UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Bioquímica

HELDER TAKASHI IMOTO NAKAYA

Identificação e Análise de Expressão de RNAs

Intrônicos Não Codificadores Humanos

São Paulo

Data do Depósito na SPG:

09/01/2007

Nove de Janeiro de Dois Mil e Sete

HELDER TAKASHI IMOTO NAKAYA

Identificação e Análise de Expressão de RNAs

Intrônicos Não Codificadores Humanos

*Tese apresentada ao Instituto de
Química da Universidade de São Paulo
para a obtenção do Título de Doutor em
Ciências (Bioquímica)*

Orientador: Sergio Verjovski Almeida

São Paulo

2007

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO DO CONJUNTO DAS QUÍMICAS

Nakaya, Helder Takashi Imoto

Identificação e Análise de Expressão de RNAs Intrônicos Não Codificadores Humanos/ Helder Takashi Imoto Nakaya; orientador Sergio Verjovski Almeida

---São Paulo, 2007

146 p.

Tese (Doutorado – Programa de Pós-Graduação em Bioquímica. Área de Concentração: Bioquímica) – Instituto de Química da Universidade de São Paulo.

1. Íntron. 2. *Microarrays*. 3. RNA não-codificador. 4. Transcriptoma. 5. *Splicing* alternativo. 6. Andrógeno.

**"Identificação e análise de expressão de RNAs intrônicos
não codificadores humanos"**

HELDER TAKASHI IMOTO NAKAYA

Tese de Doutorado submetida ao Instituto de Química da Universidade de São Paulo como parte dos requisitos necessários à obtenção do grau de Doutor em Ciências - Área: Bioquímica.

Aprovado(a) por:

Prof. Dr. SERGIO VERJOVSKI DE ALMEIDA
(Orientador e Presidente)

Prof. Dr. HUGO AGUIRRE ARMELIN
IQ - USP

Profa. Dra. CARLA COLUMBANO DE OLIVEIRA
IQ - USP

Prof. Dr. FERNANDO LUIZ DE LUCCA
FMRP - USP

Prof. Dr. LUIS FERNANDO LIMA REIS
LUDWIG

SÃO PAULO
09 DE MARÇO 2007

À minha família pelo total apoio e incentivo

AGRADECIMENTOS

Ao Prof. Dr. Sergio Verjovski Almeida pela orientação, críticas construtivas, ensinamentos e, principalmente, pela liberdade e oportunidades recebidas nesta etapa importante de minha formação.

Ao Prof. Dr. Eduardo Reis, meu co-orientador, pelas enormes contribuições neste trabalho e por seu grande apoio ao longo de meu doutoramento.

Aos excelentes cientistas que participaram deste trabalho: Paulo Paiva, Rodrigo Louro, Fernanda Festa, Aline Maria da Silva, Ricardo de Marco, dentre outros.

À FAPESP e ao CNPq pela bolsa de pesquisa e auxílio financeiro.

À família que eu amo. Agradeço minha mãe, Lourdes, pelos joguinhos de química, microscópio de brinquedo, livros e outros incentivos que me fizeram escolher esta carreira.

Ao meu pai, Tacaioshi, pelo orgulho e pelo patrocínio em fases importantes da minha vida.

Às minhas irmãs, Lyna e Tânia e, suas respectivas famílias, pelo enorme carinho e suporte.

Ao meu avô Hideo (*in memoriam*) pelo orgulho do neto (biólogo) que ele achava que era médico da USP. À minha avó Sônia pelas suas sugestões científicas engraçadas e aos outros membros da família pelos bons momentos.

À Daniela Soltys por ser minha paduska.

À minha segunda família. Aos meus irmãos do tempo de colégio que ainda fazem parte da minha vida: Dumans e Careca. Aos irmãos do tempo de graduação, Blanes, Paulinho, Monalisa e Bianca, aos meus novos irmãozinhos, Karlinha e PCR, aos PDBs (Mineiro, Rita, Camila, DiVerde, Julio, Murilo, Flô, Pererê, Guilherme), aos biólogos (Gandhi, Teté, MBS, Gustavinho) e aos amigos de verdade do laboratório (Minhoca, Felipe e Thiago) por toda serotonina e risadas proporcionadas em inúmeras baladas, viagens, congressos, cafés da tarde e bandeijões.

Agradeço também os meus colegas de laboratório (Julio, Yuri, Kátia, Ana Cláudia, Ana Paula, Maria Luisa, Denise, Adriana, Renato, Giulliana, Glauber, Ângela, Tarik, Lauren, Ana Tahira, Jefferson, Esther, Ana Ayupe e Anna) pela ótima convivência e boas trocas de idéias.

Ao Departamento de Bioquímica e ao Instituto de Química da USP pela infra-estrutura, auxílio e ótimos seminários.

The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful. If nature were not beautiful, it would not be worth knowing, and if nature were not worth knowing, life would not be worth living.

Jules Henri Poincaré

RESUMO

Nakaya, H.I. **Identificação e Análise de Expressão de RNAs Intrônicos Não Codificadores Humanos**. 2007. 146 p. Tese (Doutorado). Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo 2007.

Neste trabalho, nós mostramos estudos em larga-escala de RNAs não codificadores antisenso que são transcritos em regiões intrônicas de genes humanos. Alguns destes transcritos intrônicos possuem níveis de expressão correlacionados ao grau de diferenciação tumoral de câncer de próstata, apontando para uma relevância biológica destas mensagens em doenças complexas como o câncer. Nós também avaliamos a existência de um mecanismo comum de regulação de transcrição, compartilhado por mRNAs codificadores de proteína e RNAs intrônicos, através de análises de perfis de expressão de uma linhagem tumoral de próstata estimulada por andrógeno. A análise de ESTs e mRNAs depositados em bancos públicos de seqüências revelou mais de 55 mil RNAs Totalmente Intrônicos Não-codificadores (TIN), transcritos dos íntrons de 74% de todos os genes RefSeq únicos. Guiados por esta informação, nós desenhamos uma plataforma de oligonucleotídeos contendo sondas senso e antisenso para cada um de 7.520 transcritos TIN selecionados aleatoriamente, além de sondas para os genes codificadores de proteína correspondentes. Nós identificamos assinaturas intrônicas e exônicas de expressão tecido-específicas em fígado, próstata e rim. Os RNAs TIN antisenso mais altamente expressos eram transcritos de íntrons de genes codificadores de proteína enriquecidos na categoria “Regulação da transcrição”. A inibição da RNA Polimerase II resultou num aumento de expressão de uma fração dos RNAs intrônicos em células em cultura, sugerindo que outras RNA Polimerases possam estar envolvidas em sua biossíntese. Um subconjunto das assinaturas intrônicas e exônicas localizadas nos mesmos

loci genômicos possuíram padrões de expressão correlacionados, sugerindo que RNAs intrônicos regulem a abundância ou o padrão de uso de éxons de mensagens codificadoras de proteína. Nós identificamos diversos padrões de expressão de RNAs intrônicos, indicando que eles possam ter papéis regulatórios. Esta estratégia orientada pelo gene, que combina um *microarray* intrônico/exônico deve permitir análises comparativas futuras de transcrição intrônica sob várias condições fisiológicas e patológicas, avançando assim em nosso conhecimento sobre as funções biológicas destes RNAs não codificadores.

Palavras-chave: *Microarrays*, Transcriptoma, *Splicing* alternativo, Andrógeno, α -amanitina

ABSTRACT

Nakaya, H.I. **Identification and Expression Analysis of Human Intronic Noncoding RNAs**. 2007. 146 p. PhD Thesis. – Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo.

In this work, we show large-scale studies of antisense noncoding RNAs transcribed from intronic regions of human genes. The correlation of expression levels of some intronic transcripts to the degree of tumor differentiation in prostate cancer points to the biological relevance of these messages in complex diseases such as cancer. We also evaluated the existence of a common mechanism of regulation of transcription shared by protein-coding mRNAs and intronic RNAs by measuring the effect of androgen on the transcriptional profile of a prostate cancer cell line. Survey of mRNA and EST public databases revealed more than 55,000 Totally Intronic Noncoding (TIN) RNAs transcribed from the introns of 74% of all unique RefSeq genes. Guided by this information, we designed an oligoarray platform containing sense and antisense probes for each of 7,520 randomly-selected TIN transcripts plus probes for the corresponding protein-coding genes. We identified exonic and intronic tissue-specific expression signatures for human liver, prostate and kidney. The most highly expressed antisense TIN RNAs were transcribed from introns of protein-coding genes enriched in the “Regulation of transcription” class. RNA Polymerase II inhibition resulted in increased expression of a fraction of the intronic RNAs in cell cultures, suggesting that other RNA polymerases may be involved in their biosynthesis. A subset of intronic and protein-coding signatures transcribed from the same genomic loci has correlated expression patterns, suggesting that intronic RNAs regulate the abundance or the pattern of exon usage in protein-coding messages. We have identified diverse intronic RNA expression patterns, indicating that they may have regulatory roles. This

gene-oriented approach, using a combined intronic/exonic microarray should permit further comparative analysis of intronic transcription under various physiological and pathological conditions, thus advancing current knowledge about the biological functions of these noncoding RNAs.

Keywords: Microarrays, Transcriptome, Alternative splicing, Androgen, α -amanitin

LISTA DE ABREVIATURAS E SIGLAS

ADAR	<i>Adenosine Deaminase Acting on RNA</i>
ANOVA	Análise de variância
AR	Receptor de Andrógeno
ARE	Elemento de Resposta a Andrógeno
AS	Antisenso
ATCC	<i>American Type Culture Collection</i>
BLAST	<i>Basic Local Alignment Search Tool</i>
BLAT	<i>BLAST-Like Alignment Tool</i>
CAGE	<i>Cap Analysis of Gene Expression</i>
CAGE	Cooperação para Análise da Expressão Gênica
CDD	<i>Conserved Domain Database</i>
cDNA	DNA complementar
ChIP	Imunoprecipitação da cromatina
cRNA	RNA complementar amplificado
Ct	<i>Cycle threshold</i>
DNA	Ácido desoxirribonucléico
DNase	Desoxirribonuclease
EDTA	Ácido etilenodiaminotetracético
ES	<i>Exon Skipping</i>
EST	<i>Expressed Sequence Tag</i>
FDR	<i>False Discovery Rate</i>
GO	<i>Gene Ontology</i>
GS	<i>Gleason Score</i>

HCGP	Projeto Genoma Humano do Câncer
IgG	Imunoglobulina G
kb	Quilobase
miRNA	<i>micro RNA</i>
MPSS	<i>Massively Parallel Signature Sequencing</i>
mRNA	RNA mensageiro
N ₂	Nitrogênio
ncRNA	RNA não-codificador de proteína
NMD	<i>Nonsense-mediated decay</i>
nt	Nucleotídeos
Oligo	Oligonucleotídeo
ORESTES	<i>ORF ESTs</i>
pb	Pares de bases
PBS	Solução salina de tampão de fosfato
PCR	Reação em Cadeia da Polimerase
PIN	Parcialmente Intrônico Não-codificador
PMSF	Fluoreto de fenilmetilsulfonila
pre-mRNA	RNA mensageiro prematuro
RACE	<i>Rapid Amplification of cDNA Ends</i>
RefSeq	<i>Reference Sequence</i>
RISC	<i>RNA-induced silencing complex</i>
RNA	Ácido ribonucleico
RNAi	RNA de interferência
RNAP	RNA Polimerase
RNase	Ribonuclease

rpm	Rotações por minuto
RPMI	<i>Roswell Park Memorial Institute</i>
rRNA	RNA ribossômico
RT	Transcrição reversa
S/AS	Senso/Antisenso
SAGE	<i>Serial Analysis of Gene Expression</i>
SAM	<i>Significance Analysis of Microarrays</i>
SDS	Dodecilsulfato de sódio
siRNAs	<i>small interfering RNAs</i>
snoRNA	<i>small nucleolar RNAs</i>
SNR	<i>Signal-to-Noise Ratio</i>
snRNAs	<i>small nuclear RNAs</i>
TFBS	<i>Transcription Factor Binding Sites</i>
TIN	Totalmente Intrônico Não-codificador
Tm	<i>Melting Temperature</i>
Tris	Tris-Hidroximetil aminometano
tRNA	RNA transportador
U	Unidade de atividade enzimática

SUMÁRIO

1. INTRODUÇÃO	19
1.1. DNA não-codificador e Complexidade Biológica	19
1.2. Funções dos Íntrons	20
1.3. Complexidade do Transcriptoma Humano	24
1.4. Função dos RNAs Não-codificadores	26
1.5. Biogênese dos RNAs Não Codificadores	29
2. OBJETIVOS	31
3. MATERIAIS E MÉTODOS	32
3.1. RNAs Intrônicos Correlacionados ao Grau de Diferenciação Tumoral	32
3.1.1. Construção do <i>Microarray</i> 4k Intrônico	32
3.1.2. Extração de RNA de Próstata	33
3.1.3. Marcação do RNA e Hibridização	34
3.1.4. Análise dos Resultados de <i>Microarrays</i>	35
3.1.5. Caracterização dos Transcritos Intrônicos	37
3.1.5.1. Ensaios de RACE	37
3.1.5.2. Northern blot fita-específica	37
3.1.5.3. Transcrição Reversa Fita-Específica	38
3.1.5.4. PCR Quantitativo em Tempo Real	40
3.2. RNAs Intrônicos Responsivos à Andrógeno	41
3.2.1. Tratamento das Células e Obtenção e Marcação de RNA Total	41
3.2.2. Experimentos de <i>Microarrays</i>	43
3.2.2.1. Limite de Detecção e Normalizações	43
3.2.2.2. Análises de Expressão por SAM	44
3.2.3. Caracterização de RNAs Intrônicos	45

3.2.4. Imunoprecipitação da Cromatina.....	46
3.3. Identificação e Caracterização <i>in silico</i> de RNAs Intrônicos	48
3.3.1. Cruzamento das Coordenadas Genômicas de Transcritos de Diferentes Conjuntos de Sequências	48
3.3.2. Cálculo da Frequência de <i>Skipping</i> de Éxon	49
3.4. Estudo da Biogênese e Especificidade de Expressão por Tecido dos ncRNAs Intrônicos.....	50
3.4.1. Obtenção de Amostras de Tecidos Humanos	50
3.4.2. Experimentos de Células LNCaP tratadas com α -amanitina.....	50
3.4.3. Marcação das Amostras e Experimentos de Hibridização	51
3.4.4. Análises Estatísticas	52
3.4.5. Análise de Enriquecimento de Categorias de <i>Gene Ontology</i> (GO)	53
4. RESULTADOS E DISCUSSÃO	54
4.1. RNAs Intrônicos Associados ao Câncer de Próstata.....	54
4.1.1. Construção do <i>Microarray</i> 4k Intrônico.....	54
4.1.2. Análise dos Resultados de <i>Microarrays</i> 4k Intrônicos.....	55
4.1.3. Caracterização dos Transcritos Diferencialmente Expressos.....	58
4.1.3.1. Novos Éxons de Genes Envolvidos em Câncer	58
4.1.3.2. Transcritos Intrônicos Antisensos Não-Codificadores	60
4.2. Regulação por Andrógeno de RNAs Intrônicos	66
4.2.1. Identificação de ncRNAs Intrônicos Respondedores a Andrógeno	67
4.2.2. Caracterização dos ncRNAs Regulados por Andrógeno	71
4.2.3. Alteração dos níveis de expressão por andrógeno de ncRNAs intrônicos e os genes codificadores de proteína correspondentes.....	77

4.2.4. Ligação do Receptor de Andr�geno na Regi�o <i>Upstream</i> do RNA intr�nico Antisense do Gene MYO5A	79
4.3. Identifica�o e Caracteriza�o <i>in silico</i> de RNAs Intr�nicos	84
4.3.1. Defini�o de um Conjunto de Refer�ncia dos Genes Codificadores de Prote�nas	84
4.3.2. Identifica�o dos Transcritos Totalmente Intr�nicos Longos e <i>Unspliced</i>	87
4.3.3. Identifica�o de Transcritos Longos <i>Unspliced</i> Parcialmente Intr�nicos	91
4.3.4. A Maioria dos Genes RefSeq Possui Transcri�o Intr�nica	92
4.3.5. Transcritos TIN e PIN s�o Poss�veis Reguladores de <i>Splicing</i> Alternativo	92
4.4. Biog�nese e Express�o Tecido-Espec�fica de RNAs Intr�nicos	98
4.4.1. Constru�o do <i>Microarray</i> Intr�nico 44k	99
4.4.2. Medidas de Express�o em Larga-Escala de Transcritos Parcial e Totalmente Intr�nicos nos Tecidos Humanos	100
4.4.3. Transcritos TIN Antisense em �trons de Genes Relacionados � Regula�o da Transcri�o	107
4.4.4. Um Grande Fra�o de RNAs TIN e PIN � Insens�vel � Inibi�o da RNAP II ou � At� Estimulada por α -Amanitina	113
4.4.5. Assinatura Tecido-Espec�fica dos RNAs TIN e PIN	118
5. CONCLUS�ES	129
REFER�NCIAS BIBLIOGR�FICAS	130
LISTA DE ANEXOS	145

1. INTRODUÇÃO

1.1. DNA não-codificador e Complexidade Biológica

O projeto genoma humano teve como um objetivo central a identificação completa de todos os genes desta espécie. Com o alinhamento de milhares de seqüências de mRNAs e com o uso de predições computacionais, estima-se que das 3 bilhões de bases que constituem o genoma humano, menos de 2% dele codificam proteínas [1, 2]. Apesar das melhorias nos programas de anotação e com o sequenciamento de mais transcritos humanos, a identificação de regiões intergênicas e intrônicas não-codificadoras com atividade transcricional continua representando um desafio não desvendado do genoma humano [3].

As proteínas desempenham papéis cruciais em quase todos os processos celulares. Por isso, acreditava-se que a complexidade biológica de um ser vivo estaria relacionada com seu número de proteínas e, conseqüentemente, com o número de genes que as codificavam [4, 5]. Em 2000, apostas sobre o número de genes humanos, codificadores de proteína, giravam entre 30 a 150 mil [6]. Porém, após o sequenciamento completo do genoma humano, menos de 25 mil destes genes foram identificados [3], um número semelhante aos 25.498 genes da planta *Arabidopsis thaliana* [7], um pouco maior do que os 19.099 genes do nematóide *Caenorhabditis elegans* [8] e muito menor do que os 37.544 genes da espécie de arroz *Oryza sativa* [9]. Fica claro, portanto, que o grau de complexidade de um organismo não deve estar correlacionado ao número de genes codificadores de proteínas que ele contém.

De fato, a complexidade biológica de eucariotos superiores pode ser gerada de uma maneira independente do número de genes codificadores de proteína de um organismo. O *splicing* alternativo de pre-mRNAs, a poliadenilação alternativa e a ligação diferencial em

elementos promotores, assim como as redes de interações no controle gênico são exemplos de processos que geram tal complexidade [5]. Porém, o problema não parece ser como gerar esta complexidade mas como controlá-la de modo a seguir trajetórias ordenadas que gerem organismos altamente organizados e complexos [10].

John Mattick mostra que existe uma correlação interessante entre o tamanho da região não-codificadora de proteínas dos genomas de eucariotos e o seu grau de complexidade biológica [10]. Genomas com uma maior relação entre regiões não-codificadoras/ codificadoras de proteínas estão presentes em seres mais complexos como animais e plantas em comparação aos procariotos ou eucariotos mais simples, como muitos protozoários [10]. Segundo ele, estas regiões representam uma camada escondida de regulação, composta por RNAs não-codificadores (ncRNAs) que poderiam direcionar a diferenciação e o desenvolvimento de células e tecidos, via um conjunto de mecanismos que incluem: *imprinting*, seleção de promotores, *splicing* alternativo, edição de RNA, estabilidade de mRNA e tradução de mRNA [10, 11].

1.2. Funções dos Íntrons

Em eucariotos, os genes codificadores de proteína são quase sempre compostos por éxons e íntrons. Éxons são as seqüências representadas no RNA mensageiro maduro (mRNA) que, quando unidos, irão compor a seqüência traduzida em um peptídeo. Os íntrons são as seqüências intervenientes, removidas quando o transcrito primário (pre-mRNA) é processado (*splicing*) para dar origem ao mRNA.

A análise do genoma humano revelou que os íntrons constituem mais de 30% do genoma e cerca de 95% da seqüência dos genes [2]. Em geral, genes humanos tendem a ter éxons pequenos (cerca de 150 pb, codificando uma média de apenas 50 códons) separados

por longos íntrons que muitas vezes excedem 10 mil bases [1]. Análises posteriores deste genoma [3] identificaram 22.287 *loci*, consistindo de 19.438 *loci* gênicos conhecidos e 2.188 preditos. Estes *loci* possuem um total de 209.380 íntrons, com uma média de 9,4 íntrons por *locus*. A transcrição originada exclusivamente nas regiões intrônicas não foi avaliada nesta análise posterior do genoma [3]. Estas longas e numerosas regiões intrônicas podem possuir algumas funções biológicas ou serem simplesmente degradadas, tornando difícil escapar da conclusão geral de que, ou o genoma humano é repleto de transcrição inútil ou estes RNAs poderiam possuir funções inesperadas.

Algumas das funções dos íntrons já são conhecidas e sua presença maciça nos genomas de eucariotos pode estar diretamente relacionada à grande complexidade biológica destes organismos [12]. No processo denominado *Nonsense-mediated decay* (NMD), se um erro de transcrição leva a um códon de parada à montante (*upstream*) a uma junção íntron-éxon próxima, aquele transcrito é marcado para degradação [13]. A presença dos íntrons poderia deste modo ser importante para a manutenção da fidelidade do transcrito [12]. Além disso, os íntrons fornecem a possibilidade de gerar novos produtos gênicos por *splicing* alternativo ou *exon shuffling*, aumentando a taxa de recombinação entre as partes das regiões codificadoras do gene e gerando proteínas com diferentes domínios e funções. Isso sem dúvida teve um importante papel na evolução de todas as plantas e animais [12].

Funções dos íntrons ainda pouco conhecidas estão sendo estudadas por vários pesquisadores. Embora se pense que os íntrons sejam simplesmente degradados após sua excisão do transcrito primário, há boas evidências de que estes RNAs intrônicos possam de fato ser processados em pequenos RNAs, com vidas-médias significativas e localizações sub-celulares específicas [14, 15]. Além disso, é sabido que, em animais, quase todos os *small nucleolar RNAs* (snoRNAs) e uma grande proporção dos microRNAs (miRNAs) são

derivados de íntrons processados ou transcritos em regiões intrônicas [16-19]. Recentemente, foi mostrado ainda que a expressão de seqüências intrônicas inteiras, derivadas do gene *CFTR*, causam mudanças específicas na transcrição de vários genes em células HeLa [20]. O mais interessante é que cada uma destas 3 seqüências intrônicas selecionadas, quando testadas, resultaram num padrão distinto de efeitos num grupo particular de genes [20].

Estudos que analisaram o processo de edição de RNA – no caso, a edição de adeninas por inosinas – pelas ADARs (*adenosine deaminases acting on RNA*), apontam os íntrons de pré-mRNAs como os maiores alvos da maquinaria de edição [21, 22]. Inosina é capaz de fazer pareamento não-canônico com citosina e é interpretado pela maquinaria de tradução como uma guanina. A edição de A-I tem sido mostrada em mRNAs traduzidos, onde ela altera as seqüências e funções de receptores de glutamato e de serotonina [21]. Porém, esse maior número de edições em regiões intrônicas não traduzidas é, no mínimo, intrigante. Em alguns casos, a edição de um íntron pode criar um sítio acceptor de *splicing* mas para a maioria das vezes, a função por trás deste processo não é compreendida [22].

Outro dado intrigante é a descoberta de várias seqüências intrônicas extremamente conservadas em vertebrados [23-25]. Por não codificar proteínas, os íntrons apresentam uma menor conservação de bases em relação às seqüências codificadoras, as quais estão geralmente sujeitas a uma seleção negativa e/ou a uma menor taxa de mutação. Contudo, mais de 1.000 seqüências maiores que 100 pb e ultra-conservadas em várias espécies de vertebrados foram encontradas em íntrons humanos, sugerindo que os íntrons devem possuir uma quantidade substancial de elementos essenciais, funcionalmente ativos [23, 24]. Estes elementos poderiam estar relacionados a alterações pós-transcricionais dos genes codificadores de proteínas ou à própria regulação da expressão destes genes [26].

Genes constitutivos, altamente expressos, tendem a possuir íntrons bem menores do que genes específicos de alguns tecidos ou com baixos níveis de expressão [27]. Dois modelos principais têm sido propostos para explicar esta variação de tamanho dos íntrons. Segundo o modelo de “seleção por economia”, existe uma forte pressão seletiva para que os íntrons sejam curtos em genes altamente expressos, mantendo nos íntrons somente o que é necessário para o correto *splicing* do gene. Já o modelo “genomic design” postula que íntrons longos ocorrem preferencialmente em genes com expressão específica em certos tecidos por permitirem uma regulação mais complexa e uma supressão gênica mediada pela condensação da cromatina [28].

