UNIVERSIDADE DE SÃO PAULO

INTERUNIDADES EM BIOINFORMÁTICA

INSTITUTO DE QUÍMICA

Programa de Pós-Graduação Interunidades em Bioinformática

VINICIUS RAMOS HENRIQUES MARACAJÁ COUTINHO

Caracterização *in silico* e análise de expressão de RNAs não codificadores longos no genoma de eucariotos

São Paulo

Julho de 2013

VINICIUS RAMOS HENRIQUES MARACAJÁ COUTINHO

Caracterização *in silico* e análise de expressão de RNAs não codificadores longos no genoma de eucariotos

Tese apresentada ao Programa de Pós-Graduação Interunidades em Bioinformática da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioinformática)

Orientadores:

Sergio Verjovski-Almeida

João Carlos Setubal

Durante o desenvolvimento desta tese, o estudante foi contemplado com bolsas da CAPES (6 meses) e FAPESP (48 meses)

São Paulo

2013

VINICIUS RAMOS HENRIQUES MARACAJÁ COUTINHO

Caracterização *in silico* e análise de expressão de RNAs não codificadores longos no genoma de eucariotos

Tese apresentada ao Programa de Pós-Graduação Interunidades em Bioinformática da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioinformática)

Aprovado em: 13/08/2013

Banca Examinadora:

Prof. Dr. Ariane Machado Lima

Instituição EACH - USP

Assinatura _____

Prof. Dr. Fabio Passetti

Instituição INCA - Externo

Assinatura _____

Prof. Dr. Luciana dos Reis Vasques

Instituição UNIFESP - Externo

Assinatura _____

Prof. Dr. Fabricio Martins Lopes

Instituição UTFPR - Externo

Aos meus pais Vandique e Regina,

ao meu irmão Vandique Filho

por todo apoio e incentivo

AGRADECIMENTOS

Uma jornada para conclusão de um Doutorado não é nada fácil. Afinal, são mais de cinco anos da minha vida dedicados quase que exclusivamente a minha formação científica. Cinco anos de uma vida com pouco mais de vinte e oito anos vividos... Quase 20% do que vivi até hoje. Cinco anos da época áurea da juventude, onde muitas ideias, conceitos e ideologias são formados para construção pessoal. Época marcada por momentos importantes, sejam eles barreiras difíceis e amargas, ou conquistas alegres e prazerosas. Ou seja, foram momentos que com certeza me marcarão pelos anos restantes que terei pela frente, dos quais com certeza muitos fizeram parte. São a estas pessoas que dedico e tenho um profundo agradecimento por toda participação que tiveram na minha vida e na construção deste trabalho nestes últimos anos:

Aos professores Dr. Sergio Verjovski Almeida e Dr. João Carlos Setubal, que me deram toda a oportunidade, confiança, estrutura de trabalho, discussões valiosas e orientação durante toda essa jornada.

Ao professor Dr. Eduardo Reis, pelas ricas discussões nos seminários do laboratório.

À Thaise Gambarra, pelo imenso apoio que tem me prestado em toda essa reta final do Doutorado.

Aos amigos Paulo Amaral, Rodrigo Louro,Thiago Venâncio e Helder Nakaya. Por todo o apoio e amizade desde o primeiro estágio que fiz no laboratório, através do programa Aristídes Pacheco Leão de Apoio às Vocações Científicas, da Academia Brasileira de Ciências, na época da graduação, e que foi fundamental para meu interesse em ingressar no laboratório.

Aos amigos Milton Yutaka, Diorge de Souza, Julio Levano, Yuri Moreira e Otto Cerqueira, pelas boas discussões e cervejadas nas noites das terças na velha Casa do Norte. Importantíssimas nos momentos de estresse.

À Thais Gaudêncio e Pablo Riul, amigos desde muito e parceiros valiosos em cervejadas que tinham o intuito de desopilar e gerar ricas discussões científicas desde a época da graduação em João Pessoa.

Aos amigos e Bioinformatas Artur Lopo, Thiberio Rangel e Alexandre Paschoal, parceiros de grandes discussões relacionadas às diferentes análises que todos desenvolveram durante seus trabalhos.

Aos professores Dr. Arthur Gruber e Dr. Alan Durham por terem me proporcionado uma excelente amizade e discussões importantíssimas para minha formação como Bioinformata.

À Leticia Bertoli (Galega), Raoni Vieira (Salomão), Patricio Leal (Patrócio), Felipe Maciel (Sapulha), Aninha Beatriz, Sival Guedes (Zimbábue), Daniel Vincent, Pablo Lira, Seu Cirão Bertoli, Dona Célia Bertoli e todos os agregados da “República dos Paraíbas”. Vivemos momentos de amizade divertidíssimos desbravando a “terra da garôa” desde minha chegada em 2007.

