentam pouca ou nenhuma cobertura de ESTs, mas esse fato não indica que eles não são expressos ou não são funcionais (POLLARD et al. 2006; WEILE et al. 2007). Os maiores bancos de dados que contem milhares de sequências de ncRNAs anotadas são o RNAdb (MEHLER e MATTICK 2006) e o NONCODE (LIU et al. 2005). A partir de um conjunto inicial de mais de 48.000 regiões estruturadas, PEDERSEN et al. (2006) (PEDERSEN et al. 2006) predisse 10.000 transcritos de RNAs estruturados no genoma humano. WASHIETL et al. (2005) (WASHIETL et al. 2005) estimou que 35.000 RNAs estruturados são conservados em mamíferos. Predições computacionais *de novo* de ncRNAs são difíceis, uma vez que esses tipos de transcritos não apresentam a maioria das assinaturas que tornam as predições de genes codificadores de proteínas possíveis (GRIFFITHS-JONES 2007). Entretanto, ncRNAs geram RNAs funcionais e frequentemente apresentam estrutura secundária de pareamento de bases conservada em vez de similaridade de sequências primária (GRIFFITHS-JONES et al. 2005; WEILE et al. 2007). Sabe-se que as estruturas secundárias de RNAs exercem um papel funcional importante, não somente em transcritos não codificadores, mas também no contexto de mRNAs codificadores de proteínas (RYMARQUIS et al. 2008). O pareamento de bases na estrutura secundária é mantido por mutações de base compensatórias e essas alterações podem ser usadas como evidência estatística de pressão evolutiva para manter os pares de bases naquelas posições (WASHIETL et al. 2005; PEDERSEN et al. 2006; GRIFFITHS-JONES 2007; WEILE et al. 2007). Existem diversas abordagens de sucesso para

predizer ncRNAs baseadas na ideia de que estruturas de RNAs significativamente funcionais são conservadas em espécies relacionadas, mesmo quando a sequências primária não é (MACHADO-LIMA et al. 2008). Evidências de estrutura secundária e conservação de sequências em nível de RNA podem ser combinadas em análises de bioinformática, resultando em perfis de alinhamentos múltiplos de sequências de ncRNAs que podem ser capturados por modelos estatístico (GRIFFITHS-JONES et al. 2005; WEILE et al. 2007).

Das sequências parcialmente exônicas, identificamos 38 candidatos a ncRNAs e das sequências não-exônicas, identificamos 1.040 ncRNAs putativos. Destas, 58 sequências possuem uma ORF putativa, mas também apresentam baixo potencial codificador e estrutura secundária e 982 sequências não apresentam ORFs putativas. Portanto, aproximadamente 28% das sequências analisadas podem representar ncRNAs (1.078 de 3.834) (Figura 3, Tabela 1).



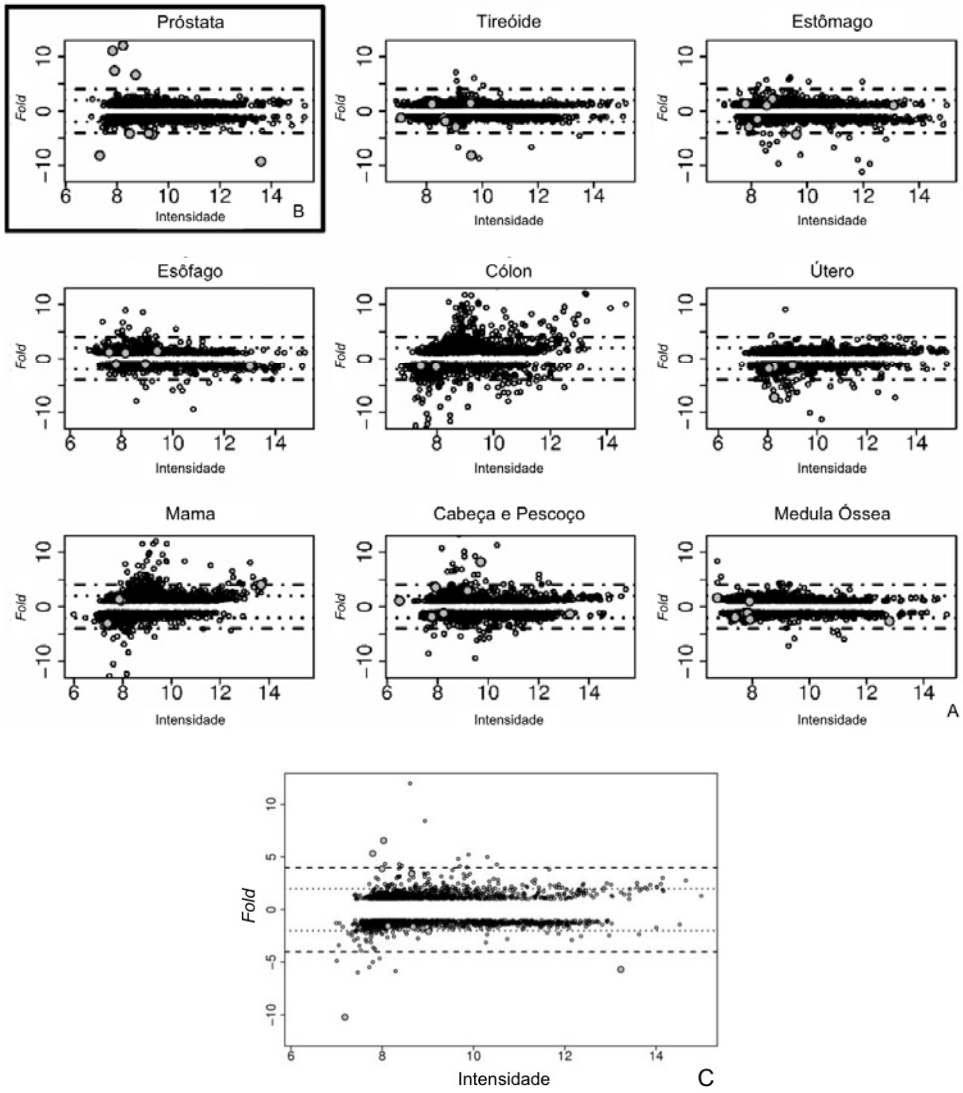
Legenda: 3.834 sequências mapeadas uma vez no genoma humano, de acordo com as bases de dados KnownGene e RefSeq (hgdownload.cse.ucsc.edu/goldenPath/hg18/database) foram separadas em três grupos: totalmente exônicas (131), parcialmente exônicas (166) e não-exônicas (3.537). Para cada grupo, foram separadas as sequências com e sem uma ORF putativa (programa getorf, www.ebi.ac.uk/Tools/emboss; e verificação manual). Foi utilizado o programa CPC (KONG et al. 2007) para refinar a predição inicial de ORFs (sequências classificadas como não codificadoras e fracamente não codificadoras foram agrupadas, bem como sequências classificadas como codificadoras e fracamente codificadoras). Finalmente, foram feitas buscas por alinhamentos múltiplos inter-espécies (genome.ucsc.edu) que sobrepunham regiões das sequências ORESTES e foi utilizado o programa RNAz (WASHIETL et al. 2005) detectar evidências de conservação de estrutura secundária. Foram consideradas ncRNAs putativos sequências que apresentaram as seguintes características: mapeamento parcialmente exônico ou não-exônico, predição do programa CPC de potencial não codificador e evidência de conservação de estrutura secundária de acordo com o programa RNAz. ***Setas preenchidas***- ncRNAs putativos.

Figura 3 - Passos seguidos para a identificação de candidatos a ncRNAs estruturados.

5.2 ANÁLISES DE EXPRESSÃO DOS DADOS DE MICROARRANJOS DE CDNA

Uma reanálise dos dados de expressão gênica obtidos durante o mestrado (MELLO 2007) foi realizada, buscando diferenças de expressão em relação as amostras normais e tumorais. Foram construídos gráficos MA mostrando, para cada sequência, a diferença de expressão em *fold* e a mediana de intensidade de sinal em duas comparações: tecido tumoral *versus* tecido normal para cada tecido estudado (Figura 4*a*) e próstata tumoral *versus* todos os tecidos normais (Figura 4*b*). Na Figura 4, é possível observar um grande número de sequências diferencialmente expressas entre amostras tumorais e normais de todos os tecidos estudados, sendo esta diferença de pelo menos duas vezes (aproximadamente 28% das sequências intrônicas e intergênicas mapeadas apenas uma vez no genoma). Esse dado ilustra o potencial dos nossos resultados na exploração de marcadores moleculares para câncer em diversos tecidos. O número total de sequências diferencialmente expressas, com diferenças em *fold* entre amostras normais e tumorais de pelo menos duas vezes em um ou mais tecidos diferentes, e uma concordância em relação a essas sequências estarem super- ou sub-expressas em todos os tecidos nos quais elas são diferencialmente expressas, foi de 1.007, sendo 111 de 335 sequências exônicas, 885 de 3.178 sequências intrônicas e 111 de 359 sequências intergênicas. O valor de *fold* de duas vezes foi adotado, uma vez que a maioria dos trabalhos envolvendo microarranjos de cDNA que não apresentam um número amostral suficiente para se fazer um teste estatístico, definem sequências diferencialmente expressas através desse valor maior ou igual a dois (SCHENA et al. 1995; SCHENA et al. 1996;

QUACKENBUSH 2001; YANG et al. 2002). Sequências que estavam super-expressas em alguns tecidos e sub-expressas em outros não foram consideradas nos nossos resultados como sequências diferencialmente expressas.



Legenda: Gráficos MA mostrando, para cada elemento, a diferença de expressão e a mediana de intensidade de sinal para as comparações: *a*- Amostras tumorais *versus* as amostras normais. *b*- Próstata tumoral *versus* próstata normal. *c*- Próstata tumoral *versus* todos os tecidos normais estudados. ***Círculos cinza***- sequências diferencialmente expressas em próstata selecionadas para a validação por PCR em tempo real. ***Linha pontilhada***- delimitação do valor de expressão igual a 2 ou -2. ***Linha tracejada***- delimitação do valor de expressão igual a 4 ou -4.

Figura 4 - Análises de expressão dos dados de microarranjos de cDNA (MELLO 2007).

Muitos ncRNAs estão relacionados a funções biológicas específicas, apresentando um padrão de expressão tecido- ou condição-específico. Então, considerando o mesmo critério de sequências diferencialmente expressas descrito acima, 291 transcritos foram classificados como ncRNAs putativos diferencialmente expressos, pela nossa abordagem. Quatro por cento desses putativos marcadores tumorais não codificadores estão presentes no banco de dados NONCODE (LIU et al. 2005) ou são preditos como pares antisenso por (GALANTE et al. 2007), novamente corroborando com a validade da nossa abordagem, mas nunca foram descritos anteriormente como sendo diferencialmente expressos em tumores. Uma dessas sequências é a ORESTES CV372409, que mostrou-se sub-expressa, nos nossos experimentos de microarranjos de cDNA, em três tumores distintos em relação aos seus tecidos normais. Na Tabela 1 é possível observar o número de tecidos onde os ncRNAs putativos estão diferencialmente expressos. Pelo menos cinco putativos marcadores tumorais não codificadores estão super-expressos em quatro ou mais tumores diferentes comparados com seus tecidos normais (AW803984, BE161676, CV358552, AW814925 e AW935941), ilustrando resultados promissores na busca de marcadores tumorais.

Tabela 1 - ncRNAs putativos e sua distribuição em relação a expressão diferencial.

	Sequências parcialmente exônicas					**Sequências não-exônicas**					
	ORF+	ORF-				ORF+		ORF-			
Número de ncRNAs putativos	0	38				58		982			
Número de tecidos □□□□□onde os ncRNAs putativos estão diferencialmente expressos	0	1	2	3	4	1	2	1	2	3	4	5
Número de ncRNAs putativos super-expressos	0	1	2	3	1	3	1	92	40	13	3	1
Número de ncRNAs putativos sub-expressos	0	5	1	0	0	6	0	103	15	1	0	0
Número de marcadores tumorais não codificadores putativos	0	13				10		268				

O tecido de próstata foi escolhido para a etapa de validação por PCR em tempo real uma vez que o câncer de próstata é o sexto tipo de câncer mais comum no mundo e o mais prevalente em homens, representando cerca de 10% do total de câncer (Ministério da Saúde 2008) e além disso, existe uma necessidade de novos marcadores moleculares para complementar o teste de PSA, uma vez que a maior parte dos novos marcadores moleculares de câncer de próstata não apresentam sensibilidade e especificidade satisfatórias para serem aplicados na rotina clínica de diagnóstico e prognóstico do câncer de próstata.

Então, focamos nas análises de diferença de expressão em próstata (Figuras 4*b* e 4*c*) para selecionar candidatos a validação. Adotamos como critérios de seleção, a expressão de pelo menos 4 ou -4 vezes em tumor de próstata em relação ao tecido normal (Figura 4*b*) e de pelo menos 2 ou -2 vezes em tumor de próstata em relação a todos os tecidos normais estudados (Figura 4*c*). Assim, nove sequências (sete intrônicas e duas intergênicas) diferencialmente expressas em tumor de próstata foram selecionadas como candidatas a validação por PCR em tempo real (Tabela 2)

sendo, em geral, diferencialmente expressas somente em próstata e não nos demais tecidos (Figura 4, *pontos cinzas*).

5.3 VALIDAÇÃO E DETERMINAÇÃO DA EXPRESSÃO DIFERENCIAL DAS REGIÕES TRANSCRITAS NÃO MAPEADAS EM ÉXONS ANOTADOS POR PCR EM TEMPO REAL

5.3.1 Iniciadores e gene normalizador

Iniciadores foram desenhados para nove sequências diferencialmente expressas entre próstata tumoral e normal (Tabela 2) utilizando os programas Primer3 (frodo.wi.mit.edu), Primer Express (Applied Biosystems, US), levando em consideração a existência de regiões repetitivas dentro das sequências (RepeatMasker Web Server, www.repeatmasker.org) e, portanto, evitando-as. Também foram analisadas a existência de estruturas secundárias e a formação de dímeros nos iniciadores usando o programa Oligo Tech (www.oligosetc.com/analysis.php).

O gene HPRT foi utilizado como normalizador nos experimentos de PCR em tempo real. A escolha desse normalizador foi baseada na literatura, uma vez que o gene HPRT foi identificado como sendo o melhor gene normalizador único a ser utilizado em experimentos de PCR em tempo real (DE KOK et al. 2005). Nesse estudo, realizado por (DE KOK et al. 2005), foram analisados 13 genes normalizadores em cinco tecidos diferentes, resultando na clara observação da expressão do gene HPRT refletindo acuradamente a expressão media de diversos genes normalizadores comumente utilizados.

Tabela 2 - Sequências provenientes das análises de próstata, selecionadas para a validação por PCR em tempo real.

Número de acesso	Tamanho do produto amplificado	Tamanho da ORESTES	Iniciador FW	Iniciador RV	Cromossomo
BQ373258	150	700	CCAGCTGAGACCTAATGCAA	CTTCACAAAAGCAGCTGGAA	9
CV398755	74	429	ACAGCCTGACCATCAGACAA	GCACTAGTGGGAGCCAGTCT	6
CV374350	81	201	CACCAAAAGGCAACAAGTGA	ATCACCCACTTCAAGCACAA	6
AW849290	108	346	TGAACACTTTTCGTGCTGAA	ACTATGCCTGGCTTGATTTG	2
BE144456	100	551	CCTGAGAACAAAGCACCTACGA	TGATTTCTCCAATGACCTTGACCTA	22
AW793062	74	658	TTCAATGTGTATGGGAGAATGA	GGTGGGGCAAATAGTAGTGA	6
BF910617	71	306	GGCCAGTGGTTAATCATCCT	TTTACAGAAGAGAAACGAGAGACAA	9
CV400462	100	589	CCGTGTTGTTGTGGCTCCTT	CCCTCACCCTCGGATTCCT	7
BF365844	100	273	GAGAAAAATGCAAGGGACAGAAG	CTCACATCTGCAGAGGTTTTGTCT	12

5.3.2 RNAs e síntese de cDNAs

Foram obtidos RNAs de três linhagens celulares de câncer de próstata (PC-3, DU 145 e LNCaP) para a construção de uma coleção de RNAs, usado na otimização das reações de PCR em tempo real. Todas as linhagens apresentaram RNA total íntegro (RIN $\geq$ 5, *RNA Integrity Number*), tanto antes como após o tratamento com DNase (Tabela 3). Quantidades equivalentes de RNAs de cada uma das linhagens foram misturadas a fim de se obter essa coleção.

Tabela 3 - Linhagens celulares de próstata que compõem a coleção de RNAs usada na otimização dos experimentos de PCR em tempo real.

Linhagem	Origem	RIN pré-DNase	RIN pós-DNase
PC-3	metástase óssea de adenocarcinoma de próstata grau IV	9,2	7,8
DU 145	metástase cerebral de carcinoma de próstata	9,2	7,3
LNCaP	metástase linfonodal de carcinoma de próstata	9,1	8,4

Para a validação das sequências selecionadas, foram utilizadas sete amostras pareadas de próstata (tecido tumoral e normal adjacente do mesmo paciente), todas com *Gleason Score* 7 (Tabela 4). Previamente a extração de RNA, essas amostras passaram pelas etapas de confirmação do diagnóstico histológico e semi-microdissecção manual, a fim de certificar que as amostras incluídas no trabalho eram compostas por pelo menos 70% de células normais ou tumorais (FREEMAN et al. 1999; DE SOUZA et al. 2000), respectivamente para os tecidos normais ou tumorais a serem testados. Como esperado, ao contrario das linhagens celulares, os RNAs provenientes dos pacientes apresentaram uma qualidade inferior, principalmente após o tratamento com DNase (Tabela 4). A avaliação da integridade

do RNA (RIN) é um primeiro passo crítico para se obter dados de expressão, uma vez que a integridade do RNA pode ter um impacto significativo, especialmente em dados de expressão relativa (nos valores de CTs) (FLEIGE e PFAFFL 2006). Recomenda-se o uso de RNAs com RINs homogêneos e maiores do que 5. Entretanto, uma vez que todos os RNAs de pacientes apresentaram um mesmo padrão de integridade, apesar da qualidade inferior, entendemos que a validação por PCR em tempo real não seria comprometida, pois a comparação da expressão relativa das sequências selecionadas seria feita em amostras comparáveis.

Tabela 4 - Amostras pareadas de próstata usadas nos experimentos de PCR em tempo real.

Amostra	Diagnóstico	TNM	RIN pré-DNase	RIN pós-DNase
1T	adenocarcinoma	T2N0M0	6,1	2,0
1N	tecido adjacente normal		6,5	2,0
2T	adenocarcinoma	T2N0M0	6,7	4,4
2N	tecido adjacente normal		*	4,6
3T	adenocarcinoma	T2N0M0	2,1	*
3N	tecido adjacente normal		4,4	3,1
4T	adenocarcinoma	T2N0M0	3,4	2,5
4N	tecido adjacente normal		3,4	*
5T	adenocarcinoma	T2N0M0	4,9	2,8
5N	tecido adjacente normal		2,2	2,0
6T	adenocarcinoma	T3N0M0	5,3	4,6
6N	tecido adjacente normal		5,9	2,2
7T	adenocarcinoma	T3N0M0	2,6	2,1
7N	tecido adjacente normal		*	*

* Valor não estimado pelo 2100 Bioanalyzer (Agilent Technologies, US).

A síntese de cDNA foi realizada da mesma forma para todas as amostras, a partir de 2μg de RNA e com seus respectivos controles negativos, na ausência da enzima transcriptase reversa. O cuidado em adotar esse tipo de controle negativo foi tomado, uma vez que as sequências selecionadas para validação não apresentavam evidência de *splicing* e, portanto, não seria possível distinguir o produto amplificado específico da contaminação com DNA genômico (se esta estivesse presente) pelo tamanho do produto amplificado.

A eficiência da síntese de cDNA foi avaliada através de PCR, utilizando iniciadores desenhados em diferentes éxons do gene p53. Todos os produtos de PCR, referentes aos cDNAs sintetizados na presença de transcriptase reversa, mostraram uma banda única e específica com o tamanho esperado para o produto amplificado (340pb). Os controles negativos da síntese de cDNA (sem transcriptase reversa) apresentaram ausência de bandas, indicando não haver contaminação nas amostras com DNA genômico (548pb).

5.3.3 PCR em tempo real

Métodos de verificação de dados independentes de microarranjos de cDNA e de análises de bioinformática, devem ser considerados como seguimento dos resultados obtidos (YANG e SPEED 2002).

Cada par de iniciadores foi utilizado nos experimentos de PCR em tempo real, partindo de cDNAs das amostras pareadas de próstata, DNA genômico humano como controle positivo e ausência de molde como controle negativo das reações. Em geral, as reações foram padronizadas para uma quantidade inicial de 100ng de cDNA, com as concentrações dos iniciadores variando entre 800nM e 1600nM.

Todos os pares de iniciadores tiveram seus controles positivos e negativos funcionando com sucesso (presença e ausência de amplificação, respectivamente). O sucesso da amplificação foi definido pela presença de um pico único na curva de dissociação, referente ao produto esperado, confirmado através da análise do tamanho do produto em gel de acrilamida 8% corado com prata.

Foram validadas como transcritos expressos oito de nove sequências (Tabela 5). A ORESTES CV398755 não teve seus valores de expressão calculados porque não conseguimos padronizar reações de PCR em tempo real utilizando seus iniciadores. Considerando uma diferença de expressão de três vezes, em pelo menos três de sete amostras, três sequências foram consideradas como potenciais marcadores moleculares de câncer de próstata (Tabela 5). O gene AMACR, utilizado como controle positivo das reações de PCR em tempo real, também foi validado, segundo os mesmos critérios descritos acima. Este marcador molecular foi previamente descrito como apresentando alta sensibilidade e especificidade para detectar cânceres de próstata de diferentes tipos e graus, sendo seu mRNA super-expresso em cerca de 30% (microarranjos) a 60% (PCR em tempo real) dos tumores de próstata e expresso em baixos níveis ou indetectável no tecido normal (XU et al. 2000; JIANG et al. 2001; JIANG et al. 2004).

Tabela 5 - Resultados da validação por PCR em tempo real, nas amostras pareadas de próstata, comparados aos resultados do microarranjo de cDNA.