Ao analisar as seqüências conservadas em humanos e camundongos, Vinogradov [27] mostrou que a fração de seqüências intrônicas conservadas é maior em genes tecido-específicos. Estes dados sugerem que o tamanho dos íntrons não pode ser explicado pelo modelo de “seleção por economia”, já que, segundo este modelo, a fração de seqüências conservadas deveria ser maior em íntrons de genes constitutivos, devido à redução de partes dispensáveis e não conservadas dos íntrons [27]. Com isso, o modelo “genomic design” sugere que o maior tamanho dos íntrons em genes com vários domínios ou funções específicas deve estar relacionado a uma maior complexidade funcional nestes íntrons e que a evolução da maioria do DNA não codificador não é totalmente neutra.

É provável que a transcrição de RNAs não codificadores nos longos íntrons humanos também seja responsável por parte da complexidade funcional destas regiões. Estudos que detectaram a expressão ou que caracterizaram alguns ncRNAs, transcritos totalmente dentro de íntrons, foram realizados por poucos grupos de pesquisa e serão descritos nos itens a seguir.

1.3. Complexidade do Transcriptoma Humano

O transcriptoma consiste no conjunto de todos os RNAs que são expressos numa célula e que possuem alguma importância biológica. Diversas técnicas têm sido usadas para investigar mudanças de expressão e de estrutura dos transcritos humanos em várias condições biológicas. Além da análise em larga-escala de ESTs (*Expressed Sequence Tags*), técnicas como SAGE (*Serial Analysis of Gene Expression*), MPSS (*Massively Parallel Signature Sequencing*) e CAGE (*Cap Analysis of Gene Expression*), que usam etiquetas de expressão (*tags*) geradas das pontas dos mRNAs, assim como a técnica de *microarrays* que consiste na hibridização de populações de mRNAs contra um grande número de *probes* imobilizados num substrato são largamente utilizadas e revelam a enorme complexidade do transcriptoma humano.

As seis milhões de ESTs depositadas em bancos de dados públicos provavelmente correspondem a melhor representação do transcriptoma humano. ESTs humanas e de camundongo têm sido extensivamente usadas para (i) identificação *in silico* de novos genes [29, 30] ou novos éxons de genes codificadores de proteína [31-33], (ii) análise do perfil digital de expressão gênica em diferentes tecidos [34, 35] e (iii) descoberta de pares de transcritos senso-antiseno [36-38]. Entretanto, estes estudos estão focados, principalmente, em ESTs processadas (que sofreram *splicing*) e que se sobrepõem em pelo menos 1 dos éxons de genes codificadores humano. Nenhuma atenção foi dada para ESTs sem evidência de *splicing* que mapeavam parcial ou inteiramente em regiões intrônicas de genes codificadores de proteína.

Uma análise detalhada de uma coleção de cDNAs completos (*full-length*) de camundongo sugere que os RNAs não codificadores (ncRNAs) ocupam uma grande fração do transcriptoma [39]. Esta conclusão é baseada na anotação funcional de 60.770 cDNAs completos de camundongo, onde 15.815 eram ncRNAs, dos quais 71% não possuíam

evidência de *splicing* [39]. Estudos de expressão do cromossomo 7 humano permitiram Scherer *et al.* [40] estimarem um número de 3.700 ncRNAs humanos. Embora bancos de dados públicos contendo alguns ncRNAs de mamíferos estejam disponíveis [41, 42], a contagem e o mapeamento completo dos ncRNAs humanos ainda não foram realizados.

Plataformas de *microarray* que cobrem longas regiões genômicas sem utilizar a anotação dos genes são chamadas de *tiling arrays*. Por não possuir qualquer viés para éxons de genes codificadores de proteínas, estas plataformas permitem identificar regiões transcricionalmente ativas no genoma que ainda não foram detectadas por outras técnicas de expressão. Análises experimentais usando *tiling arrays* dos cromossomos 21 e 22 humanos detectaram a transcrição de um total de 5,3 kb de seqüências novas, transcritas dentro de regiões intrônicas de genes já bem caracterizados, das quais 2,7 kb (51%) são antisenso aos genes codificadores de proteína [43]. Além disso, *tiling arrays* que cobrem toda a região não repetitiva do genoma humano estenderam estas análises e detectaram cerca de 3 mil novas regiões transcricionalmente ativas intrônicas em fígado, sendo metade originada da fita de DNA oposta ao gene codificador de proteína (direção antisenso) [44]. Outros experimentos de *tiling arrays* também mostraram que alguns destes RNAs intrônicos são mensagens longas (400-2.000 pb) [45, 46].

A transcrição intrônica de ncRNAs também tem sido detectada usando outras abordagens experimentais. O mapeamento da ponta 3' de transcritos humanos (SAGE) revelou 575 etiquetas de expressão localizadas no interior de íntrons [47]. Apesar de alguns destes *tags* representarem éxons novos de genes conhecidos [47], uma fração deles podem ser originados de transcritos totalmente intrônicos. Outros estudos usando técnicas similares como CAGE [48] e MPSS [49] também encontraram evidências de transcrição intrônica. Ficou claro que íntrons, assim como regiões intergênicas, constituem as principais fontes de RNAs não codificadores de proteína [50].

1.4. Função dos RNAs Não-codificadores

RNAs não codificadores podem ser agrupados em classes específicas baseado no tamanho, motivos estruturais ou de seqüências, proteínas associadas ou localização sub-celular. Algumas classes de ncRNAs já bem conhecidas incluem tRNAs (RNAs transportadores) e rRNA (RNAs ribossomais) que servem como componentes essenciais para a maquinaria de síntese protéica; snRNAs (*small nuclear RNAs*), necessários para o processamento de mRNA imaturos e snoRNAs (*small nucleolar RNAs*), envolvidos na modificação de outros RNAs [19]. RNAs não codificadores com tamanhos muito curtos, os chamados siRNAs (*small interfering RNAs*) e miRNAs (*micro RNAs*) foram identificados [51, 52] e parecem desempenhar um papel regulatório importante na regulação dos genes codificadores de proteínas em diversos organismos [53]. Porém, o estudo sistemático destes RNAs que não possuem poder codificador de proteínas e seu reconhecimento como possíveis reguladores em diversos processos na biologia celular começaram apenas recentemente [11].

Os snoRNAs, em geral, possuem um comprimento de 60 a 300 nucleotídeos e guiam a modificação de nucleotídeos, de forma sítio-específica, em RNAs alvos via regiões curtas de pareamento de bases [50]. Existem duas classes principais, os snoRNAs box C/D e os snoRNAs H/ACA. Enquanto os primeiros direcionam a metilação do grupo 2'-hidroxila de ribonucleotídeos alvos (metilação da 2'-O-ribose), os segundos são os responsáveis pela conversão de uridina em pseudouridina (pseudouridilação) de rRNAs e snRNAs [54]. Embora quase todos os snoRNAs sejam derivados de íntrons de genes codificadores de proteínas ou genes não-codificadores, a presença do *cap* 5' em alguns transcritos sugere que parte dos snoRNAs C/D humanos sejam transcritos independentes [50]. Recentemente

[55], um snoRNA foi descrito como capaz de regular o *splicing* alternativo do receptor de serotonina que está situado num *locus* diferente ao do pequeno RNA (*trans-acting*).

Os miRNAs suprimem a tradução via um pareamento imperfeito com a região 3' não traduzida (3' UTR) dos mRNAs alvos [56] ou, no caso de uma complementaridade perfeita com o sítio alvo – fenômeno conhecido como RNA de interferência (RNAi) – causam a degradação do RNA alvo através do complexo RISC [50]. Até o momento, somente 457 miRNAs foram identificados em nosso genoma [57] e estima-se que aproximadamente um terço dos 25 mil genes codificadores de proteínas humanos sejam controlados por miRNAs [58].

A expressão de muitos miRNAs é regulada e seus papéis são centrais para uma ampla variedade de processos biológicos, como proliferação celular, apoptose, desenvolvimento temporal, metabolismo de gordura, expressão gênica neuronal, diferenciação muscular e divisão de células tronco [50]. Além disso, uma alteração na expressão de um miRNA também pode nos fornecer pistas sobre seu possível papel em muitas doenças genéticas, incluindo o câncer. Por exemplo, O'Donnell *et al.* [59] mostraram que o proto-oncogene c-Myc ativa a expressão de 2 miRNAs que, por sua vez, regulam negativamente a expressão do fator de transcrição E2F1, um ativador da progressão do ciclo celular. Vários outros estudos mostraram que a expressão de miRNAs está desregulada em células cancerosas [60-62] e que o perfil de expressão destes ncRNAs pequenos pode ser usado como ferramenta para a classificação tumoral [63, 64].

Alguns ncRNAs funcionais, muito maiores que os miRNAs e snoRNAs, também foram identificados. Estes longos RNAs podem possuir de centenas a milhares de nucleotídeos e desempenhar papéis regulatórios, como *imprinting* de genes codificadores de proteína, inativação do cromossomo X ou regulação de *splicing* alternativo [65].

O exemplo mais estudado de ncRNA longo, com 17 kb de comprimento, é o gene *Xist* (*X-inactive specific transcript*) que controla a inativação do cromossomo X em mamíferos [66]. Durante a diferenciação das células, o gene *Xist* pára de ser expresso em machos, assim como no cromossomo X ativo das células de fêmeas. O ncRNA *Xist* é então capaz de se ligar no cromossomo X que o transcreve e promover sua inativação, talvez por permitir o recrutamento de proteínas silenciadoras [67]. A expressão do gene *Xist*, por sua vez, parece ser regulada pelo ncRNA longo, *Tsix*, que possui direção oposta (antisense) ao gene *Xist* [67].

Um outro exemplo de ncRNA longo funcional é o gene *Air* envolvido no *imprinting* de 3 genes codificadores de proteínas humanos, expressos apenas no cromossomo materno (*Igf2r*, *Slc22a2*, *Slc22a3*) [68]. Através da inserção de sinais de poliadenilação que truncam o transcrito *Air*, Sleutels *et al.* [68] mostraram que este ncRNA antisense é necessário para o silenciamento dos genes codificadores no alelo paterno. Uma evidência adicional sobre silenciamento gênico e metilação de ilhas CpG mediado por ncRNA longo também foi reportada. Imamura *et al.* [69] mostraram que um transcrito parcialmente intrônico de 1.290 pb (*KHPS1a*), antisense ao gene *SPHK1*, regula a desmetilação de uma ilha CpG de forma tecido-específica, provavelmente pela interação em *cis* do ncRNA antisense e a sequência de DNA. Estes resultados indicam que os ncRNAs longos possuem um papel ativo no silenciamento gênico e *imprinting* genômico.

O *splicing* alternativo afeta mais de 70% dos genes humanos codificadores de proteínas [70], dos quais o *skipping* de éxon é o evento mais freqüente [71]. A correta excisão dos íntrons de um pré-mRNA e a junção dos éxons são dirigidas pelo Spliceossomo que reconhece sítios de *splicing* nas junções íntron/éxon ou elementos regulatórios localizados dentro de regiões intrônicas ou exônicas do pré-mRNA [72]. Foi demonstrado que RNAs naturalmente expressos [71] ou artificialmente introduzidos [73,

74] em células humanas, com a orientação antisense a estes sítios de *splicing* ou a elementos regulatórios do pré-mRNA, era capaz de promover o *skipping* de éxons. Um dado recente na literatura identificou um transcrito longo *unspliced* antisense que foi chamado de gene *Saf*. Este gene era um transcrito parcialmente intrônico não codificador antisense situado no primeiro íntron de *Fas*, um gene que codifica para uma proteína pró-apoptótica [75]. A super-expressão de *Saf* em células Jurkat induziu a expressão de diferentes isoformas alternativas de *Fas*. Esta super-expressão causou a retenção do éxon alternativo que possuía sobreposição ao *Saf* e o *skipping* de éxons 5' adjacentes, apontando para uma função *cis*-regulatória de RNAs parcialmente intrônicos antisense [75].

1.5. Biogênese dos RNAs Não Codificadores

As moléculas de RNA são transcritas no núcleo das células humanas por 3 RNA polimerases (RNAP). A RNAP I possui 14 subunidades e é responsável pela transcrição dos cerca de 400 genes ribossomais (rRNAs) humanos [76]. Os RNAs codificadores de proteína ou mRNAs e muitos genes que codificam os snoRNAs são transcritos pela RNAP II [77]. Por fim, a RNAP III transcreve uma coleção eclética de genes cuja principal característica em comum são que eles codificam RNAs catalíticos ou estruturais (tRNAs, rRNA 5S, U6, dentre outros) e que são, em regra, menores que 400 pb [78].

A maioria dos RNAs não codificadores caracterizados funcionalmente são os *micro RNAs* [58] e *small nucleolar RNAs* [54]. Os miRNAs são transcritos pela RNAP II como miRNAs primários (pri-microRNAs), com comprimentos que variam de centenas a milhares de nucleotídeos [79-81] e estão envolvidos, principalmente, no controle da tradução dos mRNAs [57]. Os snoRNAs são moléculas muito estáveis, localizadas no interior do nucléolo e que participam dos diferentes passos na biossíntese do ribossomo

[54]. Estes pequenos RNAs são geralmente liberados após a degradação, por exonucleotidasas, das pontas dos íntrons excisados de genes específicos, transcritos pela RNAP II [82].

Uma parcela considerável dos ncRNAs longos parece ser transcrita pela RNAP II já que muitos destes transcritos compartilham características estruturais com mRNAs, como a poliadenilação na cauda 3', *splicing* e *capping* 5' [83, 84]. Além disso, experimentos de imunoprecipitação da cromatina e *tiling-arrays* demonstraram que muitos sítios de ligação de fatores de transcrição da RNAP II estão localizados *upstream* a regiões transcricionalmente ativas intrônicas e intergênicas [85-88]. No entanto, o completo envolvimento da RNAP II na transcrição destes abundantes e diversos RNAs permanece desconhecido [88].

Recentemente, foi descrito que a transcrição de alguns RNAs mensageiros humanos e de roedores também pode ser mediada por uma RNA polimerase IV nuclear que ainda não era conhecida [89]. Esta proteína seria um produto de um *splicing* alternativo do mRNA da RNA polimerase mitocondrial. Neste artigo, Kravchenko *et al.* [89] também viram que a RNA polimerase IV nuclear é resistente à α -amanitina, um potente inibidor de RNA polimerase II, demonstrando que o tratamento de células em cultura com esta droga poderia ser de grande valia na identificação de transcritos regulados por esta nova RNA polimerase. Mostraram também que, ao contrário do esperado, alguns transcritos tinham sua expressão aumentada na presença de α -amanitina, e que a inibição da ação da RNA Polimerase IV com siRNA específico abole esta ativação [89]. No entanto, nenhum estudo foi realizado descrevendo a fração de RNAs não codificadores que são transcritos por esta RNA polimerase IV nuclear.

2. OBJETIVOS

Este trabalho procurou identificar e caracterizar RNAs não codificadores de proteína originados de íntrons humanos. Para isso, estudos em larga-escala destes transcritos intrônicos foram realizados usando diferentes técnicas experimentais e análises computacionais.

A análise do perfil de expressão dos RNAs intrônicos em diferentes tecidos, condições fisiológicas ou patológicas, foi usada para nos fornecer evidências sobre as possíveis funções dos transcritos intrônicos.

De forma complementar, a identificação e caracterização *in-silico* de milhares de RNAs total ou parcialmente intrônicos a genes codificadores de proteína visou catalogar os sítios de transcrição intrônica de todo o genoma humano e revelar novos candidatos funcionais de RNAs não codificadores.

Este trabalho também teve como objetivo estudar a biogênese dos RNAs intrônicos identificados em nossas análises. Experimentos de *microarrays* associados com o tratamento de andrógeno em células humanas ou com a inibição da RNA Polimerase II foram realizados para se identificar a fração destas mensagens intrônicas possivelmente transcrita pela RNA Polimerase II ou, alternativamente, por uma recém descoberta RNA Polimerase IV nuclear.

3. MATERIAIS E MÉTODOS

3.1. RNAs Intrônicos Correlacionados ao Grau de Diferenciação Tumoral

3.1.1. Construção do *Microarray* 4k Intrônico

Todos os clones ORESTES selecionados tiveram seu DNA amplificado, usando oligonucleotídeos iniciadores (*primers*) complementares ao vetor de clonagem. Os produtos de PCR foram purificados e concatenados em placas de 384-poços e, em seguida, depositados na lâmina de vidro utilizando o equipamento *GenIII microarray spotter robot* (Molecular Dynamics).

O *chip* construído no laboratório possui 12 *sub-arrays*, com cerca de 384 fragmentos de DNA (ou sondas) imobilizados em cada *sub-array*. Todo este conjunto de 12 *sub-arrays* está depositado em duplicata (lado esquerdo e direito da lâmina). Dos 4.608 elementos de cDNA depositados em duplicata (totalizando 9.216 elementos), aproximadamente 1.920 representam genes sabidamente envolvidos em câncer, e 619 representam regiões intrônicas transcricionalmente ativas. As sondas restantes foram reservadas para servir de controle do *chip*. Genes constitutivos (*housekeeping genes*) e genes conhecidamente expressos em câncer de pulmão e próstata serviram de controle positivo. Um gene que codifica para uma proteína de cloroplasto de planta (PORA), sem similaridade com o genoma humano e de bactéria, foi usado como controle negativo da hibridização. Ainda, um conjunto de cDNAs de uma placa de 384-poços do *kit Lucidea Microarray ScoreCard* (Molecular Dynamics) também foi depositada. Além de adicionar mais controles positivos e negativos à lâmina, este *kit* contém fragmentos de DNA correspondentes a regiões intergênicas de levedura, que não hibridizam com nenhum cDNA humano, só hibridizando com uma mistura de mRNA complementar contido no *kit* e servindo, portanto, de um controle adicional da marcação da sonda. Os controles

presentes no *kit Scorecard* permitem determinar a faixa de detecção linear da intensidade do sinal de hibridização (*dynamic range controls*) e a eficiência relativa da marcação do RNA com nucleotídeos ligados a fluoróforos (*ratio controls*).

3.1.2. Extração de RNA de Próstata

As amostras de tecido normal e tumoral, obtidas de pacientes que sofreram a remoção cirúrgica da próstata, são provenientes de um convênio com o Hospital do Câncer/ RJ do Instituto Nacional de Câncer do Rio de Janeiro e com o Hospital Sírio Libanês/ SP. Todos os pacientes assinaram um Consentimento Informado, e a pesquisa foi aprovada pelos respectivos Comitês de Ética dos Hospitais envolvidos. Uma amostra da próstata é congelada em nitrogênio líquido até 10 minutos após ter sido removida do paciente. Antes do processamento para extração de RNA, a amostra é macro-dissecada por um patologista a fim de remover tecidos normais adjacentes e garantir que as amostras de tumor de próstata analisadas contenham no mínimo 80% de células malignas.

Inicialmente o tecido congelado em N₂ líquido é macerado no mortar. O RNA mensageiro (mRNA) é obtido utilizando-se o *kit MACS mRNA isolation* (Miltenyi Biotec) segundo o protocolo do fabricante. Resumidamente, as células são lisadas e passadas num filtro para retenção dos restos celulares. Em seguida oligos-dT ligados a partículas magnéticas são hibridizados com as moléculas de mRNA. Um campo magnético é aplicado sobre o filtro por onde é passada a solução, isto retém os mRNAs que estão associados ao oligo-dT. Em seguida, são feitas lavagens para eliminação de DNA e RNA total e, ao final, o mRNA é recuperado através da passagem pelo filtro de água sem RNase aquecida, que destrói a ligação com o oligo-dT.

Para se determinar a qualidade e a quantidade do mRNA e RNA total extraídos, as amostras foram analisadas no equipamento *Bioanalyser 2100* (Agilent Technologies). No *Bioanalyser* as amostras de RNA total ou mensageiro são submetidas a uma eletroforese microfluídica e os fragmentos são identificados por fluorescência, de acordo com o tempo de migração, por comparação a um *ladder* de RNA (composto por 6 transcritos de RNA de tamanhos bem definidos). Nos ensaios de RNA total, os picos referentes às bandas ribossomais 18S e 28S são detectados dinamicamente baseados na corrida do *ladder*. A visualização do resultado pode ser feita através da reconstituição virtual da imagem de um gel representativo ou de um eletroferograma.

3.1.3. Marcação do RNA e Hibridização

Os mRNAs extraídos de amostras tumorais e normais são marcados durante a transcrição reversa, quando análogos de nucleotídeos ligados a fluoróforos, que emitem em diferentes comprimentos de onda (Cy3-dCTP e Cy5-dCTP), são incorporados à primeira fita do cDNA. Durante a reação de marcação é feita a adição de uma mistura de mRNAs controles contidos no *kit Lucidea Microarray ScoreCard*.

Cada amostra (2 µg de mRNA) foi marcada separadamente tanto com Cy3-dCTP quanto com Cy5-dCTP, para que a intensidade medida pelo mesmo mRNA não sofresse um viés devido às diferenças de incorporação e emissão destes fluoróforos.

As amostras marcadas foram purificadas e hibridizadas contra nosso *chip* de cDNA. Os níveis de expressão relativa, para cada gene representado na lâmina, foram determinados quando tal lâmina é varrida por um *scanner*, em comprimentos de onda de 550 nm (Cy3) e 650 nm (Cy5).

3.1.4. Análise dos Resultados de *Microarrays*

Após o *scanning* das lâminas, os dados são extraídos de cada imagem utilizando-se o *software ArrayVision 6.0* (Imaging Research Inc.). Os dados brutos correspondentes a emissão de fluorescência de Cy3 e Cy5 (descontado o *background*) para cada sonda são, então, importados para o *software Lucidea Microarray ScoreCard* (Amersham Bioscience). Este *software* usa os dados de alguns genes controles, depositados no *chip* e os que foram hibridizados com o mRNA do *kit*, para reportar a qualidade da hibridização.

A normalização dos dados é feita primeiramente, usando o programa implementado com base no algoritmo LOWESS (*locally weighted linear regression algorithm*), utilizando o pacote R (<http://www.r-project.org/>). Esta etapa visa corrigir o viés sistemático dos dados originados pela diferença de incorporação e emissão de fluorescência de Cy3 e Cy5.

Após esta normalização, temos, para cada amostra marcada e hibridizada, valores de fluorescência em quadruplicata, 2 referentes ao mRNA marcado com Cy3 (lado esquerdo e direito) e 2 referentes a marcação do mesmo mRNA com Cy5 normalizado, também em duplicata. Foi observado que a taxa de erro dos dados (desvio-padrão em relação à média) varia de acordo com a faixa de intensidade de sinal dos genes. Com o objetivo de detectar e eliminar os dados espúrios ou pouco reprodutíveis, o programa *ArrayStat* (Imaging Research Inc.) foi utilizado. Este programa usa um modelo estatístico para eliminar dados com desvios acima do desvio-padrão observado, para a faixa de intensidade do gene analisado, através de uma curva de erro ajustada.

Algumas outras normalizações foram incluídas na análise para melhorar a comparação de nossos dados. A normalização por *sub-array* foi usada para corrigir possíveis diferenças de intensidade, emitida pelas sondas, dentro de uma mesma lâmina, devido a uma hibridização irregular ou presença de manchas. Esta normalização usa o

valor das réplicas das intensidades de genes constitutivos, presentes em todos os *sub-arrays*, assumindo que não devam variar entre um *sub-array* e outro. O uso de médias internas, excluindo os 40% extremos, permitiu uma melhor normalização das intensidades médias, entre os pacientes, pois excluiu uma porcentagem de pontos espúrios (20% dos pontos com maior sinal e 20% dos pontos de menor sinal), que acabavam afetando a média não normalizada. Desta forma, estas normalizações corrigiram as diferenças entre as intensidades individuais de cada sonda e globais das lâminas dos pacientes analisados, para que estas se tornem mais comparáveis.

O coeficiente de Pearson dos genes foi calculado para se determinar a existência de correlação entre o padrão de expressão gênica e o Grau de Gleason (GS) do tumor. Para se determinar uma significância estatística da correlação encontrada, foram realizadas 50.000 permutações das classes de GS de cada amostra, e após cada permutação o coeficiente de Pearson foi re-calculado [90]. Através de um método de classificação supervisionada, foi gerada uma lista contendo os genes que, através de sua diferença de expressão, mais discriminassem tumores com GS baixo, geralmente associados a um prognóstico favorável (tumores GS 5 e 6), de tumores com GS alto, geralmente casos com mau prognóstico (tumores GS 9 e 10). Para isso, usou-se 11 amostras de tumor (5 de GS alto e 6 de GS baixo) e calculou-se o coeficiente de Pearson entre a expressão gênica e o GS. Foram encontrados nesta análise, 56 genes diferencialmente expressos, num intervalo de confiança maior que 99,9 %.