Mais uma vez ao Helder Nakaya, Felipe Beckedorff, Eliezer Stefanelo, Valdir Blasios, Murilo Amaral, Rodrigo Panda (agregado eterno), Renato Grempel, André 01, Mauricio Liesen, Cícero Batata, Itácio Padilha, Letícia Bertoli (Galega), Rino Cazé e Daniel Vincent. Companheiros e grandes amigos de diferentes repúblicas que morei nestes cincos anos importantíssimos para minha formação. Tenho certeza que vivi momentos únicos, incríveis e fundamentais para minha vida profissional e pessoal com cada um deles.

A todos os amigos que fizeram parte do laboratório ao longo destes cinco anos Lab, em especial a Aninha Tahira, Felipe Beckedorff, Angela Fachel, Ana Ayupe, Helder Nakaya, Thiago Venâncio, Rodrigo Louro, Katia Oliveira, Lauren Camargo, Santiago Arias, Rodrigo Borges, Otto Cerqueira e Julio Levano.

À Ana Paula Vidal, que sem ela todos os trabalhos burocráticos junto a USP e a FAPESP estariam enrolados até hoje.

À Patrícia Martorelli, por toda a amizade e “quebra-galhos” nos assuntos junto a CPG da Bioinfo.

Ao Colégio Visão, que me forneceu toda a formação básica escolar e Guilherme Stanford, professor de Biologia por três anos seguidos que tanto me estimulou para seguir nas Ciências Biológicas.

À UFPB por toda a base, aprendizado e estigação para seguir na pesquisa científica no ramo das Ciências Biológicas. Em especial aos professores Rómulo Llamoca-Zárate e Demetrius de Araújo; orientadores, parceiros e amigos durante toda minha jornada de graduação e iniciação científica.

À toda minha familia por todo o apoio nos mais variados momentos ao longo de toda a minha vida.

À FAPESP e à CAPES por fornecerem todo o apoio financeiro para que este trabalho fosse realizado.

O problema são problemas demais
Se não correr atrás da maneira certa de solucionar
(Chico Science)

RESUMO

Maracaja-Coutinho, V. **Caracterização in silico e análise de expressão de RNAs não codificadores longos no genoma de eucariotos**. 2012. 147 p. Tese (Doutorado). Programa de Pós-Graduação Interunidades em Bioinformática. Instituto de Química, Universidade de São Paulo, São Paulo, 2012.

Diversos estudos têm revelado que a maior parte das mensagens transcritas nos organismos superiores é composta por RNAs não codificadores de proteínas (ncRNAs). Os estudos finais do projeto ENCODE mostram que cerca de 75% das bases do genoma humano são transcritas. Nosso grupo vem estudando os RNAs longos (> 200 nt) não codificadores (RNAs TINs, RNAs totalmente intrônicos não codificadores e RNAs PIN, parcialmente intrônicos não codificadores) que potencialmente apresentam papéis regulatórios importantes nos genomas eucariotos. Embora vários estudos já tenham sido realizados a fim de elucidar suas funções, ainda há uma enorme escassez de informações ligadas à biologia destes transcritos, sendo necessária uma análise mais extensa e detalhada acerca de como eles são regulados e quais os seus papéis no universo celular. Neste trabalho realizamos uma extensa busca *in silico* em bancos de dados de ESTs públicas a fim de atualizarmos toda a evidência de atividade transcricional intrônica em 21 diferentes organismos eucariotos. Além disso, realizamos um estudo de expressão intrônica em fígado humano a partir de dados publicados na literatura, obtidos por duas técnicas de expressão gênica em larga-escala: microarranjos de cDNA com sondas intrônico-exônicas 44k e *Massively Parallel Signature Sequencing* (MPSS). Estes estudos de expressão permitiram identificar milhares de transcritos longos (> 200 nt) parcial ou totalmente intrônicos expressos no fígado humano. Posteriormente realizamos uma completa anotação destes RNAs baseada em homologia e estruturas secundárias evolutivamente conservadas e termodinamicamente estáveis, identificando quais deles seriam possíveis precursores de novas ou já conhecidas classes de ncRNAs. Também verificamos que RNAs não codificadores longos (lncRNAs) intrônicos e intergênicos humanos estão expressos em tecidos pancreáticos normais ou neoplásicos. Estes lncRNAs são transcritos em *loci* genômicos que estão enriquecidos com marcas de cromatina associadas a regiões promotoras, são conservados em outras espécies e apresentam estruturas secundárias conservadas e estáveis. Além disso, re-analizamos todo o repertório de ESTs públicas do parasita humano prostostomado *Schistosoma mansoni*, identificando uma variedade de potenciais novos ncRNAs e novos genes ainda não descritos para este organismo. Ainda desenvolvemos uma ferramenta *web:* NRDR – *Noncoding RNA Databases Resource*, um repositório *online* para recuperação de bancos de dados de ncRNAs disponíveis na rede. Também apresentamos o IntromeDB, uma base de dados pública referente a dados da literatura sobre expressão e caracterização deste tipo de ncRNAs intrônicos em organismos eucariotos, obtidos a partir dos mais variados tecidos e condições biológicas. Por fim, desenhamos duas novas plataformas customizadas Agilent de *microarray*