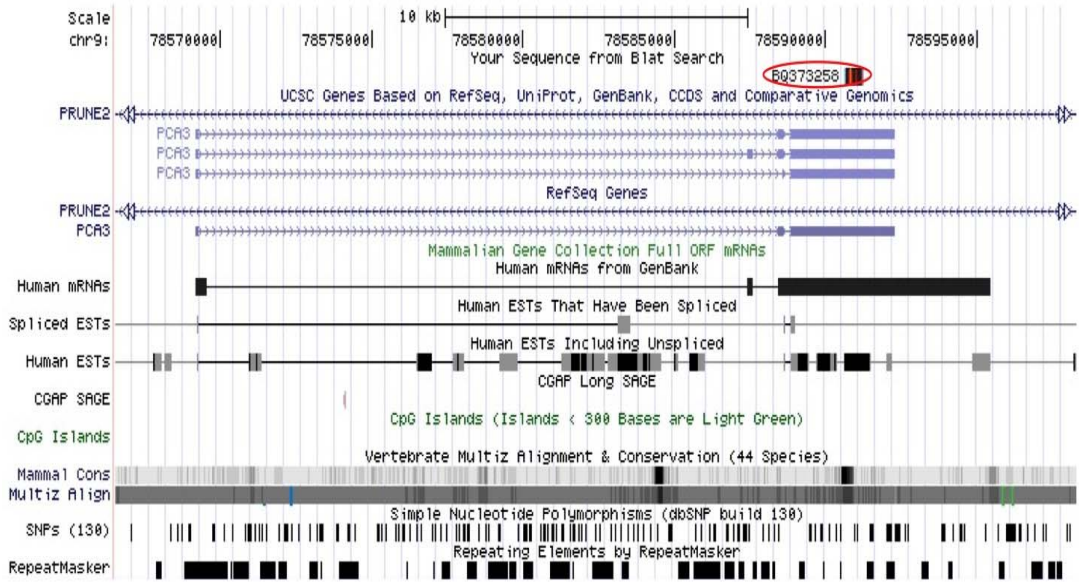
Número de acesso	Par 1	Par 2	Par 3	Par 4	Par 5	Par 6	Par 7	Expressão média (PCR em tempo real)	Expressão (microarranjos de cDNA)
AMACR	-1,16	-0,33	5,51	1,52	3,13	-0,40	6,50	2,11	-
BQ373258	-0,95	-0,42	100%*	100%*	100%*	4,62	5,09	5,50	7,41
CV398755	-	-	-	-	-	-	-	-	-4,14
CV374350	0,03	-0,97	-0,01	0,12	-1,49	0,97	-0,62	-0,28	-4,32
AW849290	0,18	-0,20	100%*	3,89	1,75	0,21	0,56	1,05	6,67
BE144456	0,71	-0,73	0,86	2,60	-0,36	0,56	1,31	0,70	-8,21
AW793062	0,21	-0,10	100%*	100%*	100%*	2,15	100%*	6,03	12,02
BF910617	0,54	-0,74	-0,25	4,56	100%*	0,86	3,05	2,57	11,06
CV400462	-0,76	0,05	-0,11	0,85	0,01	-2,67	-0,85	-0,50	-4,11
BF365844	0,14	-0,92	2,95	2,84	1,80	0,30	-0,88	0,89	-9,28

* Os valores representados por 100% indicam expressão somente nas amostras tumorais (nenhum sinal detectado em amostras normais) e foram convertidos para 10 vezes nos cálculos de média de expressão.

** Todos os valores representam $\log_2$ dos valores de expressão, considerando tumor/normal.

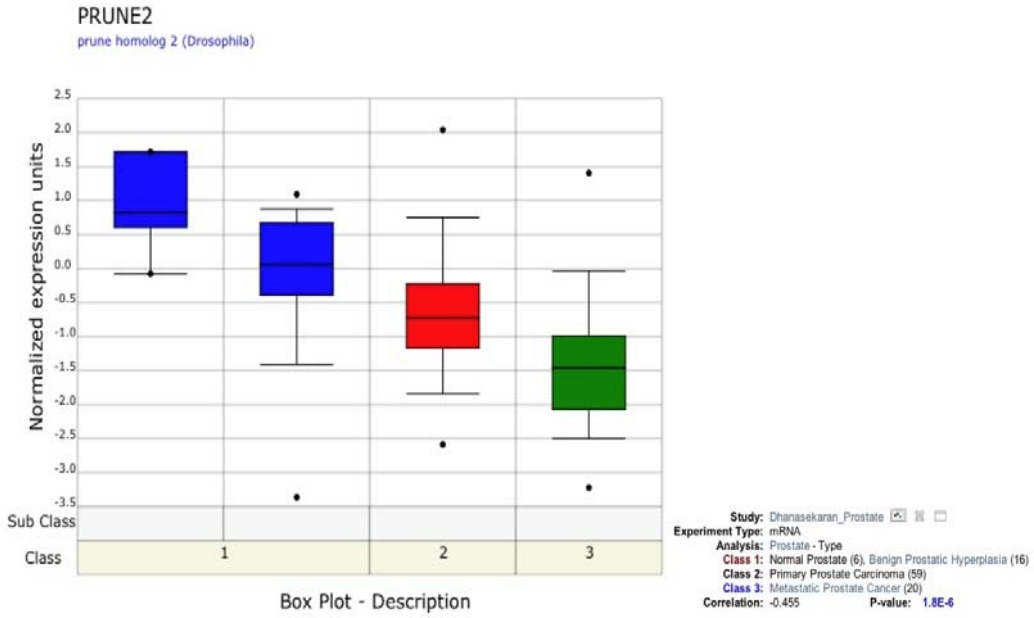
A validação da ORESTES BQ373258 ilustra o valor da nossa abordagem para identificar marcadores moleculares em regiões transcritas não mapeadas em éxons descritos, uma vez que esta sequências esta mapeada no último éxon de um ncRNA super-expresso em tumores de próstata já descrito (PCA3) (BUSSEMAKERS et al. 1999) (Figura 5). Sua expressão diferencial foi confirmada em cinco de sete amostras pareadas de próstata, sendo super-expressa em tumor de próstata, o que também serviu como controle positivo para os nossos experimentos de PCR em tempo real. Um número incomum de códons de parada foi identificado ao longo de toda a extensão da sequências de cDNA de PCA3 (BUSSEMAKERS et al. 1999; SCHALKEN et al. 2003), o que adicionalmente a ausência de uma ORF e a análises de proteínas putativas de pequenas ORFs preditas, resultou na classificação

de PCA3 como um ncRNA poliadenilado (HESSELS et al. 2003; SCHALKEN et al. 2003; TINZL et al. 2004). Sua função é desconhecida, apesar de existirem especulações de que PCA3 possa agir na regulação da expressão genica ou participar de processos de *splicing* (SCHALKEN et al. 2003). O mapeamento de PCA3 indica que esse ncRNA é um transcrito antisenso, localizado dentro de um íntron do gene PRUNE2 (Figura 5). Em análises realizadas através do Oncomine Research (www.oncomine.org), encontramos quatro conjuntos de dados (DHANASEKARAN et al. 2001; LAPOINTE et al. 2004; YU et al. 2004; VARAMBALLY et al. 2005) mostrando uma diminuição da expressão de PRUNE2 em amostras tumorais de próstata, principalmente em amostras de tumor metastático, quando comparadas a amostras normais (Figura 6), sendo essa relação de expressão diferencial inversa a observada para a ORESTES BQ373258 (PCA3) nos nossos dados. Tanto nossos dados de microarranjos de cDNA (7, 41 vezes), quanto os de PCR em tempo real (em media 5,50 vezes), além dos dados da literatura (60 vezes) (BUSSEMAKERS et al. 1999; DE KOK et al. 2002; HESSELS et al. 2003; DERAS et al. 2008), mostraram PCA3 super-expresso em câncer de próstata quando comparado ao tecido normal. Com base nesses dados, nossa hipótese é de que a super-expressão de PCA3 em câncer de próstata poderia estar inibindo a expressão de PRUNE2 ou favorecendo alguma variante específica desse gene, através de interferência no mecanismo de *splicing*, de acordo com a progressão tumoral.



Legenda: Mapeamento genomico da ORESTES submetida mostrando sua localização cromossômica, predições gênica feitas pela UCSC Genome Bioinformatics, genes da base de dados RefSeq, mRNAs com ORFs completas da coleção genica de mamíferos, mRNAs humanos da base de dados GenBank, ESTs humanas com ou sem *splicing*, sequências provenientes do CGAP-SAGE, ilhas de CpG, regiões conservadas entre mamíferos e alinhamentos múltiplos inter-espécies, polimorfismos de nucleotídeos únicos e regiões de sequências repetitivas. ***Circulo vermelho***- ORESTES BQ373258; ***linhas***- íntrons; ***caixas***- éxons; ***setas***- indicação da direção da transcrição genica.

Figura 5 - Alinhamento genômico, obtido pela ferramenta BLAT (UCSC Genome Bioinformatics, genome.ucsc.edu) da ORESTES BQ373258.

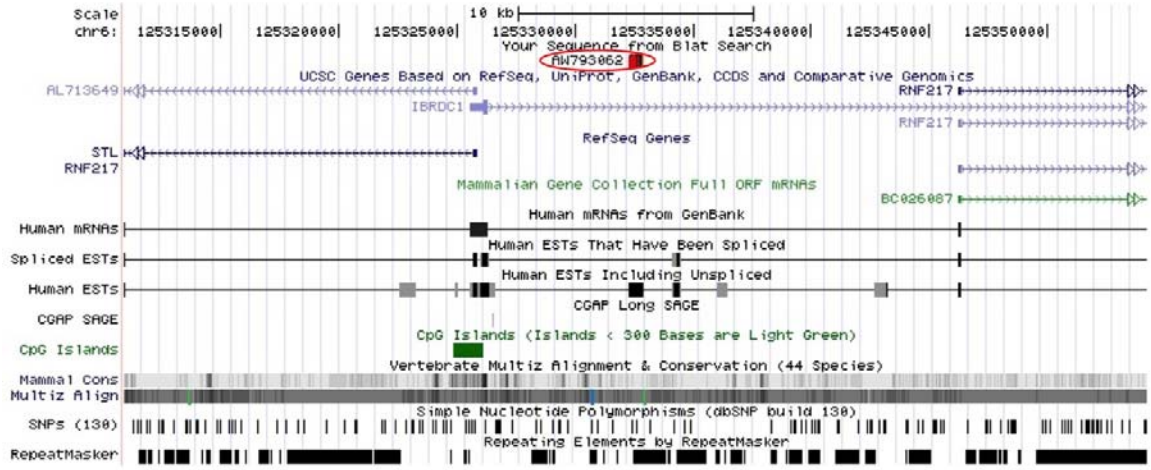


Legenda: Gráficos *box plot* de um dos quatro conjuntos de dados que mostram uma diminuição da expressão do gene PRUNE2 em amostras tumorais de próstata, principalmente em amostras de tumor metastático, quando comparadas a amostras normais (DHANASEKARAN et al. 2001). ***Classe 1 (azul)***- próstata normal (n = 6) e hiperplasia prostática benigna (n = 16); ***classe 2 (vermelho)***- tumor primário (n = 59); ***classe 3 (verde)***- tumor metastático (n = 20).

Figura 6 – Análise de expressão do gene PRUNE2 proveniente do Oncomine Research (www.oncomine.org).

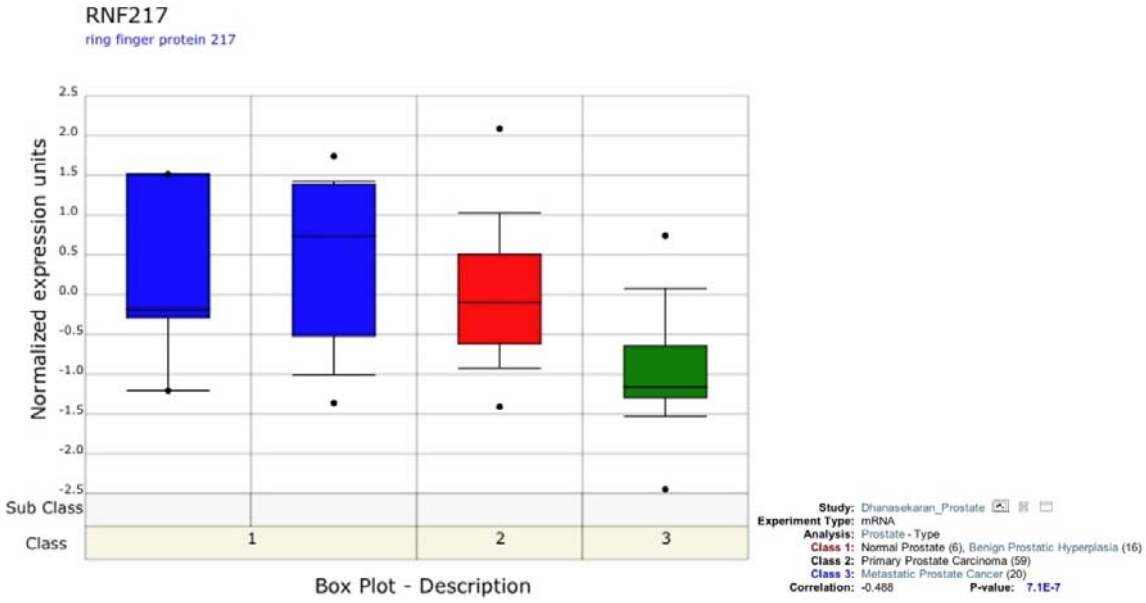
Outra ORESTES que teve sua super-expressão em tumor de próstata confirmada, com valores de *fold* altos, em quatro amostras pareadas, é a AW793062. Seu mapeamento genômico mostrou estar alinhada ao primeiro íntron de uma variante predita do gene RNF217 e próxima a um ncRNA antisenso já descrito (AL713649), que compartilham uma mesma ilha de CpG (Figura 7). De acordo com análises realizadas através do Oncomine Research (www.oncomine.org), a expressão diferencial da ORESTES AW793062 na comparação tecido tumoral *versus* tecido normal de próstata foi oposta a expressão diferencial do gene RNF217 na comparação tecido normal *versus* tumor primário *versus* tumor metastático de próstata (DHANASEKARAN et al. 2001). Nessas análises o gene RNF217 apresentou-se mais expresso em amostras de próstata

normal em relação a amostras tumorais, principalmente em relação a amostras de tumor metastático de próstata, mostrando uma diminuição de sua expressão de acordo com a agressividade tumoral (Figura 8). Levando em consideração os dados de PCR em tempo real e o mapeamento genômico dessa ORESTES, além dos dados das análises do Oncomine Research para o gene RNF217, hipotetizamos que talvez a ORESTES AW793062 possa estar relacionada ao gene RNF217, por exemplo representando um éxon alternativo de uma variante de *splicing* desse gene, ou um ncRNA antisenso, ou ainda estar relacionada ao ncRNA antisenso localizado próximo a ela, talvez agindo como um promotor alternativo e apresentando algum papel na regulação do gene RNF217.



Legenda: Mapeamento genomico da ORESTES submetida mostrando sua localização cromossômica, predições gênica feitas pela UCSC Genome Bioinformatics, genes da base de dados RefSeq, mRNAs com ORFs completas da coleção genica de mamíferos, mRNAs humanos da base de dados GenBank, ESTs humanas com ou sem *splicing*, sequências provenientes do CGAP-SAGE, ilhas de CpG, regiões conservadas entre mamíferos e alinhamentos múltiplos inter-espécies, polimorfismos de nucleotídeos únicos e regiões de sequências repetitivas. ***Circulo vermelho***- ORESTES AW793062; ***linhas***- íntrons; ***caixas***- éxons; ***setas***- indicação da direção da transcrição genica.

Figura 7 – Alinhamento genômico, obtido pela ferramenta BLAT (UCSC Genome Bioinformatics, genome.ucsc.edu) da ORESTES AW793062.

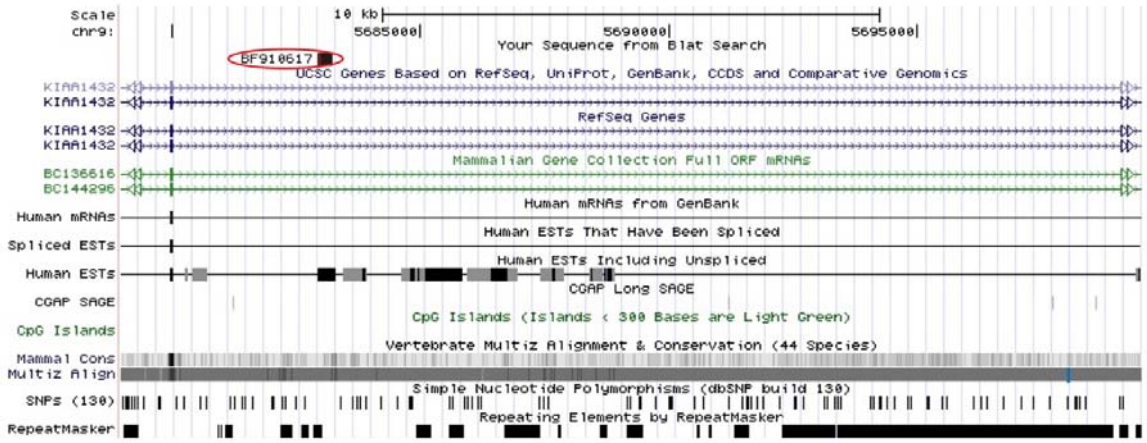


Legenda: Gráficos *box plot* mostrando uma diminuição da expressão do gene RNF217 de acordo com a tumorigênese e a agressividade do tumor de próstata (DHANASEKARAN et al. 2001). ***Classe 1 (azul)***- próstata normal (n = 6) e hiperplasia prostática benigna (n = 16); ***classe 2 (vermelho)***- tumor primário (n = 59); ***classe 3 (verde)***- tumor metastático (n = 20).

Figura 8 – Análise de expressão do gene RNF217 proveniente do Oncomine Research (www.oncomine.org).

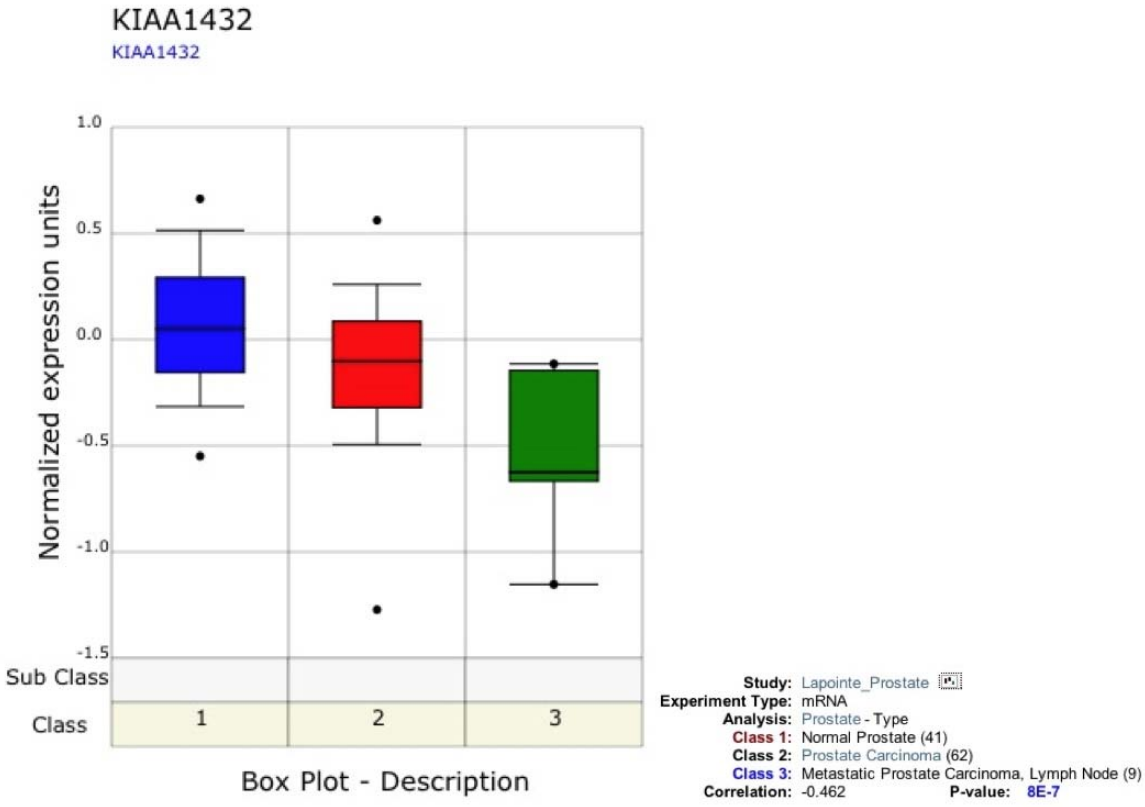
Já a ORESTES BF910617, foi validada em três amostras pareadas e mostrou-se super-expressa em câncer de próstata. Esta é um sequência intrônica ao gene KIAA1432. (Figura 9). De acordo com análises realizadas a partir do conjunto de dados de (LAPOINTE et al. 2004), através do Oncomine Research (www.oncomine.org), foi observado que a expressão diferencial da ORESTES BF910617 foi oposta a expressão diferencial do gene KIAA1432, na comparação tecido tumoral *versus* tecido normal de próstata. Nessas análises o gene KIAA1432 apresentou-se altamente expresso em amostras de próstata normal, com sua expressão diminuindo de acordo com a progressão da agressividade do tumor de próstata (Figura 10). Com isso, nossa hipótese é de que a ORESTES BF910617 pode

estar exercendo algum papel na regulação do gene KIAA1432, inibindo sua expressão no câncer de próstata, quando expressa em altos níveis.



Legenda: Mapeamento genomico da ORESTES submetida mostrando sua localização cromossômica, predições gênica feitas pela UCSC Genome Bioinformatics, genes da base de dados RefSeq, mRNAs com ORFs completas da coleção genica de mamíferos, mRNAs humanos da base de dados GenBank, ESTs humanas com ou sem *splicing*, sequências provenientes do CGAP-SAGE, ilhas de CpG, regiões conservadas entre mamíferos e alinhamentos múltiplos inter-espécies, polimorfismos de nucleotídeos únicos e regiões de sequências repetitivas. ***Circulo vermelho***- ORESTES BF910617; ***linhas***- íntrons; ***caixas***- éxons; ***setas***- indicação da direção da transcrição genica.

Figura 9 – Alinhamento genômico, obtido pela ferramenta BLAT (UCSC Genome Bioinformatics, genome.ucsc.edu) da ORESTES BF910617.