3.1.5. Caracterização dos Transcritos Intrônicos

3.1.5.1. Ensaios de RACE

Ensaios de RACE (*Rapid Amplification of cDNA Ends*) utilizando-se cDNAs comerciais de diversos tecidos (*Marathon-Ready cDNA*, Clontech), contendo adaptadores em cada uma das extremidades, foram realizados para determinar o tamanho real de transcritos. Uma biblioteca de cDNAs clonados direcionalmente em vetor lambda de bacteriófagos, com adaptadores específicos para as extremidades 5' e 3' dos transcritos, extraídos de um adenocarcinoma colo-retal (*5'-Stretch PLUS cDNA Library*, Clontech) também foi utilizada. Nesta biblioteca direcional de fagos é possível confirmar, por RACE, a sequência e a direção do transcrito, visto que só há uma combinação certa entre o *primer* específico do gene e o *primer* do adaptador, no caso, vetor.

3.1.5.2. Northern blot fita-específico

Membranas de nitrocelulose contendo mRNA de três a cinco linhagens de células foram construídas com a ajuda da doutoranda Fernanda Festa (orientada pela Dra. Mari Sogayar). Todas as linhagens de células possuem a morfologia de tecido epitelial prostático e sua cultura seguiu rigorosamente as recomendações de seu fornecedor (ATCC – *American Type Culture Collection*). Uma membrana comercial, contendo mRNA de múltiplos tecidos (*Multiple Tissue Northern Blot*, Clontech) também foi utilizada nos experimentos de Northern blot, descritos a seguir.

Oligonucleotídeos de fita simples foram marcados radioativamente e hibridizados contra estas membranas. Utilizamos para a marcação uma enzima quinase (*T4 Polynucleotide Kinase*, Invitrogen) que é capaz de transferir o fosfato radioativo do γ -³²P-ATP para a ponta 5' do RNA ou DNA. Este oligo que possuía sua sequência idêntica à fita

do mRNA de *RASSF1*, porém situado na região intrônica correspondente ao clone ORESTES, depositado no *chip*, iria hibridizar assim, contra um transcrito antisenso contido na membrana comercial utilizada (*Multiple Tissue Northern Blot*, Clontech).

Para aumentar a eficiência de marcação da sonda e melhorar o tempo de exposição e a temperatura de hibridização, usou-se o *kit SP6/ T7 Transcription* (Roche) e o protocolo citado no artigo de Yelin *et al.* [37]. O inserto correspondente à sequência intrônica, contida no *chip*, era clonado em um vetor pGEM-T (*pGEM-T Easy Vector System I*, Promega) que possui um promotor para T7 RNA polimerase de *E.coli*. A direção do transcrito clonado era determinada e o fragmento de DNA com o promotor e o inserto era linearizado, através de uma amplificação por PCR, utilizando-se *primers* específicos do vetor (situado numa região *upstream* ao sítio de ligação da T7 RNA polimerase) e do inserto. A síntese de uma sonda de RNA fita simples, marcada, era feita a partir do fragmento purificado e da adição de T7 RNA polimerase, ribonucleotídeos frios (rATP, rCTP, rGTP e rUTP) e α -³²P-rCTP radioativo. Esta sonda, complementar a um possível transcrito antisenso, era hibridizada contra a membrana construída com as linhagens de células. Com este método, foi possível observar a hibridização de um transcrito antisenso para o gene da PPHLN1, com um tamanho parecido ao obtido nos experimentos de RACE, em três linhagens testadas. Além disso, o transcrito referente à fita senso do gene TLE3, correspondente ao éxon novo, também foi detectada nestes ensaios de Northern blot.

3.1.5.3. Transcrição Reversa Fita-Específica

Para validar a hipótese de que os RNAs com cauda poli-A do nosso classificador eram transcritos antisensos, fizemos uma transcrição reversa fita-específica usando *primers* para a região do clone, no nosso *chip* de DNA, seguido de uma amplificação de uma região

interna por PCR (RT-PCR fita-específica). O mRNA de diferentes amostras tumorais era tratado com DNase livre de RNase (*kit RQ1 RNase-Free DNase*, Promega), conforme o protocolo do fabricante. Para cada gene candidato, eram realizadas 4 reações de RT-PCR usando o mesmo mRNA (400 ng de mRNA por reação), obtido de linhagem ou tecido. Na primeira reação, apenas o *primer* convencionalmente denominado *reverse* (com a sequência complementar ao intron da fita senso do gene conhecido) estava presente. Nesta reação, esperava-se medir a quantidade de RNA imaturo presente na preparação, já que apenas a fita contendo o intron seria reversamente transcrita. Na segunda reação, foi adicionado apenas o *primer* convencionalmente denominado *forward* (com a sequência idêntica ao intron da fita senso do gene conhecido) para transcrever reversamente apenas transcritos antisensos intrônicos. A terceira e a quarta reações foram usadas como controle negativo da transcrição reversa fita-específica. Na terceira, eram adicionados todos os reagentes menos a enzima transcriptase reversa (controle de contaminação genômica). Para saber se existia um auto-anelamento (*self-annealing*) do mRNA na produção de cDNA, na quarta reação foi adicionada a transcriptase reversa porém não foi adicionado qualquer *primer*. Foi usado o protocolo do *kit* (*SuperScript First-Strand Synthesis System for RT-PCR*, Invitrogen Life Technologies) voltado para evitar a formação de estrutura secundária do mRNA na transcrição reversa. Após a síntese deste cDNA específico, um PCR quantitativo em tempo real (*Real-Time PCR*) era realizado, utilizando-se pares de *primers* internos. A sequência do produto foi confirmada e os controles negativos mostraram que não havia contaminação com DNA genômico nem auto-anelamento do mRNA de forma significativa.

3.1.5.4. PCR Quantitativo em Tempo Real

O mRNA é transcrito reversamente em um cDNA, que será amplificado por PCR. O ciclo do PCR em que a intensidade de fluorescência do *SybrGreen* (*kit SybrGreen PCR Core Reagent*, Applied Biosystems), ligado à dupla fita de DNA amplificada é detectada, determina o Ct do gene (*Cycle threshold*). Este Ct representa, de certa forma, a quantidade inicial de moléculas de cDNA presentes para aquele gene. Quanto menor o Ct, mais expresso seria o gene. Para normalizar o efeito da quantidade total de cDNA da amostra, usa-se também, como referência, o Ct de um gene constitutivo (β -*tubulina*). O delta-Ct (Δ Ct), ou seja, a diferença entre o Ct do gene em estudo, em cada uma das 3 réplicas, e o Ct médio da tubulina, é calculado. O valor calculado como $2^{-\Delta Ct}$ nos dá a diferença de expressão relativa do transcrito em relação à tubulina. O erro padrão da média é calculado pela divisão do desvio-padrão das expressões relativas sobre a raiz quadrada do número de amostras.

Desenhou-se pares de *primers* específicos para cada gene escolhido, usando o *software Primer Express v2.0* (Applied Biosystems). α -*Actina* e β -*tubulina* foram escolhidas como genes de referência. A padronização dos *primers* foi realizada, visando otimizar a melhor concentração destes *primers* em cada experimento. Como a quantidade de *primers* altera a velocidade de amplificação do cDNA, foi possível, com este experimento, determinar a menor concentração de *primers* necessária para saturar a reação (concentração de *primer* não-limitante).

3.2. RNAs Intrônicos Respondedores à Andrógeno

3.2.1. Tratamento das Células e Obtenção e Marcação de RNA Total

Células de adenocarcinoma de próstata metastática, LNCaP, foram cultivadas com o meio e condições recomendados pela ATCC (*American Type Culture Collection*). Cerca de 2×10^6 células foram plaqueadas (placas de 10 cm²) e após um dia, trocou-se o meio de cultura para um meio RPMI suplementado com 10% de soro fetal bovino, tratado com carvão ativado gelado, por 1 hora sob agitação constante e seguido de diversas filtrações. O objetivo deste tratamento seria de retirar do meio de cultura hormônios que pudessem interferir na resposta ao andrógeno sintético. Após um período de 24 horas, as células foram expostas à dose de 1 nM de hormônio Methyltrienolone (R1881), um agonista sintético do andrógeno, ou veículo (etanol 0,01%) por 6, 9, 12, 18, 24 e 48 horas em um novo meio também tratado com carvão ativado. Este processo foi realizado principalmente pela aluna de doutorado da Profa. Mari Sogayar, Fernanda Festa.

O RNA total de cada condição (com ou sem tratamento com andrógeno) nos diversos tempos experimentais (6, 9, 12, 18, 24 e 48 horas) e no tempo zero foi extraído com o *kit RNeasy* (Qiagen), de acordo com o protocolo do fabricante. Extraíu-se também o RNA total de 3 linhagens de próstatas (LNCaP, PC3 e DU145), sem qualquer tratamento, para servir de referência nos experimentos de *microarrays*. O RNA obtido das diversas extrações foi agrupado de acordo com sua condição e tempo experimental.

Para retirar uma possível contaminação genômica nas amostras de RNA extraídas, fizemos um tratamento com DNase I, seguida de uma purificação do RNA tratado, através do *kit On-Column DNase Digestion* (Qiagen) e segundo o protocolo do fabricante. Para nos certificarmos que o tratamento havia funcionado, reações controle-negativo de PCR com as amostras de RNA, tratadas ou não com DNase I, foram realizadas utilizando-se um

par de *primers* para uma região genômica de um intron de tubulina. A qualidade do RNA total foi também analisada através da leitura da absorbância nos comprimentos de onda 260 nm e 280 nm (razão 260/280 próxima de 2) e da intensidade das bandas referentes ao 18 S e 28 S em eletroforese em gel de agarose 1,2%.

Após o tratamento com DNase I, 10 µg de RNA total de cada amostra foi transcrito reversamente usando análogos de nucleotídeos ligados a fluoróforos, que emitem em diferentes comprimentos de onda (normalmente Cy3-dCTP e Cy5-dCTP) e que são incorporados a primeira fita do cDNA. O RNA obtido das linhagens tratadas com R1881 ou etanol foi marcado com Cy5-dCTP, segundo o protocolo do fabricante do *kit CyScribe first-strand cDNA labeling kit* (Amersham Bioscience - GE Healthcare) e utilizando-se oligo-dT para a síntese do cDNA marcado. O mesmo processo de marcação, porém com Cy3-dCTP, foi usado para marcar o RNA de referência (contendo o RNA extraído das 3 linhagens de próstata) nos experimentos de *microarrays*. Junto com cada amostra de RNA e logo antes da marcação, mRNAs controles, contidos no *kit Lucidea Microarray ScoreCard* (Amersham Bioscience - GE Healthcare) foram adicionados para servir de controle de qualidade da marcação e hibridização dos experimentos. A marcação e hibridização do RNA de referência com Cy3-dCTP, assim como a marcação e hibridização com Cy5-dCTP do RNA de cada uma das condições e tempo de tratamento foram realizadas duas vezes, em dois *chips* diferentes, nos fornecendo portanto, uma réplica-técnica para cada ponto experimental.

Após o *scanning* dos *chips*, os dados são extraídos de cada imagem utilizando-se o *software ArrayVision 6.0* (Imaging Research Inc.). Os dados brutos correspondentes a emissão de fluorescência de Cy3 e Cy5 (descontado o *background*) para cada sonda são, então, importados para o *software Lucidea Microarray ScoreCard* (Amersham Bioscience - GE Healthcare). Este *software* usa os dados de alguns genes controles, depositados no

chip e os que foram hibridizados com o mRNA do *kit*, para reportar a qualidade da hibridização.

3.2.2. Experimentos de *Microarrays*

3.2.2.1. Limite de Detecção e Normalizações

Para definir se um determinado fragmento de DNA (sonda) depositado no *chip* estava ou não expresso nas células LNCaP em qualquer condição ou tempo de tratamento, usamos como critério seu valor bruto de expressão comparado a um *background* local. O valor do *background* local foi estabelecido para cada um dos 12 *sub-arrays* e para cada um dos *dyes* (Cy3 ou Cy5) como: a intensidade bruta (sem a diminuição da intensidade de fundo para a sonda) do controle negativo do *scorecard* (*putative transmembrane protein Glp*, um gene de planta) mais 1 desvio-padrão (referente à variação nas intensidades dos *pixels* da sonda). Este valor foi então comparado à intensidade bruta de cada uma das sondas daquele *sub-array* e marcados com o mesmo *dye*.

Com o objetivo de minimizar a variação técnica encontrada neste tipo de experimento de larga escala, usamos a média interna excluída de 40% das intensidades extremas como base no fator de correção. Este método assume que apenas uma pequena fração de genes deve variar significativamente entre os diversos períodos de tempo de tratamento e que a grande maioria permanece com o nível de expressão constante entre os *chips*.

O uso do RNA de referência que foi marcado com Cy3-dCTP e hibridizado em todas as lâminas não se mostrou um bom método de normalização. A análise por coeficiente de variação mostrou que as lâminas réplicas corrigidas pelo sinal do RNA de referência não eram corrigidas tão bem quanto o método citado no parágrafo anterior. A pior

normalização também resultou num número menor de genes diferencialmente expressos com significado estatístico, após o tratamento com andrógeno (descrito no item a seguir). Por isso, todas as análises de expressão foram feitas usando apenas os valores referentes ao RNA teste marcado com Cy5-dCTP das células tratadas com andrógeno ou veículo, e os dados normalizados entre si, usando a média interna das intensidades de todos os genes de cada lâmina, com exclusão de 20% das sondas com maior e menor intensidades.

3.2.2.2. Análises de Expressão por SAM

Para detectar se houve uma variação significativa de expressão de um determinado gene após o tratamento com andrógeno, foi utilizado o algoritmo SAM (*Significance Analysis of Microarrays*) [91]. Este algoritmo, implementado em um programa disponível no site <http://www-stat.stanford.edu/~tibs/SAM/>, confere uma pontuação (*score*) para cada gene baseado na mudança de expressão deste gene em relação ao desvio-padrão de diversas medições repetidas. Para os genes com *score* maior do que um valor mínimo ajustável (*threshold*), o programa usa repetidas medidas de várias permutações para estimar a porcentagem, ou FDR (*false discovery rate*), destes genes que seriam identificados ao acaso. Cada gene encontrado também está associado a um valor de q que mede a sua significância de forma parecida com o valor de p , mas adaptado para a análise de um grande número de genes e permutações. Na verdade, este valor de q representa o menor FDR para o qual este gene ainda é chamado de significativo.

Para a seleção de genes diferencialmente expressos foi feito o seguinte procedimento: (I) separamos os 6 tempos experimentais (6, 9, 12, 18, 24 e 48 horas) em 5 grupos de 3 tempos consecutivos (6h + 9h + 12h ou 9h + 12h + 18h, etc.) para as células que receberam tratamento (classe 1) e as que não receberam (classe 2); (II) selecionou-se

para a análise por SAM, apenas os genes expressos em pelo menos uma das condições, ou seja, genes com uma expressão maior do que o *background* local em metade ou mais da metade das lâminas, em uma das 2 classes (com ou sem tratamento), de um grupo de 3 tempos consecutivos; (III) o SAM foi executado utilizando-se os seguintes parâmetros: *two-class response (paired data)*, 1.000 permutações, *K-Nearest Neighbors Imputer*, *fold-change* $\geq 1,5$ e *FDR* $< 5\%$.

3.2.3. Caracterização de RNAs Intrônicos

Os experimentos de RT-PCR fita-específica foram realizados com uma quantidade de 4 μg de RNA total de células tratadas ou não com andrógeno e conforme o protocolo descrito no item 3.1.5.3.

Para os ensaios de RACE utilizou-se uma biblioteca comercial de cDNA dupla-fita de tecido de próstata (*Marathon-Ready cDNA*, Clontech) e o protocolo descrito no item 3.1.5.1.

A membrana utilizada nos experimentos de Northern blot fita-específico com o mesmo protocolo do item 3.1.5.2. (*kit SP6/ T7 Transcription*) contém 2 μg de mRNA poli-A+ de diversos tecidos (Human MTN Blot II).

Experimentos de PCR quantitativo em tempo real (protocolo semelhante ao item 3.1.5.4.) também foram utilizados. Para calcular a diferença de expressão do mesmo transcrito em diferentes condições (com ou sem tratamento com andrógeno) foi usada a diferença entre o ΔCt das células tratadas e o ΔCt das células não tratadas, o chamado " $\Delta\Delta\text{Ct}$ ". O valor calculado como $2^{-\Delta\Delta\text{Ct}}$ é a diferença de expressão do transcrito em uma condição comparado com sua expressão em outra. O erro padrão da média é calculado pela

divisão do desvio-padrão das expressões relativas sobre a raiz quadrada do número de amostras. Nos gráficos que mostram o erro, apenas a barra acima da média é exibida.

3.2.4. Imunoprecipitação da Cromatina

Ao meio de cultura das células LNCaP (tratadas com andrógeno ou veículo) foi adicionado formaldeído 1% e as placas foram mantidas à temperatura ambiente por 10 minutos, a fim de promover ligações cruzadas para fixação das proteínas associadas à cromatina. Em seguida, o meio foi incubado com glicina 125 mM por 5 min, para interromper a formação de ligações cruzadas, e, em seguida, removido. As células foram lavadas por 2 vezes com PBS gelado contendo os inibidores de proteases PMSF 1 mM, aprotinina 1 µg/mL e pepstatina A 1 µg/mL (Sigma). Após a adição de 1 mL de PBS com os inibidores, com as placas sobre gelo, as células foram raspadas e transferidas para tubos de 1,5 mL. As suspensões foram centrifugadas a 3.000 rpm por 5 min a 4 °C e o sobrenadante foi removido. O sedimento de células foi ressuspensionado em 600 µL de tampão de lise (SDS 1%; EDTA 10 mM; Tris 50 mM, pH 8,1; contendo uma mistura comercial (Upstate) de inibidores de protease) para uma concentração aproximada de 2×10^6 células/100 µL de tampão. Para a fragmentação da cromatina, os extratos celulares foram submetidos a 15 pulsos de 10 s na intensidade de 11% do aparelho *Sonic Dismembrator Model 500* (Fischer Scientific), tomando-se o cuidado de evitar a formação de espuma, que revelaria a desnaturação das proteínas, não desejável. Esta condição gera fragmentos de DNA genômico entre 200 e 1000 pb, faixa de tamanho adequada para realização da ChIP [92].

Para a realização da imunoprecipitação da cromatina, foi utilizado o *kit EZ ChIP* (Upstate), seguindo protocolo recomendado pelo fabricante. Em resumo, após a lise celular

e fragmentação da cromatina, o procedimento inclui as seguintes etapas: (a) eliminação de ligantes de proteína G inespecíficos, (b) incubação com o anticorpo de interesse, (c) precipitação do complexo anticorpo-cromatina com proteína G, (d) lavagens do imunoprecipitado com soluções salinas de estringência crescente e (e) eluição da cromatina. Segue-se a reversão das ligações cruzadas, digestão das proteínas da cromatina com Proteinase K, purificação e detecção dos fragmentos de DNA co-imunoprecipitados [92].

Para a imunoprecipitação do Receptor de Andrógeno (AR), foi utilizado o anticorpo policlonal IgG H-280 (Santa Cruz), específico para a região N-terminal do AR. Aliquotas de 100 µL do extrato celular (obtido após sonicação de células tratadas com andrógeno ou veículo) foram diluídas em 1 mL de tampão de diluição e utilizadas para cada imunoprecipitação. Uma fração de 10 µL (0,01 %) das amostras diluídas foi separada para servir como controle da quantidade inicial de DNA presente (*DNA input*), e as imunoprecipitações foram realizadas por 14 h de incubação com 3 µg do anticorpo H-280 ou com 1 µg de IgG normal de camundongo, utilizado como controle negativo. O DNA resultante da imunoprecipitação e o *DNA input* foram purificados e a imunoprecipitação do complexo AR-cromatina foi confirmada por PCR com *primers* específicos flanqueando o ARE III (*Androgen Response Element III*) presente no *enhancer* do gene PSA (*Prostate Specific Antigen*) (controle positivo). Para os outros fragmentos de interesse, a análise por PCR foi realizada utilizando-se *primers* específicos, em reações de 50 µL contendo 2 µL do produto imunoprecipitado. Como controle da inespecificidade na imunoprecipitação da cromatina, utilizamos *primers* para amplificação de uma sequência interna do gene da tubulina sabidamente não ativado por andrógeno (controle negativo).

3.3. Identificação e Caracterização *in silico* de RNAs Intrônicos

3.3.1. Cruzamento das Coordenadas Genômicas de Transcritos de Diferentes Conjuntos de Seqüências

O conjunto de seqüências analisadas é composto por todas as seqüências humanas de RefSeq, mRNAs e ESTs das quais as coordenadas genômicas foram obtidas do Genome Browser (<http://genome.ucsc.edu>) (hg17; NCBI Build 35). Primeiro, seqüências com baixa qualidade de alinhamento (cobertura menor que 0.70 e identidade menor que 0.90) ou que mapeavam em mais de uma região genômica foram removidas. Segundo, nós descartamos seqüências com padrões complicados de arranjo como os receptores de células T e genes de imunoglobulinas. ESTs e mRNAs que alinhavam a 2 ou mais RefSeq não sobrepostos foram filtradas como suspeitas de seqüências quimeras. Erros de sequenciamento de transcritos alinhados à seqüência genômica poderiam levar a espaços que poderiam ser interpretados como íntrons por nossos programas. Para evitar estes chamados falsos íntrons, nós juntamos os éxons adjacentes de todo “íntron” menor que 30 bases.

Seqüências de ESTs que possuíam éxons sobrepostos foram agrupadas em *clusters* de ESTs, usando suas coordenadas genômicas. Seqüências de RefSeqs e mRNAs foram processadas separadamente e divididas de acordo com a fita de DNA genômico que eles mapeavam e depois, sub-divididas em mensagens *spliced* ou *unspliced*. Seqüências de uma mesma fita de DNA em cada grupo de mensagens *spliced* e *unspliced* foram agrupadas em uma unidade transcricional quando eles possuíam éxons sobrepostos num mesmo *locus* genômico. O alinhamento deste conjunto de mRNAs contra o conjunto de RefSeqs foi realizado para identificar variantes adicionais de *splicing*, transcritos intrônicos e transcritos antisensos representados apenas na coleção de mRNAs.

Resultando da combinação das seqüências de RefSeqs e mRNAs descrita acima, um conjunto de dados de referência foi definido compreendendo as 15.783 unidades transcricionais de RefSeq não redundantes *spliced* e as evidências adicionais de variantes de *splicing* obtidas por cada unidade transcricional de todas as seqüências de mRNAs que mapeavam no mesmo *locus*. Este conjunto de referência foi comparado com o conjunto total de *clusters* de ESTs com o objetivo de definir quais *clusters* de ESTs eram exônicos ou totalmente intrônicos aos genes de nossa referência.

Seqüências *contigs* de ESTs parcialmente intrônicas foram identificadas num passo posterior, ao procurar por evidência de 2 ou mais ESTs *unspliced* que se sobrepunham a um éxon e se estendiam em mais de 30 bases de cada lado na região intrônica que flanqueava o éxon.

3.3.2. Cálculo da Frequência de *Skipping* de Éxon

Para cada éxon de um gene do conjunto de RefSeq nós contamos o número de vezes que este mapeava em um éxon (# em éxon) ou em um íntron (# em íntron) em todas as seqüências de mRNA de um mesmo subgrupo (mRNAs e RefSeqs de um mesmo *locus* e de uma mesma fita de DNA). A frequência de *skipping* de éxon (ES) é dado por:

$$ES = 1 - [\# \text{ em íntron} / (\# \text{ em íntron} + \# \text{ em éxon})].$$

3.4. Estudo da Biogênese e Especificidade de Expressão por Tecido dos ncRNAs Intrônicos

3.4.1. Obtenção de Amostras de Tecidos Humanos

RNA total foi purificado de 2 *pools* de tecidos de fígado, cada um contendo amostras de 5 indivíduos e de 4 *pools* de tecidos de rim, obtidos de 17 pacientes. Além disso, 2 amostras de tumor de próstata foram usadas. Todas as amostras foram obtidas de pacientes que assinaram um Consentimento Informado de doação e receberam a aprovação de comitês de ética dos respectivos hospitais. A purificação do RNA total foi feita utilizando Trizol (Life Technologies), de acordo com as instruções do fabricante e seguidas de um tratamento com DNase I, utilizando-se o protocolo “on-column DNase digestion” do *kit Qiagen RNeasy* (Qiagen) para remover uma possível contaminação genômica de DNA. A pureza de todas as amostras de RNA foi checada pelo espectrofotômetro *ND-1000* (NanoDrop Technologies) e sua integridade avaliada por eletroforese no aparelho *BioAnalyzer 2100* (Agilent Technologies).