contendo 244 mil sondas para transcritos intrônicos, exônicos e intergênicos humanos. Os resultados obtidos deixam claro que os transcritos intrônicos apresentam diversos sinais biológicos que apontam para um tipo molecular funcional, e que as análises computacionais, em conjunto com estudos de expressão gênica em larga escala (p.ex. microarranjos, MPSS e sequenciadores de última geração), são importantes para o arsenal de ferramentas utilizadas na caracterização biológica destes transcritos.

ABSTRACT

Maracaja-Coutinho, V. **In silico characterization and expression analysis of long non-coding RNAs in eukaryotic genomes**. 2012. 147 p. PhD Thesis. Graduate Program in Bioinformatics. Instituto de Química, Universidade de São Paulo, São Paulo, 2012.

Evidence suggesting that non-coding RNAs (ncRNAs) act as key components for the fine regulation of cellular processes has been accumulating in recent years. The final studies of the ENCODE project showed that around 75% of the bases of the human genome are transcribed. Our group has been studying long (> 200 nt) non-coding RNAs (RNAs TINS, totally intronic non-coding RNAs and PIN RNAs, partially intronic non-coding) that potentially have important regulatory roles in eukaryotic genomes. Although a number of studies have been conducted to elucidate their functions, there is still a huge lack of information related to the biology of these transcripts, and a more extensive and detailed analysis on how they are regulated and what are their actual roles in the cells is warranted. In this work, we performed an extensive in silico search of public ESTs databases in order to update all the evidence of intronic transcriptional activity in 21 different eukaryotic organisms. Furthermore, we conducted an expression analysis of intronic transcripts in human liver based on published datasets obtained with two techniques of large-scale gene expression measurements: intronic-exonic 44k microarrays and Massively Parallel Signature Sequencing (MPSS). These expression studies have identified thousands of long (> 200 nt) partially or totally intronic transcripts expressed in human liver. Subsequently we performed a complete annotation of these RNAs based on homology and on evolutionarily conserved secondary structures and thermodynamic stability, identifying which of them would be possible precursors of new or already known classes of ncRNAs. We also found that human long intronic and intergenic non-coding RNAs (lncRNAs) are expressed in normal or neoplastic pancreatic tissues. These lncRNAs are transcribed from genomic *loci* enriched with chromatin marks associated with promoter regions and showed to be originated from regions of the genome conserved in other species, as well as to be folded into stable and conserved RNA secondary structures. Moreover, we re-analyzed all public ESTs of the human parasite *Schistosoma mansoni*, identifying a variety of potential new ncRNAs and new genes not yet described for this organism. We also developed a web tool named NRDR - Noncoding RNA Databases Resource, an online repository for ncRNAs dabatases retrieval; and we present the IntromeDB, a public database related to data from the literature on expression and characterization of intronic ncRNAs of eukaryotic organisms obtained from a variety of tissues and biological conditions. Finally, we designed two new Agilent 244k intronic-exonic-intergenic microarray platforms for human transcripts. The results obtained clearly show that intronic transcripts have different biological signals, which indicate molecular functionalities, and suggest that computer analysis coupled to large-scale

experiments (eg. microarrays, MPSS, NGS) is an important strategy in the arsenal of tools used in the biological characterization of these transcripts.