Legenda: Gráficos *box plot* mostrando o gene KIAA1432 mais expresso em amostras de próstata normal, com sua expressão diminuindo de acordo com a progressão da agressividade do tumor de próstata (LAPOINTE et al. 2004). ***Classe 1 (azul)-*** próstata normal (n = 41); ***classe 2 (vermelho)-*** tumor próstata (n = 62); ***classe 3 (verde)-*** linfonodo de tumor metastático de próstata (n = 9).

Figura 10 – Análise de expressão do gene KIAA11432 proveniente do Oncomine Research (www.oncomine.org).

Essas três ORESTES consideradas potenciais marcadores moleculares de câncer de próstata tiveram seus clones originais sequênciados e, suas correspondências às sequências esperadas foram confirmadas.

5.4 ESTUDOS ADICIONAIS DOS POTENCIAIS MARCADORES MOLECULARES DE CÂNCER DE PRÓSTATA

Uma vez que os potenciais marcadores moleculares de câncer de próstata identificados nesse estudo se mostraram interessantes em relação ao seus mapeamentos genômicos, além de suas expressões diferenciais, demos continuidade aos seus estudos.

5.4.1 Avaliação da possível regulação do gene PRUNE2 pelo ncRNA PCA3

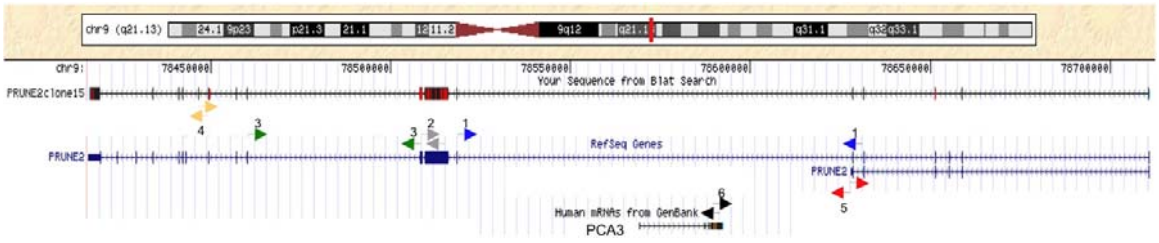
Nós quisemos avaliar a nossa hipótese de que a super-expressão de PCA3 em câncer de próstata poderia estar inibindo a expressão de PRUNE2 ou favorecendo alguma variante específica desse gene, de acordo com a progressão tumoral. Essa proposta foi realizada em colaboração com o GU Medical Oncology Department (Arap/Pasqualini Laboratory) do MD Anderson Cancer Center.

Para isso, experimentos de PCR em tempo real foram realizados em um conjunto de 20 amostras pareadas de próstata (provenientes do biobanco do Hospital A. C. Camargo) de diferentes características clínicas (Tabela 6), utilizando iniciadores desenhados no ncRNA PCA3, em diferentes regiões do gene PRUNE2 (NM_015225.2) e em duas variantes adicionais deste gene. Ao todo, 6 pares de iniciadores foram utilizados para avaliar a expressão de PCA3 e PRUNE2 (um para PCA3 e cinco para PRUNE2) (Tabela 7; Figura 11). Dentro da variante principal de PRUNE2 (NM_015225.2) avaliamos três regiões (próximo a terminação 5', central e próximo a terminação 3') e adicionalmente avaliamos mais duas variantes deste gene, uma identificada no Arap/Pasqualini Laboratory (MD Anderson Cancer Center), com um novo éxon próximo a extremidade 3' do gene e outra predita por análises de BLAT (UCSC Genome Bioinformatics, genome.ucsc.edu).

Tabela 6 - Características clínicas das amostras tumorais e dos pacientes que compreendem as 20 amostras pareadas de próstata usadas na avaliação da expressão de PCA3 e PRUNE2 e as 28 amostras pareadas de próstata usadas na avaliação da presença e frequência da retenção intrônica entre os éxons 3 e 4 de PCA3.

Amostra	TNM	*Gleason Score*	Recidiva bioquímica	Outros tumores
B2	T2bN0M0	(3+3) 6	não	carcinoma bicelular de cavidade oral
B3	T2aN0M0	(3+3) 6	sim	não
B10	T3aN0M0	(4+4) 8	?	mieloma múltiplo
B12	T2bN0M0	(3+3) 6	não	não
B29	T2bN0M0	(3+3) 6	não	carcinoma basocelular/ de células escamosas
B32	T3aNxM0	(3+3) 6	?	não
B36*	T3aN0M0	(3+3) 6	sim	não
B37	T2bNxM0	(4+3) 7	?	não
B38	T2bNxM0	(3+4) 7	?	Adenocarcinoma tubular de estômago
B39	T2bN1M0	(4+5) 9	?	não
B40	T2bN0M0	(4+4) 8	não	não
B42	T2aN0M0	(3+3) 6	não	não
B50	T2bN0M0	(3+3) 6	não	não
B51*	T2bN0M0	(3+3) 6	não	não
B54	T2aN0M0	(3+3) 6	não	não
B57*	T3aN0M0	(3+3) 6	não	não
B58*	T2bNxM0	(3+3) 6	?	não
B59	T2bNxM0	(3+3) 6	?	não
B60	T2bN0M0	(3+3) 6	não	não
B61	T2bN0M0	(3+3) 6	não	não
B65*	T2bN0M0	(4+5) 9	?	não
B66*	T3aN0M0	(3+3) 6	não	não
B68	T3aN0M0	(3+3) 6	não	não
B69*	T3bN1M0	(5+4) 9	?	não
B74	T3bN1M0	(4+5) 9	não	não
B76	T2cN0M0	(3+3) 6	?	não
B77	T3aN0M0	(3+4) 7	?	não
B80*	?	(4+3) 7	?	não

* amostras utilizadas somente na avaliação da presença e frequência da retenção intrônica entre os éxons 3 e 4 de PCA3.



Legenda: ***Setas azuis (1)-*** par de iniciadores próximo a terminação 5' de PRUNE2 (PRUNE2 5'); ***setas cinza (2)-*** par de iniciadores localizado na região central de PRUNE2 (PRUNE2 central); ***setas verdes (3)-*** par de iniciadores próximo a terminação 3' de PRUNE2 (PRUNE2 3'); ***setas amarelas (4)-*** par de iniciadores para a variante de PRUNE2 identificada no Arap/Pasqualini Laboratory (MD Anderson Cancer Center) (PRUNE2clone15), com um novo éxon (PRUNE2 novo éxon); ***setas vermelhas (5)-*** par de iniciadores para a variante de PRUNE2 predita por análises de BLAT (UCSC Genome Bioinformatics, genome.ucsc.edu); ***setas pretas (6)-*** par de iniciadores para PCA3.

Figura 11 – Localização de cada par de iniciadores utilizados na avaliação da expressão de PCA3 e PRUNE2.

Tabela 7 - Iniciadores utilizados na avaliação da expressão de PCA3 e PRUNE2.

Identificação do iniciador	Tamanho do produto amplificado	Iniciador FW	Iniciador RV
PRUNE2 5'	179	GGAGACCCAGTTCAGTGCTC	TGTAAATGCTTTCAAGTCACTGGT
PRUNE2 central	81	CGTTTATTTGCCGGTAGGAG	GCTCAGGCTCTTTGGTAGGA
PRUNE2 3'	152	GGGAAATGCTTTCACCACAG	CTCTTCAAAGGGGATGTCCA
PRUNE2 novo éxon	94	CATCGAGCCCTACAGGAGAG	CCTCATTCTTTCCAGGGTCA
PRUNE2 predito	163	GATTTCCCACAATTGCAGTC	AGGTTTGGCATTTTTCATTG
PCA3	150	CCAGCTGAGACCTAATGCAA	CTTCACAAAAGCAGCTGGAA

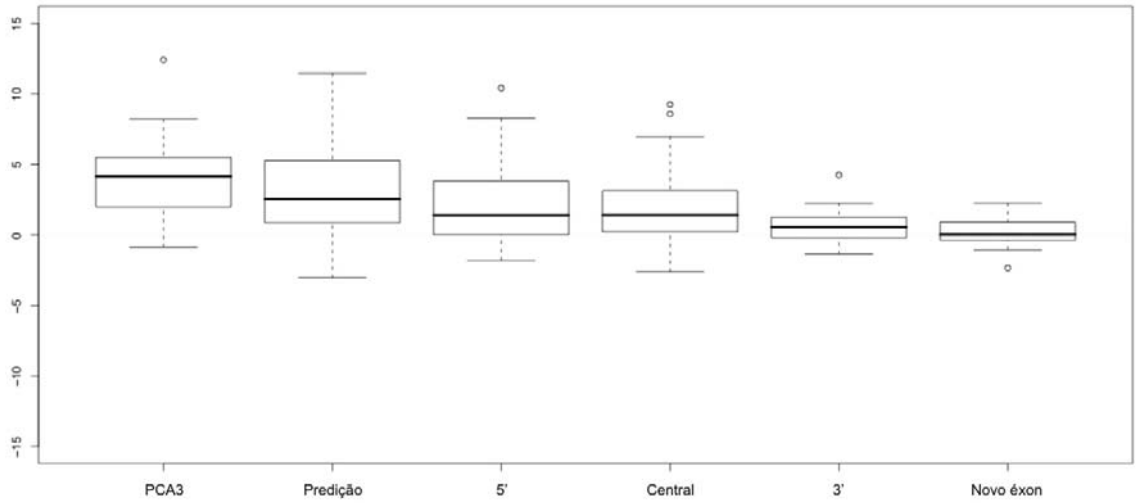
Todas etapas metodológicas foram realizadas da mesma forma que descrito anteriormente (Materiais e Métodos). Entretanto, neste caso em que os iniciadores foram desenhados em éxons distintos, possibilitando a diferenciação da amplificação do produto esperado da amplificação de DNA genômico contaminante, a síntese de cDNA não foi realizada com controles negativos, na ausência da enzima transcriptase reversa. Além disso, nesta etapa utilizamos iniciadores pentadecâméricos randômicos (pd(N)15) (STANGEGAARD et al. 2006), e a enzima

SuperScript™ III Reverse Transcriptase (Invitrogen, US). A ausência de contaminação genômica foi confirmada da mesma forma que descrita anteriormente (amplificação de um fragmento do gene p53). As reações de PCR em tempo real foram padronizadas usando a linhagem celular de tumor de próstata LNCaP e as curvas de eficiência foram obtidas a partir da diluição seriada (1:2, 10 concentrações) do cDNA desta mesma linhagem celular, uma vez que apenas as linhagens celulares de tumor de próstata LNCaP e 22Rv1 expressam PCA3 (BUSSEMAKERS et al. 1999; SCHALKEN et al. 2003). As reações foram padronizadas para uma quantidade inicial de 100ng de cDNA, com as concentrações dos iniciadores variando entre 660nM e 800nM. Foram utilizados três genes normalizadores (HPRT1, GUSB e TBP) nas reações de PCR em tempo real.

Primeiro, as razões de expressão relativa obtidas a partir da utilização dos três genes normalizadores de forma independente nos cálculo, foram próximas, indicando a reprodutibilidade dos nossos dados. Com a utilização do programa geNorm (medgen.ugent.be/~jvdesomp/genorm/), um algoritmo que determina o gene de referencia mais estável dentro de um conjunto de candidatos em um dado painel de amostras (VANDESOMPELE et al. 2002), o gene endógeno HPRT1 foi indicado apresentar a transcrição mais estável dentre os três genes normalizadores utilizados (HPRT1: M = 0,858; GUSB: M = 0,878; TBP: M = 1,087; M = medida de estabilidade do gene), corroborando com os dados da literatura (DE KOK et al. 2005), e por isso foi selecionado para ser considerado como normalizador nos cálculos de expressão relativa.

O principal resultado desses experimentos mostrou que quando PCA3 está altamente expresso em amostras tumorais de próstata, as regiões de PRUNE2 mais

próximas a terminação 5’ parecem estar super-expressas em relação as regiões mais próximas a terminação 3’ (Figura 12). Além disso, parece haver uma expressão preferencial da variante predita de PRUNE2 (Figura 12). Portanto, talvez a super-expressão de PCA3 em câncer de próstata possa estar influenciando na transcrição de PRUNE2, atrapalhando sua progressão em direção a terminação 3’, e então favorecendo variantes menores, compostas principalmente por éxons próximos a terminação 5’. Isso implica na perda do domínio BCH C-terminal de PRUNE2, o qual interage com proteínas Rho, resultando no potencial de inibir a proliferação celular (SOH e LOW 2008). Mais especificamente, a interação do domínio BCH ocorre com as proteínas RhoA, que participa da transformação celular e RhoC, que promove migração celular e metástase (WANG et al. 2003; RIDLEY 2004).

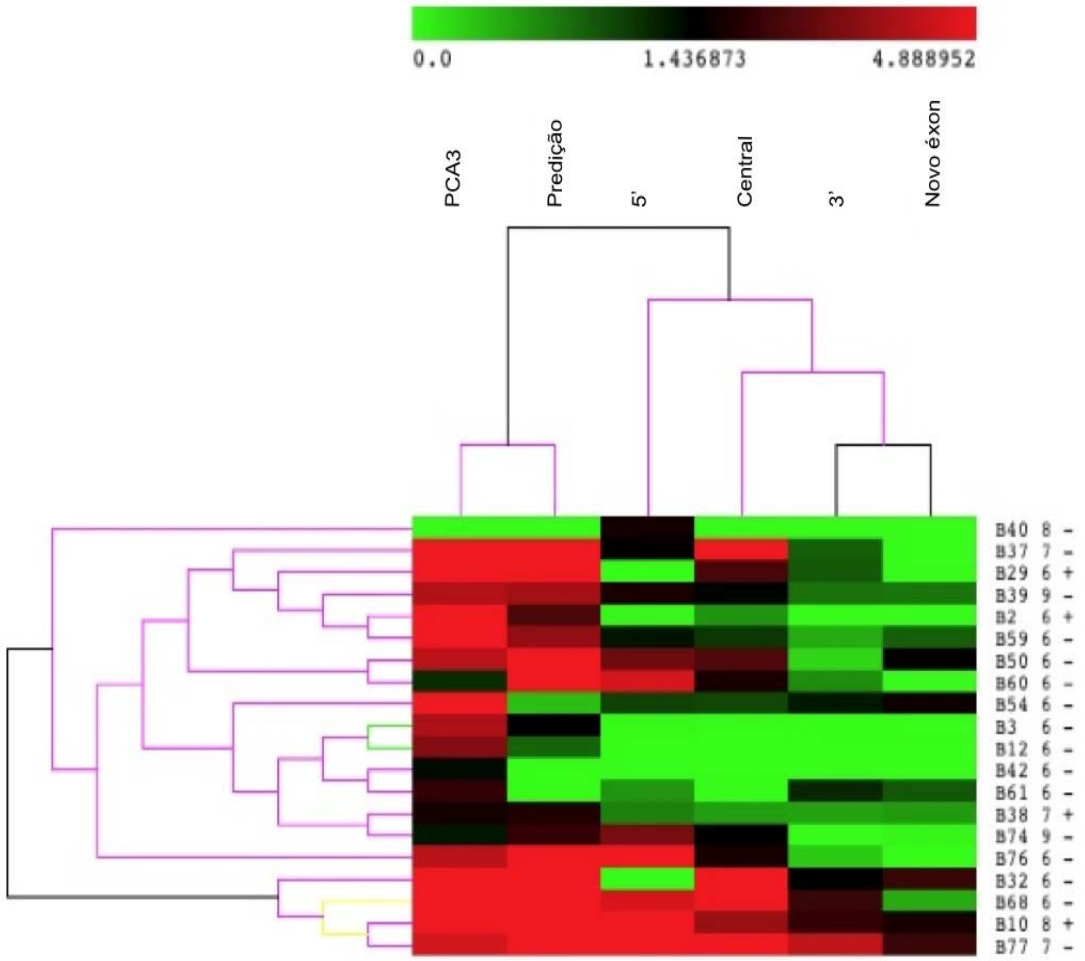


Legenda: Esses gráficos mostram uma diminuição da expressão das regiões de PRUNE2 mais próximas a terminação 3’ quando comparadas as mais próximas a terminação 5’, quando PCA3 esta expresso em altos níveis. ***Vertical-*** *fold*; ***horizontal-*** identificação dos iniciadores utilizados na avaliação da expressão de PCA3 e PRUNE2.

Figura 12 – Gráficos *box plot* mostrando a distribuição das amostras pareadas de próstata em relação a expressão diferencial de PCA3 e PRUNE2, de acordo com cada par de iniciadores utilizados nos experimentos de PCR em tempo real.

Nós também fizemos análises de agrupamento hierárquico não supervisionado, com o intuito de observar alguma correlação dos nossos dados de expressão com as características clínicas das amostras tumorais, entretanto só pudemos fazer esse tipo de análise observando o *Gleason Score* das amostras e a ocorrência de outros tumores , uma vez que apenas um paciente apresentou recidiva bioquímica evidenciada pelo aumento do PSA e todos os pacientes apresentaram estadiamento TNM próximos (T2 ou 3, apenas três N1 e todos M0) (Figura 13). O agrupamento hierárquico é uma abordagem aglomerativa baseada em uma matriz de distância par a par, a partir da qual as amostras mais semelhantes são reunidas duas a duas, sucessivamente, formando uma única árvore hierárquica, sendo que os braços terminais representam espécimes intimamente relacionados. Entretanto, a análise de agrupamentos, em que genes situados em um mesmo grupo precisam dividir algum tipo de elemento comum, não dá uma resposta absoluta, é somente uma técnica que permite a exploração das relações entre perfis de expressão gênica através de diversos espécimes (QUACKENBUSH 2001; PUSZTAI et al. 2003). Na análise não supervisionada a busca de relação entre sequências ou amostras é realizada sem informação adicional, ou seja, baseada apenas na expressão gênica dos dados obtidos no experimento (QUACKENBUSH 2001; RAMASWAMY e GOLUB 2002; PUSZTAI et al. 2003). O método não supervisionado apresenta menos tendências e é uma técnica melhor para revelar subtipos de tumores previamente irreconhecíveis dentro de um conjunto de cânceres morfologicamente similares, que podem ou não refletir diferenças biológicas ou relevância clínica, baseada em perfis de expressão gênica global (RAMASWAMY e GOLUB 2002; PUSZTAI et al. 2003). Nenhuma correlação entre os nossos dados de expressão e os dados clínicos das amostras foi

observada (Figura 13). Entretanto, nós observamos uma separação, em relação a expressão de PCA3 e das regiões de PRUNE2 estudadas, de dois grupos: um composto por PCA3 e as regiões mais próximas a terminação 5'do gene PRUNE2 (principalmente PCA3 e a variante predita de PRUNE2), super-expresso na maioria das amostras tumorais e o segundo, composto pelas regiões mais próximas a terminação 3'do gene PRUNE2, com uma tendência a apresentarem-se menos expressas em relação ao primeiro grupo. Além disso, a distribuição dessas regiões no agrupamento seguem a mesma ordem que a sua distribuição ao longo da sequências do gene PRUNE2 (5'-3') e de suas variantes (Figura 11).



Legenda: Não observamos nenhuma correlação entre os dados de expressão e o *Gleason Score* das amostras ou a ocorrência de outros tumores (***lateral-*** identificação das amostras, seus *Gleason Score* e ocorrência, + ou ausência, - de outros tumores). Entretanto, dois grupos foram separados, um composto por PCA3 e as regiões mais próximas a terminação 5'do gene PRUNE2 e outro composto pelas regiões mais próximas a terminação 3'do gene PRUNE2 (***superior-*** identificação dos iniciadores utilizados na avaliação da expressão de PCA3 e PRUNE2). ***Barra superior-*** equivalência dos valores de expressão em *fold* a representação de cores.

Figura 13 – Agrupamento hierárquico, utilizando distancia de correlação Euclidiana media dos dados de expressão de PCA3 e PRUNE2 em relação ao *Gleason Score* das amostras e a ocorrência de outros tumores.

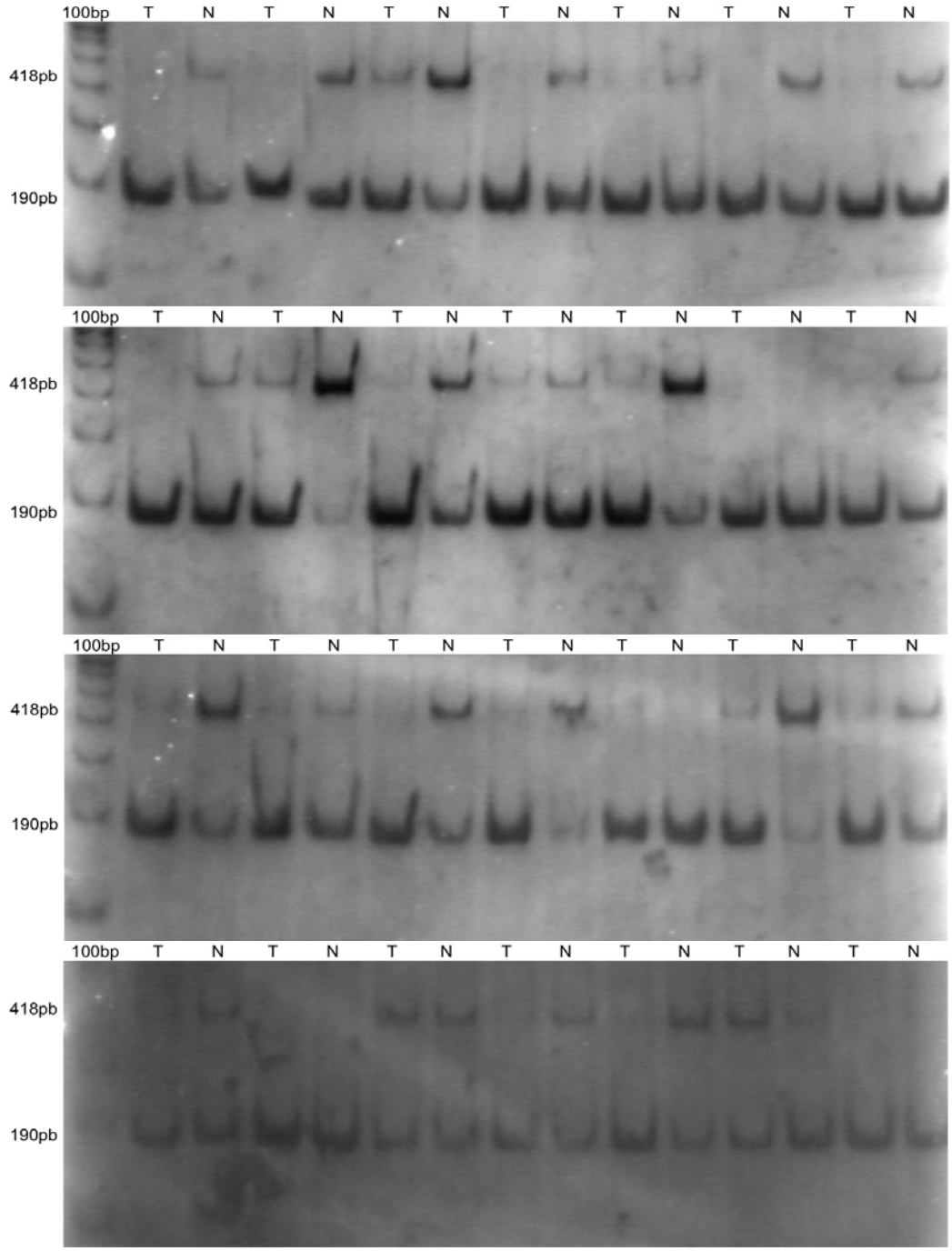
5.4.2 Retenção intrônica de PCA3

A sequências genômica de PCA3 foi inicialmente descrita sendo composta por quatro éxons, com o ultimo dividido em três partes (1, 2, 3, 4a, 4b e 4c) (BUSSEMAKERS et al. 1999). A variante mais frequente é composta pelos éxons 1, 3, 4a e 4b.