3.4.2. Experimentos de Células LNCaP tratadas com α -amanitina

Células de câncer de próstata, LNCaP, foram obtidas da ATCC e mantidas em meio RPMI 1640 (Gibco), suplementada com 10% (vol/vol) de FCS (fetal calf serum), 3 mM L-Glutamina, 100 μ g/ml estreptomicina e 100 U/ml penicilina, a 37°C e 5% CO₂. Para os experimentos de inibição da RNAP II, 8 x 10⁵ células foram plaqueadas em placas p60 e mantidas em cultura por 2 dias quando o meio foi substituído por um meio fresco com 50 μ g/ml α -amanitina (Roche) ou com veículo. Após 24h, as células foram lavadas em PBS gelado, coletadas e o *pellet* foi mantido a -80°C. RNA total foi isolado usando o *kit Qiagen RNeasy* e tratado com DNase I, seguindo o protocolo “on-column DNase digestion”

(Qiagen). A qualidade do RNA foi analisada como descrito acima. Duas réplicas biológicas foram processadas separadamente.

3.4.3. Marcação das Amostras e Experimentos de Hibridização

A amplificação do RNA poli-A foi feita por uma T7-RNA polimerase partindo de 300 ng de RNA total. Para se obter o cRNA marcado com Cy3 ou Cy5 nós utilizamos o *kit Agilent Low RNA Input Fluorescent Linear Amplification* (Agilent Technologies). A marcação do cRNA amplificado por T7-polimerase substitui com grande vantagem a marcação de cDNA por transcriptase reversa (RT) por garantir que a fita original do mRNA molde seja preservada. A marcação por RT pode eventualmente gerar uma segunda fita de cDNA complementar e marcar, de forma artefactual, um alvo que possui uma direção oposta à mensagem original. Para as amostras de células LNCaP (tratadas ou não com α -amanitina) 500 ng de RNA total foi usado junto com um *spike* de mRNAs sintetizados *in vitro* (*Agilent RNA Spike-In kit*) para a amplificação e marcação. Para amostras de rim, os 4 *pools* de indivíduos foram considerados como réplicas e 2 foram marcados com Cy3 e os outros 2 com Cy5. Para cada amostra de *pool* de fígado, amostra de tumor de próstata ou amostra de células LNCaP a marcação foi feita separadamente em réplica com Cy3 e Cy5. A hibridização de 750 ng de cRNA marcado com Cy3 ou Cy5 foi realizada com o *kit Agilent In situ Hybridization kit-plus* e seguindo as recomendações do fabricante (Agilent Technologies). Um total de 6 *chips* de *microarray* intrônico de 44k foi utilizado. Os slides foram lavados e processados de acordo com o protocolo *Agilent 60-mer Oligo Microarray Processing protocol* e lidos no *GenePix 4000B scanner* (Molecular Devices). Os dados numéricos foram extraídos das imagens com o programa *ArrayVision 8.0* (Imaging Research Inc.). Os dados de intensidade de Cy3 e Cy5 de uma mesma amostra foram

corrigidos por uma função lowess [93] implementada num pacote R (www.r-project.org). A normalização entre os diferentes experimentos com tecidos humanos foi feita utilizando-se a média interna (removidos os 40% extremos) das intensidades de todas as sondas em cada *slide* que estavam acesas. Estas eram definidas como as sondas cujas intensidades estavam acima da média mais 2 desvios-padrão de 1.198 controles negativos. Para os experimentos com células LNCaP (tratadas ou não com α -amanitina), a normalização foi feita usando a média interna (removidos os 40% extremos) das intensidades de 300 sondas controles que hibridizaram com os alvos marcados originados dos *spikes* de mRNA.

3.4.4. Análises Estatísticas

Para o perfil de expressão tecido-específico, o método *Significance Analysis of Microarray* (SAM) foi empregado usando como parâmetros: *multi-class response*, 1.000 permutações, *K-Nearest Neighbors Imputer* e *false discovery rate* (FDR) $< 0,01\%$. Análise de variância (ANOVA) implementada no programa *SpotFire Decision Site for Functional Genomics* (SpotFire Inc.) com um *cutoff* de valor de p menor que 0,001 também foi utilizada. Conjuntos de genes identificados por SAM e ANOVA foram combinados para se identificar um grupo mais restrito de sondas que possuísem mudanças estatísticas mais significativas nos tecidos.

Para os experimentos de inibição da RNAP II, o método de *Significance Analysis of Microarray* (SAM) foi empregado, usando como parâmetros: *two-class unpaired response*, teste t , 1.000 permutações, *K-Nearest Neighbors Imputer* e *false discovery rates* (FDR) indo de 0,2% até 3,5%. Paralelamente, uma análise de razão sinal-ruído (SNR) com 10 mil permutações ($p < 0,05$) foi realizada. Conjuntos de genes identificados por SAM e SNR

foram combinados para se identificar um grupo mais restrito de sondas que possuísem mudanças estatísticas mais significativas após o tratamento com α -amanitina.

3.4.5. Análise de Enriquecimento de Categorias de *Gene Ontology* (GO)

Foi utilizado o programa *BiNGO, the Biological Network Gene Ontology plug-in tool* [94] do pacote Cytoscape (<http://www.cytoscape.org/>). Nós usamos os seguintes parâmetros: *Hypergeometric statistical test, Benjamini & Hochberg's FDR multiple testing correction* e nível de significância de 0,05. Como população de referência nós usamos todos os genes (Gene_IDs) presentes no *microarray* intrônico 44k. Para as sondas de RNA TIN e PIN analisadas foram usados os Gene_IDs dos genes codificadores de proteína que as continham.

4. RESULTADOS E DISCUSSÃO

4.1. RNAs Intrônicos Associados ao Câncer de Próstata

Com o objetivo de identificar RNAs intrônicos possivelmente envolvidos no desenvolvimento do câncer, foi construída uma plataforma de *microarray* contendo centenas de ESTs mapeadas em íntrons de genes humanos. Esta plataforma foi usada para comparar o perfil de expressão dos RNAs intrônicos em tumores de próstata de diferentes estágios de diferenciação.

A escolha das sondas, o desenho e a construção do *microarray* intrônico, as hibridizações e análises dos experimentos fizeram parte do projeto do Dr. Eduardo M. Reis, naquela época pós-doutorando do grupo, e atualmente professor do IQ-USP. As validações dos transcritos intrônicos fazem parte do presente trabalho de tese. Todos os dados foram publicados no artigo de Reis *et al.* (anexo 1).

4.1.1. Construção do *Microarray* 4k Intrônico

O projeto CAGE (Cooperação para Análise da Expressão Gênica) de Câncer, coordenado pelo Dr. Sergio Verjovski-Almeida, teve por objetivo analisar a expressão diferencial de genes codificadores e não codificadores de proteínas que pudessem estar envolvidos em câncer. Para isso, este grupo depositou num *chip* de DNA cerca de 4.200 ESTs (*Expressed Sequence Tags*) que representavam regiões transcritas intrônicas, exônicas e intergênicas do genoma humano. As sondas foram selecionadas através de uma análise informática feita em nosso grupo, a partir de 1.016.501 seqüências de ESTs obtidas de diversos tipos de câncer, geradas em 2000 - 2001, no Projeto Genoma Humano do Câncer (HCGP), financiado pela FAPESP e Instituto Ludwig, usando a técnica ORESTES [95]. As ESTs de genes codificadores de proteína relacionadas a câncer foram selecionadas

através do cruzamento de informações entre a anotação destas ESTs, com listas de genes já descritos na literatura como associados a câncer de próstata e pulmão, além de genes envolvidos na tumorigênese, metabolismo e progressão do câncer. Para se construir o *microarray* enriquecido em seqüências intrônicas (Materiais e Métodos item 3.1.1.), foram escolhidas ao acaso 619 ESTs, que mapeavam inteiramente em regiões intrônicas de genes conhecidos, para serem depositadas no *array*.

4.1.2. Análise dos Resultados de *Microarrays* 4k Intrônicos

Para se verificar que fração dos transcritos intrônicos estava expressa em próstata, amostras de mRNA de tumor de próstata e tecido normal adjacente de 27 pacientes foram hibridizadas contra o *chip* de cDNA intrônico (Materiais e Métodos itens 3.1.2. e 3.1.3.). Observamos que as frações de transcritos expressos em próstata que mapeavam em regiões intrônicas e exônicas de RefSeq (*Reference Sequence*) eram similares (Reis *et al.*, anexo 1). Estes dados confirmam que as ESTs *unspliced* intrônicas, depositadas em banco de seqüências públicas, devem representar transcritos verdadeiros e não possíveis contaminantes contidos nas bibliotecas de cDNA, como sugerido por Sorek e Safer [96].

Nosso próximo passo foi procurar transcritos intrônicos que pudessem ter algum papel no tumor de próstata ou ser bons candidatos a marcadores de prognóstico. Assim, comparamos o nível de expressão destes transcritos com o grau de diferenciação dos tumores de próstata. A determinação do grau de diferenciação tumoral é um importante fator de prognóstico. Atualmente, isto é feito principalmente através do *Gleason grading system* que se baseia em critérios histológicos. O *Gleason score* (GS) do tumor varia de 2 a 10 e é obtido pela análise dos aspectos celulares, como tamanho do núcleo e grau de diferenciação, além do número de células. Porém, esta classificação é subjetiva e os

valores de GS variam dentro de uma mesma amostra de acordo com o corte histológico observado e do patologista que o está classificando [97]. Geralmente, enquanto tumores de pacientes classificados como GS 2 a 6 tem pouca agressividade biológica, tumores GS 8 a 10 são agressivos e provavelmente levarão o paciente à morte por câncer. A maioria dos tumores retirados de pacientes com câncer de próstata, no entanto, são GS 7, cuja agressividade é intermediária e o prognóstico indefinido.

Para procurar transcritos intrônicos cuja expressão estivesse correlacionada ao grau de diferenciação tumoral (*Gleason score*, GS), nós primeiramente analisamos o perfil de expressão de amostras de tumor de próstata com valores de GS baixo (5 e 6) e alto (8 a 10). Diversos programas foram utilizados para normalizar, comparar e, finalmente, obter um classificador tumoral capaz de prever, com certa confiabilidade (valor de $p < 0,001$), a progressão de um tumor de GS intermediário ou de prognóstico indefinido (tumores GS 7). Um conjunto de 56 genes cuja expressão distingue perfeitamente as duas classes de tumor (GS baixos e altos) (Reis *et al.*, anexo 1) foram os alvos da caracterização descrita em detalhe a seguir. As análises estatísticas envolvidas nestas etapas estão descritas nos métodos (Materiais e Métodos item 3.1.4.).

O método de validação cruzada pela retirada de um paciente (*Patient Leave-One-Out Cross-Validation*) foi utilizado para aumentar a robustez do classificador e diminuir assim, o *set* de transcritos a serem caracterizados. Esta análise consiste na retirada sistemática de cada um dos 11 pacientes, usados previamente para montar o classificador, e a obtenção de um novo conjunto de genes, significativamente correlacionados com as amostras restantes de GS alto ou baixo. Os genes que apareciam em mais de 60% dos classificadores tumorais, obtidos por este método, estão apresentados na Figura 1. Desta forma, os 56 genes capazes de classificar o tumor, como sendo correlacionados com amostras de GS

alto (mau prognóstico) ou de GS baixo (bom prognóstico) foram reduzidos para um conjunto de 12 genes a serem caracterizados.

Figura 1. Transcritos correlacionados ao grau de diferenciação tumoral. Painei A. Conjunto dos 12 genes mais correlacionados com o grau de GS que aparecem em mais de 60% dos classificadores, identificados pelo método de validação cruzada pela retirada de um paciente, como indica a escala no alto do painel. O nome do gene onde o transcrito está inserido está situado no lado esquerdo do Painei A. Barras verdes claras representam transcritos intrônicos e as barras verdes escuras correspondem a éxons conhecidos. Painei B. Nível de expressão dos transcritos (linhas) em cada amostra de tumor (colunas) de GS baixo (Baixo GS) e GS alto (Alto GS). O nível de expressão de cada gene é representado

pelo número de desvios-padrão acima (vermelho) ou abaixo (azul) do valor médio, para cada gene, ao longo das diferentes amostras.

4.1.3. Caracterização dos Transcritos Diferencialmente Expressos

O mapeamento genômico dos 56 clones ORESTES do classificador revelou um número muito grande de transcritos (41%) mapeados inteiramente em íntrons de genes conhecidos. Dos 12 melhores genes classificadores foi possível constatar que 7 alinhavam em íntrons de genes codificadores de proteína (Figura 1A, verde claro) e apenas 5 dos clones depositados no *chip* representavam éxons conhecidos (Figura 1A, verde escuro).

4.1.3.1. Novos Éxons de Genes Envolvidos em Câncer

O primeiro passo na caracterização destes clones ORESTES, revelados por *microarrays* de cDNA, era descobrir se tais transcritos, mapeados num íntron, representavam éxons novos ainda não descritos. Mais de 120 *primers* foram desenhados e encomendados para estes testes. Para demonstrar que o clone ORESTES representava um éxon novo do gene, *primers* que alinhavam no éxon conhecido e no clone foram utilizados em ensaios de amplificação por PCR. Os fragmentos que continham o novo éxon, um íntron e o éxon conhecido foram confirmados por seqüenciamento, tamanho no gel e alinhamento no genoma humano (Figura 2).

Dos clones presentes no *chip* que possuíam uma intensidade de expressão acima do limite de detecção (3 desvios-padrão em relação ao *background* da lâmina), definido nas análises de *microarrays*, em mais de 70% das lâminas, 3 foram confirmados como sendo possíveis éxons novos, ainda não descritos, dos seguintes genes: *VLDLR2* (*Very Low Density Lipoprotein Receptor 2*), *PCNT1* (*Pericentrin 1*) e *SIRT7* (*Sirtuin 7*). Nenhum

deles está entre os genes significativamente correlacionados com o grau de malignidade mostrados na Figura 1.

Figura 2. Tentativa de amplificação por PCR dos clones que mapeavam em região intrônica, utilizando-se *primers* em éxons conhecidos externos e no clone. Painel A. Esquema mostrando a localização relativa dos *primers* na estrutura genômica do gene-alvo. Éxons conhecidos, íntrons e a região intrônica onde poderia haver um novo éxon estão exibidas esquematicamente conforme legenda. As linhas contínuas com setas indicam o segmento de cDNA que seria amplificado caso exista o novo éxon. Os tamanhos esperados para a confirmação de um novo éxon para o gene *TLE3* estão mostrados por estas retas. Painel B. Foto de um gel de agarose 1,2 %, corado com brometo de etídeo, mostrando fragmentos amplificados do gene *TLE3*, com duas combinações de *primers* diferentes, citadas no Painel A.

O caso mais interessante de éxon novo, descoberto pelas análises de *microarrays* e pelos ensaios de PCR foi o do gene *Transducin-Like Enhancer of Split 3 (TLE3)*. Esta proteína faz parte de uma família de co-repressores humanos, descrita em diversos organismos. Como co-repressora, esta proteína não se liga diretamente ao DNA mas é recrutada pelo repressor, através um domínio C-terminal chamado WD-40. Além de estar presente no conjunto dos 56 genes do classificador tumoral, este clone também aparece na lista dos genes mais correlacionados nas análises de validação cruzada pela retirada de um paciente (Figura 1). Utilizando-se uma ferramenta gráfica (ORF Finder, <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>), capaz de encontrar a fase aberta de leitura e traduzir a seqüência de nucleotídeos desejada, descobriu-se que este éxon novo do gene TLE3, situado entre os éxons 6 e 7, introduz um códon de parada na tradução desta proteína. O mRNA deste gene que possui este éxon novo foi confirmado por experimentos de RACE e de Northern blot (Figura 3) e parece produzir uma proteína truncada, onde os dois domínios da proteína original (TLE_N e WD40) são separados. O domínio TLE_N, também conhecido como domínio “Q”, parece estar envolvido com a tetramerização da proteína e, quando separado do domínio WD-40, pode regular negativamente a proteína não-truncada, possivelmente por sequestrá-la na tetradimerização [98]. O aumento da expressão desta forma truncada de mRNA, em tumores mais agressivos (alto valor de GS) pode causar uma ausência de repressão dos diversos fatores de transcrição (alguns relacionados com o câncer) que se ligam a TLE3.

4.1.3.2. Transcritos Intrônicos Antisenso Não-Codificadores

Apesar de inúmeras combinações de *primers* e ensaios de PCR terem sido utilizados na tentativa de encontrar éxons novos para os clones do *chip*, presentes em nosso

classificador, somente para o gene *TLE3* este dado foi confirmado. Postulou-se então que estes transcritos poderiam ser RNAs não codificadores e que possuiriam uma orientação contrária ao gene conhecido que o envolvia, sendo, portanto, um transcrito antisense. Estas considerações foram feitas baseadas no reduzido poder codificador da sequência do clone quando se usava um algoritmo preditor [99] e nos diversos trabalhos da literatura que descreviam esta nova classe de transcritos antisensos. Em seguida, a caracterização dos transcritos intrônicos com a expressão correlacionada ao grau de diferenciação tumoral foi realizada usando técnicas experimentais diferentes, a seguir relacionadas. O tamanho real dos ncRNAs intrônicos foi obtido através de experimentos de RACE (Materiais e Métodos item 3.1.5.1.). Estes experimentos revelaram que estas mensagens são relativamente longas (600-1.100 pb) quando comparadas aos microRNAs ou snoRNAs (Figura 3). A existência de um transcrito intrônico antisense e com o tamanho próximo ao mostrado por RACE também pôde ser confirmada através de experimentos de Northern blot fita-específico (Figura 3) (Materiais e Métodos item 3.1.5.2.).

Figura 3. Caracterização de transcritos intrônicos relacionados ao câncer. O mapeamento de cada clone de cDNA ORESTES original que foi depositado no *chip* está representado em azul e as coordenadas genômicas estão indicadas. Paineis A, C e E. Forma alternativa do gene *TLE3* confirmado por RACE-PCR (bloco vermelho). Paineis B, D e F. Um transcrito de 5,53 kb detectado por Northern blot em 3 diferentes linhagens de célula de câncer de próstata, usando uma sonda complementar a fita senso do novo éxon de *TLE3*. Paineis C e E. Caracterização de transcritos antisense intrônicos dos genes *RASSF1* e *PPHLN1*. A seta vermelha mostra um transcrito obtido por RACE-PCR de biblioteca direcionada de cDNA de cólon. Os blocos amarelos mostram transcritos obtidos por RACE-PCR não direcionado usando uma biblioteca de cDNA de pulmão. Paineis D e F. Detecção dos transcritos antisense intrônicos de *RASSF1* (Painel D) e *PPHLN1* (Painel F) por Northern blot fita específico.

Para confirmar a hipótese de que os RNAs com cauda poli-A do nosso classificador eram transcritos antisenso, fizemos uma transcrição reversa (RT) fita-específica usando *primers* para a região do clone, no nosso *chip* de DNA, seguido de uma amplificação de uma região interna por PCR (Materiais e Métodos item 3.1.5.3.). Uma reação de transcrição reversa sem qualquer *primer* foi incluída como controle negativo de um possível auto-anelamento (*self-annealing*) do mRNA na produção de cDNA. A sequência de cada produto foi confirmada e os controles negativos mostraram que não havia contaminação com DNA genômico nem auto-anelamento do mRNA de forma significativa (Figura 4). Surpreendentemente, a Figura 4 também mostra que foram detectados transcritos tanto da fita senso quanto da antisenso, na região intrônica, da maioria dos genes candidatos à transcritos antisenso intrônicos. Este ensaio foi repetido para o clone que continha éxons do gene *TMEPAI*, como um controle negativo, para demonstrar que a especificidade da fita detectada é real. Neste caso, somente a banda correspondente à amplificação do cDNA da fita senso pôde ser visualizada.

Figura 4. RT-PCR fita-específica para os genes candidatos a transcritos antisense intrônicos. Mensagens intrônicas tanto da fita senso quanto da antisense são detectadas para a maioria dos genes. Disposição das colunas: (I) Somente o *primer reverse* presente; (II) Somente o *primer forward* presente; (III) *Primer forward* presente mas sem a enzima transcriptase reversa; (IV) Sem *primers* na transcrição reversa.

Experimentos de PCR quantitativo em tempo-real (*Real-Time PCR*) foram feitos com o objetivo de validar a diferença de expressão vista nas análises de *microarrays* (Materiais e Métodos item 3.1.5.4.). Escolhemos como alvo principal desta validação, o gene *RASSF1* e seus possíveis transcritos intrônicos senso e antisense. Pela Figura 5, podemos ver que tanto o gene *RASSF1* (valor de $p \leq 0,009$), quanto os transcritos senso (valor de $p \leq 0,010$) e antisense (valor de $p \leq 0,017$) intrônicos estão diminuídos em tumores de GS alto ou de mau prognóstico. Estes dados corroboram a diferença de expressão detectada nos experimentos de *microarrays* de cDNA e levantam hipóteses sobre um possível papel deste ncRNA intrônico no desenvolvimento do câncer.

Figura 5. Nível de expressão do gene e dos transcritos intrônicos senso e antisenso de *RASSF1*. A fração relativa é calculada pela média de expressão relativa dos mRNA em todas as amostras. Painel A. Nível de expressão de transcritos antisenso (quadrados hachurados) e senso (quadrados cheios) do gene *RASSF1*, em 3 tumores de GS baixo (linha cinza) e 3 tumores de GS alto (linha preta). Painel B. Nível de expressão dos éxons conhecidos do gene, medidos nas mesmas amostras do Painel A.

Com os dados obtidos neste estudo mostramos pela primeira vez uma possível implicação de um conjunto de transcritos intrônicos antisensos em uma doença humana. Contudo, ainda não é sabido se a expressão de ncRNAs intrônicos é um mecanismo verdadeiro de regulação ou apenas reflete erros no reconhecimento de promotores ou no

processo de início da transcrição. No entanto, a descoberta de um conjunto de mensagens intrônicas antisenso correlacionadas ao grau de malignidade de tumores de próstata tem um impacto no diagnóstico molecular do câncer em geral. Nosso achado argumenta a favor da inclusão de transcritos não codificadores intrônicos dentro do arsenal de ferramentas usadas para o diagnóstico molecular, que no momento é exclusivamente dominado por *microarrays* que contêm apenas éxons de genes codificadores de proteína.

4.2. Regulação da Síntese de RNAs Intrônicos por Andrógeno

Após a identificação de transcritos intrônicos diferencialmente expressos em tumores de diferentes graus de malignidade, começamos a investigar o mecanismo de regulação da transcrição destes RNAs não codificadores intrônicos. Para isso, células de linhagem de próstata LNCaP foram tratadas com andrógeno e os níveis de expressão de todos os transcritos, localizados em íntrons (intrônicos) ou representando os éxons (exônicos) de genes codificadores de proteína, foram analisados. Estes experimentos visaram descobrir se ncRNAs intrônicos e mRNAs podem compartilhar um mesmo mecanismo de regulação. Esta regulação poderia ser mediada pela ligação do receptor de andrógeno e outras proteínas do complexo, a sítios específicos de DNA, presentes na região promotora destes transcritos intrônicos.

Este conjunto de experimentos sobre a regulação dos transcritos intrônicos foi realizado em conjunto com o aluno Rodrigo Louro que escreveu o artigo anexado. Rodrigo também é aluno de doutorado orientado pelo Dr. Sergio Verjovski-Almeida.

4.2.1. Identificação de ncRNAs Intrônicos Respondedores a Andrógeno

A plataforma de *microarrays* enriquecida de transcritos intrônicos foi usada para identificar os RNAs intrônicos e exônicos respondedores a andrógeno. Células de linhagem de próstata LNCaP foram tratadas com um andrógeno sintético, R1881, por 6 a 48 horas e o RNA total de cada um dos 6 tempos experimentais foi extraído, tratado com DNase, marcado e hibridizado contra o *microarray* 4k intrônico (Materiais e Métodos item 3.2.1.). Genes como *KLK3* e *TMEPAI*, identificados como genes com expressão aumentada por este hormônio tanto nas análises de *microarray* quanto em experimentos de PCR quantitativo em tempo real (Figura 6), já foram descritos como regulados positivamente por andrógeno [100, 101]. Estes resultados indicam que tanto as análises de *microarrays* quanto o tratamento com andrógeno das células LNCaP foram realizados de forma apropriada.

Figura 6. Nível de expressão de 2 genes sabidamente regulados por andrógeno. O nível de expressão dos genes *KLK3* e *TMEMAI* foram medidos por PCR quantitativo em tempo real para confirmar a efetividade do tratamento com andrógeno. A diferença de expressão (*Relative Fold Change*) foi calculada pela razão entre cada tempo experimental e o tempo zero em células tratadas (barras pretas) ou não tratadas (barras brancas) com andrógeno. As barras de erro refletem o número de desvios-padrão da média da diferença de expressão em cada tempo, determinado por duas réplicas técnicas e usando o mesmo cDNA.

As análises de *microarrays* revelaram ainda um conjunto de 168 genes codificadores de proteína e 39 RNAs intrônicos diferencialmente expressos após o tratamento por andrógeno (Materiais e Métodos itens 3.2.2.1. e 3.2.2.2.). A análise por ESTScan [99] confirmou que nenhuma destas 39 seqüências depositadas no *array* possuem um potencial

codificador. A Figura 7 mostra o perfil de expressão dos 39 ncRNAs intrônicos nos 6 diferentes tempos experimentais.