LISTA DE ABREVIATURAS E SIGLAS

AS	Antisenso
AR	Receptor de andrógeno
BKPM	*Bases per Kilobases per Million mapped bases*
BLAST	*Basic Local Alignment Search Tool*
BLAT	*BLAST-like Alignment Tool*
CAGE	*Cap-analysis gene expression*
cDNA	DNA complementar
CGI	*Common Gateway Interface*
ChIP	Imunoprecipitação da cromatina
ChIP-seq	Imunoprecipitação da cromatina seguida de sequenciamento
CNE	*Conserved Non-coding Elements*
CPC	*Coding Potential Calculator*
dbEST	*EST database do NCBI*
DNA	Ácido desoxirribonucleico
ENCODE	*Encyclopedia of DNA Elements*
EST	*Expressed Sequence Tag*
FDR	*False Discovery Rate*
GEO	*Gene Expression Omnibus*
GO	Ontologia Gênica
H3K4me3	Histona H3 tri-metilada na lisina 4
HTML	*HyperText Markup Language*
HUGO	*Human Genome Organization*
IT	Intrônico
kb	Quilobases
KS	*Kolmogorov-Smirnov*
lncRNA	RNA longo não codificador de proteína
miRNA	microRNA
MPSS	*Massively Parallel Signature Sequencing*
NCBI	*National Center for Biotechnology Information*
NATs	*Natural Antisense Transcripts*
ncDNA	DNA não codificador de proteína

ncRNA	RNA não codificador de proteína
nt	Nucleotídeos
ORF	*Open Read Frame*
PDAC	Adenocarcinoma ductal pancreático
PHP	*Hypertext Preprocessor*
piRNA	*Piwi interacting RNA*
RefSeq	*NCBI Reference Sequence Database*
RNA	Ácido ribonucleico
RNA-seq	Sequenciamento de RNAs
rRNA	RNA ribossomal
siRNA	*small interfering RNA*
snoRNA	*small nucleolar RNA*
SQL	*Structured Query Language*
SRP RNA	*Signal Recognition Particle RNA*
TERC	*Telomerase RNA component*
TPM	*Transcripts per million*
tRNA	RNA transportador
TSS	Lugar de início de transcrição
UCSC	*University of California, Santa Cruz*
UTR	*Untranslated regions*

LISTA DE FIGURAS

Figura 30 – Desenho esquemático da distribuição das sondas desenhadas em nosso novo oligoarray 244k intrônico-exônico-intergênico.

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Introdução

1. INTRODUÇÃO

1.1. O transcriptoma e a complexidade biológica

Há exatos 35 anos, Williamson (Williamson, 1977) sugeria pela primeira vez a existência das regiões intrônicas nos genomas eucarióticos em um comentário na revista Nature. No mesmo ano, Phillip Sharp e colaboradores publicaram um artigo mostrando o *splicing* de um gene (Berget et al, 1977), definindo a existência de éxons e íntrons. Entretanto, durante mais de duas décadas, o estudo de transcritos originados nas regiões intrônicas por transcrição independente foi posto de lado pelos grandes grupos de pesquisa pelo fato de regiões intrônicas não serem capazes de codificar proteínas. Esta classe de RNAs não codificadores fugia do pressuposto por Francis Crick, em meados do último século, o chamado “dogma central da Biologia Molecular”, de que toda informação genética nos seres vivos seguiria um único fluxo: DNA→RNA→Proteína (Crick, 1970). Nesta visão, os RNAs atuariam como principais intermediários na produção das proteínas, o que levou grande parte das pesquisas genômicas a focar seus estudos apenas nos RNAs codificadores de proteína, compostos de éxons concatenados pelo mecanismo de *splicing*.

Após o sequenciamento completo do genoma humano (Lander et al, 2001), e sua anotação detalhada (Human Genome Sequencing, 2004), identificou-se entre 20 e 25 mil genes codificadores de proteínas presentes no genoma. Estima-se que das 3 bilhões de bases que constituem o genoma humano menos de 2% codificam proteínas. Em paralelo, com a finalização de inúmeros outros projetos genomas, foi possível obter uma comparação entre a quantidade de genes codificadores de proteínas dentre os mais diversos organismos eucariotos, evidenciando uma enorme semelhança no número de genes codificadores entre esses organismos. A Tabela 1 lista o número de genes codificadores de proteínas bem anotados (genes RefSeq) atualmente para seis espécies evolutivamente distintas, de acordo com o NCBI (*National Center for Biotechnology Information*).

Estes números semelhantes fizeram surgir uma discussão que aponta para o fato de que o grau de complexidade dos organismos pode não estar correlacionado apenas com o número de genes codificadores de proteínas que ele contém, como previsto anteriormente. A complexidade ao longo da evolução estaria diretamente associada com a expansão dos elementos não codificadores de proteínas do genoma.

Uma comparação entre o crescimento no número de regiões codificadoras de proteínas, em relação a regiões não codificadoras na história evolutiva de espécies dos mais diversos ramos evolutivos, revela claramente que a expansão do DNA não codificador (ncDNA), especialmente em regiões intrônicas, é bem maior que a expansão das regiões codificadoras (Mattick, 2004).