Durante os experimentos de PCR em tempo real citados acima, na etapa de padronização de um par de iniciadores para PCA3, desenhados nos éxons 3 e 4a, foi identificada uma retenção entre esses dois éxons não descrita na literatura. Essa retenção intrônica foi identificada no conjunto de linhagens celulares de próstata (PC-3, DU 145 e LNCaP), através de experimento de RT-PCR padrão, realizado de acordo com as instruções para o uso da enzima Platinum Taq DNA Polymerase (Invitrogen, US), com posterior visualização do produto em gel de poliacrilamida 8%, corado com prata. Essa retenção intrônica de PCA3 foi confirmada por sequênciamento do produto de PCR. O produto de PCR foi clonado utilizando o InsTAclon PCR Cloning Kit (Fermentas, CA) e as colonias resultantes foram utilizadas como molde em reações de PCR, como descrito anteriormente, assim como os passos seguintes (Materias e Métodos).

Quisemos então avaliar a presença dessa retenção nas amostras de tecidos de próstata. Para isso, repetimos o mesmo procedimento utilizado na identificação dessa retenção, em 28 amostras pareadas de câncer de próstata (provenientes do biobanco do Hospital A. C. Camargo) (Tabela 6). A retenção intrônica entre os éxons 3 e 4 de PCA3 foi confirmada em 36 de 56 amostras (evidenciada pela banda de 418pb) e foi observada ocorrendo preferencialmente em amostras normais de próstata (24 amostras normais *versus* 12 amostras tumorais) (Figura 14). Esses dados, apesar de

reais e interessantes, são preliminares, uma vez que optamos por fazer essa avaliação através de RT-PCR convencional. Essa opção foi feita porque os nossos iniciadores não eram aplicáveis a experimentos de PCR em tempo real, uma vez que detectariam a presença de ambos os transcritos (com e sem a retenção) sem distinção.



Legenda: Todas as 56 amostras avaliadas estão representadas nessa figura. Lado a lado estão amostras tumoral e normal do mesmo paciente, seguindo a ordem apresentada na Tabela 8. ***T*** amostras tumorais; ***N***- amostras normais; ***100bp***- 100bp DNA Ladder (Invitrogen, US); ***418pb***- tamanho esperado dos produtos na presença da retenção intrônica ; ***190pb***- tamanho esperado dos produtos na ausência da retenção intrônica.

Figura 14 – Avaliação da presença e frequência da retenção intrônica entre os éxons 3 e 4 de PCA3 através do fracionamento de produtos de PCR em gel de poliacrilamida 8%, corado com prata.

Portanto, com experimentos adicionais de PCR de baixa ciclagem ou quantificação em tempo real de um número maior de amostras normais e tumorais de próstata, talvez possamos descrever a retenção intrônica entre os éxons 3 e 4 de PCA3, possivelmente relacionada ao estado normal/tumoral de próstata, pela primeira vez. Possivelmente essa retenção intrônica nunca foi descrita na literatura pois a caracterização de PCA3 e a maioria dos trabalhos relacionados a esse ncRNA utilizaram a linhagem celular LNCaP, uma vez que somente esta e a linhagem 22Rv1 expressam PCA3 (BUSSEMAKERS et al. 1999; SCHALKEN et al. 2003). Um trabalho recente mostrou uma complexidade muito maior dos transcritos de PCA3 do que previamente descrito (CLARKE et al. 2009). Nesse trabalho foram identificados quatro novos sítios de início de transcrição, quatro sítios de poliadenilação e dois éxons com *splicing* diferencial em uma variante estendida de PCA3 (CLARKE et al. 2009). Esses novos transcritos de PCA3 são diferencialmente expressos em câncer de próstata e apresentam acurácias distintas na classificação de amostras normais e tumorais de próstata (CLARKE et al. 2009).

5.4.3 ORESTES AW793062

A ORESTES AW793062 apresentou uma super-expressão de 12 vezes em tumor de próstata comparado ao tecido normal, nos nossos dados de micrarranjos de cDNA e uma confirmação com valores de *fold* altos em quatro de sete amostras pareadas por PCR em tempo real, além de ter apresentado expressão diferencial significativa apenas em mais dois outros tecidos (7 vezes menos expressa em câncer de útero e 3 vezes mais expressa em câncer de cabeça e pescoço, quando comparados aos seus respectivos tecidos normais) nos nossos dados de microarranjos de cDNA.

O aparente fato dessa ORESTES ser um transcrito preferencialmente expresso em tumor de próstata, agregado com o interessante mapeamento dessa sequências no genoma, nos levaram a querer investigar melhor essa sequências.

Experimentos de PCR em tempo real foram realizados em 32 amostras pareadas de próstata adicionais de diferentes características clínicas e 6 amostras de NIP (provenientes do biobanco do Hospital A. C. Camargo) (Tabelas 8 e 9), com todas etapas metodológicas realizadas da mesma forma que descrito anteriormente (Materiais e Métodos), com as mesmas modificações realizadas na avaliação da possível regulação do gene PRUNE2 pelo ncRNA PCA3. Os iniciadores utilizados para avaliar a expressão da ORESTES AW793062 foram os mesmos utilizados na primeira validação. Também foram realizados experimentos de PCR em tempo real para o ncRNA PCA3, com os mesmos iniciadores utilizados na primeira validação, como controle positivo das nossa reações. Uma vez que o gene HPRT1 foi indicado apresentar a transcrição mais estável, tanto pela literatura (DE KOK et al. 2005) quanto pela utilização do programa geNorm (medgen.ugent.be/~jvdesomp/genorm/), na avaliação da possível regulação do gene PRUNE2 pelo ncRNA PCA3, este gene foi novamente selecionado para ser considerado como normalizador nos cálculos de expressão relativa.

Tabela 8 - Características clínicas das amostras tumorais e dos pacientes que compreendem as 32 amostras pareadas de próstata usadas nos experimentos de PCR em tempo real e resultados da validação da ORESTES AW793062 e de PCA3.

Amostra	TNM	*Gleason Score*	Recidiva bioquímica	Outros tumores	Expressão AW793062	Expressão PCA3
B81	T4NxM0	?	?	?	1.80	3.39
B82	T3bN0M0	(3+3) 6	sim	não	3.26	-0.06
B83	T3bN0M0	(3+3) 6	não	não	3.43	5.54
B89	T2bN0M0	(3+3) 6	não	não	0.88	4.26
B96	T3bNxM0	(3+3) 6	?	?	1.68	6.41
B98	T2bNxM0	(3+4) 7	?	?	0%*	0.94
B100	T2NxM0	(3+3) 6	?	?	4.15	6.62
B105	T2bNxM0	(3+3) 6	?	?	2.65	4.13
B107	T2bN0M0	(3+3) 6	não	não	-2.53	-0.24
B109	T2bN0M0	(3+3) 6	não	não	-1.23	3.66
B111	T2aN0M0	(3+3) 6	sim	não	0.04	2.71
B117	T2bN0M0	(3+3) 6	não	não	3.69	6.53
B122	T2aN0M0	(3+3) 6	não	não	1.52	1.40
B131	T2aN0M0	(3+3) 6	não	não	0.97	1.26
B137	T3aN0M0	(3+3) 6	não	não	-6.38	-2.55
B145	T3aN0M0	(3+3) 6	sim	carcinoma epidermóide	2.21	1.06
B148	T3aN0M0	(3+3) 6	não	carcinoma oral	10.22	13.41
B155	T2bN0M0	(3+3) 6	não	carcinoma de bexiga	3.69	4.31
B2	T2bN0M0	(3+3) 6	não	carcinoma basocelular de cavidade oral	0.02	5.80
B29	T2bN0M0	(3+3) 6	não	carcinoma basocelular/ de células escamosas	2.86	4.61
B38	T2bNxM0	(3+4) 7	?	adenocarcinoma tubular de estômago	-1.03	-1.16
B39	T2bN1M0	(4+5) 9	?	não	-0.57	1.88
B40	T2bN0M0	(4+4) 8	não	não	-3.17	3.12
B50	T2bN0M0	(3+3) 6	não	não	3.74	3.73
B59	T2bNxM0	(3+3) 6	?	não	2.84	5.89
B60	T2bN0M0	(3+3) 6	não	não	0.26	-0.41
B68	T3aN0M0	(3+3) 6	não	não	100%*	7.54
B74	T3bN1M0	(4+5) 9	não	não	6.66	0.45
B76	T2cN0M0	(3+3) 6	?	não	100%*	3.66
B139	?	(3+3) 6	?	não	-0.04	6.29037
B149	T2bN0M0	(3+3) 6	sim	sim	-0.83	-0.42
B151	T3aN0M0	(3+2) 5	sim	carcinoma gástrico	1.49	2.19

* Os valores representados por 0% e 100% indicam expressão somente nas amostras normais e tumorais, respectivamente (nenhum sinal detectado nas amostras de comparação).

** Todos os valores representam $\log_2$ dos valores de expressão, considerando tumor/normal.

Considerando uma diferença de expressão de pelo menos duas vezes entre amostras tumorais e normais, a ORESTES AW793062 apresentou-se super-expressa em 14 e sub-expressa em três amostras tumorais em relação aos seus pares normais (20 e um pares, respectivamente na quantificação de PCA3) (Tabela 8). Nas amostras de NIP, essa ORESTES apresentou uma alta super-expressão (em geral, maior que seis vezes) em cinco de seis amostras, enquanto PCA3 apresentou-se super expresso em cinco amostras e sub-expresso em uma (Tabela 9).

Tabela 9 - Resultados da validação da ORESTES AW793062 e de PCA3 utilizando amostras de NIP.

Amostra	Expressão AW793062	Expressão PCA3
B81	2,82	3,59
B88	6,01	2,80
B132	6,78	4,62
B138	6,89	6,74
B149	-1,16	4,32
B156	8,61	-6,45

** Todos os valores representam $\log_2$ dos valores de expressão, considerando NIP/amostra calibradora (controle positivo das reações de PCR em tempo real).

Com relação aos dados clínicos, ao contrário do esperado, todos os pacientes que apresentaram recidiva bioquímica (cinco pacientes) também apresentaram um *Gleason Score* baixo (5 ou 6). Dois desses casos apresentaram uma super-expressão da ORESTES mas não de PCA3, e em outros dois casos ocorreu o contrário, PCA3 apresentou-se super-expresso mas a ORESTES não. No quinto caso não houve expressão diferencial de nenhuma das duas sequências. Interessantemente, um paciente de cada caso (com super-expressão só da ORESTES ou só de PCA3)

tiveram também outros tumores, além de próstata. A recidiva bioquímica é considerada ocorrida quando um aumento persistente do PSA acima do nível detectado imediatamente após o tratamento é detectado (revisado por (BICKERS e AUKIM-HASTIE 2009), ocorrendo em cerca de 30% a 40% dos pacientes tratados com prostatectomia radical (AMLING 2006), indicando uma alta probabilidade de recorrência do tumor de próstata e progressão da doença e metástase (POUND et al. 1999). Na avaliação da efetividade de um novo marcador molecular, frequentemente é levada em consideração a recidiva bioquímica, como indicador de falha de tratamento (revisado por (BICKERS e AUKIM-HASTIE 2009). Dos pacientes com linfonodo positivo (dois), a ORESTES detectou um deles, enquanto PCA3 não detectou nenhum. Considerando os oito casos de ocorrência de outros tumores, a ORESTES apresentou-se super-expressa em quatro deles, sendo um não detectado por PCA3 (cinco e dois, respectivamente na quantificação de PCA3) (Tabela 8).

Esses dados indicam que talvez a ORESTES AW793062 possa ser utilizada como um novo marcador molecular, complementando os resultados de PCA3, uma vez que pelos dados mostrados aqui, apesar de não apresentarem significância estatística por causa do número amostral de cada subgrupo de amostras, parece haver uma tendência dessas duas sequências apresentarem-se diferencialmente expressas e em diferentes tipos de amostras com acurácias distintas.

A avaliação dos níveis de PSA continuará sendo uma ferramenta importante na prática clínica no câncer de próstata, entretanto com limitações pela baixa especificidade do PSA, enquanto eficientes novos marcadores não são descobertos. Alterações genéticas comuns, como hipermetilação de ilhas de CpG e fusões genicas, vem sendo identificadas em pacientes com câncer de próstata, bem como marcadores

em nível de RNA e proteína, entretanto ainda com uma sensibilidade e especificidade insatisfatórias para serem traduzidos para a rotina clínica. Futuramente, a maior probabilidade é de que seja utilizado na prática clínica um painel de novos marcadores associados àqueles já estabelecidos, como o PSA, melhorando a detecção e classificação do câncer de próstata, prognóstico e determinação do tratamento. Neste sentido, acreditamos que pudemos contribuir com este trabalho, sendo talvez a ORESTES AW793062 um potencial novo marcador a ser explorado num painel de marcadores de câncer de próstata.

6 CONCLUSÕES

- Identificamos como possíveis ncRNAs 28% das sequências analisadas (1.078 de 3.834 sequências).
- Mil e sete sequências foram identificadas como diferencialmente expressas entre os tecidos tumorais e normais analisados.
- Duzentos e noventa e um prováveis ncRNAs foram identificados como diferencialmente expressos entre os tecidos tumorais e normais analisados.
- Identificamos pelo menos cinco putativos marcadores tumorais não codificadores super-expressos em quatro ou mais tumores diferentes comparados com seus tecidos normais.
- Validamos três potenciais marcadores tumorais de próstata em amostras de tecidos normais e tumorais de próstata por PCR em tempo real.
- Observamos um possível papel do ncRNA PCA3 na regulação do gene no qual ele está mapeado, PRUNE2, na progressão do câncer de próstata.
- Identificamos a existência de uma retenção intrônica não descrita na sequência de PCA3, aparentemente mais frequente em amostras normais de próstata.
- Contribuímos com análises iniciais de um potencial novo marcador de câncer de próstata a ser explorado para a complementação de marcadores já existentes, como o PCA3 e o PSA.

Assim, com base nesses resultados, nós comprovamos o valor da nossa abordagem para identificar marcadores moleculares não-caracterizados, ilustrado pelo grande numero de sequências diferencialmente expressas entre amostras normais e tumorais de todos os tecidos estudados. Nossos resultados contém muito mais sequências de regiões ativamente transcritas do genoma humano, não mapeadas em éxons anotados, ainda não exploradas e, provavelmente representando novos genes, variantes de *splicing*, NATs ou ncRNAs, os quais poderão ser pesquisados como marcadores moleculares para outros cânceres.

7 REFERÊNCIAS BIBLIOGRÁFICAS

Adler HL, McCurdy MA, Kattan MW, et al. Elevated levels of circulating interleukin-6 and transforming growth factor-beta1 in patients with metastatic prostatic carcinoma. **J Urol** 1999; 161:182-7.

Amling CL. Biochemical recurrence after localized treatment. **Urol Clin North Am** 2006; 33:147-59.

Babak T, Blencowe BJ, Hughes TR. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. **BMC Genomics** 2005; 5:6-104.

Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling arrays. **Science** 2004; 306:2242-6.

Bickers B, Aukim-Hastie C. New molecular biomarkers for the prognosis and management of prostate cancer--the post PSA era. **Anticancer Res** 2009; 29:3289-98.

Bill-Axelson A, Holmberg L, Filen F, et al. Radical prostatectomy versus watchful waiting in localized prostate cancer: the Scandinavian prostate cancer group-4 randomized trial. **J Natl Cancer Inst** 2008; 100:1144-54.

Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. **Clin Pharmacol Ther** 2001; 69:89-95.

Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **Nature** 2007; 447:799-816.

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

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

Brito GC, Fachel AA, Vettore AL, et al. Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. **Mol Carcinog** 2008; 47:757-67.

Bussemakers MJ, van Bokhoven A, Verhaegh GW, et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. **Cancer Res** 1999; 59:5975-9.

Calin GA, Ferracin M, Cimmino A, et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. **N Engl J Med** 2005; 353:1793-801.

Calin GA, Croce CM. MicroRNA signatures in human cancers. **Nat Rev Cancer** 2006; 6:857-66.

Camargo AA, Samaia HP, Dias-Neto E, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci USA** 2001; 98:12103-8.

Camargo AA, de Souza SJ, Brentani RR, et al. Human gene discovery through experimental definition of transcribed regions of the human genome. **Curr Opin Chem Biol** 2002; 6:13-6.

Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. **Science** 2005; 309:1559-63.

Chan JM, Stampfer MJ, Giovannucci E, et al. Plasma insulin-like growth factor-I and prostate cancer risk: a prospective study. **Science** 1998; 279:563-6.

Charrier JP, Tournel C, Michel S, et al. Differential diagnosis of prostate cancer and benign prostate hyperplasia using two-dimensional electrophoresis. **Electrophoresis** 2001; 22:1861-6.

Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. **Science** 2005; 308:1149-54.

Clarke RA, Zhao Z, Guo AY, et al. New genomic structure for prostate cancer specific gene PCA3 within BMCC1: implications for prostate cancer detection and progression. **PLoS One** 2009; 4:e4995.

Clement JQ, Qian L, Kaplinsky N, et al. The stability and fate of a spliced intron from vertebrate cells. **RNA** 1999; 5:206-20.

Clement JQ, Maiti S, Wilkinson MF. Localization and stability of introns spliced from the Pem homeobox gene. **J Biol Chem** 2001; 276:16919-30.

Collins FS, Lander ES, Rogers J, et al. Finishing the euchromatic sequence of the human genome. **Nature** 2004; 431:931-45.

Cooper CS, Campbell C, Jhavar S. Mechanisms of Disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer. **Nat Clin Pract Urol** 2007; 4:677-87.

Darson MF, Pacelli A, Roche P, et al. Human glandular kallikrein 2 (hK2) expression in prostatic intraepithelial neoplasia and adenocarcinoma: a novel prostate cancer marker. **Urology** 1997; 49:857-62.

de Kok JB, Verhaegh GW, Roelofs RW, et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. **Cancer Res** 2002; 62:2695-8.

de Kok JB, Roelofs RW, Giesendorf BA, et al. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. **Lab Invest** 2005; 85:154-9.

de Souza SJ, Camargo AA, Briones MR, et al. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. **Proc Natl Acad Sci USA** 2000; 97:12690-3.

Demichelis F, Fall K, Perner S, et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. **Oncogene** 2007; 26:4596-9.

Deras IL, Aubin SM, Blase A, et al. PCA3: a molecular urine assay for predicting prostate biopsy outcome. **J Urol** 2008; 179:1587-92.

DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. **Nat Genet** 1996; 14:457-60.

Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. **Nature** 2001; 412:822-6.

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

Fleige S, Pfaffl MW. RNA integrity and the effect on the real-time qRT-PCR performance. **Mol Aspects Med** 2006; 27:126-39.

Florea L. Bioinformatics of alternative splicing and its regulation. **Brief Bioinform** 2006; 7:55-69.

Fonseca RS. **Avaliação das ORESTES NO MATCH geradas pelo Projeto Genoma Humano do Câncer (LICR/FAPESP-HCGP).** São Paulo; 2005. [Dissertação de Mestrado-Fundação Antônio Prudente].

Fonseca RD, Carraro DM, Brentani H. Mining ORESTES no-match database: can we still contribute to cancer transcriptome? **Genet Mol Res** 2006; 5:24-32.

Fradet Y, Saad F, Aprikian A, et al. uPM3, a new molecular urine test for the detection of prostate cancer. **Urology** 2004; 64:311-5.

Fradet Y. Biomarkers in prostate cancer diagnosis and prognosis: beyond prostate-specific antigen. **Curr Opin Urol** 2009; 19:243-6.

Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: Pitfalls and potential. **Biotechniques** 1999; 26:112-25.

Frith MC, Pheasant M, Mattick JS. The amazing complexity of the human transcriptome. **Eur J Hum Genet** 2005; 13:894-7.

Galante PAF, Vidal DO, de Souza JE, et al. Sense-antisense pairs in mammals: functional and evolutionary considerations. **Genome Biology** 2007; 8:R40.

Garzon R, Fabbri M, Cimmino A, et al. MicroRNA expression and function in cancer. **Trends Mol Med** 2006; 12:580-7.

Gaylis FD, Keer HN, Wilson MJ, et al. Plasminogen activators in human prostate cancer cell lines and tumors: correlation with the aggressive phenotype. **J Urol** 1989; 142:193-8.

Ginzinger DG. Gene quantification using real-time quantitative PCR: An emerging technology hits the mainstream. **Exp Hematol** 2002; 30:503-12.

Goessl C, Krause H, Muller M, et al. Fluorescent methylation-specific polymerase chain reaction for DNA-based detection of prostate cancer in bodily fluids. **Cancer Res** 2000; 60:5941-5.

Gonzalgo ML, Nakayama M, Lee SM, et al. Detection of GSTP1 methylation in prostatic secretions using combinatorial MSP analysis. **Urology** 2004; 63:414-8.

Goodrich JA, Kugel JF. Non-coding-RNA regulators of RNA polymerase II transcription. **Nat Rev Mol Cell Biol** 2006; 7:612-6.

Gopalkrishnan RV, Kang DC, Fisher PB. Molecular markers and determinants of prostate cancer metastasis. **J Cell Physiol** 2001; 189:245-56.