Os RNAs intrônicos regulados por andrógeno puderam ser agrupados em 3 diferentes *clusters* baseados nos seus perfis temporais de expressão após a exposição ao andrógeno (Figura 7, Painel B). O *cluster* I contém RNAs cujos níveis de expressão diminuíram 2 a 3 vezes dentro de 24h de tratamento com hormônio e em geral foram restituídos com 48h. Os *clusters* II e III incluem RNAs intrônicos que foram regulados positivamente por andrógeno dentro de 24h de tratamento com andrógeno. O *cluster* II agrupa RNAs que aumentaram seus níveis de expressão em 2 a 3 vezes e o *cluster* III em 3 a 6 vezes.

Figura 7. *Cluster* hierárquico e perfil temporal dos transcritos intrônicos revelados por análise dos experimentos de *microarrays*. Painel A. *Cluster* hierárquico (UPGMA, distância Euclidiana) dos 39 transcritos intrônicos diferencialmente expressos em pelo menos 3 tempos consecutivos após o tratamento com andrógeno. Cada coluna representa um tempo diferente e a intensidade da cor representa a diferença de expressão dos ncRNAs intrônicos (linhas), calculada como o log₂ da razão entre a média de expressão em cada tempo de tratamento e a média de expressão de seu controle (células não tratadas) no tempo correspondente. Os quadrados vermelhos indicam que a expressão está aumentada; quadrados verdes indicam uma diminuição da expressão e os quadrados pretos mostram

que não houve alteração no nível de expressão do transcrito após o tratamento das células com andrógeno em relação às tratadas com veículo controle. As barras verticais I, II e III representam os grupos de transcritos intrônicos com padrão similar de expressão. Painel B. Perfil temporal dos 3 grupos indicados no *cluster* do painel anterior, obtido pelo método K-means *clustering* (*Data Centroid Based Search*, distância Euclidiana).

É interessante notar também que vários RNAs não codificadores, identificados em nossas análises, são intrônicos para genes codificadores de proteínas ou genes pertencentes a famílias gênicas reguladas por andrógeno e identificados nos artigos de Nelson *et al.* [100] e DePrimo *et al.* [102], como *KDELR2*, *ITGA5*, *ACTN1* e *RAB3*.

4.2.2. Caracterização dos ncRNAs Regulados por Andrógeno

Com o objetivo de revelar a orientação dos transcritos intrônicos em relação aos genes codificadores de proteínas que os contêm, experimentos de RT-PCR fita-específicos foram realizados para 13 ncRNAs intrônicos regulados por andrógeno (Figura 8). A necessidade destes experimentos (Materiais e Métodos item 3.2.3.) reside no fato de que em nosso *chip* foram depositados produtos de PCR, referentes às seqüências de ESTs ORESTES que por serem clones de DNA dupla-fita, não nos informam qual das fitas hibridizou contra o cDNA marcado. Foi possível detectar um ncRNA intrônico antisenso em 10 dos 13 transcritos testados (Figura 8). Para seis dos RNAs antisensos (*DNAJC3*, *SAP18*, *KDELR2*, *FLJ10154*, *MYO5A* e *ATF2*) foi também detectada simultaneamente uma mensagem senso intrônica transcrita na fita oposta a eles. Para quatro (*GAS6*, *RAB25*, *ITGA6* e *ADD3*) ou três (*ACTN4*, *STARD13* e *ERG*) RNAs foi detectada somente a orientação antisenso ou senso, respectivamente (Figura 8). Os transcritos intrônicos senso

de *ACTN4*, *STARD13* e *ERG* não possuem potencial codificador quando analisado por algoritmo preditor [99]. Outros experimentos devem ser realizados para saber se estes RNAs intrônicos representam novos éxons de formas alternativas de *splicing* ou se são ncRNAs originados de uma unidade transcricional independente localizada na fita senso. Esta última possibilidade é corroborada por dados recentes de *tiling-arrays* mostrando transcritos intrônicos senso [44, 45] e por experimentos de RACE-PCR (descritos adiante).

Figura 8. Produtos de PCR quantitativo em tempo real de cDNAs gerados de forma fita-específica. Painel A. Perfil de expressão obtido por análise de *microarrays* de 13 mensagens intrônicas reguladas por andrógeno e selecionadas para os experimentos de RT-PCR fita-específicos. Painel B. O nome do gene codificador de proteína que contém a mensagem intrônica a ser validada está no início de cada linha e a presença (+) ou ausência (-) da transcriptase reversa (RT) ou *primer* nas colunas acima. A primeira coluna (Antisense) e a segunda (Sense) mostram produtos de PCR gerados a partir de um cDNA

correspondente a um transcrito intrônico antisenso ou intrônico senso de um RNA, respectivamente.

A maioria dos snoRNAs e uma grande fração dos microRNAs são derivados de seqüências intrônicas de genes codificadores e não codificadores humanos. Para investigar a possibilidade de que alguns destes ncRNAs intrônicos, regulados por andrógeno, pudessem ser íntrons processados que depois iriam gerar os pequenos RNAs destas classes, nós comparamos as seqüências dos 39 ncRNAs intrônicos com as seqüências dos 346 snoRNAs (<http://www-snoRNA.biotoul.fr/>) e 383 microRNAs [103]. Nenhuma similaridade foi encontrada nessas análises levantando a hipótese de que estes RNAs intrônicos sejam transcritos longos.

Experimentos de RACE-PCR com cDNA de tumor de próstata foram realizados com o objetivo de se obter a seqüência completa de um conjunto de RNAs intrônicos selecionados (Materiais e Métodos item 3.2.3.). Os experimentos revelaram que os RNAs intrônicos antisensos de *DNAJC3*, *MYO5A* e *ITGA6* eram transcritos longos (500 a 800 pb) e não processados (*unspliced*) que mapeavam totalmente dentro de regiões intrônicas do gene codificador de proteína correspondente (Figura 9, painéis A a C). Um transcrito senso, *unspliced* e totalmente intrônico ao gene *STARD13*, com aproximadamente 1.400 pb, foi obtido (Figura 9, Painel D). A análise pelo algoritmo preditor ESTScan [99] deste longo transcrito intrônico senso indica que ele pode gerar um polipeptídeo de 74 aminoácidos. Porém, a análise por BLASTp [104] não encontrou nenhuma similaridade deste pequeno polipeptídeo com proteínas conhecidas do *GenBank* ou com domínios protéicos conservados no CDD [105].

Os ensaios de RACE para o transcrito intrônico antisenso do gene *GAS6* mostrou que este fazia parte de um mRNA antisenso intrônico já depositado no *GenBank*

(AK123478) por outro laboratório (Figura 9, Painei E). Este dado foi confirmado por experimento de Northern blot fita-específico (Figura 9, Painei F) que detectou um transcrito antisenso de aproximadamente 5 kb, expresso em diferentes tecidos humanos. A análise *in silico* usando-se ESTscan [99] dos longos transcritos intrônicos para os genes *GAS6*, *DNAJC3*, *ITGA6* e *MYO5A* não detectou nenhuma fase aberta de leitura maior que 50 aminoácidos.

Figura 9. Ensaio de RACE-PCR e Northern blot dos ncRNAs intrônicos respondedores a andrógeno. Painéis A a E. O mapeamento de cada clone de cDNA ORESTES original que foi depositado no *chip* está representado em azul e as coordenadas genômicas estão indicadas. Os painéis mostram a extensão por RACE-PCR (painéis A a D, blocos vermelhos) ou PCR (Painel E, bloco amarelo) de 5 transcritos intrônicos selecionados. Os genes codificadores de proteína de cada *locus* estão representados por setas verdes e o transcrito antisense intrônico já depositado no GenBank é mostrado por uma seta preta (número de acesso indicado abaixo da seta). Painel F. Detecção do transcrito antisense intrônico de *GAS6* por Northern blot fita específico.

4.2.3. Alteração dos níveis de expressão por andrógeno de ncRNAs intrônicos e os genes codificadores de proteína correspondentes

Ensaio de PCR quantitativo em tempo real usando RNA de células tratadas ou não tratadas com andrógeno foram realizados com o objetivo de se verificar a modulação por andrógeno dos níveis de ncRNAs intrônicos e genes codificadores situados num mesmo *locus*. Estes dados poderiam fornecer evidências de uma possível regulação do transcrito intrônico no gene codificador de proteína que o contém. Nós testamos 2 RNAs intrônicos que só possuíam a orientação antisense (*GAS6* e *ITGA6*) e 2 mensagens intrônicas na qual a transcrição senso e a antisense foram detectadas (*DNAC3* e *KDELR2*) (Figura 10, painéis “intronic” A a D). Em cada um dos casos, nós quantificamos também a expressão dos éxons dos genes codificadores correspondentes (Figura 10, painéis “exonic” A a D).

Uma diminuição da expressão do transcrito intrônico antisense do gene *GAS6* após a exposição ao andrógeno foi confirmada (Figura 10, Painel “intronic” A). Um dado interessante é que foi detectado um aumento do nível de expressão do mRNA contendo os éxons 4 e 5 deste gene (Figura 10, Painel “exonic” A), mostrando uma modulação inversa causada por andrógeno entre o ncRNA intrônico e o gene codificador de proteína que o contém. O mesmo resultado foi obtido entre a abundância relativa dos ncRNAs intrônico senso e antisense do gene *DNAJC3* (Figura 10, Painel “intronic” B). O aumento de expressão observado para o RNA intrônico senso de *DNAJC3* e a diminuição do RNA intrônico de *GAS6*, após o tratamento por andrógeno estão de acordo com os resultados dos experimentos de *microarray*.

Os dados de PCR quantitativo em tempo real também confirmam o que foi visto nos experimentos de *microarrays* para os RNAs intrônicos de *KDELR2* e *ITGA6* (Figura 10, Painel “intronic” C e D). No entanto, no *locus* do gene *ITGA6* não foi detectada nenhuma

mudança na abundância relativa dos transcritos codificadores de proteína deste gene (Figura 10, Painel “exonic” C e D).

Figura 10. Transcrição modulada por andrógeno de mensagens codificadoras de proteína e ncRNAs intrônicos. Painéis A a D. Níveis de expressão de transcritos intrônicos senso e antisenso (painéis “intronic”) e das mensagens codificadoras de proteína (painéis “exonic”) dos *loci* *GAS6* (A), *DNAJC3* (B), *KDELR2* (C) ou *ITGA6* (D) em células tratadas com andrógeno (barras pretas) ou controle (barras brancas). Os mRNAs completos de referência mapeados em cada *locus* estão mostrados como setas pretas e o ncRNAs

intrônic *unspliced* como setas cinzas. A diferença relativa de expressão (*Relative fold-change*) indica o aumento ou diminuição dos níveis dos transcritos após o tratamento com andrógeno em relação à sua abundância nas células não tratadas controle.

4.2.4. Ligação do Receptor de Andrógeno na Região *Upstream* do RNA intrônico Antisense do Gene MYO5A

Para podermos inferir, com mais confiança, sobre os possíveis mecanismos comuns de regulação, entre os transcritos intrônicos e os genes codificadores regulados por andrógeno, foi feita uma procura por domínios de ligação do receptor de andrógeno na região *upstream* aos transcritos intrônicos identificados nas análises de *microarrays*. Estes domínios de ligação, chamados motivos ARE (*Androgen Response Element*) já foram caracterizados experimentalmente para 7 genes humanos [100]. Suas seqüências são compostas, geralmente, por 2 elementos assimétricos de 6 bases e separadas por 3 nucleotídeos quaisquer.

Como as extremidades 5' destes transcritos intrônicos ainda não eram conhecidas (pois possuímos apenas um clone de EST do transcrito e não sua seqüência completa) e assumindo que estes transcritos teriam uma direção contrária (antisense) aos éxons que os flanqueiam, foram extraídas para nossas análises, as 3.000 bases das seqüências genômicas que antecedem o início de cada seqüência que foi depositada no *chip*. Um programa de alinhamento de seqüências (*cross-match*) foi usado para procurar o consenso AGAACAnnnTGTTCT (obtida do site: <http://transfac.gbf.de/TRANSFAC/>) nestas regiões possivelmente *upstream* aos transcritos intrônicos antisensos. Um outro programa, desenvolvido em nosso laboratório, selecionou apenas seqüências candidatas com alto grau de identidade (9 das 12 bases idênticas ao consenso) conforme definido por Nelson *et al.*

[100]. Com isso, identificou-se um conjunto de candidatos a motivos ARE para vários destes transcritos intrônicos regulados por andrógeno (Figura 11).

GenBank Accession (EST)	Locus Name	ARE Sequence	Position	Genome Location	% Match
CK327134	N A	A G G A C A G A A T G T T T T	-2,056	2p23.2	83
BF333281	NIBP	A G A A G G T G C T G T T C A	-621	8q24.3	75
BF879099	GAS6	G G A A A A T C A T G T T C A	-1,528	13q34	75
AW860958	DST	A G A A C A T C T T T T T G C	-799	6p12.1	75
BF881469	ALMS1	G G A C C A G T T T T T T C T	-1,091	2p13.1	75
CK327189	P2RY14	A G A A C A G C A G G C T C T	-1,837	3q25.1	83
BF368584	DNAJC3	A G A A T A C C A T G T T G C	-1,608	13q32.1	75
BF768459	ACTN4	G G A A C A T G A A T T T C T	-785	19q13.2	75
BF768642	N A	T G A A C T G G G T T T T C T	+170	Xq27.1	75
BF882783	ACTN4	G G A A C C A C T G G T T C T	-1,839	19q13.2	75
CK327094	ATF2	T G A A C A G A C A C T T C T	-920	2q31.1	75
BE087892	STARD13	A G A T C A T G T G G T T C T	-2,239	13q13.1	83
CK327184	ITGA6	A G A A C A G G T T T T T G A	-2,788	2q31.1	75
BF926454	ADD3	T G A C C A T G C T G G T C T	-2,285	10q25.1	75
BF848956	MYO5A	A G T A C A T A A T T T A C T	-812	15q21.2	75
BF805741	PMF1	A G A A A T T C T T G C T C T	-2,104	1q22	75
AW805635	ERG	A G A A C T G C T T T G T C T	+246	21q22.2	75
BF350758	KDELR2	A G A A C A G C T G C T T C T	+219	7p22.1	83
BE156190	N A	A G A C C A G T T T T A T C T	-181	4q32.3	75
BE062809	UBE2V1	A G A A C A A F A T G T T A A	-2,786	20q13.13	83
AW819863	PPP2R2A	A G A G C A G G A A G C T C T	-881	17q23.3	75
BF332494	SAP18	A G A C C A G C C T G G T C A	-414	13q12.11	75
BF364131	N A	T G A C C A G G C T G G T C T	-2,481	16q23.2	75
CK327190	ZNF644	A G A A T C T A C T G T T C C	-283	1p22.2	75

ARE Consensus Sequence

Figura 11. Candidatos a motivos ARE (*Androgen Response Element*) identificados em regiões *upstream* de 24 possíveis transcritos intrônicos antisense. As letras pintadas correspondem às bases idênticas ao consenso utilizado na procura por bioinformática. O melhor candidato a motivo ARE, encontrado para cada um dos 24 transcritos intrônicos, está representado, assim como o nome do mRNA exônico que o contém, a localização no cromossomo e a porcentagem de bases idênticas ao consenso. A coluna “Position” mostra a posição genômica relativa ao início da EST supostamente com direção antisense (“+”

downstream e "-" *upstream*). O consenso para esses 24 candidatos a motivos ARE (Painel inferior) foi gerado no *software* WEBLOGO (<http://www.bio.cam.ac.uk/seqlogo/>) e corresponde ao motivo ARE consenso AGAACAnnnTGTCT.

Para se confirmar uma ação direta do receptor de andrógeno nas regiões *upstream* dos transcritos intrônicos, foram realizados ensaios de imunoprecipitação da cromatina (ChIP, do inglês *Chromatin immunoprecipitation*) em células LNCaP tratadas ou não com o andrógeno. Estes ensaios, realizados pelo mestrando Paulo Paiva Amaral, consistem, basicamente, numa precipitação diferencial de fragmentos de DNA ligados ao receptor de andrógeno e seu hormônio ligante, utilizando-se um anticorpo anti-receptor de andrógeno (Materiais e Métodos item 3.2.4.).

A ligação do receptor de andrógeno (AR) em um motivo ARE, já descrito anteriormente como importante na regulação da transcrição do gene *PSA*, foi testado em paralelo como um controle positivo do ensaio de imunoprecipitação da cromatina (Figura 12, Painel B “*PSA ARE*”). O gene *TUBA6* que não contém motivos ARE em sua região promotora foi usado como controle negativo de uma possível imunoprecipitação não específica de fragmentos de DNA (Figura 12, Painel B “*TUBA6*”). Estes ensaios revelaram um aumento da ligação (~2,5 vezes) do AR após o tratamento com andrógeno (Figura 12, Painel B e C “*asMYO5A ARE*”), em um motivo ARE identificado em nossa análise (Figura 11) e localizado na posição -812 na região *upstream* de um transcrito intrônico antisenso do gene *MYO5A*. Este resultado está de acordo com o aumento de expressão do RNA intrônico antisenso de *MYO5A* detectado por experimentos de *microarray* (Figura 8) e confirmado por RT-PCR fita-específico seguida de PCR quantitativo em tempo real (Figura 12, Painel A).

Figura 12. Ligação do receptor de andrógeno a um motivo ARE do transcrito intrônico antisense do gene *MYO5A*. PAINEL A. Níveis de expressão do transcrito intrônico antisense do gene *MYO5A* em células tratadas com andrógeno (barra preta) ou controle (barra branca). A diferença relativa de expressão (Relative fold-change) indica o aumento ou diminuição dos níveis dos transcritos após o tratamento com andrógeno em relação à sua abundância nas células não tratadas controle. PAINEL B. Lisados de células tratadas (+) ou não (-) com andrógeno foram usados nos ensaios de imunoprecipitação da cromatina na presença de um anticorpo específico ao receptor de andrógeno (anti-AR). Reações de PCR semi-quantitativo foram usadas na detecção do DNA imunoprecipitado contendo o motivo ARE situado na região *upstream* do transcrito intrônico antisense de *MYO5A* (asMYO5A ARE) em células tratadas com andrógeno (+) ou controle (-). O motivo ARE do gene *PSA* (*PSA ARE*) e do gene não-ativado por andrógeno *TUBA6* (*TUBA6*) foram usados como controles positivos e negativos, respectivamente. Um anticorpo não relacionado (*mouse IgG*) foi usado como um controle negativo para uma imunoprecipitação não específica. Volumes iguais de lisados de células tratadas com andrógeno ou controles (1/100 do total) foram tratados em paralelo e usados para normalizar a quantidade inicial de DNA (input) em cada ensaio de imunoprecipitação da cromatina. PAINEL C. O enriquecimento de AR

ligado ao motivo ARE do ncRNA intrônico de *MYO5A* em células tratadas com andrógeno em relação às células controle é mostrado usando os dados brutos (*Raw*) ou normalizado pela quantidade inicial de DNA da amostra (*Normalized*).

Os resultados descritos até aqui confirmaram que RNAs não codificadores intrônicos e mRNAs podem compartilhar um mesmo mecanismo de regulação da transcrição. Usando como modelo experimental células de próstata tratadas com andrógeno, mostramos que a expressão de alguns ncRNAs intrônicos pode ser regulada direta ou indiretamente por este hormônio.

Evidências adicionais de que ncRNAs intrônicos podem ser regulados pelos mesmos sinais fisiológicos que normalmente agem em genes codificadores de proteína, como os hormônios, podem ser fornecidas pelos estudos de Cawley *et al.* [85] e Euskirchen *et al.* [87]. Nestes estudos, eles mostraram um grande número de sítios de ligação para fatores de transcrição (TFBS, do inglês *transcription factor binding sites*) de proteínas bem caracterizadas como Sp1, c-Myc, p53 e Creb situados *upstream* a regiões intrônicas transcricionalmente ativas de genes [85, 87].

Além disso, o caso do transcrito intrônico do gene *STARD13* se mostrou interessante. Experimentos de PCR quantitativo em tempo real de um cDNA obtido por transcrição reversa fita-específica mostraram a presença apenas da fita senso e confirmou a diferença de expressão detectada nas análises de *microarrays*. Parecia se tratar de um novo éxon ainda não descrito deste gene. Porém, experimentos de RACE desta EST resultaram numa sequência intrônica, sem evidência de processamento, de cerca de 1.400 pb sem nenhuma sobreposição com éxons de mRNAs neste *locus* gênico. Este seria um exemplo semelhante ao transcrito intrônico senso, de 1.125 pb e poli-adenilado, situado no primeiro íntron do gene p53, encontrado por Reisman *et al.* [106]. Com isso, torna-se claro que RNAs

intrônicos senso também podem estar envolvidos na complexidade do transcriptoma humano.

Estas novas descobertas sugerem que a expressão de transcritos ainda não descritos, intrônicos, regulados direta ou indiretamente por andrógeno, podem possuir um papel importante na fisiologia da próstata.

4.3. Identificação e Caracterização *in silico* de RNAs Intrônicos

Experimentos de análise de expressão em larga-escala como *tiling-arrays* ou SAGE apontam algumas evidências de transcrição intrônica. Porém, o pequeno número de regiões intrônicas identificadas nestes estudos não parece refletir a quantidade real de RNAs intrônicos transcritos no genoma humano. Por isso, nós usamos o conjunto de milhões de ESTs humanas depositadas em bancos públicos de seqüências para identificar e caracterizar *in silico* um grande conjunto de regiões intrônicas transcricionalmente ativas do genoma humano. A caracterização compreende uma análise de distribuição das longas mensagens intrônicas ao longo dos íntrons dos genes codificadores de proteínas e a identificação de uma maior freqüência de éxons que sofrem *skipping* em genes que possuem transcrição intrônica.

4.3.1. Definição de um Conjunto de Referência dos Genes Codificadores de Proteínas

Com o objetivo de analisar em larga-escala a complexa distribuição das regiões transcricionalmente ativas do genoma, nós começamos mapeando no genoma humano o conjunto dos 22.458 transcritos de referência (RefSeq) da versão de Março/2005 do genome browser. Deste total, foram excluídos 1.184 RefSeq não processados (*unspliced*) e

601 RefSeq que eram totalmente intrônicos a outro RefSeq. Após o agrupamento dos transcritos RefSeq processados (*spliced*) que mapeavam no mesmo *locus* e originados da mesma fita do DNA, nós identificamos um conjunto de 15.783 unidades de RefSeq *spliced* não redundantes. Deste modo, um total de 4.890 RefSeq que representavam isoformas de genes foram agrupados dentro destas unidades. Adicionalmente, o conjunto de 161.993 mRNAs humanos do *GenBank* foi mapeado no genoma com o objetivo de documentar todas as possíveis variações de *splicing* presentes em nosso conjunto de referência. Inicialmente, estes mRNAs foram agrupados em um total de 45.137 unidades transcricionais de acordo com suas localizações genômicas e origens da fita de DNA (Tabela 1). Uma análise detalhada do mapeamento destes *clusters* de mRNAs revelou que um total de 11.361 *clusters* de mRNAs *spliced* e *unspliced* mapeavam fora das regiões do conjunto de RefSeq não-redundante, representando transcritos humanos não caracterizados. Como esperado, a maioria dos *clusters* de mRNAs (14.575) eram *spliced* e mapeavam em éxons de genes RefSeq na direção senso (Tabela 1). Além disso, 2.559 *clusters* de mRNAs *spliced* mapeavam na direção oposta (antisense) ao conjunto não redundante de RefSeq, mostrando que 16% dos genes RefSeq possuíam evidência de transcritos naturais *spliced* antisense. Dentre estas mensagens antisense, um total de 1.414 estava anotado como transcritos RefSeq. Esta organização genômica de pares senso-antisense parece ser conservada durante a evolução dos vertebrados [36, 37, 107, 108]. Após a inclusão dos *clusters* de mRNAs *unspliced*, nós encontramos um total de 4.231 mensagens antisense com sobreposição a éxons de genes RefSeq, indicando que cerca de 27% dos genes RefSeq possuíam um par antisense. Isto está de acordo com a predição de que cerca de 20% dos transcritos humanos podem formar pares senso-antisense [38].

Tabela 1. Evidência de transcrição intrônica nos genes humanos.