Tabela 1 – Número de genes codificadores de proteínas para algumas espécies pertencentes a diferentes ordens evolutivas. Dados obtidos em julho de 2012, levando em conta os genes RefSeq não redundantes do NCBI. Apenas os NM_RefSeq foram considerados na contagem.

Espécie	Genes codificadores
Trypanosoma cruzi (protozoário)	22.570
Caenorhabditis elegans (verme)	20.603
Drosophila melanogaster (mosca-das-frutas)	14.399
Takifugu rubripes (peixe)	22.089
Mus musculos (camundongo)	26.996
Homo sapiens (espécie humana)	23.299

Todas estas informações fizeram emergir a hipótese de que a regulação das funções complexas é mediada a partir de mecanismos sofisticados envolvendo RNAs não codificadores de proteínas. Provavelmente, isto não foi evidenciado anteriormente devido em parte ao pensamento conservador existente na biologia molecular, focando suas atenções exclusivamente nas proteínas como principais moduladores funcionais da célula.

O aperfeiçoamento das técnicas de análise de expressão gênica em larga escala fez com que as atenções de inúmeros grupos de pesquisas em todo o mundo se voltassem mais recentemente para a determinação de toda a atividade transcricional dos organismos, por meio de ferramentas que permitem medir a transcrição ao longo de todo o genoma. O acúmulo recente de grandes quantidades de dados de sequenciamento de transcritos expressos (Core et al, 2008; Djebali et al, 2012; Eswaran et al, 2012; Guffanti et al, 2009; Mortazavi et al, 2008; Oliver et al, 2009; Prensner et al, 2011), e de experimentos utilizando *tiling arrays* (Johnson et al, 2005; Perez et al, 2008; Wilhelm et

al, 2008) ou *oligoarrays* intrônicos (Baratti et al, 2010; Nakaya et al, 2007; Secco et al, 2009), evidenciaram uma extensa atividade transcricional disseminada fora dos éxons dos genes codificadores, especialmente nos eucariotos superiores. Como consequência disso tudo, o número de artigos relacionados à descrição e ao estudo de ncRNAs tem crescido consideravelmente nos últimos anos (Mattick, 2009).

A etapa piloto do projeto ENCODE (*Encyclopedia of DNA Elements*), por exemplo, desenvolveu uma exaustiva caracterização da atividade transcricional em um trecho selecionado de apenas 1% do genoma humano, e observou que mais de 90% do genoma naquela região apresenta alguma atividade transcricional (Birney et al, 2007a). Os dados levaram os autores a estimarem que, se esta observação fosse extrapolada para o restante do genoma, mais de 90% da parte não repetitiva do genoma humano seria transcrita, sendo a grande maioria destes transcritos correspondente a RNAs não codificadores de proteínas originados dentro de íntrons de genes (Birney et al, 2007a). Em setembro do ano passado, a segunda etapa do projeto ENCODE foi concluída (Consortium et al, 2012). Esta nova caracterização envolveu um extensivo estudo dos elementos genéticos de todo o genoma humano, o que resultou na publicação de 30 artigos científicos nas revistas Nature, Genome Research, Genome Biology e BMC Genetics (Consortium et al, 2012); os dados confirmam que pelo menos 75% do genoma humano é transcrito, em uma ou mais de uma entre as 15 linhagens celulares estudadas (Djebali et al, 2012).

Grande parte destas mensagens é de sequências senso ou antisenso de regiões intrônicas do genoma humano (Birney et al, 2007b; Djebali et al, 2012; Galante et al, 2007; Nakaya et al, 2007). Outros estudos também evidenciaram este alto nível transcricional em diferentes organismos eucariotos complexos, como *C. elegans*, onde 70% do seu genoma é transcrito (Gerstein et al, 2010), e *D. melanogaster*, onde 85% do seu genoma é transcrito (Graveley et al, 2011).

Estes estudos deixam claro que os íntrons podem não ser apenas meros “lixos” genômicos como antes eram considerados. A possibilidade de identificar um novo mecanismo modulador da regulação de toda a complexidade gênica motivou o nosso grupo a ingressar na caracterização das mensagens existentes neste tipo molecular, um campo de importância ímpar para a elucidação do real fluxo da informação genética, porém, ainda com números escassos de informações relevantes.