Griffiths-Jones S, Moxon S, Marshall M, et al. Rfam: annotating non-coding RNAs in complete genomes. **Nucleic Acids Res** 2005; 33:D121-4.

Griffiths-Jones S. Annotating noncoding RNA genes. **Annu Rev Genomics Hum Genet** 2007; 8:279-98.

Groskopf J, Aubin SM, Deras IL, et al. APTIMA PCA3 molecular urine test: development of a method to aid in the diagnosis of prostate cancer. **Clin Chem** 2006; 52:1089-95.

Gustincich S, Sandelin A, Plessy C, et al. The complexity of the mammalian transcriptome. **J Physiol** 2006; 575:321-32.

Hammond SM. MicroRNAs as oncogenes. **Curr Opin Genet Dev** 2006; 16:4-9.

Harman SM, Metter EJ, Blackman MR, et al. Serum levels of insulin-like growth factor I (IGF-I), IGF-II, IGF-binding protein-3, and prostate-specific antigen as predictors of clinical prostate cancer. **J Clin Endocrinol Metab** 2000; 85:4258-65.

Hessels D, Gunnewiek JMTK, van Oort I, et al. DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. **Eur Urol** 2003; 44:8-15.

Hienert G, Kirchheimer JC, Pfluger H, et al. Urokinase-type plasminogen activator as a marker for the formation of distant metastases in prostatic carcinomas. **J Urol** 1988; 140:1466-9.

Hillier L, Lennon G, Becker M, et al. Generation and analysis of 280,000 human expressed sequence tags. **Genome Res** 1996; 6:807-28.

Hoque MO, Topaloglu O, Begum S, et al. Quantitative methylation-specific polymerase chain reaction gene patterns in urine sediment distinguish prostate cancer patients from control subjects. **J Clin Oncol** 2005; 23:6569-75.

Huppi K, Volfovsky N, Mackiewicz M, et al. MicroRNAs and genomic instability. **Semin Cancer Biol** 2007; 17:65-73.

Iacoangeli A, Lin Y, Morley EJ, et al. BC200 RNA in invasive and preinvasive breast cancer. **Carcinogenesis** 2004; 25:2125-33.

Ilyin SE, Belkowski SM, Plata-Salaman CR. Biomarker discovery and validation: technologies and integrative approaches. **Trends Biotechnol** 2004; 22:411-6.

Jeronimo C, Usadel H, Henrique R, et al. Quantitation of GSTP1 methylation in non-neoplastic prostatic tissue and organ-confined prostate adenocarcinoma. **J Natl Cancer Inst** 2001; 93:1747-52.

Ji P, Diederichs S, Wang WB, et al. MALAT-1, a novel noncoding RNA, and thymosin beta 4 predict metastasis and survival in early-stage non-small cell lung cancer. **Oncogene** 2003; 22:8031-41.

Jiang Z, Woda BA, Rock KL, et al. P504S - A new molecular marker for the detection of prostate carcinoma. **Am J Surg Pathol** 2001; 25:1397-1404.

Jiang Z, Woda BA, Wu CL, et al. Discovery and clinical application of a novel prostate cancer marker - alpha-Methylacyl CoA racemase (P504S). **Am J Clin Pathol** 2004; 122:275-89.

Kampa D, Cheng J, Kapranov P, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. **Genome Res** 2004; 14:331-42.

Kan ZY, Rouchka EC, Gish WR, et al. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. **Genome Res** 2001; 11:889-900.

Kanety H, Madjar Y, Dagan Y, et al. Serum insulin-like growth factor-binding protein-2 (IGFBP-2) is increased and IGFBP-3 is decreased in patients with prostate cancer: correlation with serum prostate-specific antigen. **J Clin Endocrinol Metab** 1993; 77:229-33.

Keer HN, Gaylis FD, Kozlowski JM, et al. Heterogeneity in plasminogen activator (PA) levels in human prostate cancer cell lines: increased PA activity correlates with biologically aggressive behavior. **Prostate** 1991; 18:201-14.

Khan MA, Partin AW, Rittenhouse HG, et al. Evaluation of proprostate specific antigen for early detection of prostate cancer in men with a total prostate specific antigen range of 4.0 to 10.0 ng/ml. **J Urol** 2003; 170:723-6.

Kong L, Zhang Y, Ye ZQ, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. **Nucleic Acids Res** 2007; 35:W345-9.

Kurek R, Nunez G, Tselis N, et al. Prognostic value of combined "triple"-reverse transcription-PCR analysis for prostate-specific antigen, human kallikrein 2, and prostate-specific membrane antigen mRNA in peripheral blood and lymph nodes of prostate cancer patients. **Clin Cancer Res** 2004; 10:5808-14.

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

Lapointe J, Li C, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. **Proc Natl Acad Sci USA** 2004; 101:811-6.

LaTulippe E, Satagopan J, Smith A, et al. Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. **Cancer Res** 2002; 62:4499-506.

Laxman B, Tomlins SA, Mehra R, et al. Noninvasive detection of TMPRSS2:ERG fusion transcripts in the urine of men with prostate cancer. **Neoplasia** 2006; 8:885-8.

Lilja H, Ulmert D, Bjork T, et al. Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. **J Clin Oncol** 2007; 25:431-6.

Liu C, Bai B, Skogerbo G, et al. NONCODE: an integrated knowledge database of non-coding RNAs. **Nucleic Acids Res** 2005; 33:D112-5.

Liu W, Mao SY, Zhu WY. Impact of tiny miRNAs on cancers. **World J Gastroenterol** 2007; 13:497-502.

Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. **Nature** 2005; 435:834-8.

Machado-Lima A, del Portillo HA, Durham AM. Computational methods in noncoding RNA research. **J Math Biol** 2008; 56:15-49.

Mao M, Fu G, Wu JS, et al. Identification of genes expressed in human CD34(+) hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. **Proc Natl Acad Sci USA** 1998; 95:8175-80.

Marks LS, Fradet Y, Deras IL, et al. PCA3 molecular urine assay for prostate cancer in men undergoing repeat biopsy. **Urology** 2007; 69:532-5.

Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. **Bioessays** 2003; 25:930-9.

Mattick JS. RNA regulation: a new genetics? **Nat Rev Genet** 2004; 5:316-23.

Mattick JS, Makunin IV. Non-coding RNA. **Hum Mol Genet** 2006; 15:R17-29.

Maxam AM, Gilbert W. A new method for sequencing DNA. **Proc Natl Acad Sci USA** 1977; 74:560-4.

McCabe NP, Angwafo FF, 3rd, Zaher A, et al. Expression of soluble urokinase plasminogen activator receptor may be related to outcome in prostate cancer patients. **Oncol Rep** 2000; 7:879-82.

Mehler MF, Mattick JS. Non-coding RNAs in the nervous system. **J Physiol** 2006; 575:333-41.

Mello BP. **Identificação de novos transcritos humanos através da exploração racional do banco de dados do projeto genoma do câncer humano (HCGP).** São Paulo; 2007. [Dissertação de Mestrado-Fundação Antônio Prudente].

Mello BP, Abrantes EF, Torres CH, et al. No-match ORESTES explored as tumor markers. **Nucleic Acids Res** 2009; 37:2607-17.

Mendes Soares LM, Valcarcel J. The expanding transcriptome: the genome as the 'Book of Sand'. **EMBO J** 2006; 25:923-31.

Ministério da Saúde. Secretaria de Atenção à Saúde. Instituto Nacional de Câncer. **Estimativa 2008: incidência de câncer no Brasil**. Rio de Janeiro: INCA; 2007.

Mistry K, Cable G. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. **J Am Board Fam Pract** 2003; 16:95-101.

Miyake H, Hara I, Yamanaka K, et al. Elevation of serum levels of urokinase-type plasminogen activator and its receptor is associated with disease progression and prognosis in patients with prostate cancer. **Prostate** 1999; 39:123-9.

Nakashima J, Tachibana M, Horiguchi Y, et al. Serum interleukin 6 as a prognostic factor in patients with prostate cancer. **Clin Cancer Res** 2000; 6:2702-6.

Nakaya HI, Amaral PP, Louro R, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. **Genome Biol** 2007; 8:R43.

Numata K, Kanai A, Saito R, et al. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. **Genome Res** 2003; 13:1301-6.

Osada H, Takahashi T. MicroRNAs in biological processes and carcinogenesis. **Carcinogenesis** 2007; 28:2-12.

Pajares MJ, Ezponda T, Catena R, et al. Alternative splicing: an emerging topic in molecular and clinical oncology. **Lancet Oncol** 2007; 8:349-57.

Pang KC, Stephen S, Dinger ME, et al. RNAdb 2.0-an expanded database of mammalian non-coding RNAs. **Nucleic Acids Res** 2007; 35:D178-82.

Panzitt K, Tschernatsch MMO, Guelly C, et al. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. **Gastroenterology** 2007; 132:330-42.

Paul B, Dhir R, Landsittel D, et al. Detection of prostate cancer with a blood-based assay for early prostate cancer antigen. **Cancer Res** 2005; 65:4097-100.

Pedersen JS, Bejerano G, Siepel A, et al. Identification and classification of conserved RNA secondary structures in the human genome. **Plos Comput Biol** 2006; 2:251-62.

Penn SG, Rank DR, Hanzel DK, et al. Mining the human genome using microarrays of open reading frames. **Nat Genet** 2000; 26:315-8.

Petrovics G, Zhang W, Makarem M, et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. **Oncogene** 2004; 23:605-11.

Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. **Nucleic Acids Res** 2001; 29:e45.

Pollard KS, Salama SR, Lambert N, et al. An RNA gene expressed during cortical development evolved rapidly in humans. **Nature** 2006; 443:167-72.

Pound CR, Partin AW, Eisenberger MA, et al. Natural history of progression after PSA elevation following radical prostatectomy. **Jama** 1999; 281:1591-7.

Pryor MB, Schellhammer PF. The pursuit of prostate cancer in patients with a rising prostate-specific antigen and multiple negative transrectal ultrasound-guided prostate biopsies. **Clin Prostate Cancer** 2002; 1:172-6.

Pusztai L, Ayers M, Stec J, et al. Clinical application of cDNA microarrays in oncology. **Oncologist** 2003; 8:252-8.

Quackenbush J. Computational analysis of microarray data. **Nat Rev Genet** 2001; 2:418-27.

Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. **J Clin Oncol** 2002; 20:1932-41.

Ravasi T, Suzuki H, Pang KC, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. **Genome Res** 2006; 16:11-9.

Reis EM, Nakaya HI, Louro R, et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. **Oncogene** 2004; 23:6684-92.

Reis EM, Ojopi EPB, Alberto FL, et al. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. **Cancer Res** 2005; 65:1693-9.

Rhodes DR, Sanda MG, Otte AP, et al. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. **J Natl Cancer Inst** 2003; 95:661-8.

Ridley AJ. Rho proteins and cancer. **Breast Cancer Res Treat** 2004; 84:13-9.

Rogic S, Mackworth AK, Ouellette FBF. Evaluation of gene-finding programs on mammalian sequences. **Genome Res** 2001; 11:817-32.

Roupret M, Hupertan V, Yates DR, et al. Molecular detection of localized prostate cancer using quantitative methylation-specific PCR on urinary cells obtained following prostate massage. **Clin Cancer Res** 2007; 13:1720-5.

Rubie C, Kempf K, Hans J, et al. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. **Mol Cell Probes** 2005; 19:101-9.

Rymarquis LA, Kastenmayer JP, Huttenhofer AG, et al. Diamonds in the rough: mRNA-like non-coding RNAs. **Trends Plant Sci** 2008; 13:329-34.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci USA** 1977; 74:5463-7.

Schalken JA, Hessels D, Verhaegh G. New targets for therapy in prostate cancer: Differential display code 3 (DD3(PCA3)) a highly prostate cancer-specific gene. **Urology** 2003; 62:34-43.

Schena M, Shalon D, Davis RW, et al. Quantitative Monitoring of Gene-Expression Patterns with A Complementary-Dna Microarray. **Science** 1995; 270:467-70.

Schena M, Shalon D, Heller R, et al. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. **Proc Natl Acad Sci USA** 1996; 93:10614-9.

Scholzova E, Malik R, Sevcik J, et al. RNA regulation and cancer development. **Cancer Lett** 2007; 246:12-23.

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

Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. **Mol Cell** 2005; 19:1-13.

Shariat SF, Andrews B, Kattan MW, et al. Plasma levels of interleukin-6 and its soluble receptor are associated with prostate cancer progression and metastasis. **Urology** 2001; 58:1008-15.

Shariat SF, Lamb DJ, Kattan MW, et al. Association of preoperative plasma levels of insulin-like growth factor I and insulin-like growth factor binding proteins-2 and -3 with prostate cancer invasion, progression, and metastasis. **J Clin Oncol** 2002; 20:833-41.

Shariat SF, Canto EI, Kattan MW, et al. Beyond prostate-specific antigen: new serologic biomarkers for improved diagnosis and management of prostate cancer. **Rev Urol** 2004; 6:58-72.

Shariat SF, Roehrborn CG, McConnell JD, et al. Association of the circulating levels of the urokinase system of plasminogen activation with the presence of prostate cancer and invasion, progression, and metastasis. **J Clin Oncol** 2007; 25:349-55.

Shariat SF, Karam JA, Margulis V, et al. New blood-based biomarkers for the diagnosis, staging and prognosis of prostate cancer. **BJU Int** 2008; 101:675-83.

Shoemaker DD, Schadt EE, Armour CD, et al. Experimental annotation of the human genome using microarray technology. **Nature** 2001; 409:922-7.

Soh UJ, Low BC. BNIP2 extra long inhibits RhoA and cellular transformation by Lbc RhoGEF via its BCH domain. **J Cell Sci** 2008; 121:1739-49.

Srebrow A, Kornblihtt AR. The connection between splicing and cancer. **J Cell Sci** 2006; 119:2635-41.

Srikantan V, Zou ZQ, Petrovics G, et al. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. **Proc Natl Acad Sci USA** 2000; 97:12216-21.

Stamey TA, Yang N, Hay AR, et al. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. **N Engl J Med** 1987; 317:909-16.

Stangegaard M, Dufva IH, Dufva M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. **Biotechniques** 2006; 40:649-57.

Steiner MS, Barrack ER. Transforming growth factor-beta 1 overproduction in prostate cancer: effects on growth in vivo and in vitro. **Mol Endocrinol** 1992; 6:15-25.

Sterky F, Lundeberg J. Sequence analysis of genes and genomes. **J Biotechnol** 2000; 76: 1-31.

Sun H, Skogerbo G, Chen RS. Conserved distances between vertebrate highly conserved elements. **Hum Mol Genet** 2006; 15:2911-22.

Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. **N Engl J Med** 2004; 350:2239-46.

Thompson IM, Ankerst DP, Chi C, et al. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. **Jama** 2005; 294:66-70.

Tinzl M, Marberger M, Horvath S, et al. DD3(PCA3) RNA analysis in urine - A new perspective for detecting prostate cancer. **Eur Urol** 2004; 46:182-7.

Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. **Science** 2005; 310:644-8.

Truong LD, Kadmon D, McCune BK, et al. Association of transforming growth factor-beta 1 with prostate cancer: an immunohistochemical study. **Hum Pathol** 1993; 24:4-9.

Twillie DA, Eisenberger MA, Carducci MA, et al. Interleukin-6: a candidate mediator of human prostate cancer morbidity. **Urology** 1995; 45:542-9.

van Gils MP, Hessels D, van Hooij O, et al. The time-resolved fluorescence-based PCA3 test on urinary sediments after digital rectal examination; a Dutch multicenter validation of the diagnostic performance. **Clin Cancer Res** 2007; 13:939-43.

Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. **Genome Biol** 2002; 3(7):RESEARCH0034.

Varambally S, Dhanasekaran SM, Zhou M, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. **Nature** 2002; 419: 24-9.

Varambally S, Yu J, Laxman B, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. **Cancer Cell** 2005; 8:393-406.

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

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

Wang L, Yang L, Luo Y, et al. A novel strategy for specifically down-regulating individual Rho GTPase activity in tumor cells. **J Biol Chem** 2003; 278:44617-25.

Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. **Proc Natl Acad Sci USA** 2005; 102:2454-9.

Washietl S, Pedersen JS, Korbel JO, et al. Structured RNAs in the ENCODE selected regions of the human genome. **Genome Res** 2007; 17:852-64.

Weile C, Gardner PP, Hedegaard MM, et al. Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes. **BMC Genomics** 2007; 8:244.

Wright JL, Lange PH. Newer potential biomarkers in prostate cancer. **Rev Urol** 2007; 9:207-13.

Xu JC, Stolk JA, Zhang XQ, et al. Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. **Cancer Res** 2000; 60:1677-82.

Yang IV, Chen E, Hasseman JP, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. **Genome Biol** 2002; 3(11):research0062.

Yang YH, Speed T. Design issues for cDNA microarray experiments. **Nat Rev Genet** 2002; 3:579-88.

Yousef GM, Diamandis EP. The new human tissue kallikrein gene family: structure, function, and association to disease. **Endocr Rev** 2001; 22:184-204.

Yu YP, Landsittel D, Jing L, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. **J Clin Oncol** 2004; 22:2790-9.

ANEXOS

Anexo 1 - Putativos marcadores moleculares não codificadores preditos pelas nossas análises.

Número de acesso	Tecido de expressão	Valores de expressão em *fold* (microarranjos de cDNA, tumor/normal)	Valores de intensidade (microarranjo de cDNA)
Seqüências não-exônicas, ORF-			
AW804665	medula óssea	2,90317	7,76673
AW797359	cólon	7,30477	13,67830
AW806217	mama	-2,24939	7,80961
AW809330	cólon	2,10881	9,87428
AW809378	mama	3,50007	10,08850
AW811238	mama	-2,60652	8,57584
AW811637	cólon	-2,44167	8,66035
AW812194	mama	-2,88070	8,13919
AW814379	útero	-2,11757	8,93306
AW814971	cabeça e pescoço	-2,81128	8,08834
AW817453	mama	4,20702	10,66020
AW835371	cólon	2,15018	8,56450
AW837360	cólon	-2,68826	8,07966
AW840674	cólon	2,46204	10,32620
AW845902	cólon	2,06276	10,74920
AW846011	mama	2,09669	9,10348
AW850216	próstata	2,18544	7,20612
AW850237	cólon	-2,02737	7,85084
AW850314	cólon	2,18142	9,37860
AW852423	cabeça e pescoço	-2,00903	7,91732
AW857472	mama	-2,05880	8,11624
AW858632	mama	2,04235	8,69758
AW859951	próstata	2,52202	8,57999
AW859955	cólon	2,16814	9,31282
AW860065	mama	2,45762	9,80616
AW861131	mama	-2,12648	9,14478
AW869047	mama	-2,04606	8,38269
AW879469	próstata	2,03529	9,42615
AW880994	mama	-2,52629	8,04103
AW885528	cólon	-2,19554	8,35132
AW888846	medula óssea	-2,01870	7,71754
AW889573	cólon	-2,39469	7,40342
AW894787	mama	2,01302	9,57799
AW896066	medula óssea	2,13106	7,97765
AW904473	cólon	-2,32732	7,63810
AW935934	medula óssea	-2,93809	7,89279
AW983872	cólon	2,40333	8,25078
AW993403	cabeça e pescoço	-2,02180	9,06354
AW996886	mama	2,38358	9,32697

AW996969	próstata	2,64669	8,77114
AW997279	cabeça e pescoço	-2,30520	9,12091
BE002422	cólon	-4,00675	7,66038
BE003186	cólon	-11,58220	8,35069
BE009835	cólon	5,22760	9,70547
BE061138	cabeça e pescoço	2,06887	9,61832
BE062646	cólon	2,01204	8,34832
BE063379	cólon	2,87298	9,52144
BE066432	mama	-2,79931	7,74204
BE066884	cabeça e pescoço	3,07172	8,43594
BE067707	cabeça e pescoço	-2,58771	8,98790
BE069408	cólon	2,22804	9,55046
BE070154	cabeça e pescoço	-3,79416	10,90090
BE071253	cólon	-3,05364	11,55310
BE072933	medula óssea	2,07614	7,82148
BE082747	mama	-2,74479	8,12146
BE095195	útero	-3,53796	8,15201
BE145046	mama	2,00710	8,25881
BE146595	esôfago	-5,39439	8,56753
BE147100	mama	2,31321	10,53930
BE148254	mama	2,31500	9,12867
BE148859	cabeça e pescoço	3,18272	8,98425
BE152488	mama	3,79156	9,89168
BE153028	cólon	3,82248	10,30390
BE155036	cólon	-2,17511	8,06006
BE170535	cólon	-2,69758	7,97129
BE181922	mama	-2,23839	8,62184
BF325688	cólon	-2,04823	8,05443
BF328142	cólon	-2,23664	8,91518
BF328285	cólon	-2,30428	8,87875
BF355813	cólon	2,57464	9,93625
BF355905	mama	-2,05493	8,78166
BF357651	esôfago	2,06025	8,87977
BF358517	mama	-28,43530	8,02444
BF362002	cólon	-2,43823	6,99114
BF365582	útero	-2,74074	8,21042
BF366002	cólon	-2,08759	8,47242
BF367794	cólon	-2,58923	7,67616
BF367838	útero	-16,22330	7,90775
BF368582	mama	-2,17974	8,22808
BF370376	cabeça e pescoço	3,00326	8,35659
BF370980	medula óssea	2,73534	8,04814

BF373821	útero	-3,80464	7,96629
BF374290	útero	-2,85271	8,04006
BF375447	cólon	-10,20610	7,62052
BF378954	tireoide	2,24916	8,22036
BF749440	cabeça e pescoço	2,40422	9,40504
BF768459	útero	-2,05374	7,78626
BF768558	estômago	2,02665	9,35269
BF803450	mama	-2,55340	8,16051
BF811956	cólon	3,39597	10,04190
BF812324	útero	2,07849	8,38850
BF813229	mama	2,99715	10,23060
BF816786	útero	-2,57227	8,44274
BF817568	cólon	3,26998	9,84772
BF818064	cólon	2,65032	9,29112
BF821521	mama	4,22155	14,59400
BF829136	cólon	-2,70739	9,56735
BF870545	cólon	-2,04903	8,15852
BF874212	cólon	2,78815	9,90236
BF874715	mama	-2,88329	9,29549
BF875190	cólon	3,10890	10,34440
BF877960	esôfago	2,84690	9,43047
BF879265	mama	-2,03210	8,63252
BF881688	cólon	2,66173	9,58884
BF885107	esôfago	8,57749	9,25560
BF892717	estômago	-4,21636	8,60783
BF894361	cólon	3,52482	9,55759
BF896657	mama	-4,62751	8,08434
BF904458	mama	2,08298	9,10899
BF907086	esôfago	-7,90259	8,05623
BF907379	medula óssea	2,54012	9,00855
BF911476	mama	-2,09129	8,29773
BF930910	cólon	3,74251	10,37890
BF935710	cólon	-2,93656	12,31380
BF943152	útero	-2,12417	7,59723
BF943480	cólon	-4,56911	8,30871
BF945767	mama	2,47729	11,07340
BF949071	cabeça e pescoço	-3,77176	8,77444
BF949465	cólon	2,74753	9,50059
BF951294	mama	2,89739	10,36180
BF951965	esôfago	3,23608	8,50098
BF957343	cabeça e pescoço	-2,00394	7,56713
BF959540	cólon	2,58312	9,20225
BF959617	útero	-2,57178	8,88090
BF962798	cólon	-3,63555	10,35740
BF962902	cólon	-2,98768	10,63530