	Clusters de mRNAs com sobreposição a éxons de RefSeq ^a		Clusters de mRNAs totalmente intrônicos a RefSeq		Clusters de mRNAs mapeando fora de RefSeqs	Total
	Direção antisense	Direção sense	Direção antisense	Direção sense		
Clusters de mRNAs <i>spliced</i> ^b	2,559 (1,414) ^c	14,575 (14,369) ^d	1,049 (378)	780 (223)	4181 (0)	23,144 (16,384)
Clusters de mRNAs <i>unspliced</i> ^b	1672 (26)	7463 (87)	1,456 (56)	4,222 (87)	7180 (927)	21,993 (1,183)
Total	4,231 (1,440)	22,038 (14,456)	2,505 (434)	5,002 (310)	11361 (927)	45,137 (17,567)

a O conjunto de dados não redundante é composto por 15.783 unidades RefSeq *spliced*. Isto foi definido pelo mapeamento no genoma humano de um total de 22.458 seqüências RefSeq do *GenBank*, excluindo-se 1.184 RefSeq *unspliced* e 601 RefSeq que mapeavam totalmente dentro de íntrons de outro RefSeq e pelo agrupamento das 20.673 seqüências RefSeq *spliced* restantes que mapeiam no mesmo *locus* dentro de 15.783 unidades RefSeq não redundante *spliced*

b Clusters de mRNA foram obtidos pelo mapeamento no genoma humano de um total de 161.993 seqüências de mRNA seguido de um agrupamento das seqüências que possuíam sobreposição de éxons (Materiais e Métodos item 3.3.1.), que resultou num conjunto não redundante de 45.137 clusters de mRNA. Este conjunto foi alinhado ao conjunto não redundante RefSeq e cada cluster de mRNA foi classificado como exônico, totalmente intrônico ou mapeando em regiões externas a qualquer unidade RefSeq *spliced* não redundante. A orientação senso/antisense foi anotada.

c Para cada classe, o número de clusters de mRNA contendo pelo menos 1 RefSeq é mostrado em parênteses.

d Excluindo do conjunto não redundante de 15.783 RefSeq um total de 1.414 RefSeq que mapeiam numa direção antisense ao respectivo RefSeq

Também foram encontrados 7.507 *clusters* de mRNAs *spliced* e *unspliced* que eram totalmente intrônicos ao conjunto não redundante de genes RefSeq. Enquanto 5.002 (67%) possuíam a direção senso e por isso podiam representar novos éxons destes genes, 2.505 (33%) mapeavam totalmente em regiões intrônicas dos genes RefSeq na direção antisenso. Estas mensagens poderiam representar pares senso-antisenso ainda não descritos devido à natureza fragmentada do dado. Entretanto muitas destas mensagens possivelmente representam uma categoria separada de transcritos longos antisenso *unspliced* o que se torna mais evidente quando dados do repositório público de EST são adicionados, como descrito no item a seguir.

4.3.2. Identificação dos Transcritos Totalmente Intrônicos Longos e *Unspliced*

Nós realizamos uma extensiva busca por evidências adicionais de transcrição intrônica, utilizando para isso a coleção de ESTs humanas dbEST do *GenBank* que compreendia 5.340.464 ESTs. Após filtrar as seqüências de ESTs que mapeavam em regiões ambíguas no genoma (Materiais e Métodos item 3.3.1.), as coordenadas genômicas das seqüências de ESTs que possuíam sobreposição foram usadas para agrupar as 4.762.523 ESTs num conjunto de 332.946 *clusters* de ESTs não redundantes (Tabela 2). Para evitar seqüências derivadas de uma possível contaminação genômica, eventualmente presente no conjunto de ESTs, 210.181 *singlets* de EST foram excluídos das análises futuras. Somente 34.398 *clusters* de ESTs *spliced* e 88.367 *clusters* de ESTs *unspliced* foram considerados (Tabela 2). Para cada *cluster* de ESTs, uma seqüência consenso do *contig* foi obtida da seqüência genômica na qual este *cluster* alinhava (Figura 1). Como esperado, a maioria das ESTs (3.616.644) foram agrupadas em 16.241 *contigs* de ESTs *spliced* que mapeavam em éxons de nosso conjunto de referência RefSeq (Tabela 2). Um

número menor de *clusters* de ESTs *spliced* adicionais mapeavam em íntrons de genes RefSeq e podiam ser fragmentos de novos éxons destes genes, já que a mediana do tamanho dos éxons destes *contigs* é 233 pb, similar à mediana do tamanho dos éxons do conjunto de referência (141 pb).

O achado mais interessante foi a identificação de 55.139 *contigs* de ESTs *unspliced* formados por 190.583 ESTs que mapeavam em regiões totalmente intrônicas do conjunto de referência (Tabela 2). Uma característica marcante destes *contigs* de ESTs totalmente intrônicos *unspliced* é que eles possuem um baixo potencial codificador. Uma análise *in silico* do potencial codificador usando o *score* normalizado do programa ESTScan2 [99] mostrou que 98% dos *contigs* de EST *unspliced* totalmente intrônicos parecem ser transcritos não codificadores de proteína, dando apoio à idéia de que eles representam uma classe separada de RNAs não codificadores. Com o objetivo de checar se o programa ESTScan2 era capaz de prever corretamente o potencial codificador de um conjunto de dados fragmentados como esse, nós criamos um conjunto de dados virtual composto exclusivamente de 55.139 fragmentos exônicos de genes RefSeq, exatamente com os mesmos tamanhos dos 55.139 *contigs* de ESTs totalmente intrônicos. O programa ESTScan2 conseguiu prever corretamente que 70% dos fragmentos exônicos virtuais de genes RefSeq possuíam potencial codificador. Este achado dá apoio à idéia de que é real a pequena fração de 2% dos *contigs* de ESTs totalmente intrônicos preditos pelo programa ESTScan2 como possuindo potencial codificador, e provavelmente reflete o fato de que na realidade a maioria dos RNAs desta classe (98%) são mensagens não codificadoras.

	<i>Clusters</i> de ESTs com sobreposição à éxons de RefSeq ^a	<i>Clusters</i> de ESTs totalmente intrônicos a RefSeq	<i>Clusters</i> de ESTs mapeando fora de RefSeq	Total
<i>Contigs</i> de ESTs <i>spliced</i>	16,241	8,013	10,144	34,398
Número de éxons de <i>contigs</i> de ESTs <i>spliced</i> (mediana)	10	2	3	
Número total de ESTs <i>spliced</i> nos <i>contigs</i>	3,616,644	162,841	241,049	4,020,534
Número de ESTs <i>spliced</i> por <i>contig</i> (mediana)	91	3	4	
<i>Contigs</i> de ESTs <i>unspliced</i>	4,030	55,139	29,198	88,367
Número total de ESTs <i>unspliced</i> nos <i>contigs</i>	56,752	190,583	140,091	387,426
Número de ESTs <i>unspliced</i> por <i>contig</i> (mediana)	4	2	2	
<i>Singlets</i> de EST <i>spliced</i>	1,053	6,205	6,631	13,889
<i>Singlets</i> de EST <i>unspliced</i>	3,539	121,091	71,662	196,292
Total de <i>clusters</i> de EST não redundante (<i>contigs</i> + <i>singlets</i>)	24,863	190,448	117,635	332,946
Total de ESTs	3,677,988	480,720	459,433	4,618,141

Tabela 2. Classificação dos *clusters* de ESTs em relação às sequências RefSeq *spliced*

a O conjunto de referência (RefSeq) é composto por 15.783 unidades *spliced* não redundantes de RefSeq. Variações adicionais de *splicing* oriundas de sequências de mRNAs foram acrescentadas a cada unidade transcricional de referência mapeada no mesmo *locus*.

Uma curva de distribuição de tamanho dos *contigs* de ESTs totalmente intrônicos foi feita (Figura 13) revelando mensagens com tamanho superior a 1.000 pb. A mediana do tamanho destas mensagens (573 pb) é 4,1 vezes maior do que o tamanho mediano dos

éxons de nossa referência (141 pb). Baseados nisso, nós passamos a chamar estas unidades transcricionais de transcritos longos “Totalmente Intrônicos Não codificadores” ou “TIN”.

Para descobrir se os transcritos TIN poderiam gerar pequenos RNAs, nós comparamos as coordenadas genômicas das seqüências TIN com as coordenadas genômicas dos 346 snoRNAs (<http://www-snorna.biotoul.fr/>) e 383 microRNAs [103] descritos. Usando esta estratégia, nós encontramos 98 snoRNA ou microRNA (14%) mapeando em 86 *contigs* de ESTs TIN, envolvendo apenas uma pequena fração (0,2%) dos 55.139 RNAs TIN de nosso conjunto de dados. Essa pequeníssima sobreposição sugere que os RNAs TIN sejam transcritos intrônicos longos e não precursores dos microRNAs ou snoRNAs já conhecidos.

Figura 13. Distribuição por tamanho de éxons de RefSeq e transcritos não codificadores parcialmente intrônicos (PIN) e totalmente intrônicos (TIN). As curvas mostram a distribuição por tamanho de 3 diferentes classes de transcritos reconstruídas a partir do mapeamento genômico e agrupamento de RefSeq e ESTs do *GenBank*. Curvas de distribuição por tamanho dos éxons de genes RefSeq codificadores de proteína (linha vermelha), *contigs* TIN (linha preta) e PIN (linha azul) são mostradas. *Contigs* TIN e PIN foram montados agrupando ESTs *unspliced* do *GenBank* (em dourado) de um mesmo *locus* genômico, como mostrado no esquema acima das curvas.

4.3.3. Identificação de Transcritos Longos *Unspliced* Parcialmente Intrônicos

Um conjunto de *contigs* de ESTs *unspliced* Parcialmente Intrônicos e Não codificadores (PIN) foi identificado. Um *contig* PIN foi definido como sendo um *contig* que se sobrepunha a um éxon de um gene RefSeq e o estendia por pelo menos 30 bases em

ambas as extremidades (Figura 13). Um total de 12.592 *contigs* de ESTs PIN foram identificados, dos quais 5.992 possuíam evidência de ser antisense ao gene RefSeq correspondente, determinado pela sobreposição genômica deste *contigs* PIN com seqüências de ESTs de bibliotecas direcionadas de alta qualidade [36], regiões transcricionalmente ativas (TARs) em experimentos de *tiling-arrays* do genoma [44] ou com o conjunto público de mRNA *full-length unspliced* do *GenBank*. Verificamos que estes *contigs* PIN eram longos (tamanho mediano de 719 pb). Análise do programa ESTScan 2 estimou que 90% dos transcritos PIN não possuíam potencial codificador.

4.3.4. A Maiorira dos Genes RefSeq Possui Transcrição Intrônica

Nós descobrimos que pelo menos 11.679 genes RefSeq correspondendo a 74% de todos os genes *spliced* humanos na nossa referência possuem regiões intrônicas transcricionalmente ativas onde *contigs* de EST TIN ou PIN foram mapeados. Se nós considerarmos *singlets* de EST TIN ou PIN, a fração de genes RefSeq tendo evidência de transcrição intrônica sobe para 86% de todos os genes RefSeq.

4.3.5. Transcritos TIN e PIN são Possíveis Reguladores de *Splicing* Alternativo

Descobrimos que a frequência média de *skipping* de éxon nos genes de nossa referência que possuem evidência de transcritos PIN é 0,23 e a frequência média de *skipping* de éxons que estão imediatamente 3' de transcritos TIN é 0,22 (Materiais e Métodos item 3.3.2.). Estas frequências de *skipping* de éxons são significativamente maiores ($p < 0,0001$) do que a frequência de *skipping* de todos os éxons (0,14) dos genes RefSeq.

Foram examinados tanto a distribuição de frequências de *skipping* de éxon dentre os diferentes éxons dos genes codificadores de proteína (Figura 14, Painel A) quanto a abundância de *contigs* de EST *unspliced* TIN dentre os diferentes íntrons dos mesmos genes (Figura 14, Painel B). Uma frequência maior de *skipping* de éxon foi detectada próxima à ponta 5' dos genes codificadores de proteínas (Figura 14, Painel A) e uma simultânea maior abundância de *contigs* de ESTs *unspliced* TIN foi detectada nos primeiros 2 íntrons destes genes (Figura 14, Painel B). O tamanho mediano do primeiro íntron da maioria dos genes com transcritos TIN (60-63%) é similar ao tamanho mediano de todos os outros íntrons nos mesmos genes (Figura 14, Painel B). Esperaríamos uma distribuição aleatória de *contig* de ESTs TIN dentre os diferentes íntrons de um gene se a transcrição TIN fosse devido a uma transcrição espúria da RNA Polimerase II. Pelo contrário, nós encontramos uma distribuição desigual, sugerindo que a transcrição TIN pudesse ser frequentemente influenciada pela proximidade de um promotor gênico e que esta transcrição pudesse ser regulada e dirigida por um mecanismo ainda não caracterizado que favorece os primeiros íntrons. Vale a pena notar que uma pequena fração (29-33%) dos genes com transcrição TIN possui o primeiro íntron maior do que os outros, resultando numa distribuição significativamente correlacionada de *contigs* TIN em função do tamanho dos íntrons (Figura 15). A hipótese é que haveria maior espaço para uma possível regulação nestas longas regiões intrônicas destes genes em particular.

A **B**

Figura 14. Frequência de *skipping* de éxon e abundância de transcrição intrônica nos genes RefSeq. Painel A. Distribuição de frequências de *skipping* de éxon ao longo dos genes RefSeq com 7, 8, 9 e 10 éxons. Quadrados cheios indicam a frequência média de *skipping* de éxon por éxon de genes que possuem evidência de RNAs TIN mapeando em seus íntrons. Quadrados abertos indicam a frequência média de *skipping* de éxon para genes sem evidência de RNAs TIN mapeando em seus íntrons. Uma frequência significativamente maior ($p < 0,002$) foi observada para genes RefSeq com transcrição TIN. Painel B. Distribuição de transcritos TIN dentre os íntrons de seqüências RefSeq com 7, 8, 9 e 10 íntrons que foram selecionados do *GenBank* como estando fora do intervalo de confiança de 95% (não correlacionado) numa análise de correlação por Pearson entre a abundância de *contigs* TIN por íntron e o tamanho do íntron (em nt). Barras indicam o tamanho mediano dos íntrons por íntron (em nt) de um conjunto selecionado de genes. Triângulos indicam o número de *contigs* TIN por íntron de genes RefSeq de um mesmo conjunto.

Figura 15. Abundância de transcrição intrônica nos genes RefSeq. Distribuição de transcritos TIN dentre os íntrons de seqüências RefSeq com 7, 8, 9 e 10 íntrons que foram selecionados do *GenBank* como possuindo uma correlação entre a abundância de *contigs* TIN por íntron e o tamanho do íntron (em nt). Uma análise de correlação de Pearson e um intervalo de confiança de 95% foram usados. Barras indicam o tamanho mediano dos íntrons por íntron (em nt) de um conjunto selecionado de genes. Triângulos indicam o número de *contigs* TIN por íntron de genes RefSeq do mesmo conjunto.

Nossas evidências de mensagens *unspliced* que mapeavam totalmente (TIN) ou parcialmente (PIN) em regiões intrônicas vieram dos 55.139 e 12.592 *clusters* de ESTs encontrados, respectivamente. Nós encontramos que a maioria (74%) dos genes definidos como seqüências RefSeq tinham transcrição intrônica. Enquanto uma fração das ESTs *unspliced* intrônicas pode representar éxons de formas alternativas de genes ainda não conhecidas, uma quantidade significativa de transcritos senso e antisense deste conjunto deve ser derivada de unidades transcricionais independentes. Este dado é apoiado pelo fato

do baixo potencial codificador e pelo longo comprimento dos *contigs* de ESTs TIN e PIN (mediana de 573 pb e 719 pb, respectivamente), tipicamente bem maior do que o comprimento dos éxons de genes codificadores de proteínas que verificamos ter um tamanho mediano de 141 pb.

A imensa maioria dos *contigs* de ESTs TIN (99.8%) não possui similaridade com os snoRNAs ou microRNAs. Nós mostramos anteriormente por experimentos de RACE e Northern-blot fita-específico a existência de ncRNAs intrônicos acima de 600 pb de comprimento (Resultados e Discussão itens 4.1.3.2. e 4.2.2.). Estes dados são parecidos com a distribuição por tamanho dos ncRNAs *unspliced* em camundongo que possui um pico acima de 1.000 pb [48], dos ncRNAs poliadenilados identificados em *Drosófila* com tamanhos entre 313 a 2.222 pb [109]. RACE de transcritos não anotados de 10 cromossomos mostraram uma média de comprimento de 680 pb (indo de 173 a 4.650 pb) [110]. Estes números estão de acordo com o comprimento mediano dos RNAs TIN identificados em nossa análise. Mais estudos serão necessários para verificar se esta vasta quantidade de ncRNAs longos, incluindo as mensagens TIN, são precursores de pequenos RNAs ainda não descobertos ou se representam ncRNAs longos maduros.

Não foi encontrada nenhuma correlação entre o tamanho dos íntrons e a abundância de *contigs* de ESTs TIN para a maioria (~60%) dos genes, sugerindo que a transcrição intrônica não ocorre ao acaso. Por outro lado, a consistente correlação observada para aproximadamente 30% dos *contigs* TIN e o tamanho do íntron apóia a hipótese do “genomic design” [27] no sentido de que longos íntrons poderiam carregar mais informações regulatórias como observado para os genes específicos de certos tecidos ou específicos de desenvolvimento [27].

Transcritos intrônicos podem regular a expressão gênica em *trans* ou em *cis* [26][50]. Aqui nós mostramos evidências da influência em *cis* de transcritos intrônicos no *splicing*

alternativo. O *splicing* alternativo afeta mais do que 70% dos genes codificadores de proteínas humanos [70], dos quais o *skipping* do éxon é o evento mais freqüente [71]. Nós encontramos uma forte correlação entre a transcrição TIN e PIN e o *skipping* de éxons de mensagens codificadoras de proteínas de um mesmo *locus*. Um dado interessante é que observamos uma maior incidência de transcrição dos ncRNAs nos primeiros íntrons junto com uma maior freqüência de *skipping* dos primeiros éxons. Exemplos de *skipping* de éxons envolvendo RNAs antisense naturalmente expressos [75][71] ou artificialmente introduzidos [73, 74] em células humanas já foram demonstrados. No entanto, a maior incidência de transcrição intrônica no primeiro íntron, próximo ao promotor do gene poderia ter outras implicações funcionais tais como interferência transcricional do gene [111, 112] ou até mesmo na regulação do promotor gênico usado [113] ou na regulação do início de transcrição pela RNAP II [114].

4.4. Biogênese e Expressão Tecido-Específica de RNAs Intrônicos

Conforme descrito anteriormente, nós utilizamos a tecnologia de *microarrays* de cDNA para estudar o perfil de expressão de transcritos intrônicos em tumores e linhagens de células humanas. Porém, a necessidade de validação, por RT-PCR fita específica, da orientação dos transcritos intrônicos (senso ou antisense) e o número relativamente pequeno de mensagens intrônicas que estão representadas em nosso *chip* de cDNA 4k tornam o processo lento, caro e limitado. Para aumentar nosso poder de análise de expressão, desenhamos uma nova plataforma de *microarrays* contendo sondas senso e antisense para milhares de RNAs TIN e PIN. Esta nova plataforma foi utilizada para identificar transcritos intrônicos com assinaturas de expressão tecido-específicas e revelar possíveis candidatos a reguladores de *splicing*. Além disso, esta ferramenta também nos

permitiu avaliar a fração destes RNAs intrônicos que não parecem ser transcritos pela RNA Polimerase II.

4.4.1. Construção do *Microarray* Intrônico 44k

Sondas de oligonucleotídeos foram desenhadas para as fitas senso e antisenso de 7.715 RNAs TIN e 5.289 RNAs PIN escolhidos de forma aleatória de uma lista de *contigs* de ESTs *unspliced* com as maiores quantidades de ESTs, em cada um dos tipos de transcritos intrônicos. Para cada RNA PIN ou TIN, nós selecionamos todas as possíveis seqüências de 60 pb que satisfaziam uma série de condições [115], descritas a seguir: as sondas não poderiam possuir 8 ou mais bases derivadas de regiões repetitivas do genoma nem possuir seqüências homopoliméricas de 7 ou mais bases (baixa complexidade); as sondas deveriam possuir um conteúdo GC de 35-55% e uma temperatura de melting (T_m) de 68-76°C. Para reduzir a hibridização cruzada, cada seqüência de 60 pb foi submetida a uma procura por BLAST contra um banco de dados contendo todas as regiões genômicas humanas que possuíam qualquer evidência de transcrição por mRNA ou EST. As seqüências de 60 pb que possuíam um segundo *hit* contra o banco de dados com um *bit-score* menor que 42.1 (correspondente a um alinhamento exato máximo em outra região de 21 bases sem *gap*) foram consideradas específicas e levadas ao próximo passo. Os oligonucleotídeos específicos de 60 pb são mapeados de volta ao PIN ou TIN alvo e uma sonda para o *microarray* é selecionada na ponta 3' deste alvo, tomando-se o RNA PIN e RNA TIN mapeado na direção antisenso ao gene codificador de proteína.

Para cada TIN ou PIN alvo, uma segunda sonda senso na fita de DNA complementar à sonda antisenso é selecionada, gerando um par senso-antisenso para cada transcrito intrônico no *microarray*. Para os RNAs PIN, esta sonda senso corresponde ao éxon do

gene onde o RNA PIN se sobrepõe. Para medir o nível transcricional dos genes codificadores de proteínas onde os RNAs PIN e TIN mapeiam, nós incluímos em nosso *array* 14.393 elementos exônicos correspondentes às 7.464 sondas para éxons desenhadas pela empresa Agilent e contidas no conjunto do “Whole Human Genome Oligo *Microarray* set”. Também foram incluídas 2.256 sondas controle da Agilent (IS-44290-1-V1_eQC-V1) para servir de controles positivos e negativos. Nosso *microarray* de oligonucleotídeos 44k intrônico personalizado foi produzido pela empresa Agilent Technologies sob encomenda. O esquema do desenho das sondas pode ser visto na Figura 16.

Figura 16. Visão esquemática da disposição das sondas do *oligoarray* intrônico 44k. Sonda 1 é para os transcritos PIN (seta azul escura). Sondas 3 e 4 para os pares de transcritos TIN antisense e senso (setas preta e hachureada, respectivamente). Sondas exônicas senso 2 (par do RNA PIN antisense) e 5 são para os transcritos codificadores de proteína (bloco e seta vermelhos) de mesmo *locus*.

4.4.2. Medidas de Expressão em Larga-Escala de Transcritos Parcial e Totalmente Intrônicos nos Tecidos Humanos

Os perfis de expressão gênica dos tecidos de próstata, rim e fígado foram medidos usando o *oligoarray* intrônico de 44k. Os *arrays* foram hibridizados com cRNA

amplificado e marcado com Cy3 e Cy5 obtido por uma amplificação linear *in vitro* dos RNAs que contém poli-A (Materiais e Métodos itens 3.4.1. e 3.4.3.). A Figura 17 (Painel A) mostra o número de sondas referentes aos genes codificadores de proteína, TIN e PIN cujos sinais estavam acima do *background* médio mais 2 desvios-padrão de 1.118 controles negativos do *chip* em pelo menos 1 dos 3 tecidos estudados, ou seja, fígado, próstata ou rim (Figura 17, Painel A, coluna “Um”) ou em cada tecido, separadamente (Figura 17, Painel A, colunas “F” - fígado, “P” - próstata ou “R” - rim). É possível observar que enquanto 74% das mensagens codificadores de proteína são expressas, somente 31% dos TIN antisense e 49% dos transcritos PIN antisense são expressos em pelo menos 1 tecido. Uma fração similar de 36% foi observada para a transcrição TIN sense, corroborando o achado da literatura de que ocorre a transcrição natural de unidades transcricionais sense e antisense [44, 45].

Olhando para cada tecido individualmente, podemos notar que 50-69% dos transcritos codificadores de proteína foram expressos, enquanto que 15-33% dos TIN antisense e sense e 21-46% dos PIN antisense foram detectados. Isto revela uma menor abundância da transcrição intrônica em relação às mensagens codificadoras de proteína, tanto em termos de diversidade de mensagens por tecido (Figura 17, Painel A) quanto em termos de distribuição relativa de sinais de intensidades (Figura 17, Painel B). De fato, a intensidade mediana das sondas de transcritos codificadores de proteína era 351, enquanto que para as sondas TIN antisense era 134, para as sondas PIN antisense 136 e para as TIN sense 128. Estes resultados corroboram as observações prévias de que os ncRNAs geralmente ocorrem em baixos níveis de abundância [43, 116].

B

Figura 17. Detecção de transcritos TIN e PIN expressos nos tecidos humanos. Painel A. A expressão de cada classe de transcritos exônicos (barra vermelha), TIN antisense (barra preta), PIN antisense (barra azul) ou TIN senso (barra hachureada) é mostrada. O número total de sondas no *microarray* para cada classe é mostrado (M) junto com o número de sondas detectadas em pelo menos 1 tecido (Um) e sondas exclusivamente detectadas em cada um dos 3 tecidos: Fígado (F), Próstata (P) ou Rim (R). A porcentagem de sondas detectadas relativas ao total de número de sondas no *array* está indicada no topo de cada barra. Painel B. Distribuição da intensidade de sinal das sondas dos TIN antisense (linha sólida preta), TIN senso (linha tracejada), PIN antisense (linha azul) ou exônica (linha vermelha). Intensidades médias de 6 diferentes hibridizações com 3 diferentes tecidos humanos estão mostradas. Somente sondas detectadas acima da média mais 2 desvios-padrão dos controles negativos do *chip* foram consideradas. As intensidades de sondas abaixo do limite de detecção estão mostradas como “não expressa” (linha cinza).