Nos últimos anos o nosso laboratório identificou que mais de 70% de todos os genes humanos atualmente anotados (RefSeq) apresentam transcrição intrônica (Nakaya et al, 2007), gerando transcritos de RNAs longos não codificadores de proteínas parcialmente (RNAs PIN) ou totalmente (RNAs TIN) intrônicos. Uma ocorrência interessante neste novo tipo molecular é a existência de um padrão conservado de expressão tecido-específica em humano e camundongo (Louro et al, 2008), e o fato de apresentarem um maior padrão de expressão tecido-específica que aqueles transcritos originados a partir de RNAs mensageiros codificadores de proteínas (Birney et al, 2007a). Além disso, RNAs intrônicos também têm demonstrado possuir um perfil de expressão menos variável entre amostras de cordão umbilical e sangue de cordão do mesmo doador que aquelas obtidas a partir de doadores distintos (Secco et al, 2009), sugerindo uma especificidade doador-específica em níveis transcricionais. Outro dado interessante observado pelo nosso grupo, é que um conjunto destes longos RNAs intrônicos tem demonstrado importância no câncer humano, apresentando um padrão de expressão com uma elevada correlação com o grau de diferenciação dos tumores em câncer de próstata (Reis et al, 2004), como também diversos deles foram identificados com expressão diferencial em câncer renal de células claras (Brito et al, 2008) e em cancer de pâncreas (Tahira et al, 2011).

Este enorme número de transcritos não codificadores aponta para uma arquitetura transcricional extremamente complexa nos organismos eucariotos, que inclui em média 12 formas alternativas de processamento (*splicing*) para cada mensagem codificadora de proteína (Djebali et al, 2012), além de uma transcrição antisenso, intergênica e intrônica disseminada (Figura 1), que provavelmente possui importantes funções biológicas, mas com um número muito pequeno delas elucidadas. Devido a todas estas descobertas recentes em relação à transcrição dos eucariotos, a definição do que é um gene está requerendo uma redefinição (Djebali et al, 2012). Uma possível definição é: “O gene é uma união de sequências genômicas que originam um coerente grupo de produtos funcionais” (Gerstein et al, 2007). A Figura 1 exemplifica alguns tipos transcricionais já identificados no genoma de organismos complexos.

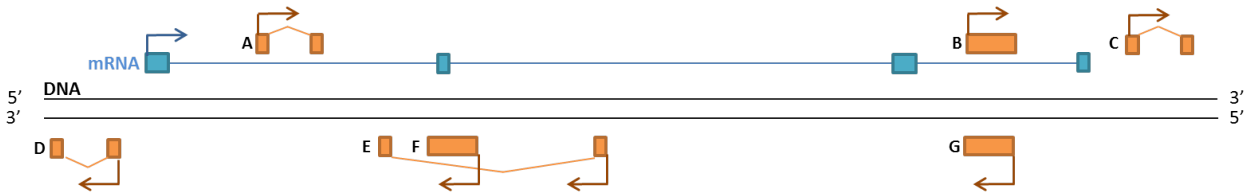


Figura 1 – Representação gráfica dos mais variados tipos de transcritos existentes em genomas de eucariotos. Em azul está representado o mRNA, enquanto que em laranja representamos os RNAs não codificadores de proteínas. (A) e (B) são exemplos de transcritos intrônicos totalmente intrônicos com e sem evidências de *splicing*, respectivamente. (C) e (D) exemplificam transcritos intergênicos. (E), (F) e (G) representam transcritos antissenso a um gene codificador de proteína; (F) e (G) representam transcritos antissenso totalmente e parcialmente intrônicos, respectivamente.

1.2. Regulação da expressão dos RNAs não codificadores longos

O controle preciso do processo de transcrição é fundamental para fazer com que as informações contidas no genoma gerem os eventos biológicos funcionais, envolvendo desde a proliferação, até a diferenciação e o desenvolvimento celular. As células precisam integrar informações intrínsecas e ambientais, e coordená-las em múltiplos mecanismos regulatórios, em diversos níveis, para poder exercer suas funções corretamente.

Muito pouco se sabe acerca dos processos que atuam regulando a transcrição destes longos RNAs não codificadores. Dois trabalhos pioneiros (Cawley et al, 2004; Euskirchen et al, 2004) forneceram algumas evidências que sugerem que a modulação da expressão destes transcritos de RNAs intrônicos pode ser regulada pelos mesmos sinais fisiológicos que normalmente agem em genes codificadores de proteína, como os hormônios. Em seus estudos, eles demonstraram a existência de 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. Estudos mais recentes associaram diversos lncRNAs a assinaturas particulares de marcas de modificação da cromatina que representam regiões genômicas transcricionalmente ativas (Dindot et al, 2009; Guttman et al, 2009). Nestes trabalhos, os autores não deram atenção a marcas ativas em regiões intrônicas de genes codificadores de proteínas.