BF963970	cólon	-3,79326	8,21711
BF986630	mama	2,58077	8,49667
BF996341	cabeça e pescoço	2,15514	8,53797
BF998252	cabeça e pescoço	2,91057	8,62033
BG000268	cólon	2,59750	10,24340
BG000273	cólon	3,92460	12,97800
BG001374	cólon	-2,06572	8,03803
BG001557	cólon	2,26918	9,32465
BG002507	estômago	5,78036	9,79865
BG004492	cólon	2,69155	13,09120
BG005515	cólon	2,74268	9,90683
BG005979	cabeça e pescoço	2,18192	8,68702
BG006748	cólon	-2,50483	8,18315
BG010297	cólon	-2,48824	7,33089
BG011405	cólon	2,13753	8,68324
BI002033	útero	-2,09668	8,26602
BI006701	cólon	-3,38104	7,74204
BI013540	cólon	-5,36947	8,23774
BI014374	cabeça e pescoço	-2,00087	8,68229
BI014621	cólon	2,50342	8,90000
BI051761	mama	2,27444	9,36252
BI054285	mama	-2,47539	7,78622
BI058423	cólon	-2,23836	7,79150
BQ312823	cólon	2,04302	10,04780
BQ325928	cólon	-2,27611	10,06260
BQ352054	próstata	-2,17215	7,91001
BQ370498	mama	-2,32328	8,27897
BQ376265	cólon	2,52629	11,04210
CK327003	cólon	-4,68750	11,20570
CV311407	cólon	2,68384	9,57314
CV313796	medula óssea	2,23095	8,15062
CV314051	cólon	-2,27902	8,00080
CV320622	cólon	-3,14805	8,88152
CV321350	cólon	2,59172	9,35034
CV338191	útero	-2,64658	7,83053
CV340653	cólon	2,95648	9,36272
CV342108	cabeça e pescoço	-2,01024	8,75695
CV344090	cólon	-2,11444	9,84612
CV344927	cabeça e pescoço	-2,16813	9,31265
CV345289	cólon	4,03325	8,51081
CV346264	cólon	2,22834	9,04562
CV346299	mama	2,90477	9,11856
CV346777	útero	-2,29546	8,21316

CV348400	cólon	2,72468	9,81164
CV351750	mama	-2,27584	7,90425
CV355210	cólon	2,72028	12,96490
CV355528	mama	-3,30378	8,23103
CV355633	cólon	-3,86793	7,93206
CV355657	mama	-12,28790	7,83259
CV356875	cólon	-2,31708	8,11554
CV369138	cabeça e pescoço	-2,29971	7,29611
CV375182	mama	2,15302	8,87036
CV375954	cólon	3,23673	10,11010
CV384859	mama	-2,33088	9,02145
CV393957	cabeça e pescoço	-2,43967	8,98528
CV394705	cabeça e pescoço	-2,36624	9,04144
CV394843	útero	2,08504	8,50281
CV396139	mama	2,29745	9,66274
CV398755	próstata	-4,14208	8,18968
CV403110	mama	-26,75650	6,65948
CV403923	cólon	2,95030	9,71988
CV404842	mama	-2,01188	7,75596
CV405108	mama	-2,10215	8,50520
CV405788	mama	-3,91407	8,27050
CV406673	cólon	2,83836	9,75551
CV411074	mama	-5,46763	8,36718
CV411420	cólon	2,07476	9,22946
CV411911	cabeça e pescoço	-2,02486	8,65734
CV414536	mama	-2,07367	8,07448
AW803985	cólon	7,61604	11,95960
	cabeça e pescoço	2,21776	14,58990
AW816582	mama	2,63288	9,31265
	cólon	2,98126	9,49782
AW817082	mama	2,13192	9,13684
	cólon	2,59989	9,39543
AW821033	mama	2,17014	9,36546
	cólon	2,87912	9,50041
AW834818	mama	-2,00716	8,12752
	cólon	-2,07550	7,96852
AW839072	mama	4,69714	10,11550
	cólon	10,42060	10,34270
AW850516	esôfago	2,25440	8,14651
	cabeça e pescoço	4,11477	7,99714
AW850948	medula óssea	-2,09607	7,32891
	próstata	-2,64627	8,02781
AW854429	cólon	14,10790	12,60670

	útero	3,94930	12,56920
AW886320	mama	2,05833	9,51624
	cólon	2,63281	9,69151
AW994530	cabeça e pescoço	2,83753	8,24688
	próstata	2,21853	8,12319
BE066976	mama	2,42580	9,46338
	cólon	2,86165	9,55341
BE074306	mama	6,01388	9,26855
	cólon	5,73789	9,91277
BE082405	mama	2,37887	9,46214
	cólon	2,49480	9,67894
BE145328	mama	-3,21122	8,21208
	cólon	-2,03698	7,84500
BE155922	mama	2,03846	11,22390
	cólon	13,27990	11,67190
BE718698	mama	3,49327	9,06405
	medula óssea	2,02861	8,63699
BF330234	mama	-2,79097	7,33697
	cólon	-3,39465	7,02549
BF335344	mama	2,73615	11,32950
	cólon	4,36221	11,60350
BF335366	esôfago	2,46318	9,27411
	estômago	2,09579	9,38457
BF335758	mama	3,01922	9,28149
	cólon	3,34439	9,40803
BF355290	esôfago	-3,01516	7,36847
	útero	-4,87587	7,90763
BF353228	esôfago	-2,29276	7,76034
	útero	-4,32662	7,83724
BF356213	cólon	-4,18810	11,70210
	cabeça e pescoço	-2,09040	11,21460
BF360278	mama	-2,38949	8,04880
	cólon	-2,22350	7,92417
BF734931	cabeça e pescoço	3,67655	9,08163
	próstata	2,16871	9,41891
BF757688	próstata	3,62969	9,90896
	útero	2,52312	9,07538
BF768515	mama	7,40933	10,16630
	cólon	9,14189	10,37750
BF801844	mama	3,15417	9,38597
	cólon	2,37731	9,32812
BF804886	cólon	6,58585	12,65580
	útero	2,56515	12,49170
BF819462	mama	3,19640	9,96080
	cólon	3,50542	10,19170

BF826265	mama	2,30694	9,67469
	cólon	2,06996	9,59390
BF874259	mama	2,71849	9,50184
	cólon	2,48263	9,58143
BF874760	mama	2,47433	9,07094
	cólon	3,77286	9,89168
BF875228	mama	-2,23940	7,67722
	cólon	-2,01618	8,11054
BF915357	mama	3,12674	9,74230
	cólon	4,03262	10,07700
BF920074	esôfago	2,33035	8,91821
	cabeça e pescoço	2,66463	9,09324
BF957966	cólon	-3,78793	10,77990
	estômago	-2,03017	10,54960
BF960441	mama	2,49913	9,41215
	medula óssea	2,26718	9,83348
BF999400	cólon	-2,95804	8,89014
	útero	-5,07159	8,57199
BG003529	mama	-2,56313	8,52123
	cólon	-2,53391	7,98152
BI001426	mama	2,08148	9,89817
	cólon	4,20081	9,75302
BI003899	mama	2,75357	9,09932
	cólon	2,31909	9,21300
BI013426	cólon	2,17485	9,92642
	estômago	2,49478	10,84100
BI045936	mama	2,06247	9,39265
	cólon	2,30019	9,45739
BI046981	cabeça e pescoço	2,01329	13,67480
	útero	2,16121	13,82640
CV315237	mama	-2,54637	8,52854
	próstata	-2,04010	8,54660
CV383909	mama	3,18341	9,59186
	cólon	3,33978	9,58571
CV391597	cólon	-2,28731	10,09380
	cabeça e pescoço	-2,21409	10,42340
CV392131	mama	2,81509	9,37939
	cabeça e pescoço	2,97009	9,94166
CV393615	mama	3,09618	11,50050
	cólon	3,49214	11,58830
CV396141	mama	2,26125	10,85000
	cólon	2,22582	11,15180
CV408606	mama	-2,18959	8,23447
	cólon	-2,63579	7,75888

CV412615	cabeça e pescoço	2,88946	8,79908
	útero	2,01749	9,00169
CV426134	mama	2,37362	9,60829
	cólon	2,75796	9,88894
AW176131	mama	2,47614	13,03690
	cólon	7,40803	13,23760
	útero	3,57530	13,76100
AW815921	mama	2,06613	11,60950
	cabeça e pescoço	2,79590	12,62350
	útero	2,46039	13,21060
AW848211	mama	4,74779	14,88330
	cólon	23,89840	15,10800
	cabeça e pescoço	2,11747	14,66790
BE069036	mama	4,12180	13,74300
	cólon	3,36233	14,23140
	próstata	2,09104	14,46150
BE093626	mama	18,03900	10,86750
	cólon	9,99020	10,78970
	cabeça e pescoço	2,38731	11,99440
BF375442	mama	27,44650	11,46120
	cólon	8,86377	11,65410
	medula óssea	2,92623	11,45730
BF840620	mama	3,47525	10,17400
	cólon	4,90882	10,38070
	medula óssea	2,38838	9,94352
BF852175	mama	2,33876	12,87310
	cólon	17,43330	13,03530
	útero	3,99099	13,23730
BF886211	mama	2,69528	9,15323
	cólon	3,43367	9,29322
	cabeça e pescoço	2,08873	10,24550
BF994093	mama	3,05087	9,77680
	cólon	3,30113	9,82907
	estômago	2,11253	10,04540
BG955873	mama	4,51801	9,49377
	cólon	5,89408	9,87572
	esôfago	2,16968	10,37660
CV372409	mama	-2,07718	10,01920
	cólon	-3,41332	10,73210
	útero	-2,74228	10,64760
CV390692	mama	2,28964	8,94648
	cólon	2,04370	8,84126
	próstata	2,37683	10,17500

CV374743	cólon	4,15102	9,53550
	esôfago	2,33032	9,35078
	medula óssea	2,08319	8,52881
AW803984	mama	5,35497	9,90961
	cólon	4,00022	9,95226
	cabeça e pescoço	2,03404	11,27370
	medula óssea	4,36160	9,53433
BE161676	mama	7,83474	8,93278
	esôfago	2,94504	8,71852
	cabeça e pescoço	2,51798	9,02069
	útero	2,63766	8,92570
CV358552	cólon	3,54354	11,76040
	estômago	2,26725	11,76110
	tireoide	2,07998	12,60750
	útero	2,03145	12,48880
AW814925	mama	2,51213	10,80680
	cólon	2,27676	11,33920
	esôfago	2,06340	11,51560
	medula óssea	2,04974	10,75790
	útero	2,12023	11,74170
Seqüências não exônicas, ORF+			
AW813395	mama	-2,29381	7,67291
AW817253	cólon	-2,00000	8,18926
AW834561	cólon	2,00000	10,90560
BE063341	tireoide	-2,03599	8,17816
BE064251	mama	-2,91713	7,80926
BF358650	cólon	-2,15919	7,52851
BF911779	mama	3,39827	9,71802
BF951247	mama	-2,67028	8,61779
CV387019	estômago	4,03627	9,07360
BI018419	mama	2,98892	9,78112
	cólon	5,49105	9,74972
Seqüências parcialmente exônicas, ORF-			
BE002737	mama	-2,33435	7,85334
BF357203	cólon	-2,30712	7,89941
BF360657	esôfago	-2,60842	8,17201
BF753917	cólon	3,52002	13,56700
BF929797	cabeça e pescoço	-3,49347	8,20672
CV397827	mama	-2,25366	8,33592
BF800826	cabeça e pescoço	2,54315	8,59221
	próstata	2,70735	8,75489
BF986350	cólon	2,65634	12,19510
	útero	2,59000	12,35680
CV370122	mama	-2,63125	8,58991

	cólon	-2,42184	8,27866
AW814005	mama	2,83808	9,58279
	cólon	3,20794	9,70781
	esôfago	2,19079	9,11511
BG002503	mama	3,65550	9,92818
	cólon	5,22184	9,84428
	útero	2,68816	10,66640
CV344043	mama	2,02907	8,82670
	cabeça e pescoço	2,18400	9,16125
	estômago	2,19398	9,66693
AW935941	mama	2,37640	8,24269
	cólon	3,10382	8,63851
	tireoide	2,09253	9,45889
	útero	2,12653	9,84027

Published online 6 March 2009

Nucleic Acids Research, 2009, Vol. 37, No. 8 ***2607–2617***
doi:10.1093/nar/gkp074

No-match ORESTES explored as tumor markers

Barbara P. Mello[1], Eduardo F. Abrantes[1], César H. Torres[1], Ariane Machado-Lima[2], Rogério da Silva Fonseca[1], Dirce M. Carraro[1], Ricardo R. Brentani[1], Luiz F. L. Reis[1] and Helena Brentani[1,*]

[1]Hospital A. C. Camargo, Rua Prof. Antônio Prudente 211, São Paulo, SP, 01509-900 and [2]IME/IPq-USP - Rua do Matão, 1010, São Paulo, SP, 05508-090, Brazil

Received October 2, 2008; Revised December 23, 2008; Accepted January 27, 2009

ABSTRACT

Sequencing technologies and new bioinformatics tools have led to the complete sequencing of various genomes. However, information regarding the human transcriptome and its annotation is yet to be completed. The Human Cancer Genome Project, using ORESTES (open reading frame EST sequences) methodology, contributed to this objective by generating data from about 1.2 million expressed sequence tags. Approximately 30% of these sequences did not align to ESTs in the public databases and were considered no-match ORESTES. On the basis that a set of these ESTs could represent new transcripts, we constructed a cDNA microarray. This platform was used to hybridize against 12 different normal or tumor tissues. We identified 3421 transcribed regions not associated with annotated transcripts, representing 83.3% of the platform. The total number of differentially expressed sequences was 1007. Also, 28% of analyzed sequences could represent noncoding RNAs. Our data reinforces the knowledge of the human genome being pervasively transcribed, and point out molecular marker candidates for different cancers. To reinforce our data, we confirmed, by real-time PCR, the differential expression of three out of eight potentially tumor markers in prostate tissues. Lists of 1007 differentially expressed sequences, and the 291 potentially noncoding tumor markers were provided.

INTRODUCTION

Understanding the genetic basis of human development and the mechanisms implicated in the physiopathology of diseases has improved dramatically after the disclosure of the human genome sequence, and its encoded genes (1–3). It is now widely accepted that, in mammals, there is no linear correlation between the number of genes, transcripts, and functionally diverse proteins. In the human transcriptome, a myriad of controlling mechanisms involving alternative splicing and a diversity of 5′ and 3′ ends contribute to, a yet unknown universe of transcripts (4). It is known that most of the genome is transcribed in complex patterns of interacting and overlapping transcripts from both strands (5–9), and most mammalian genes also have antisense transcripts (7,9–11). We currently have a great deal of information (4,5,12–14) arising from modern technologies, such as tiling arrays, that confirm the genome to be pervasively transcribed, and that the noncoding regions, such as the introns and intergenic regions, play an important role in human genome regulation by *cis*-acting at the transcriptional level (4,15,16). These approaches have resulted in the discovery of many novel transcribed sequences, and provide a new perspective on the number and extent of transcripts.

Noncoding RNAs (ncRNAs) are emerging as key players in transcriptional and translational control, and represent a new level of complexity (17,18). Available data shows that the ratio of noncoding versus coding RNAs increases from prokaryotes to mammals (6,19). Furthermore, ncRNAs appear to have cell- or condition-restricted expression, and at lower levels compared with the well-characterized coding genes (20–22). In addition, although cross-species conservation of many ncRNA transcribed regions is weak, promoters of these transcripts are generally much more evolutionarily conserved, and the conserved regions extend further than in the promoters of protein coding RNAs (5 kb versus 500 bp) (5,22,23). In recent years, the use of bioinformatics tools allied to experimental studies, particularly for the whole genome, has become a common and promising means to predict and screen novel ncRNAs and antisense RNAs (10,14,22,24,25).

Although sequencing efforts based on generating cDNA fragments had a major impact on gene discovery, the unspliced human transcripts that map exclusively to introns, and with no similarity to known expressed genes from any organism, were not fully appreciated.

*To whom correspondence should be addressed. Tel: +55-11-2189-5000-1134; Fax: +55-11-2189-5163; Email: helena@lbhc.hcancer.org.br.

Most investigators selected transcripts with evidence of splicing, or ESTs only where both a polyadenylation signal and a poly(A) tail were present (18). It is now accepted that only a small fraction of the sequences generated through EST methods represent mitochondrial transcripts, reverse transcribed copies of rRNA, bacterial contaminants or immature mRNA molecules (26,27). Large fractions of what were, until recently, considered 'junk' DNA are indeed transcribed, and may play a fundamental role in understanding genomes (5,15,28). In addition, the results presented by Ravasi *et al*. (29) show that most of the cloned, noncoding sequences in the RIKEN cDNA collection, are expressed and are not, on the whole, derived from genomic, or pre-mRNA (premature mRNA), contamination.

A large contribution toward identifying ESTs was the Human Cancer Genome Project (HCGP) (3,26,27,30), performed by the ORESTES (open reading frame EST sequences) methodology. ORESTES is a technique to generate ESTs encompassing midpoints of genes, unlike conventional EST methodologies (5′ and 3′) that cover the ends of transcripts. This characteristic results from the cDNA synthesis using arbitrarily selected, nondegenerate primers under low-stringency conditions, that permits sequence analysis of less abundant gene transcripts, and therefore, lead us to access genes with lower levels of expression (26). Thus, the HCGP, through ORESTES methodology, generated 1 190 044 open reading frame EST sequences using RNA extracts from 24 types of normal or tumor tissues (3,27). From this total, almost 30% (341 680 sequences) showed no similarity with known transcripts and were considered no-match ORESTES (27). With the aim to explore the potential of ORESTES with no similarities with ESTs in the public databases as tumor markers, we constructed a cDNA microarray. This platform, containing ORESTES with a high probability of representing actively transcribed regions not associated with annotated transcripts, was hybridized against 12 different normal and tumor human tissues. The differential expression observed among distinct tissues or pathological conditions demonstrates that this strategy was very useful for identifying tissue-specific, or tumor-specific RNAs that do not correspond to previously annotated transcripts. These hitherto-uncharacterized transcripts may represent new human genes, splice variants, ncRNAs or natural antisense transcripts (NATs) with a restricted pattern of gene expression. As prostate tumor is the most prevalent cancer in the Brazilian male population (http://www.inca.gov.br), we have explored some of these sequences as potential prostate tumor markers.

MATERIALS AND METHODS

Selection of ORESTES and genome mapping

To construct the array, 4356 ORESTES with higher probability to represent actively transcribed regions of the human genome not associated with annotated transcripts, were randomly selected from the data generated by Fonseca *et al*. (31), resulted from the exploration of the 341 680 ORESTES from the Human Cancer Genome Project that showed no similarity to known transcripts (27). In this work, a bioinformatics pipeline was constructed for the sequences mapped on the human genome that were annotated as no-match in the Human Cancer Genome Project, starting with the removal of sequences derived from libraries containing genomic DNA or immature mRNA contamination, according to Sorek & Safer, 2003 (32), followed by selection of clusters containing at least one no-match sequence derived from prostate or breast tissues and that were formed by ESTs originating from at least two distinct libraries, and the singletons that showed gaps upon genomic alignment. Also, clusters aligned with full-length transcripts or ESTs of other projects were removed.

Genome mapping was done through a local database composed of data downloaded from the UCSC Genome Bioinformatics database (http://genome.ucsc.edu). ORESTES were classified according to their mapping on the human genome using three different gene tracks (Ensembl, KnownGene and RefSeq), and sequences mapped once on the genome were further classified as exonic, intronic and intergenic sequences.

cDNA microarrays

Glass arrays with 4356 elements were prepared in our lab with the aid of the Flexys Robot (Genomic Solutions, Ann Arbor, MI, USA), as described by Brentani *et al*., 2005 (33). Microarray data are deposited at Gene Expression Omnibus (GEO) under accession number GSE12737. Detailed information is provided in Supplementary Data.

RNA extraction and amplification

The institutional research ethics committee approved the current study (REC number 970/07), which was performed in accordance with the principles expressed in the Declaration of Helsinki. All samples kept in the A.C. Camargo Hospital BioBank, have signed informed consent for use in research, provided and approved by patients.

Total RNA derived from 56 normal or tumor tissues, obtained from the A.C. Camargo Hospital BioBank, was extracted with TRIzol (Invitrogen, Carlsbad, CA, USA) (Supplementary Data, Table S1). As a reference, we used a pool of RNAs obtained from 15 distinct human cell lines (Table S2). RNA samples were linearly amplified using a T7-based protocol (34,35). cDNA was prepared with aminoallyl-dUTP (Sigma–Aldrich, St. Louis, MO, USA) (36). Detailed information is provided in Supplementary Data.

Labeling, hybridization and data extraction

cDNA samples were submitted to indirect labeling (36) using Alexa Fluor 555 or Alexa Fluor 647 labels (Invitrogen). Hybridizations were performed in duplicate using the dye-swap method (35,37) in the GeneTAC Hybridization Station (Genomic Solutions). Slides were scanned on a confocal laser scanner (ScannArray Express, PerkinElmer, Waltham, MA, USA), using identical parameters for all slides and data was extracted with ScanArray Express software (PerkinElmer).

The histogram method was used to estimate signal and local background intensities. Detailed information is provided in Supplementary Data.