Análise de distribuição ao longo dos cromossomos dos números de RNAs TIN expressos em fígado (Figura 18, barras cinzas) está de acordo com a distribuição revelada pela análise informática baseada no conjunto total de ESTs do *GenBank* (Figura 18, barras pretas). Ambas as distribuições seguem geralmente o número de genes RefSeq em cada cromossomo (Figura 18, barras vermelhas). Como algumas exceções, nós notamos que o cromossomo 10 parece conter uma fração maior de RNAs TIN do que os genes RefSeq codificadores de proteínas, e o cromossomo 19 e X parecem ter uma fração menor da transcrição intrônica em relação ao número de genes codificadores de proteínas.

Figura 18. Distribuição genômica dos RNAs intrônicos. Tamanhos relativos dos cromossomos (barras azuis) e a fração de genes RefSeq (barras vermelhas) mapeados por cromossomo são mostrados. A distribuição ao longo dos cromossomos dos *contigs* TIN (barras pretas) e a distribuição dos RNAs intrônicos expressos (barras cinzas) em fígado humano detectados por *microarray* também são mostrados. Os números no eixo Y se referem a fração de genes de cada cromossomo.

Considerando apenas os transcritos TIN senso e antisense 10% mais expressos em cada tecido e fazendo uma análise de coincidência verificamos que apenas 2–7% da expressão TIN senso e antisense foram detectadas simultaneamente nas 2 fitas de DNA de um mesmo íntron de um gene codificador de proteína. A Figura 19 mostra a distribuição dos transcritos TIN senso e antisense que foram expressos simultaneamente em um mesmo *locus* em função da fração de transcritos expressos em cada 1 dos 3 tecidos. Dentre os

RNAs TIN 50% mais expressos, cerca de 82–89% da transcrição intrônica é um evento específico de uma das fitas de DNA (Figura 19). Mesmo quando consideramos todos os transcritos TIN expressos (100%), 62–78% foram encontrados como exclusivos de apenas uma fita de DNA. Isto sugere que a maioria das mensagens senso e antisense são unidades transcricionais independentes. É notório que os transcritos intrônicos mais altamente expressos sejam fita-específicos, o que novamente sugere um processo celular regulado. Isto também sugere que uma grande fração de transcritos intrônicos pode agir em *trans*, talvez fazendo parte de um grande complexo RNA-proteína que regule a transcrição de um modo análogo ao descrito para os pequenos RNAs regulatórios.

Figura 19. Pares de transcritos TIN senso-antisense simultaneamente detectados em diferentes faixas de intensidade em cada um dos três diferentes tecidos humanos estudados. As porcentagens de pares de transcritos TIN simultaneamente detectados como transcritos em ambas as direções em um mesmo *locus* genômico (símbolos cheios) em diferentes faixas de intensidade são mostradas para cada um dos 3 tecidos: fígado (losangos), próstata (triângulos) e rim (quadrados). As porcentagens de mensagens TIN detectadas como transcritas, em cada tecido, em somente uma das duas fitas do DNA (senso ou antisense) são mostradas como símbolos abertos.

4.4.3. Transcritos TIN Antisenso em Íntrons de Genes Relacionados à Regulação da Transcrição

Nós selecionamos os transcritos TIN antisenso 40% mais expressos em cada tecido e identificamos os genes codificadores de proteína onde estes transcritos mapeavam. A anotação de *Gene Ontology* (GO) destes genes codificadores de proteínas foi comparada usando a ferramenta BiNGO [94] contra a lista completa de genes codificadores de proteína em nosso *chip* de *microarray* que possuíam evidência de transcrição TIN antisenso (Materiais e Métodos item 3.4.5.). A categoria GO “Regulação da Transcrição, dependente de DNA” (GO:006355) foi encontrada como significativamente enriquecida em todos os 3 tecidos. Uma análise de enriquecimento de categorias GO típica é mostrada na Figura 20, Painel A.

Figura 20. Transcritos TIN mais altamente expressos mapeiam em genes relacionados à regulação da transcrição. Dados de expressão de RNAs TIN de 3 tecidos humanos foram usados para selecionar os genes codificadores de proteína nos quais os transcritos TIN 40% mais expressos mapeavam. O programa BiNGO foi utilizado para identificar categorias de GO enriquecidas significativamente ($p < 0,05$) dentro do conjunto de genes codificadores de proteína selecionado. Painei A. Categorias de GO enriquecidas em próstata são mostradas em cores, que são correspondentes ao nível de significância, como indicado na barra de escala de cor. A categoria de GO “Regulação da transcrição, dependente de DNA” (GO:006355) é a mais significativamente enriquecida. Resultados semelhantes foram obtidos para fígado e rim (Figura 21). Painei B. Diagrama de Venn para 132 genes codificadores de proteína que pertencem a categoria GO:006355 “Regulação da transcrição, dependente de DNA”. O número de genes em cada tecido onde a transcrição intrônica foi detectada é mostrado nos parênteses. Os números de genes coincidentes e não-coincidentes entre os tecidos estão mostrados nos círculos.

Resultados similares para rim e fígado estão mostrados na Figura 21. Dentre os 742 genes codificadores de proteína que possuíam os transcritos TIN antisense 40% mais expressos em próstata, um total de 114 pertencem a “Regulação da Transcrição, dependente de DNA” (Figura 20, Painel B). Similarmente, dentre os 450 genes em fígado e 887 em rim, 78 e 128 genes, respectivamente, pertencem a esta categoria de GO. Um total de 132 genes únicos relacionados à categoria “Regulação da Transcrição, dependente de DNA” estão dentre os genes dos 3 tecidos onde os transcritos TIN antisense 40% mais expressos mapeiam. A maioria deles (75 genes, isto é 57%) foi encontrada expressa em todos os 3 tecidos (Figura 20, Painel B), enquanto que uma parte é compartilhada entre 2 tecidos e uns poucos genes são detectados exclusivamente em cada um deles. Esta categoria GO inclui um grande número de diferentes proteínas ligadoras de DNA tais como genes de fatores de transcrição, zinc finger e receptores nucleares. Análises similares com os transcritos TIN senso e PIN antisense 40% mais expressos não identificaram nenhuma categoria GO enriquecida.

Figura 21. Transcritos TIN mais altamente expressos em rim e fígado mapeiam em genes relacionados à regulação da transcrição. Dados de expressão de RNAs TIN de 3 tecidos humanos foram usados para selecionar os genes codificadores de proteína nos quais os transcritos TIN 40% mais expressos mapeavam. O programa BiNGO foi utilizado para identificar categorias de GO enriquecidas significativamente ($p < 0,05$) dentro do conjunto de genes codificadores de proteína selecionado. Painel A. Categorias de GO enriquecidas em rim são mostradas em cores, que são correspondentes ao nível de significância, como indicado na barra de escala de cor. Painel B. Categorias de GO enriquecidas em fígado são mostradas. Em ambos os casos a categoria de GO “Regulação da transcrição, dependente de DNA” (GO:006355) é a mais significativamente enriquecida.

Uma análise similar usando os genes codificadores de proteína 40% mais expressos mostrou um conjunto de categorias de GO totalmente diferente. Entre 10 e 15 categorias foram detectadas como enriquecidas significativamente ($p < 0,05$) em cada tecido e nenhuma estava relacionada à “Regulação da Transcrição” (Figura 22). As categorias mais

[illegible]

B

C

Figura 22. Análise de *Gene Ontology* com os transcritos codificadores de proteína mais expressos. Dados de expressão de 3 tecidos humanos foram usados para selecionar os genes codificadores de proteína 40% mais abundantes. O programa BiNGO foi utilizado para identificar categorias de GO enriquecidas significativamente ($p < 0,05$) dentro do conjunto de genes codificadores de proteína selecionado. Painel A. Categorias de GO enriquecidas em rim são mostradas em cores, que são correspondentes ao nível de significância, como indicado na barra de escala de cor. Painel B. Categorias de GO enriquecidas em próstata são mostradas. Painel C. Categorias de GO enriquecidas em fígado são mostradas.

4.4.4. Um Grande Fração de RNAs TIN e PIN é Insensível à Inibição da RNAP II ou é Até Estimulada por α -Amanitina

α -Amanitina é um octapeptídeo bicíclico extraído do cogumelo *Amanita phalloides* que inibe fortemente a atividade da RNA Polimerase II (RNAP II) [117]. Assume-se que uma grande quantidade de ncRNAs detectados em células humanas sejam transcritos pela RNAP II. Nós tratamos células derivadas de câncer de próstata, LNCaP, com α -amanitina por 24 horas e usamos nosso *oligoarray* para verificar seu efeito na expressão de RNAs intrônicos e codificadores de proteínas. Transcritos diferencialmente expressos (Figura 23) foram determinados com a combinação de duas abordagens estatísticas (Materiais e Métodos itens 3.4.2. e 3.4.4.), o método SAM com FDR = 0,001 [91] e a análise de razão sinal-ruído (SNR) com permutações ($p < 0,05$) [118]. Cerca de 39% (3.909) das mensagens expressas codificadoras de proteína foram significativamente afetadas pela inibição da RNAP II, enquanto que o restante não teve sua abundância alterada e pode-se presumir que sejam mRNAs mais estáveis. Como esperado, a maioria das mensagens codificadoras de

proteínas (96%) foram reguladas negativamente enquanto que 4% foram estimuladas. Nós encontramos 56 RNAs codificadores de proteínas com a expressão gênica aumentada em 2 vezes ou mais em nossos experimentos. Um número similar de RNAs codificadores de proteínas (70 transcritos) foram encontrados por Kravchenko *et al.* [89] como sendo estimulados em 2 vezes ou mais por α -amanitina em células HeLa, numa análise com um *microarray* Affymetrix representando ~20.000 transcritos codificadores de proteína [89].

Este conjunto de experimentos sobre a biogênese dos transcritos intrônicos foi realizado em conjunto com o aluno Paulo Paiva Amaral, aluno de mestrado orientado pela Dra. Aline M. da Silva.

Figura 23. Efeito do inibidor da RNA polimerase II α -amanitina na abundância dos RNAs. As linhas de cada painel representam diferentes transcritos com os níveis de expressão significativamente alterados entre células tratadas ou não com α -amanitina. Cada réplica é mostrada em uma coluna. Para cada linha, as intensidades de expressão entre 2 condições foram normalizadas e coloridas em função do número de desvios-padrão de cada réplica em relação ao valor médio de intensidade de todas as réplicas. Painel A. 3,909 transcritos codificadores de proteína significativamente alterados. Painel B. 400 transcritos TIN antisenso significativamente alterados. Painel C. 468 transcritos TIN senso significativamente alterados. Painel D. 522 transcritos PIN antisenso significativamente alterados.

Uma fração menor de transcritos TIN antisenso (16%) e senso (18%) expressos foi afetada por α -amanitina quando comparada aos 39% da fração das mensagens codificadoras de proteínas afetadas. Porém, as frações de transcritos diferencialmente expressos que foram estimulados nas células tratadas com α -amanitina foi 16% (62/400) e

13% (59/468) para os transcritos TIN antisense e senso, respectivamente. Estas frações são 3 a 4 vezes maiores do que a fração das mensagens codificadoras de proteína estimuladas (4%). Também é importante notar que os níveis dos transcritos PIN antisense exibiram um padrão diferente quando comparado aos transcritos codificadores de proteína, após inibição da RNAP II: somente 19% foram afetados, dos quais 12% (64/522) foram estimulados.

Um dado intrigante é que as mensagens intrônicas (transcritos TIN e PIN) que tiveram um aumento na transcrição, em células tratadas com α -amanitina, mapeiam em íntrons de genes codificadores de proteína que novamente estão enriquecidos de forma significativa ($p < 0,05$) na categoria de GO “Regulação da Transcrição” (Figura 24).

Figura 24. Genes com transcrição intrônica aumentada pela presença do inibidor da RNA polimerase II α -amanitina. Análise de *Gene Ontology* foi realizada com os genes codificadores de proteína que tiveram transcritos PIN e TIN aumentados após o tratamento com o inibidor da RNA polimerase II α -amanitina. Categorias de GO significativamente enriquecidas ($p < 0,05$) são mostradas em cores, que são correspondentes ao nível de significância, como indicado na barra de escala de cor.

Nós consideramos que o critério estrito que foi usado para determinar os transcritos diferencialmente expressos, onde 2 métodos estatísticos são combinados, pode ser uma abordagem conservativa. Portanto, a proporção das mensagens intrônicas estimuladas por α -amanitina pode ser até maior do que o mostrado aqui. De qualquer modo, o número de ncRNAs intrônicos insensíveis à inibição ou estimulados pelo tratamento com α -amanitina

pode representar milhares de transcritos não codificadores se levarmos em conta todos os transcritos intrônicos encontrados em células humanas. Considerando apenas os 55.139 *clusters* de ESTs totalmente intrônicos, espera-se que mais de 1.100 sejam estimulados se pelo menos 13% forem afetados pela inibição da RNAP II.

4.4.5. Assinatura Tecido-Específica dos RNAs TIN e PIN

Assinaturas tecido-específicas da expressão intrônica foram determinadas para os 3 diferentes tecidos humanos: próstata, rim e fígado. Um total de 475 transcritos TIN antisenso (Figura 25, Painel A), 625 TIN senso (Figura 25, Painel B) e 544 PIN antisenso (Figura 25, Painel C) foram detectados usando uma combinação de 2 métodos estatísticos. Nenhuma categoria de GO está significativamente enriquecida nos genes codificadores de proteína que contém as centenas de transcritos intrônicos presentes nas assinaturas (Figura 25, Painel A a C).

Uma assinatura tecido-específica contendo 2.979 transcritos codificadores de proteína foi também identificada (Figura 25, Painel D). Uma análise de enriquecimento de GO (Figura 22) revelou que em fígado, a assinatura de tecido dos genes codificadores de proteína está enriquecida em categorias de GO relacionadas ao ciclo da uréia (GO: 006594), metabolismo de cisteína (GO: 006534), biossíntese de colesterol (GO: 008203) e metabolismo de prostaglandina (GO: 006693) enquanto que em rim, está enriquecida em categorias de GO relacionadas ao transporte de íons sódio e potássio (GO: 006834 and GO: 006813, respectivamente). Em próstata, nenhuma categoria de GO estava enriquecida, no entanto genes específicos de próstata foram encontrados como *KLK3* e *TMEPAI*.

Figura 25. Assinaturas tecido-específicas de transcritos intrônicos e codificadores de proteínas. Transcritos possuindo níveis diferentes de expressão entre próstata, rim e fígado foram selecionados por análise estatística SAM combinada com ANOVA e agrupados hierarquicamente. Em cada painel os transcritos são mostrados nas linhas e as réplicas nas colunas. Para cada linha, as intensidades de expressão entre os tecidos foram normalizadas e coloridas em função do número de desvios-padrão de cada réplica em relação ao valor médio de intensidade de todas as réplicas. Painel A. Assinatura tecido-específica de 475 transcritos TIN antisense. Painel B. Assinatura tecido-específica de 625 transcritos TIN sense. Painel C. Assinatura tecido-específica de 544 transcritos PIN antisense. Painel D. Assinatura tecido-específica de 2.979 transcritos codificadores de proteína.

Nós procuramos por pares de mensagens intrônicas e codificadoras de proteína que pudessem estar co-reguladas num mesmo tecido de um mesmo *locus* genômico com o objetivo de identificar ncRNAs potencialmente envolvidos na modulação da expressão gênica de uma maneira *cis*. Para isso, nós cruzamos as listas de RNAs PIN antisense (Figura 25, Painel C) e genes codificadores de proteína (Figura 25, Painel D) que apareceram nas respectivas assinaturas tecido-específicas para ver se os éxons que possuíam sobreposição com os PIN estavam presentes (Figura 16, sonda 2). Nós encontramos 117 *loci* de genes onde o RNA PIN antisense e o éxon-alvo do PIN eram simultaneamente detectados nas duas assinaturas. Um padrão similar de expressão entre os tecidos foi encontrado para um sub-conjunto de 59 *loci* (Figura 26, Painel A, blocos esquerdo e central). Neste sub-conjunto, o éxon 3' do transcrito codificador de proteína (Figura 26, Painel A, bloco direito) segue o mesmo padrão de expressão. Devemos notar que este é o padrão predominante observado nesta assinatura tecido-específica. É possível que os níveis relativos similares de RNA PIN antisense e dos éxons dos genes

codificadores de proteína reflitam um papel funcional desempenhado pelos RNA intrônico modulando a transcrição ou a estabilidade do transcrito do gene codificador de proteína correspondente. Alternativamente, níveis similares de RNA PIN antisense e RNA codificador de proteína em cada tecido podem ser resultado de um fator comum de modulação simultânea da transcrição de ambas as mensagens do mesmo *locus*.

Num sub-conjunto menor de 12 *loci*, o éxon 3' do transcrito codificador de proteína (Figura 26, Painel B, bloco direito) não acompanha o padrão de expressão entre os tecidos do RNA PIN antisense e seu éxon alvo (Figura 26, Painel B, blocos esquerdo e central). Adicionalmente, para 7 *loci* o RNA PIN (Figura 26, Painel C, bloco esquerdo) possui um padrão de expressão inverso ao éxon que ele sobrepõe (Figura 26, Painel C, bloco central). Devemos notar que para estes 2 conjuntos, há um padrão invertido nos níveis relativos dos éxons sobrepostos por PIN e o éxon 3' (Figura 26, Painéis B e C, blocos central e direito), sugerindo que a mensagem codificadora de proteína sofre um *splicing* alternativo de uma maneira tecido-dependente. Os níveis similares observados para RNAs PIN e os éxons que eles sobrepõem na Figura 26, Painel B (blocos central e direito) sugerem que nestes casos, o RNA PIN pode estar envolvido na retenção do éxon. O padrão invertido observado na Figura 26, Painel C (blocos central e direito) sugere que o RNA PIN possa alternativamente favorecer um *skipping* do éxon que se sobrepõe ao PIN.

Figura 26. Assinaturas tecido-específicas de RNAs PIN antisenso e dos éxons dos transcritos codificadores de proteína correspondentes. Cada linha representa um *locus* genômico coberto por 3 diferentes tipos de sondas (“RNA PIN Antisenso”, “Éxon sobreposto ao PIN” e “Éxon 3’”). Para cada linha, as intensidades de expressão entre os tecidos foram normalizadas e coloridas em função do número de desvios-padrão de cada réplica em relação ao valor médio de intensidade de todas as réplicas. Um subconjunto de 78 pares de RNA PIN e do éxon que ele sobrepõe (“Éxon sobreposto ao PIN”) foi identificado dentre as assinaturas tecido-específicas mostradas na Figura 25 como possuindo padrões de expressão correlacionados das seguintes maneiras: Painel A. 59 pares foram identificados onde o éxon 3’ do transcrito codificador de proteína (painel da direita) segue um padrão similar de expressão do par RNA PIN/ PIN RNA-*overlapped* éxon (painéis da esquerda e do centro); Painel B. 12 pares foram identificados onde o éxon 3’ do transcrito codificador de proteína (painel da direita) não segue o padrão de expressão do par RNA PIN/ PIN RNA-*overlapped* éxon (painéis da esquerda e do centro); Painel C. 7 pares onde o RNA PIN (painel da esquerda) possui um padrão de expressão invertido ao PIN RNA-*overlapped* éxon (painel central).

Um cruzamento análogo entre as assinaturas tecido-específicas de mensagens intrônicas e codificadoras de proteína (Figura 25, Painel D) foi realizado desta vez usando as assinaturas dos RNAs TIN senso e antisenso (Figura 25, Painéis A e B). Nós encontramos 304 *loci* onde o RNA TIN antisenso ou senso e o éxon 3’ do gene codificador de proteína são simultâneamente detectados nas assinaturas de tecido. Um padrão de expressão entre os tecidos similar para RNA TIN antisenso e o éxon 3’ foi detectado num sub-conjunto de 33 *loci* (Figura 27, Painel A). Para 23 pares de RNA TIN antisenso e éxon 3’, um padrão de expressão invertido foi observado (Figura 27, Painel B). Padrões

similares diretos ou invertidos foram observados para 67 pares (diretos) ou 21 pares (invertido) quando cruzamos as assinaturas de tecido dos RNAs TIN senso e os éxons 3' de um mesmo *locus* (Figura 27, Painéis C e D, respectivamente). A maior proporção ($67/88 = 0,76$) de pares diretamente correlacionados de TIN senso e éxon 3' quando comparado com os pares TIN antisenso e éxon 3' ($33/56 = 0,59$) pode ser resultado de anelamento artefactual em segmentos intrônicos de mRNAs prematuros contendo seqüências de poli-A durante a amplificação do cRNA na preparação do alvo marcado. Em analogia aos RNAs PIN co-regulados, a expressão correlacionada observada para os RNAs TIN e os éxons 3' dos transcritos codificadores de proteína pode refletir num papel destes ncRNA na modulação da taxa transcricional ou na estabilidade do RNA codificador de proteína correspondente.

Figura 27. Assinaturas tecido-específicas de RNAs TIN e dos éxons dos transcritos codificadores de proteína correspondentes. Para cada linha, as intensidades de expressão entre os tecidos foram normalizadas e coloridas em função do número de desvios-padrão de cada réplica em relação ao valor médio de intensidade de todas as réplicas. O cruzamento das assinaturas de tecidos, mostradas na Figura 25, identificou subconjuntos de RNAs TIN que possuíam padrões de expressão correlacionados ao éxon 3' dos genes codificadores de proteína situados no mesmo *locus* genômico. Painel A. 33 pares foram identificados onde o éxon 3' do transcrito codificador de proteína (painel direito) possui um padrão de expressão similar ao do RNA TIN antisense (painel esquerdo). Painel B. 23 pares foram identificados onde o éxon 3' do transcrito codificador de proteína (painel direito) possui um padrão de expressão invertido ao do RNA TIN antisense (painel esquerdo). Painel C. 67 pares foram identificados onde o éxon 3' do transcrito codificador

de proteína (painel direito) possui um padrão de expressão similar ao do RNA TIN senso (painel esquerdo). Painel D. 21 pares foram identificados onde o éxon 3' do transcrito codificador de proteína (painel direito) possui um padrão de expressão invertido ao do RNA TIN senso (painel esquerdo).

Com esses dados, nós avaliamos a contribuição da RNAP II na biossíntese dos ncRNAs intrônicos em células humanas, bloqueando sua atividade com α -amanitina e medindo os níveis de expressão dos genes codificadores de proteínas e das mensagens intrônicas. Uma fração considerável (12-16%) dos transcritos total ou parcialmente intrônicos tiveram sua expressão estimulada após a inibição da RNAP II, uma fração 3-4 vezes maior do que a observada para as mensagens codificadoras de proteína (4%). Além disso, uma proporção menor dos transcritos intrônicos (16-19%) foi sensível à inibição da RNAP II quando comparada à fração dos genes codificadores de proteína (39%). O comportamento observado de expressão dos transcritos intrônicos após o tratamento com α -amanitina aumenta a possibilidade de que um sistema diferente transcricional possa estar agindo na biossíntese destes ncRNAs. Um candidato provável seria a RNAP IV nuclear, que é ativada por α -amanitina [89], embora o mecanismo envolvido não seja conhecido.

A especificidade de expressão em diferentes tecidos dos microRNAs em amostras normais e tumorais humanas foi observada [63], provendo fortes evidências de sua relevância fisiológica. Usando um *oligoarray* contendo 19 mil sondas intrônicas, nós mostramos que diferentes conjuntos de RNAs TIN e PIN são transcritos preferencialmente em cada tecido humano analisado (próstata, fígado e rim).

Uma pista da possível relevância funcional dos RNAs intrônicos é dada por nossa observação original de que os RNAs intrônicos antisense mais abundantes são transcritos em íntrons de genes relacionados à regulação da transcrição, o que fornece uma pista para

possíveis papéis funcionais destas mensagens. Regiões genômicas não codificadoras contendo segmentos intrônicos sob forte pressão seletiva têm sido identificadas por diferentes abordagens: identificação de seqüências não transcritas ultraconservadas [24], seqüências conservadas em muitas espécies [25], seqüências conservadas em vertebrados mas altamente divergentes entre humanos e chimpanzés [119], pequenos blocos de seqüências de múltiplas cópias (“pyknons”) [120] ou ausência de inserções de transposons [121]. Um dado interessante é que estas regiões do DNA genômico conservadas são altamente enriquecidas em genes envolvidos com regulação da transcrição [24, 25, 119-121]. Nossos resultados mostram pela primeira vez que os segmentos de DNA intrônico de genes envolvidos com regulação da transcrição são as fontes de transcrição dos RNAs intrônicos mais abundantes nos tecidos humanos. A possibilidade de controle destes genes regulatórios por RNAs intrônicos expressos no mesmo *locus* é atraente. Isto representaria um mecanismo adicional de “regulação dos reguladores”, em um sistema sofisticado de ajuste fino da expressão gênica dos eucariotos.