Dando suporte às evidências de que RNAs não codificadores intrônicos e RNAs mensageiros codificadores de proteínas podem compartilhar um mesmo mecanismo de regulação da transcrição, tomando como modelo experimental células de próstata humana tratadas com andrógeno, nosso grupo demonstrou que a expressão de alguns ncRNAs intrônicos pode ser regulada direta ou indiretamente por este hormônio (Louro et al, 2007). No referido estudo, foram identificados 39 RNAs intrônicos não codificadores de proteínas cujas abundâncias de expressão eram significativamente reguladas pela exposição ao andrógeno. Diversos estudos utilizando ensaios de imunoprecipitação da cromatina seguidos por sequenciamento massivo de DNA/RNA (ChIP-Seq) ou *oligoarray* (ChIP-chip) usando anticorpos para proteínas reguladoras de transcrição ou marcadores de histonas vêm sendo publicados. Exemplos podem ser observados nos trabalhos: (Guil et al, 2012; Yu et al, 2010; Zhao et al, 2010). No entanto, muito pouca atenção tem sido dada para a correlação entre estes motivos regulatórios e a transcrição intrônica.

1.3. Papéis funcionais dos RNAs não codificadores longos

Até algum tempo atrás, apenas os RNA não codificadores infraestruturais e alguns pequenos RNAs tinham suas funções celulares conhecidas. Algumas classes já bem estudadas incluem os tRNAs (RNAs transportadores) e rRNAs (RNAs ribossomais), que servem como componentes essenciais para a maquinaria de síntese proteica; os snRNAs (*small nuclear* RNAs), necessários para o processamento de mRNAs imaturos e os snoRNAs (*small nucleolar* RNAs), envolvidos na modificação de outros RNAs. RNAs não codificadores com tamanhos muito curtos (aprox. 22 a 30 nt), os chamados siRNAs (*small interfering* RNAs), miRNAs (microRNAs) e piRNAs (*piwi interacting* RNAs) foram identificados na última década e têm demonstrado desempenhar um papel regulatório importante na regulação dos genes codificadores de proteínas em diversos organismos, atuando como participantes ativos de praticamente todos os níveis de regulação gênica em eucariotos (Mattick & Makunin, 2006).

Importantes funções regulatórias, incluindo a regulação de redes que envolvem diretamente moléculas de lncRNAs vem sendo propostas nos últimos anos (Chen et al, 2010; De Lucia & Dean, 2011; Dinger et al, 2009a; Louro et al, 2009; Mattick, 2004; Nag & Jack, 2010; Reis et al, 2004; St Laurent et al, 2009). Estes longos RNAs não-*spliced*

podem possuir de centenas a milhares de nucleotídeos e desempenhar papéis funcionais distintos, como *imprinting* de genes codificadores de proteína (Sleutels, Zwart et al. 2002), inativação do cromossomo X (Lee and Lu 1999), regulação de *splicing* alternativo (Louro et al, 2007; Nakaya et al, 2007; Yan et al, 2005), reparação de danos no DNA (Francia et al, 2012), expressão alterada em câncer e outras doenças (Brito et al, 2008; Reis et al, 2004; Wapinski & Chang, 2011). A Figura 2 sumariza os principais mecanismos de atuação dos ncRNAs em organismos modelo (Brosnan & Voinnet, 2009; Chen & Carmichael, 2010; Mercer et al, 2009; Wilusz et al, 2009).