Selection of *bonafide* transcripts

After subtracting local background, data was normalized by Lowess (38). For each sample, we determined the correlation between replica hybridizations and the number of spots with signal greater than local background. We also determined, for each sample, the differences between average signal intensity for elements representing intergenic or intragenic (exonic and intronic) sequences and for exonic or intronic sequences. To define a sequence as expressed, and to minimize the risk of a false-positive call, we applied a second level of cutoff for low-intensity spots. First, we determined, for each element, the lowest background-corrected intensity value among the 112 reads (main and swap slides) in each channel. Then, for each channel, we considered, as threshold, the highest value among the 112 lowest reads in each slide. Next, we eliminated, for each channel, all elements with median intensity below this threshold. Elements that survived these criteria were considered *bonafide* transcripts. For all expression data we applied $\log_2$ to the values.

Prediction of structured ncRNA candidates

Genomic sequences corresponding to ORESTES were analyzed to predict structured ncRNAs candidates. First, we separated the sequences into three groups: fully exonic, partially exonic, and nonexonic, according to the annotation systems KnownGene and RefSeq (UCSC Genome Bioinformatics). For each group, we combined searches for three features: (i) putative ORF, (ii) coding/noncoding potential of sequences and (iii) sequence and secondary structure conservation. To determine if a sequence is entirely an ORF, we used the getorf program (EMBOSS program suite, http://www.ebi.ac.uk/Tools/emboss), which analyzes if the three reading frames of both strands of the sequence could generate a coding sequence, and checked if the longest ORF identified by this software corresponded to the whole sequence (or its trimmed version of up to 2 bases from each end). Also, we used the Coding Potential Calculator (CPC) software (39), with default parameters, which classifies sequences in coding and noncoding (weak-coding, coding, weak noncoding and noncoding), to refine our initial ORF prediction. This software takes into account six features, being three of them based on the predicted ORF extension, quality and integrity, and the other three derived from BLASTX searches (UniRef90, BLAST Assembled Genomes; http://blast.ncbi.nlm.nih.gov/Blast.cgi): the number, quality and frame of the hits. We grouped sequences classified as noncoding or weak noncoding and sequences classified as coding and weak coding. To detect sequence and secondary structure conservation, we searched for multispecies alignments (16 vertebrate genomes with human http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way) that overlapped the ORESTES sequence locations. These alignments were analyzed using the RNAz software (40) with default parameters, to detect evidence of secondary structure conservation, like compensatory base substitution.

Validation by RT-PCR

To select sequences for validation by RT-PCR, we first determined the average intensity value for each element in all slides. Using an MA plot (intensity ratios versus average intensities), we randomly selected elements with intensity 20-fold higher than the background (cutoff value of $\log_2 12$ for A, average intensities), since we intended to validate highly expressed sequences. Primers for 12 selected sequences were designed using Primer3 software (http://frodo.wi.mit.edu) (Table S3). RNAs from 23 normal or tumor tissues were obtained from the A.C. Camargo Hospital BioBank (Table S1), extracted with TRIzol (Invitrogen) and DNase treated (Illustra RNAspin Mini Isolation Kit, GE Healthcare, Buckinghamshire, ENG, UK). RT-PCR reactions were carried on Gene Amp PCR System 9700 (Applied Biosystems, Foster City, CA, USA) and the amplicons were fractionated by electrophoresis thorough a 3% NuSieve GTG (Cambrex, East Rutherford, NJ, USA) and stained with ethidium bromide. Detailed information is provided in Supplementary Data.

Differential expression analysis

To select differentially expressed sequences to be considered as tumor marker candidates we constructed MA plots showing, for each spot, fold differences and median signal intensity for tumor versus normal tissues. For these analyses, three (placenta, lung and testis) out of 12 tissues that were used in cDNA microarray experiments were discarded because we had only normal samples from them, and therefore, we could not perform differential expression analyses with the aim to identify tumor markers for these tissues.

Validation by quantitative real-time PCR

To select sequences to validate by real-time PCR, we determined, for each element, fold differences between median signal intensity for: (i) prostate tumor versus normal prostate tissue and (ii) prostate tumor versus all normal tissues analyzed on cDNA microarray experiments. Using MA plots, we selected elements expressed at least 4-fold more or 4-fold less in prostate tumor relative to normal prostate, and at least 2-fold more or 2-fold less in prostate tumor relative to all normal tissues (values converted to $\log_2$). Primers were constructed for nine sequences differentially expressed in prostate tissue, using Primer Express software (Applied Biosystems) and Oligo Tech program (http://www.oligosetc.com/analysis.php) (Table S5). Real-time PCR reactions were optimized using a pool of RNAs from three tumor prostate cell lines (PC-3, DU 145 and LNCaP), provided by the São Paulo branch of the Ludwig Institute for Cancer Research, and cultivated by the Laboratório de Investigação Médica/24 from Universidade de São Paulo. Real-time PCR validation was performed in seven paired samples from prostate (prostate adenocarcinoma and its surrounding non-neoplasic tissue), obtained

from the A.C. Camargo Hospital BioBank (Table S6), extracted with TRIzol (Invitrogen) and DNase treated (RQ1 RNase-Free DNase, Promega, Madison, WI, USA). Real-time PCR experiments were carried out in duplicate using the SYBR Green detection method (Applied Biosystems). The housekeeping gene HPRT was selected through literature review (41). We used a previously described molecular marker for prostate carcinoma (AMACR) (42) as positive control for real-time PCR reactions. Real-time PCR was performed on a 7900HT Fast Real-Time PCR System (Applied Biosystems). The relative expression ratio was calculated according to Pfaffl formula (43). For all expression data we applied $\log_2$ to the values. Detailed information is provided in Supplementary Data.

Sequencing of validated ORESTES

ORESTES validated as real-transcripts by RT-PCR had their PCR products sequenced to verify their correspondence to the immobilized sequences and differentially expressed ORESTES validated by real-time PCR had their original clones sequenced to verify their correspondence to the sequences with which we expected that they were. Sequencing was carried on the 3130 Genetic Analyzer (Applied Biosystems). Detailed information is provided in Supplementary Data.

RESULTS

Genomic mapping of the cDNA microarray sequences

An analysis comparing the genomic location of ORESTES and non-ORESTES ESTs, with respect to coordinates of coding genes, was performed. As expected, we found that both non-ORESTES ESTs, as well as ORESTES, were preferentially mapped in transcribed regions of the human genome, using three different gene tracks (RefSeq, Ensembl and KnownGene, UCSC Genome Bioinformatics; http://genome.ucsc.edu) (Figure 1A). The proportion of ORESTES sequences that overlapped annotated exons of coding genes was somewhat reduced in the ORESTES data set (Figure 1B). The preferential mapping of ORESTES to transcriptional units suggests that fully intronic ORESTES may represent valid transcripts instead of genomic DNA contamination of ORESTES libraries.

We constructed a cDNA microarray containing 4356 distinct ORESTES, selected using a previously described pipeline developed to maximize the probability of identifying new expressed sequences (31). Our data showed that most ORESTES that compounded the array was mapped to transcribed regions of the genome (Figure 1A), and had a fully intronic location (Figure 1B). Only a small fraction of spotted sequence overlapped annotated exons of coding genes or had intergenic mapping (Figure 1). For further analysis, we considered 3872 sequences that map once to the human genome. We divided these sequences into exonic (335 sequences), intronic (3178 sequences) and intergenic sequences (359 sequences), representing 8.6%, 82.1% and 9.3% of the sequences respectively. A large proportion of these ORESTES (3767) are unspliced relative to the genome.

Analysis and identification of actively transcribed regions not associated with annotated transcripts, and their evaluation as potential ncRNAs

Many low expression transcripts, splicing isoforms and ncRNAs are involved in specialized biological functions, and show a tissue-specific or even a pathological-specific expression patterns. To survey new transcripts associated with ORESTES, 24 tumor and 32 normal RNA samples from 12 different tissues (Table S1) were hybridized with the microarray platform.

Some preliminary analyses were performed to determine the overall quality of data. The Pearson correlation between two replicate slides showed a median value of 0.86, and 76% of the elements that compounded this platform had signal greater than local background. We investigated if there was any bias that could be associated with the different types of sequences immobilized on the array, according to the previous classification: exonic, intronic or intergenic sequences. Using the Wilcoxon test, there were no statistically significant differences in either case, i.e. in the comparison of average signal intensity for elements representing intergenic or intragenic sequences, as well as in the comparison of only intragenic (exonic or intronic) sequences. The spotted sequences showed no systematic bias associated with their classification, corroborating

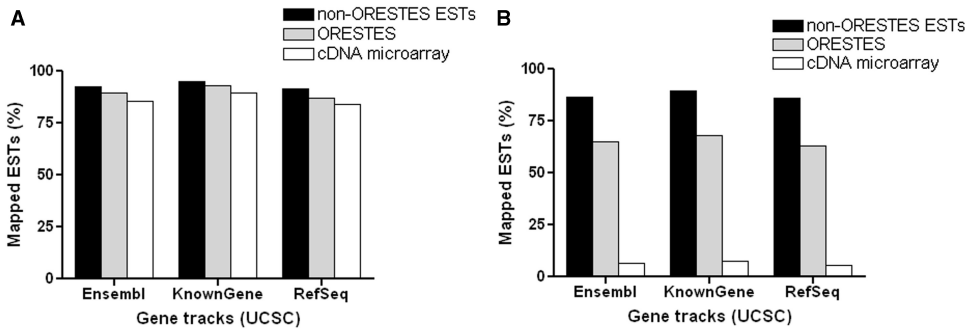


Figure 1. Mapping of ESTs on the human genome according to three different data sets. (**A**) ESTs mapped onto human transcript regions. (**B**) ESTs mapped onto human exonic regions. Black bar, ESTs; gray bar, ORESTES (open reading frame expressed sequence tags); and white bar, ORESTES that compound the cDNA microarray.

Table 1. Putative noncoding RNAs and their distribution with respect to differential expression

	Partially exonic sequences					Nonexonic sequences						
	ORF+	ORF–				ORF+		ORF–				
Number of putative ncRNAs	0	38				58		982				
Number of tissue types where putative ncRNAs were differentially expressed	0	1	2	3	4	1	2	1	2	3	4	5
Number of upregulated putative ncRNAs	0	1	2	3	1	3	1	92	40	13	3	1
Number of downregulated putative ncRNAs	0	5	1	0	0	6	0	103	15	1	0	0
Number of putative noncoding tumor markers	0	13				10		268				

ncRNAs (noncoding RNAs).

the likelihood of those sequences mapped on nonexonic regions as being transcribed sequences. To be more accurate in defining true hybridization signals we created a more stringent criterion, described in 'Materials and methods' section, with signal intensity cutoff values of 196 and 65 for channels 1 and 2, respectively. Thus, for each channel, we eliminated all elements with a median signal intensity below these thresholds. For channels 1 and 2 we had 86.6% and 91.1% of slides with more than 3000 valid elements, respectively. Therefore, the total number of actively transcribed regions not associated with annotated transcripts was 3421 (3079 out of 3178 intronic and 342 out of 359 intergenic sequences). The additional number of 319 out of 335 exonic elements identified as valid elements, corroborated the potential of our approach to identify new real, transcribed regions, since these sequences were deposited by others in public databases while this work was being performed. From this final number of valid elements (3740), 96 sequences (80 intronic, 6 intergenic and 10 exonic sequences) had intensity above our established cutoff value (20-fold higher than the background) and were eligible for RT-PCR validation. From this 96 sequences, we arbitrarily selected nine intronic sequences (roughly 11% of the total of intronic sequences), and three intergenic sequences (50% of intergenic group) and validated the existence of all of them as actively transcribed regions not associated with annotated transcripts, in RNAs derived from 10 different tissues (Tables S1 and S3, Figure S1). PCR products of validated sequences were submitted to sequencing and their correspondence to the immobilized sequences on the array was confirmed.

Evidence of secondary structures coupled with some sequence conservation at the RNA level can provide important clues that a given 'locus' is probably transcribed, and that this transcript may have a biological role (14,40,44,45). RNA secondary structures are known to play an important functional role, not only in many noncoding transcripts, but also in the context of protein-coding mRNAs (46). To analyze the proportion of spotted sequences that may represent structurally conserved putative ncRNAs, we searched for three features: (i) putative ORF, (ii) coding/noncoding potential and (iii) sequence and secondary structure conservation. For this analysis, sequences that did not overlap to known exons (intronic and intergenic sequences) were grouped together (3537 sequences) and the exonic sequences were further classified to fully exonic (131) and partially exonic (166). As for this analysis we only considered the KnownGene and RefSeq gene tracks to classify analyzed sequences, we discarded 38 sequences, previously classified as exonic according to the initial mapping, using the RefSeq, Ensembl and KnownGene gene tracks (UCSC Genome Bioinformatics) (Figure S2). We considered as putative ncRNAs sequences which presented all following features: partially exonic or nonexonic mapping, CPC software prediction of noncoding potential and evidence of secondary structure conservation according to the RNAz software. From the partially exonic sequences, we found 38 putative ncRNAs and from the nonexonic sequences, we found 1040 ncRNAs candidates (Table 1, Figure S2). It is noteworthy that some known ncRNAs possess a subsequence that is not as short as is usual, and resembles an ORF (46). In summary, about 28% (1078 of 3834) of our transcribed regions, not associated with annotated transcripts, are potential ncRNAs (Table 1, Figure S2).

Differential expression analyses and validation by quantitative real-time PCR

We constructed MA plots (intensity ratios versus average intensities) showing, for each spot, fold differences and median signal intensity for tumor versus normal tissues, for all the different tissues used in the cDNA microarray (Figure 2). We observed in all tissues, a large number of differentially expressed (at least 2-fold) sequences between tumor and normal samples, suggesting the potential to explore uncharacterized molecular markers (about 28% of the intronic and intergenic sequences mapped once on the genome). The total number of differentially expressed sequences, with fold differences between tumor and normal samples of at least two, in one or more different tissues and in agreement in respect to these sequences being up- or downregulated in all tissues in which they were expressed, were 1007, being 111 out of 335 exonic sequences, 885 out of 3178 intronic sequences and 111 out of 359 intergenic sequences (a list of all 1007 differentially expressed sequences is provided in our website, http://www.lbhc.hcancer.org.br/orestes_tumor_markers).

Considering the same criteria of differentially expressed sequences described above, 291 transcripts were classified as differentially expressed putative ncRNAs by our pipeline. Four percent of these putative noncoding tumor markers were in the NONCODE database (47), or were

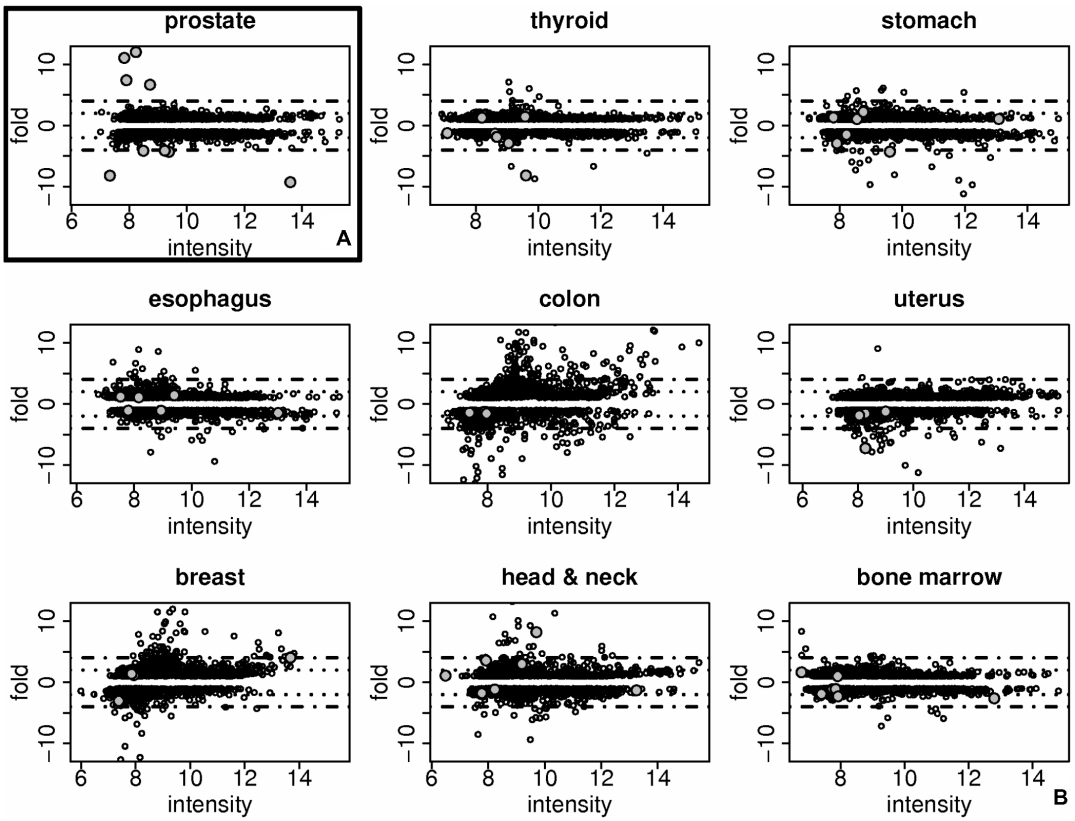


Figure 2. MA plot (intensity ratios versus average intensities) showing the fold differences and median signal intensity for tumor versus normal tissues for each spot on microarray. (**A**) Prostate tissue. (**B**) Other tissues used on cDNA microarray. Gray circles, the sequences from prostate selected for real-time PCR validation (with fold value in prostate tumor 4-fold more or 4-fold less relative to normal prostate and 2-fold more or 2-fold less, relative to all normal tissues). Dotted line, 2-fold line; dashed line, 4-fold line.

predicted as an antisense pair by Galante *et al.* (48), again corroborating the validity of our approach, but have never been identified as differentially expressed in tumors. In Table 1 we assessed whether these candidates were expressed in one or more tumor tissues and found that at least five putative noncoding tumor markers were upregulated in at least four different tumors (AW803984, BE161676, CV358552, AW814925 and AW935941), compared with normal tissues. This is a very promising result for the search for tumor markers. A list of all putative noncoding tumor markers is provided in Table S4.

We constructed MA plots to present an overview of the sequence expression distribution in prostate tissue (Figure 2A). For each spot, we observed the fold differences and median signal intensity for prostate tumor versus normal prostate (Figure 2A), and for prostate tumor versus all normal tissues (Figure S3). The nine sequences from prostate selected for validation by real-time PCR (Table S5) had at least a 4-fold variation in prostate tumor relative to normal prostate, and had at least 2-fold variation in prostate tumor relative to all normal tissues. We observed that, in general, the selected sequences were differentially expressed only in prostate when compared with other tissues (Figure 2, gray circles).

Using real-time PCR, we validated eight of the nine sequences as real transcripts. We considered valid differentially expressed sequences as those that presented a 3-fold difference in at least three out of seven paired samples. Using this criterion, three sequences were considered to be potential prostate tumor markers (Table 2). One of the potential tumor markers (BQ373258) was previously described as a ncRNA ($DD3^{PCA3}$) by Bussemakers *et al.* (49). Its differential expression was confirmed in five of our seven paired prostate samples, and was upregulated in prostate cancer, serving as a positive control for our real-time PCR experiments. The overexpression of AW793062 ORESTES in prostate tumor was confirmed in four paired tissues. Genome mapping of this sequence showed its alignment to the first intron of a putative isoform of the RNF217 gene. The sequence BF910617 was validated in three samples and showed overexpression in prostate cancer. It is an intronic sequence of the KIAA1432 gene. Considering our criteria of valid differentially expressed transcripts (3-fold difference in at least three out of seven paired samples), we validated the overexpression of the AMACR gene. This molecular marker for prostate carcinoma was previously described as having high sensitivity and specificity for prostate carcinoma from different grades and types, being its mRNA overexpressed in about 30% (microarray) to 60% (real-time PCR) of prostate tumors and is low to undetectable in normal tissues (42,50,51).

A summary of all sequences and samples sets used in each performed assay, as well as obtained results, is provided in Supplementary Data (Table S7).

Table 2. Results of quantitative real-time PCR validating paired prostate samples with cDNA microarray results

Accession number	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6	Pair 7	Real-time PCR fold mean	cDNA microarray fold
AMACR	–1.16	–0.33	5.51	1.52	3.13	–0.40	6.50	2.11	–
BQ373258	–0.95	–0.42	100%[a]	100%[a]	100%[a]	4.62	5.09	5.50	7.41
CV398755	–	–	–	–	–	–	–	–	–4.14
CV374350	0.03	–0.97	–0.01	0.12	–1.49	0.97	–0.62	–0.28	–4.32
AW849290	0.18	–0.20	100%[a]	3.89	1.75	0.21	0.56	1.05	6.67
BE144456	0.71	–0.73	0.86	2.60	–0.36	0.56	1.31	0.70	–8.21
AW793062	0.21	–0.10	100%[a]	100%[a]	100%[a]	2.15	100%[a]	6.03	12.02
BF910617	0.54	–0.74	–0.25	4.56	100%[a]	0.86	3.05	2.57	11.06
CV400462	–0.76	0.05	–0.11	0.85	0.01	–2.67	–0.85	–0.50	–4.11
BF365844	0.14	–0.92	2.95	2.84	1.80	0.30	–0.88	0.89	–9.28

[a]100% values represent expression only in tumor samples (no detectable signal in normal samples) and were converted 10-fold to calculate fold mean. All values represent $\log_2$ of expression values, considering tumor/normal ratios.

DISCUSSION

Since a significant set of ORESTES remains unassociated with annotated transcripts, and could potentially represent actively transcribed regions of the human genome, we constructed a cDNA microarray containing ORESTES with a high probability of representing actively transcribed regions of the human genome, and not associated with annotated transcripts. Most of the sequences immobilized on the array map on intronic regions and are unspliced. After hybridization using 12 different tissues, we identified 3421 actively transcribed regions not associated with annotated transcripts. With RT-PCR we validated 100% of actively transcribed regions not associated with annotated transcripts that were evaluated (12 sequences).

Based on an ORF detector program (getorf, http://www.ebi.ac.uk/Tools/emboss), only 9% of the sequences mapped once on the genome may represent coding genes, leading us to search for potential noncoding sequences. In spite of the ORESTES methodology being biased to cover transcript midpoints with high probability of representing open reading frames, our data showed that from the sequences mapped to intronic or intergenic location (nonexonic group) only 7.6% presented a putative ORF. In contrast, 47.3% of fully exonic sequences had a putative ORF (Figure S2).