A maioria dos RNAs TIN e PIN selecionados nas assinaturas tecido-específicas tiveram o éxon 3' correspondente apresentando o mesmo padrão de expressão nos diferentes tecidos estudados. Isso poderia indicar uma ação em *trans* de mensagens intrônicas onde, por exemplo, sua transcrição controlaria a arquitetura regional da cromatina, como demonstrado para ncRNAs longos [122-124]. A super-expressão de íntrons inteiros do gene CFTR afeta a expressão em *trans* de um grande número de mensagens codificadoras de proteínas, muitas das quais ligadas à função do CFTR [20]. Tem sido mostrado também que há milhares de regiões entre os genomas de humanos e camundongos que não podem ser alinhadas e que contém ncRNAs expressos e que assumem estruturas secundárias conservadas [125]. É possível que muitos dos transcritos intrônicos senso e antisense detectados neste trabalho possam agir em *trans* através de

estruturas secundárias similares a alguns ncRNAs longos [126]. Uma explicação radicalmente diferente para esta observação seria baseada na proposição feita por Chakalova *et al.* [127], de que a transcrição de grandes seções de regiões não codificadoras intergênicas fosse a fonte de energia que direcionaria as proteínas motoras da RNA Polimerase II na reorganização da cromatina e no movimento da mesma através de “fábricas nucleares transcrpcionais”. De acordo com esta hipótese, nós especulamos que a transcrição de uma fração de ncRNAs intrônicos senso/antisense iria fornecer energia adicional para ajudar a manter e estabilizar as “fábricas transcrpcionais” que estivessem processando ativamente os genes codificadores de proteína correspondentes.

Uma pequena fração de RNAs PIN, selecionados nas assinaturas tecido-específicas foi identificada como tendo um padrão de expressão nos tecidos correlacionado ao padrão dos correspondentes éxons sobrepostos pelo PIN, e que não era acompanhado pelo padrão de expressão de um éxon 3' distante, sugerindo que isoformas alternativas dos transcritos codificadores de proteína possuíam uma expressão tecido-específica. Um dado recente na literatura identificou um transcrito longo *unspliced* antisense que foi chamado de gene *Saf*. Este gene era um transcrito parcialmente intrônico situado no primeiro íntron de *Fas*, um gene que codifica para uma proteína anti-apoptótica [75]. A super-expressão de *Saf* em células Jurkat induziu a expressão de diferentes isoformas alternativas de *Fas*. Esta super-expressão causou a retenção do éxon alternativo que possuía sobreposição ao *Saf* e o *skipping* de éxons 5' adjacentes, apontando para uma função *cis*-regulatória de RNAs intrônicos antisense [75]. Uma abordagem experimental mais direta poderia revelar novos candidatos para RNA antisense reguladores de *splicing*.

5. CONCLUSÕES

Este trabalho avaliou a contribuição dos íntrons do genoma humano como uma fonte de RNAs não codificadores possivelmente envolvidos em diversos processos biológicos. As análises *in silico* das seqüências intrônicas, depositadas em bancos públicos, combinadas com medidas de expressão em larga-escala, realizadas através experimentos de *microarray*, forneceram ncRNAs intrônicos candidatos a reguladores de *splicing* alternativo ou reguladores de genes envolvidos com regulação da transcrição.

A biogênese e o mecanismo de regulação da transcrição de RNA intrônicos também foram abordados neste trabalho. Experimentos com células tratadas com hormônio e *microarrays* revelaram que mensagens intrônicas podem possuir mecanismos reguladores de transcrição semelhante aos mRNAs e por isso devem ser transcritos pela RNA polimerase II. Alternativamente, foi mostrado que o tratamento de células com α -amanitina alterou a expressão de dezenas ncRNAs intrônicos, apontando para o envolvimento de outra RNA polimerase na transcrição de algumas destas mensagens intrônicas.

Finalmente, foi descrito o desenho e o uso de uma plataforma intrônica/exônica combinada, composta por oligonucleotídeos fita-específicos, muito maior do que a usada nos estudos iniciais. Esta plataforma fornece um balanço conveniente e efetivo entre os *arrays* comumente usados, que são tendenciosos porque contêm apenas os genes codificadores de proteínas, e os chamados *tiling-arrays* contendo todo o genoma humano, que são extremamente caros e de difícil utilização e análise. Estes *oligoarrays* intrônicos/exônicos aqui descritos permitirão uma análise de expressão comparativa das regiões intrônicas sob um diferente número de condições fisiológicas e patológicas, tornando possível o avanço no conhecimento sobre os papéis dos RNAs não codificadores.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
3. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
4. Claverie JM: **Gene number - What if there are only 30,000 human genes?** *Science* 2001, **291**:1255-1257.
5. Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M: **A question of size: the eukaryotic proteome and the problems in defining it.** *Nucleic Acids Research* 2002, **30**:1083-1090.
6. Nature Editorial Board: **The nature of the number.** *Nature Genetics* 2000, **25**:127-128.
7. International Human Genome Sequencing Consortium: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
8. International Human Genome Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: A platform for investigating biology.** *Science* 1998, **282**:2012-2018.
9. Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T, et al: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793-800.

10. Mattick JS: **RNA regulation: a new genetics?** *Nature Reviews Genetics* 2004, **5**:316-323.
11. Mattick JS: **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms.** *Bioessays* 2003, **25**:930-939.
12. Roy SW, Gilbert W: **The evolution of spliceosomal introns: patterns, puzzles and progress.** *Nature Reviews Genetics* 2006, **7**:211-221.
13. Hentze MW, Kulozik AE: **A perfect message: RNA surveillance and nonsense-mediated decay.** *Cell* 1999, **96**:307-310.
14. Clement JQ, Qian L, Kaplinsky N, Wilkinson MF: **The stability and fate of a spliced intron from vertebrate cells.** *Rna-a Publication of the Rna Society* 1999, **5**:206-220.
15. Clement JQ, Maiti S, Wilkinson MF: **Localization and stability of introns spliced from the Pcm homeobox gene.** *Journal of Biological Chemistry* 2001, **276**:16919-16930.
16. Ying SY, Lin SL: **Intronic microRNAs.** *Biochemical and Biophysical Research Communications* 2005, **326**:515-520.
17. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** *Rna-a Publication of the Rna Society* 2005, **11**:241-247.
18. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A: **Identification of mammalian microRNA host genes and transcription units.** *Genome Research* 2004, **14**:1902-1910.
19. Huttenhofer A, Brosius J, Bachellerie JP: **RNomics: identification and function of small, non-messenger RNAs.** *Current Opinion in Chemical Biology* 2002, **6**:835-843.

20. Hill AE, Hong JS, Wen H, Teng LH, McPherson DT, McPherson SA, Levasseur DN, Sorscher EJ: **Micro-RNA-like effects of complete intronic sequences.** *Frontiers in Bioscience* 2006, **11**:1998-2006.
21. Blow M, Futreal PA, Wooster R, Stratton MR: **A survey of RNA editing in human brain.** *Genome Research* 2004, **14**:2379-2387.
22. Athanasiadis A, Rich A, Maas S: **Widespread A-to-I RNA editing of alu-containing mRNAs in the human transcriptome.** *Plos Biology* 2004, **2**:2144-2158.
23. Sun H, Skogerbo G, Chen RS: **Conserved distances between vertebrate highly conserved elements.** *Human Molecular Genetics* 2006, **15**:2911-2922.
24. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**:1321-1325.
25. Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N, Pozzoli U: **Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences.** *Human Molecular Genetics* 2005, **14**:2533-2546.
26. Reis EM, Louro R, Nakaya HI, Verjovski-Almeida S: **As antisense RNA gets intronic.** *Omics-a Journal of Integrative Biology* 2005, **9**:2-12.
27. Vinogradov AE: **"Genome design" model: Evidence from conserved intronic sequence in human-mouse comparison.** *Genome Research* 2006, **16**:347-354.
28. Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M: **Intron size in mammals: complexity comes to terms with economy.** *Trends In Genetics* 2007, **23**:20-24.

29. Ferguson DA, Chiang JTA, Richardson JA, Graff J: **eXPRESSION: An in silico tool to predict patterns of gene expression.** *Gene Expression Patterns* 2005, **5**:619-628.
30. Reis EM, Ojopi EPB, Alberto FL, Rahal P, Tsukumo F, Mancini UM, Guimaraes GS, Thompson GMA, Camacho C, Miracca E, et al: **Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer.** *Cancer Research* 2005, **65**:1693-1699.
31. Kan ZY, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Research* 2001, **11**:889-900.
32. Thanaraj TA, Clark F, Muilu J: **Conservation of human alternative splice events in mouse.** *Nucleic Acids Research* 2003, **31**:2544-2552.
33. Gupta S, Zink D, Korn B, Vingron M, Haas SA: **Genome wide identification and classification of alternative splicing based on EST data.** *Bioinformatics* 2004, **20**:2579-2585.
34. Wang JX, Liang P: **DigiNorthern, digital expression analysis of query genes based on ESTs.** *Bioinformatics* 2003, **19**:653-654.
35. Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel CV: **Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues.** *Cancer Immunity* 2004, **4**:2-19.
36. Shendure J, Church GM: **Computational discovery of sense-antisense transcription in the human and mouse genomes.** *Genome Biol* 2002, **3**:1-14.
37. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, et al: **Widespread occurrence of antisense transcription in the human genome.** *Nature Biotechnology* 2003, **21**:379-386.

38. Chen JJ, Sun M, Kent WJ, Huang XQ, Xie HQ, Wang WQ, Zhou GL, Shi RZ, Rowley JD: **Over 20% of human transcripts might form sense-antisense pairs.** *Nucleic Acids Research* 2004, **32**:4812-4820.
39. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420**:563-573.
40. Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, Herbrick JA, Carson AR, Parker-Katiraei L, Skaug J, Khaja R, et al: **Human chromosome 7: DNA sequence and biology.** *Science* 2003, **300**:767-772.
41. Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS: **RNAdb 2.0--an expanded database of mammalian non-coding RNAs.** *Nucleic Acids Res* 2007, **35**:178-182.
42. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Research* 2005, **33**:121-124.
43. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Research* 2004, **14**:331-342.
44. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu XW, Rinn JL, Tongprasit W, Samanta M, Weissman S, et al: **Global identification of human transcribed sequences with genome tiling arrays.** *Science* 2004, **306**:2242-2246.
45. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays.** *Genome Research* 2005, **15**:987-997.

46. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al: **The transcriptional activity of human Chromosome 22.** *Genes & Development* 2003, **17**:529-540.
47. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome.** *Nature Biotechnology* 2002, **20**:508-512.
48. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**:1559-1563.
49. Jongeneel CV, Delorenzi M, Iseli C, Zhou DX, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJG, Vasicek TJ: **An atlas of human gene expression from massively parallel signature sequencing (MPSS).** *Genome Research* 2005, **15**:1007-1014.
50. Mattick JS, Makunin IV: **Non-coding RNA.** *Human Molecular Genetics* 2006, **15**:17-29.
51. Lee RC, Feinbaum RL, Ambros V: **The C-Elegans Heterochronic Gene Lin-4 Encodes Small Rnas with Antisense Complementarity to Lin-14.** *Cell* 1993, **75**:843-854.
52. Tuschl T, Zamore PD, Lehmann R, Bartel DP, Sharp PA: **Targeted mRNA degradation by double-stranded RNA in vitro.** *Genes & Development* 1999, **13**:3191-3197.
53. Herbert A: **The four Rs of RNA-directed evolution.** *Nature Genetics* 2004, **36**:19-25.
54. Kiss T: **Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions.** *Cell* 2002, **109**:145-148.

55. Kishore S, Stamm S: **The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.** *Science* 2006, **311**:230-232.
56. Tsuchiya S, Okuno Y, Tsujimoto G: **MicroRNA: Biogenetic and functional mechanisms and involvements in cell differentiation and cancer.** *Journal of Pharmacological Sciences* 2006, **101**:267-270.
57. Sarnow P, Jopling CL, Norman KL, Schutz S, Wehner KA: **MicroRNAs: expression, avoidance and subversion by vertebrate viruses.** *Nature Reviews Microbiology* 2006, **4**:651-659.
58. Du TT, Zamore PD: **microPrimer: The biogenesis and function of microRNA.** *Development* 2005, **132**:4645-4652.
59. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT: **c-Myc-regulated microRNAs modulate E2F1 expression.** *Nature* 2005, **435**:839-843.
60. Croce CM, Calin GA: **miRNAs, cancer, and stem cell division.** *Cell* 2005, **122**:6-7.
61. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, et al: **MicroRNA gene expression deregulation in human breast cancer.** *Cancer Research* 2005, **65**:7065-7070.
62. Jiang JM, Lee EJ, Gusev Y, Schmittgen TD: **Real-time expression profiling of microRNA precursors in human cancer cell lines.** *Nucleic Acids Research* 2005, **33**:5394-5403.
63. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al: **MicroRNA expression profiles classify human cancers.** *Nature* 2005, **435**:834-838.
64. Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, Iorio MV, Visone R, Sever NI, Fabbri M, et al: **A MicroRNA signature associated with**

- prognosis and progression in chronic lymphocytic leukemia.** *New England Journal of Medicine* 2005, **353**:1793-1801.
65. Werner A: **Natural antisense transcripts.** *RNA Biology* 2005, **2**:53-62.
 66. Chureau C, Prissette M, Bourdet A, Barbe V, Cattolico L, Jones L, Eggen A, Avner P, Duret L: **Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine.** *Genome Research* 2002, **12**:894-908.
 67. Lee JT, Lu NF: **Targeted mutagenesis of *Tsix* leads to nonrandom X inactivation.** *Cell* 1999, **99**:47-57.
 68. Sleutels F, Zwart R, Barlow DP: **The non-coding Air RNA is required for silencing autosomal imprinted genes.** *Nature* 2002, **415**:810-813.
 69. Imamura T, Yamamoto S, Ohgane J, Hattori N, Tanaka S, Shiota K: **Non-coding RNA directed DNA demethylation of *Sphk1* CpG island.** *Biochemical and Biophysical Research Communications* 2004, **322**:593-600.
 70. Johnson JM, Castle J, Garrett-Engele P, Kan ZY, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**:2141-2144.
 71. Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.** *Human Molecular Genetics* 2002, **11**:451-464.
 72. Lopez AJ: **Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation.** *Annual Review of Genetics* 1998, **32**:279-305.
 73. De Angelis FG, Sthandier O, Berarducci B, Toso S, Galluzzi G, Ricci E, Cossu G, Bozzoni I: **Chimeric snRNA molecules carrying antisense sequences against the splice junctions of exon 51 of the dystrophin pre-mRNA induce exon**

- skipping and restoration of a dystrophin synthesis in Delta 48-50 DMD cells.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:9456-9461.
74. McClorey G, Fletcher S, Wilton S: **Splicing intervention for Duchenne muscular dystrophy.** *Current Opinion in Pharmacology* 2005, **5**:529-534.
 75. Yan MD, Hong CC, Lai GM, Cheng AL, Lin YW, Chuang SE: **Identification and characterization of a novel gene *Saf* transcribed from the opposite strand of *Fas*.** *Human Molecular Genetics* 2005, **14**:1465-1474.
 76. White RJ: **RNA polymerases I and III, growth control and cancer.** *Nature Reviews Molecular Cell Biology* 2005, **6**:69-78.
 77. Sim RJ, Belotserkovskaya R, Reinberg D: **Elongation by RNA polymerase II: the short and long of it.** *Genes & Development* 2004, **18**:2437-2468.
 78. Schramm L, Hernandez N: **Recruitment of RNA polymerase III to its target promoters.** *Genes & Development* 2002, **16**:2593-2620.
 79. Cai XZ, Hagedorn CH, Cullen BR: **Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.** *Rna-a Publication of the Rna Society* 2004, **10**:1957-1966.
 80. Lee Y, Kim M, Han JJ, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *Embo Journal* 2004, **23**:4051-4060.
 81. Parizotto EA, Dunoyer P, Rahm N, Himber C, Voinnet O: **In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA.** *Genes & Development* 2004, **18**:2237-2242.

82. de Turris V, Di Leva G, Caldarola S, Loreni F, Amaldi F, Bozzoni I: **TOP promoter elements control the relative ratio of intron-encoded snoRNA versus spliced mRNA biosynthesis.** *Journal of Molecular Biology* 2004, **344**:383-394.
83. Erdmann VA, Barciszewska MZ, Hochberg A, de Groot N, Barciszewski J: **Regulatory RNAs.** *Cellular and Molecular Life Sciences* 2001, **58**:960-977.
84. Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, Hayashizaki Y, Tomita M, Grp RG, et al: **Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection.** *Genome Research* 2003, **13**:1301-1306.
85. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.** *Cell* 2004, **116**:499-509.
86. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, et al: **Distribution of NF-kappaB-binding sites across human chromosome 22.** *Proceedings of the National Academy of Sciences U S A* 2003, **100**:12247-12252.
87. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, et al: **CREB binds to multiple loci on human chromosome 22.** *Molecular and Cellular Biology* 2004, **24**:3804-3814.
88. Kim TH, Barrera LO, Zheng M, Qu CX, Singer MA, Richmond TA, Wu YN, Green RD, Ren B: **A high-resolution map of active promoters in the human genome.** *Nature* 2005, **436**:876-880.

89. Kravchenko JE, Rogozin IB, Koonin EV, Chumakov PM: **Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin.** *Nature* 2005, **436**:735-739.
90. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, et al: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1**:203-209.
91. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:5116-5121.
92. Das PM, Ramachandran K, vanWert J, Singal R: **Chromatin immunoprecipitation assay.** *Biotechniques* 2004, **37**:961-969.
93. Peixoto BR, Vencio RZN, Egidio CM, Mota-Vieira L, Verjovski-Almeida S, Reis EM: **Evaluation of reference-based two-color methods for measurement of gene expression ratios using spotted cDNA microarrays.** *Bmc Genomics* 2006, **7**:35.
94. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks.** *Bioinformatics* 2005, **21**:3448-3449.
95. Neto ED, Correa RG, Verjovski-Almeida S, Briones MRS, Nagai MA, da Silva W, Zago MA, Bordin S, Costa FF, Goldman GH, et al: **Shotgun sequencing of the human transcriptome with ORF expressed sequence tags.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:3491-3496.
96. Sorek R, Safer HM: **A novel algorithm for computational identification of contaminated EST libraries.** *Nucleic Acids Research* 2003, **31**:1067-1074.

97. Carroll P, Coley C, McLeod D, Schellhammer P, Sweat G, Wasson J, Zietman A, Thompson I: **Prostate-specific antigen best practice policy - part I: Early detection and diagnosis of prostate cancer.** *Urology* 2001, **57**:217-224.
98. Chen GQ, Courey AJ: **Groucho/TLE family proteins and transcriptional repression.** *Gene* 2000, **249**:1-16.
99. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proceedings International Conference on Intelligent Systems for Molecular Biology* 1999:138-148.
100. Nelson PS, Clegg N, Arnold H, Ferguson C, Bonham M, White J, Hood L, Lin BY: **The program of androgen-responsive genes in neoplastic prostate epithelium.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:11890-11895.
101. Xu LL, Shanmugam N, Segawa T, Sesterhenn IA, McLeod DG, Moul JW, Srivastava S: **A novel androgen-regulated gene, PMEPA1, located on chromosome 20q13 exhibits high level expression in prostate.** *Genomics* 2000, **66**:257-263.
102. DePrimo SE, Diehn M, Nelson JB, Reiter RE, Matese J, Fero M, Tibshirani R, Brown PO, Brooks JD: **Transcriptional programs activated by exposure of human prostate cancer cells to androgen.** *Genome Biol* 2002, **3**:1-12.
103. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Research* 2004, **32**:109-111.
104. Altschul S, Madden T, Schaffer A, Zhang JH, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Faseb Journal* 1998, **12**:1326-1326.

105. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWweese-Scott C, Geer LY, Gwadz M, He SQ, Hurwitz DI, Jackson JD, Ke ZX, et al: **CDD: a conserved domain database for protein classification.** *Nucleic Acids Research* 2005, **33**: 192-196.
106. Reisman D, Balint E, Loging WT, Rotter V, Almon E: **A novel transcript encoded within the 10-kb first intron of the human p53 tumor suppressor gene (D17S2179E) is induced during differentiation of myeloid leukemia cells.** *Genomics* 1996, **38**:364-370.
107. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y, Grp RG, Members G: **Antisense transcripts with FANTOM2 clone set and their implications for gene regulation.** *Genome Research* 2003, **13**:1324-1334.
108. Dahary D, Elroy-Stein O, Sorek R: **Naturally occurring antisense: Transcriptional leakage or real overlap?** *Genome Research* 2005, **15**:364-368.
109. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, Celniker SE, Rubin GM: **Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:5495-5500.
110. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammanna H, Helt G, et al: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308**:1149-1154.
111. Prescott EM, Proudfoot NJ: **Transcriptional collision between convergent genes in budding yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:8796-8801.
112. Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene.** *Nature* 2004, **429**:571-574.

113. Geirsson A, Lynch RJ, Paliwal I, Bothwell ALM, Hammond GL: **Human trophoblast noncoding RNA suppresses CIITA promoter III activity in murine B-lymphocytes.** *Biochemical and Biophysical Research Communications* 2003, **301**:718-724.
114. Goodrich JA, Kugel JF: **Non-coding-RNA regulators of RNA polymerase II transcription.** *Nature Reviews Molecular Cell Biology* 2006, **7**:612-616.
115. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, et al: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nature Biotechnology* 2001, **19**:342-347.
116. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru KL, Frith MC, Gongora MM, et al: **Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.** *Genome Research* 2006, **16**:11-19.
117. Bushnell DA, Cramer P, Kornberg RD: **Structural basis of transcription: alpha-Amanitin-RNA polymerase II cocrystal at 2.8 Å resolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:1218-1222.
118. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al: **Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
119. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al: **An RNA gene expressed during cortical development evolved rapidly in humans.** *Nature* 2006, **443**:167-172.

120. Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, Platt D: **Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:6605-6610.
121. Simons C, Pheasant M, Makunin IV, Mattick JS: **Transposon-free regions in mammalian genomes.** *Genome Research* 2006, **16**:164-172.
122. Sanchez-Elsner T, Gou DW, Kremmer E, Sauer F: **Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to ultrabithorax.** *Science* 2006, **311**:1118-1123.
123. Sun BK, Deaton AM, Lee JT: **A transient heterochromatic state in *Xist* preempts X inactivation choice without RNA stabilization.** *Molecular Cell* 2006, **21**:617-628.
124. Mayer C, Schmitz KM, Li JW, Grummt I, Santoro R: **Intergenic transcripts regulate the epigenetic state of rRNA genes.** *Molecular Cell* 2006, **22**:351-361.
125. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J: **Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure.** *Genome Research* 2006, **16**:885-889.
126. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG: **A strategy for probing the function of noncoding RNAs finds a repressor of NFAT.** *Science* 2005, **309**:1570-1573.
127. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P: **Replication and transcription: Shaping the landscape of the genome.** *Nature Reviews Genetics* 2005, **6**:669-677.

LISTA DE ANEXOS

Anexos inclusos no CD-ROM suplementar a esta Dissertação

Anexo 1: Reis E.M., Nakaya H.I., Louro R., Canavez F.C., Flatschart A.V., 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

Anexo 2: Louro R., Nakaya H.I., Paquola A.C., Martins E.A., da Silva A.M., Verjovski-Almeida S., Reis E.M. (2004) RASL11A, member of a novel small monomeric GTPase gene family, is down-regulated in prostate tumors. **Biochem Biophys Res Commun** 316: 618-627.

Anexo 3: Reis E.M., Louro R., Nakaya H.I., Verjovski-Almeida S (2005) As antisense RNA gets intronic. **Omics-a Journal of Integrative Biology** 9: 2-12.

Anexo 4: Amaral, P.P. & Nakaya, H.I. (2006) DNA não-codificador: o lixo que vale ouro? **Ciência Hoje** 38: 36-42.

Anexo 5: Louro, R., Nakaya, H.I., Amaral, P.P., Festa, F., Sogayar, M.C., da Silva, A.M., Verjovski-Almeida, S., Reis, E.M. (2007) Androgen responsive intronic non-coding RNAs. **BMC Biology**. (Aceito para publicação)

Anexo 6: 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., Verjovski-Almeida, S. Genome mapping and

expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. **Genome Biology** (Submetido)

Anexo 7: Nakaya, H.I., Reis, E.M., Verjovski-Almeida, S. Concepts on microarray design for genome and transcriptome analyses. Capítulo do livro “**Nucleic acid hybridization: modern applications**”, Editores: Anton A. Buzdin e Sergey A. Lukyanov, Publisher: Springer. (No prelo)

Anexo 8: Súmula Curricular.