Our next step was to look for sequences that could be tumor, tissue or tumor/tissue associated. We observed in all tissues, a large number of differentially expressed (at least 2-fold) sequences between tumor and normal samples, suggesting the potential to explore uncharacterized molecular markers (about 28% of the intronic and intergenic sequences mapped once on the genome). The total number of differentially expressed sequences, with fold differences between tumor and normal samples of at least two and in agreement in respect to these sequences being up- or downregulated in all tissues in which they were expressed, in one or more different tissues, were 1007. We investigated the number of intronic ORESTES that mapped in a cancer gene list, compounded by 382 genes for which mutations have been causally implicated in cancer. This catalog of cancer genes is available on the Sanger Institute (Cancer Gene Census, http://www.sanger.ac.uk/genetics/CGP/Census) and it is based on a previously published review (52). We found 189 intronic ORESTES mapped to 97 cancer genes. The number of the differentially expressed ORESTES, considering the same criteria described above, located within introns of these cancer genes were 47. Using a list of cellular signal pathways curated by NCI-Nature (http://pid.nci.nih.gov), we expanded the original list of cancer genes for 1003 cancer pathway related genes. We found that 287 ORESTES mapped to 170 cancer-pathway related genes. From these 287 ORESTES related to cancer pathways, 70 were differentially expressed, considering the same criteria described above.

De novo computational prediction of ncRNA genes is difficult, since these transcripts lack most of the signatures that make protein-coding gene prediction possible (45). However, ncRNA genes produce a functional RNA rather than a translated protein, and often display a conserved, base-paired secondary structure instead of primary sequence similarity. These features can be combined in analyses and result in profiles of a multiple sequence alignment of ncRNAs that can be captured by statistical models (14,53). There are several approaches that are used to successfully predict ncRNAs based on the idea that functionally significant RNA structures will be conserved in related species, even when primary sequence is not conserved (54). The secondary structure base pairings are maintained by compensatory base mutations. These changes can be used as statistical evidence of evolutionary pressure to keep the base pairs at those positions (14,40,44,45). Pedersen *et al*. (44) predicted, from an initial set of more than 48 000 structured regions, ~10 000 structured RNA transcripts in the human genome. Washietl *et al*. (40) estimated that 35 000 structured RNAs are conserved in mammals. The annotation of ncRNAs on a genome-wide scale is currently restricted to searching for homologs of known RNA families. More than 1500 homologs of known classical RNA genes can be annotated in the human genome sequence, and automatic, homology-based methods predict up to 5000 related sequences (45). Major databases containing thousands of annotated ncRNA sequences are RNAdb (10) and

NONCODE (47). Thus, using a combination of methods (see 'Materials and methods' section), we identified about 28% (1078 of 3834) of our transcripts as potential ncRNA. These sequences showed a small overlap (4%) to sequences deposited on these ncRNA databases. One of them, CV372409 ORESTES, aligns with a sequence in the NONCODE database and was downregulated in three different tumors, compared with normal tissues in our cDNA microarray experiments. A common theme seems to be that many ncRNA genes have a very restricted expression. Often, they have low, or no, EST coverage, but this does not necessarily mean that they are not expressed and are nonfunctional (14,55).

Microarray technology has dramatically enhanced the discovery of molecular markers for cancer. Prostate cancer is the most prevalent cancer in Brazilian males (http://www.inca.gov.br) as well in men wordwide (http://www.cancer.gov), and investigators have searched for molecular markers of the disease. The first gene identified by cDNA microarray to be suitable for clinical practice, and to potentially improve the diagnosis of prostate cancer was AMACR (42). AMACR was suggested as a new molecular marker for prostate carcinoma by Xu *et al.* (42) in 2000, and confirmed by Jiang *et al.*, (51). This protein is already used clinically as an aid in distinguishing prostate cancer from benign disease (56), and discriminating different grades and types of prostate cancer (50). Another potential molecular marker for prostate cancer, identified through cDNA microarray analysis, is the polycomb gene, EZH2. The expression of EZH2 indicates poor survival, and could be used as a marker for prostate cancer progression and metastasis (57–59). Also identified as a molecular marker is the TMPRSS2-ERG gene fusion, which is involved in the development of prostate cancer (60).

Increasing evidence shows a relationship between changes in expression levels of ncRNAs and cancer (18,61–63), emphasizing the potential role of ncRNAs in tumorigenesis, and the potential of this type of transcript as a tumor molecular marker (62). For example, in breast carcinoma, BC1 is deregulated (64), and the overexpression of BC200 RNA was recently evaluated as a new molecular marker for a poor prognosis (65). In lung cancer, increased expression of the MALAT-1 gene indicates a poor clinical outcome (66), and in hepatocellular carcinoma, HULC ncRNA is one of the most upregulated genes (62). In prostate cancer, there is overexpression of PCGEM (67), and $DD3^{PCA3}$ (49) is implicated in tumorigenesis (68). These findings present a strong argument for the inclusion of noncoding transcripts into the arsenal of markers used for molecular diagnostics, which, thus far, has been almost exclusively populated by assays of protein-coding transcripts (11).

We validated three differentially expressed sequences in paired prostate samples as potential tumor markers. Validation of the BQ373258 sequence enhanced the value of our approach to identify molecular markers, since this sequence is mapped on the last exon of a described ncRNA ($DD3^{PCA3}$) (49). $DD3^{PCA3}$ has been described as highly overexpressed in prostate cancer tissue when compared with adjacent nonmalignant prostatic tissue, and its expression is restricted to the prostate (49). An unusually high density of stop codons has been identified along the entire $DD3^{PCA3}$ cDNA sequence (49,69), which, in addition to the lack of an extended open frame and, after several years of analyzing putative proteins from predicted small ORFs, has resulted in the classification of $DD3^{PCA3}$ as a polyadenylated ncRNA (69–71). Its function is unknown, although there is speculation that $DD3^{PCA3}$ functions to regulate gene expression or participates in gene splicing (69). Both our cDNA microarray and real-time PCR show that this sequence is upregulated in prostate cancer relative to normal prostate (fold mean of 5.50 for real-time PCR and 7.41 for cDNA microarray).

An interesting observation arises from the data of two ORESTES, BF910617 and AW793062. ORESTES BF910617 is aligned with an intron of the KIAA1432 gene. From the analyses performed through Oncomine Research (http://www.oncomine.org) of the Lapointe *et al.* (72) data set, we observed that, in prostate cancer relative to normal prostate, the BF910617 ORESTES has diametrically opposite expression compared with the KIAA1432 gene. In the data set provided by Lapointe *et al.* (72) using cDNA microarray, the KIAA1432 gene was highly expressed in normal prostate, decreasing as the aggressiveness of prostate cancer increased. According to this data set, it was least expressed in metastatic prostate cancer in the lymph node (72). Therefore, our hypothesis is that BF910617 ORESTES may play a role in regulating the KIAA1432 gene, inhibiting its expression in prostate cancer when it is expressed at high levels. ORESTES AW793062 was validated with high fold values in almost 70% of the paired samples. This sequence is located in the first intron of a putative isoform of the RNF217 gene. Once again, the differential expression of AW793062 in prostate cancer was opposite to that observed for the RNF217 gene, with respect to primary and metastatic prostate cancer (Oncomine Research analyses) (73).

Although the differential expression of –4.32-fold in prostate tumor, showed by cDNA microarray experiments, of the CV374350 ORESTES was not confirmed by real-time PCR, we observed that this sequence maps to the last intron of the SGK1 gene, an inducible Ser/Thr kinase activated via phosphoinositide 3-kinase (PI3K) signal pathway (74,75). It is worth to note that there is an mRNA sequence (BX649005), also mapped to the SGK1 locus, which shows an extensive intron retention that includes the SGK1 last intron. It has been suggested that SGK1 may regulate androgen receptor activity, affecting androgen-mediated prostate cancer growth through a positive-feedback mechanism (76). Oncomine Research analysis of the SGK1 gene (73) suggests that this gene is expressed in normal prostate and benign prostate hyperplasia and its expression is fairly reduced among primary prostate carcinoma samples but is significantly reduced in metastatic prostate cancer. Further analysis of metastatic tumor samples could reveal if CV374350 expression follows the pattern of SGK1 gene expression and if this sequence may represents an SGK1 intron retention event or may be associated with other gene-regulation mechanism. This ORESTES was found in the

list of cellular signal pathways related to cancer, analyzed as described above.

The power of our data to explore uncharacterized molecular markers was demonstrated with the large number of differentially expressed sequences, between tumor and normal samples from all tissues (about 28% of the intronic and intergenic sequences mapped once on the genome). Also, 291 of these differentially expressed transcripts have ncRNA potential, as predicted by our analysis. It is also very promising that at least five putative noncoding tumor markers are upregulated in at least four different tumors, compared with normal tissues. On the basis of these results, we believe in the value of our approach to identify uncharacterized molecular markers. Our data set contains a large number of actively transcribed regions of the human genome not associated with annotated transcripts not yet widely explored. These may represent new genes, splice variants, NATs or ncRNAs, which could be used as molecular markers for other cancers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Alex Carvalho for construction of the cDNA microarray. We thank Dra. Anamaria Camargo for providing the prostate cancer cell lines and Dra. Maria Mitzi Brentani and Dra. Rose Roela for cultivating them.

FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 142330/2007-8 to B.P.M.); and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; 04/11774-8, 07/55791-1 to B.P.M.; 07/01549-5 to A.M.-L.). Funding for open access charge: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Conflict of interest statement. None declared.

REFERENCES

1. Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA*, **74**, 560–564.
2. Sanger,F., Nicklen,S. and Coulson,A.R. (1992) DNA sequencing with chain-terminating inhibitors (classical article: 1977). *Biotechnology*, **24**, 104–108.
3. Brentani,H., Caballero,O.L., Camargo,A.A., da Silva,A.M., da,S.W.A. Jr, Dias,N.E., Grivet,M., Gruber,A., Guimaraes,P.E., Hide,W. *et al*. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **100**, 13418–13423.
4. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z.P., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
5. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
6. Frith,M.C., Pheasant,M. and Mattick,J.S. (2005) The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet*., **13**, 894–897.
7. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al*. (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
8. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*., **15**, R17–R29.
9. Mehler,M.F. and Mattick,J.S. (2006) Non-coding RNAs in the nervous system. *J. Physiol*., **575**, 333–341.
10. Pang,K.C., Stephen,S., Dinger,M.E., Engstrom,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0-an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res*., **35**, D178–D182.
11. Reis,E.M., Nakaya,H.I., Louro,R., Canavez,F.C., Flatschart,A.V.F., Almeida,G.T., Egidio,C.M., Paquola,A.C., Machado,A.A., Festa,F. *et al*. (2004) Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene*, **23**, 6684–6692.
12. Johnson,J.M., Edwards,S., Shoemaker,D. and Schadt,E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*., **21**, 93–102.
13. Kapranov,P., Drenkow,J., Cheng,J., Long,J., Helt,G., Dike,S. and Gingeras,T.R. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res*., **15**, 987–997.
14. Weile,C., Gardner,P.P., Hedegaard,M.M. and Vinther,J. (2007) Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes. *BMC Genomics*, **8**, 244.
15. Soares,L.M.M. and Valcarcel,J. (2006) The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J*., **25**, 923–931.
16. Seidl,C.I.M., Stricker,S.H. and Barlow,D.P. (2006) The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J*., **25**, 3565–3575.
17. Goodrich,J.A. and Kugel,J.F. (2006) Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol*., **7**, 612–616.
18. Nakaya,H.I., Amaral,P.P., Louro,R., Lopes,A., Fachel,A.A., Moreira,Y.B., El Jundi,T.A., da Silva,A.M., Reis,E.M. and Verjovski-Almeida,S. (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol*., **8**, R43.
19. Mattick,J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet*., **5**, 316–323.
20. Numata,K., Kanai,A., Saito,R., Kondo,S., Adachi,J., Wilming,L.G., Hume,D.A., Hayashizaki,Y. and Tomita,M. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res*., **13**, 1301–1306.
21. Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*., **14**, 331–342.
22. Gustincich,S., Sandelin,A., Plessy,C., Katayama,S., Simone,R., Lazarevic,D., Hayashizaki,Y. and Carninci,P. (2006) The complexity of the mammalian transcriptome. *J. Physiol*., **575**, 321–332.
23. Sun,H., Skogerbo,G. and Chen,R.S. (2006) Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet*., **15**, 2911–2922.
24. Babak,T., Blencowe,B.J. and Hughes,T.R. (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, **6**, 104.
25. Washietl,S., Pedersen,J.S., Korbel,J.O., Stocsits,C., Gruber,A.R., Hackermuller,J., Hertel,J., Lindemeyer,M., Reiche,K., Tanzer,A. *et al*. (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*., **17**, 852–864.

26. Dias,N.E., Correa,R.G., Verjovski-Almeida,S., Briones,M.R., Nagai,M.A., da,S.W. Jr, Zago,M.A., Bordin,S., Costa,F.F., Goldman,G.H. *et al.* (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **97**, 3491–3496.
27. Camargo,A.A., Samaia,H.P., Dias-Neto,E., Simao,D.F., Migotto,I.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A. *et al.* (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl Acad. Sci. USA*, **98**, 12103–12108.
28. Sironi,M., Menozzi,G., Comi,G.P., Cagliani,R., Bresolin,N. and Pozzoli,U. (2005) Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet.*, **14**, 2533–2546.
29. Ravasi,T., Suzuki,H., Pang,K.C., Katayama,S., Furuno,M., Okunishi,R., Fukuda,S., Ru,K.L., Frith,M.C., Gongora,M.M. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, **16**, 11–19.
30. de Souza,S.J., Camargo,A.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A., Andrade,L.E., Carrer,H., El Dorry,H.F. *et al.* (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **97**, 12690–12693.
31. Fonseca,R.D., Carraro,D.M. and Brentani,H. (2006) Mining ORESTES no-match database: can we still contribute to cancer transcriptome? *Genet. Mol. Res.*, **5**, 24–32.
32. Sorek,R. and Safer,H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.
33. Brentani,R., Carraro,D., Verjovski-Almeida,S., Reis,E., Neves,E., de Souza,S., Carvalho,A., Brentani,H. and Reis,L. (2005) Gene expression arrays in cancer research: methods and applications. *Crit. Rev. Oncol. Hematol.*, **54**, 95–105.
34. Vangelder,R.N., Vonzastrow,M.E., Yool,A., Dement,W.C., Barchas,J.D. and Eberwine,J.H. (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl Acad. Sci. USA*, **87**, 1663–1667.
35. Gomes,L.I., Silva,R.L.A., Stolf,B.S., Cristo,E.B., Hirata,R., Soares,F.A., Reis,L.F.L., Neves,E.J. and Carvalho,A.F. (2003) Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray. *Anal. Biochem.*, **321**, 244–251.
36. DeRisi,J. (2003) In Bowtell,D. and Sambrook,J (eds.), *DNA Microarrays: A Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press, New York, pp. 187–193.
37. Yang,Y.H. and Speed,T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.*, **3**, 579–588.
38. Yang,Y., Dudoit,S., Luu,P., Lin,D., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
39. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
40. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
41. de Kok,J.B., Roelofs,R.W., Giesendorf,B.A., Pennings,J.L., Waas,E.T., Feuth,T., Swinkels,D.W. and Span,P.N. (2005) Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab. Invest.*, **85**, 154–159.
42. Xu,J.C., Stolk,J.A., Zhang,X.Q., Silva,S.J., Houghton,R.L., Matsumura,M., Vedvick,T.S., Leslie,K.B., Badaro,R. and Reed,S.G. (2000) Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. *Cancer Res.*, **60**, 1677–1682.
43. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.*, **29**, e45.
44. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *Plos Comput. Biol.*, **2**, 251–262.
45. Griffiths-Jones,S. (2007) Annotating noncoding RNA genes. *Annu. Rev. Genomics Hum. Genet.*, **8**, 279–298.
46. Rymarquis,L.A., Kastenmayer,J.P., Huttenhofer,A.G. and Green,P.J. (2008) Diamonds in the rough: mRNA-like non-coding RNAs. *Trends Plant Sci.*, **13**, 329–334.
47. Liu,C., Bai,B., Skogerbo,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.
48. Galante,P.A.F., Vidal,D.O., de Souza,J.E., Camargo,A.A. and de Souza,S.J. (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol.*, **8**, R40.
49. Bussemakers,M.J., van Bokhoven,A., Verhaegh,G.W., Smit,F.P., Karthaus,H.F., Schalken,J.A., Debruyne,F.M., Ru,N. and Isaacs,W.B. (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.*, **59**, 5975–5979.
50. Jiang,Z., Woda,B.A., Wu,C.L. and Yang,X.M.J. (2004) Discovery and clinical application of a novel prostate cancer marker - alpha-Methylacyl CoA racemase (P504S). *Am. J. Clin. Pathol.*, **122**, 275–289.
51. Jiang,Z., Woda,B.A., Rock,K.L., Xu,Y.D., Savas,L., Khan,A., Pihan,G., Cai,F., Babcook,J.S., Rathanaswami,P. *et al.* (2001) P504S - a new molecular marker for the detection of prostate carcinoma. *Am. J. Surg. Pathol.*, **25**, 1397–1404.
52. Futreal,P., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
53. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
54. Machado-Lima,A., del Portillo,H.A. and Durham,A.M. (2008) Computational methods in noncoding RNA research. *J. Math. Biol.*, **56**, 15–49.
55. Pollard,K.S., Salama,S.R., Lambert,N., Lambot,M.A., Coppens,S., Pedersen,J.S., Katzman,S., King,B., Onodera,C., Siepel,A. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443**, 167–172.
56. Cooper,C.S., Campbell,C. and Jhavar,S. (2007) Mechanisms of Disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer. *Nat. Clin. Pract. Urol.*, **4**, 677–687.
57. LaTulippe,E., Satagopan,J., Smith,A., Scher,H., Scardino,P., Reuter,V. and Gerald,W.L. (2002) Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.*, **62**, 4499–4506.
58. Varambally,S., Dhanasekaran,S.M., Zhou,M., Barrette,T.R., Kumar-Sinha,C., Sanda,M.G., Ghosh,D., Pienta,K.J., Sewalt,R.G.A.B., Otte,A.P. *et al.* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624–629.
59. Rhodes,D.R., Sanda,M.G., Otte,A.P., Chinnaiyan,A.M. and Rubin,M.A. (2003) Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. *J. Natl Cancer Inst.*, **95**, 661–668.
60. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.W., Varambally,S., Cao,X.H., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
61. Reis,E.M., Ojopi,E.P.B., Alberto,F.L., Rahal,P., Tsukumo,F., Mancini,U.M., Guimaraes,G.S., Thompson,G.M.A., Camacho,C., Miracca,E. *et al.* (2005) Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. *Cancer Res.*, **65**, 1693–1699.
62. Panzitt,K., Tschernatsch,M.M.O., Guelly,C., Moustafa,T., Stradner,M., Strohmaier,H.M., Buck,C.R., Denk,H., Schroeder,R., Trauner,M. *et al.* (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as non-coding RNA. *Gastroenterology*, **132**, 330–342.
63. Brito,G.C., Fachel,A.A., Vettore,A.L., Vignal,G.M., Gimba,E.R., Campos,F.S., Barcinski,M.A., Verjovski-Almeida,S. and Reis,E.M. (2008) Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. *Mol. Carcinog.*, **47**, 757–767.

64. Chen,W., Bocker,W., Brosius,J. and Tiedge,H. (1997) Expression of neural BC200 RNA in human tumours. *J. Pathol.*, **183**, 345–351.
65. Iacoangeli,A., Lin,Y., Morley,E.J., Muslimov,I.A., Bianchi,R., Reilly,J., Weedon,J., Diallo,R., Bocker,W. and Tiedge,H. (2004) BC200 RNA in invasive and preinvasive breast cancer. *Carcinogenesis*, **25**, 2125–2133.
66. Ji,P., Diederichs,S., Wang,W.B., Boing,S., Metzger,R., Schneider,P.M., Tidow,N., Brandt,B., Buerger,H., Bulk,E. *et al.* (2003) MALAT-1, a novel noncoding RNA, and thymosin beta 4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, **22**, 8031–8041.
67. Srikantan,V., Zou,Z.Q., Petrovics,G., Xu,L., Augustus,M., Davis,L., Livezey,J.K., Connell,T., Sesterhenn,I.A., Yoshino,K. *et al.* (2000) PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proc. Natl Acad. Sci. USA*, **97**, 12216–12221.
68. Petrovics,G., Zhang,W., Makarem,M., Street,J.P., Connelly,R., Sun,L., Sesterhenn,I.A., Srikantan,V., Moul,J.W. and Srivastava,S. (2004) Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene*, **23**, 605–611.
69. Schalken,J.A., Hessels,D. and Verhaegh,G. (2003) New targets for therapy in prostate cancer: Differential display code 3 (DD3(PCA3)) a highly prostate cancer-specific gene. *Urology*, **62**, 34–43.
70. Hessels,D., Gunnewiek,J.M.T.K., van Oort,I., Karthaus,H.F.M., van Leenders,G.J.L., van Balken,B., Kiemeney,L.A., Witjes,J.A. and Schalken,J.A. (2003) DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur. Urol.*, **44**, 8–15.
71. Tinzl,M., Marberger,M., Horvath,S. and Chypre,C. (2004) DD3(PCA3) RNA analysis in urine - a new perspective for detecting prostate cancer. *Eur. Urol.*, **46**, 182–187.
72. Lapointe,J., Li,C., Higgins,J.P., van de Rijn,M., Bair,E., Montgomery,K., Ferrari,M., Egevad,L., Rayford,W., Bergerheim,U. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.
73. Dhanasekaran,S.M., Barrette,T.R., Ghosh,D., Shah,R., Varambally,S., Kurachi,K., Pienta,K.J., Rubin,M.A. and Chinnaiyan,A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.
74. Kobayashi,T. and Cohen,P. (1999) Activation of serum- and glucocorticoid-regulated protein kinase by agonists that activate phosphatidylinositide 3-kinase is mediated by 3-phosphoinositide-dependent protein kinase-1 (PDK1) and PDK2. *Biochem. J.*, **339** (Pt 2), 319–328.
75. Park,J., Leong,M., Buse,P., Maiyar,A., Firestone,G. and Hemmings,B. (1999) Serum and glucocorticoid-inducible kinase (SGK) is a target of the PI 3-kinase-stimulated signaling pathway. *EMBO J.*, **18**, 3024–3033.
76. Shanmugam,I., Cheng,G., Terranova,P., Thrasher,J., Thomas,C. and Li,B. (2007) Serum/glucocorticoid-induced protein kinase-1 facilitates androgen receptor-dependent cell survival. *Cell Death Differ.*, **14**, 2085–2094.