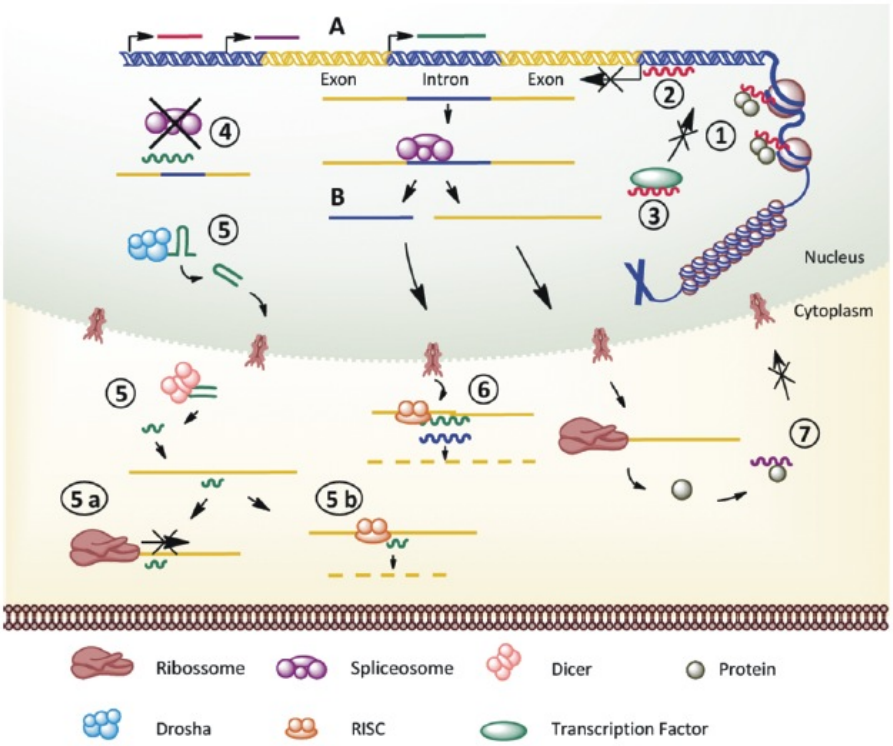


Figura 2 – Principais mecanismos de ação dos ncRNAs longos não-*spliced* já descritos na célula. Mecanismos pré-transcricionais: (1) atuando em complexos de proteínas que estão envolvidos com remodelagem da cromatina, levando à regulação local da expressão gênica; (2) formando *triplex* com o DNA na região promotora de genes, inibindo sua transcrição; (3) interagindo com fatores de transcrição, atuando como um co-ativador ou co-repressor da transcrição; (4) interagindo com o complexo *spliceossomo*, interferindo no processamento do *splicing*; (5) lncRNAs sendo processados em miRNAs e, (5a) inibindo a tradução do mRNA, ou (5b) degradando o mRNA alvo pelo complexo RISC; (6) atuando como *interfering* RNAs endógenos. Mecanismos pós-transcricionais: (7) lncRNAs interagindo com proteínas alterando sua função ou localização celular. Biossíntese dos lncRNAs: (A) podem ser gerados como transcritos independentes, com seu próprio mecanismo de regulação, ou (B) a partir de regiões intrônicos retiradas no processamento do *splicing*. Figura retirada de Oliveira e colaboradores (Oliveira et al, 2011).

O exemplo mais estudado de ncRNAs longos, é o transcrito não codificador *Xist* (*X-inactive specific transcript*), que atua como principal componente no processo de inativação do cromossomo X em mamíferos (Plath et al., 2002). *Xist* é transcrito com 32 kb e sofre *splicing* gerando um transcrito com 19 kb de comprimento em humanos; *Xist* é transcrito a partir do cromossomo X, no qual, posteriormente, é capaz de se ligar, desencadeando todos os processos que levam a sua inativação (Plath et al., 2002), provavelmente através do recrutamento de proteínas silenciadoras. A expressão do gene *Xist*, por sua vez, parece ser regulada por outro ncRNA longo não-*spliced* com 40 kb, *Tsix*, localizado na fita oposta (antisenso) a 15 kb *downstream* do *Xist* (Lee et al., 1999).

Outros casos bem conhecidos são os ncRNAs *HOTAIR* e *p21*. O transcrito longo intergênico *HOTAIR* é transcrito a partir dos *clusters* gênicos *HOX*, e liga-se a proteínas do grupo Polycomb (PRC2) e LSD1, acarretando um silenciamento do gene *HOXD* (Tsai et al, 2010), a partir de modificações na cromatina para um estado repressivo (Gupta et al., 2010). Já o *p21*, é um transcrito longo intergênico responsivo ao p53, que se liga à proteína hnRNP-K e media uma repressão da expressão gênica global e apoptose na via do p53 (Huarte et al, 2010).

Analisando o perfil da transcrição tecido-específica de RNAs intrônicos não codificadores humanos em três tecidos distintos (fígado, rim e próstata), nosso grupo verificou que os transcritos que se apresentavam mais expressos eram totalmente intrônicos a genes relacionados à regulação da transcrição (Nakaya et al, 2007), e que 74% dos genes codificadores de proteinas humanos apresentam transcrição intrônica (Nakaya et al, 2007). Outros indícios que apontam para a importância que estes transcritos intrônicos possuem na regulação de toda a arquitetura e complexidade dos sistemas eucarióticos são a identificação de lncRNAs cuja expressão está significativamente correlacionada com o grau de diferenciação dos tumores, em câncer de próstata (Reis et al, 2004) e a identificação de uma assinatura de expressão em carcinoma de células renais que inclui lncRNAs (Brito et al, 2008); além disso, nosso grupo mostrou que o padrão de expressão tecido-específica de certos lncRNAs é conservado em humano e camundongo (Louro et al, 2008). De fato, a expressão de lncRNAs intrônicos é mais tecido-específica do que a dos mRNAs codificadores de proteínas (Birney et al, 2007a).

1.4. Métodos para identificação de RNAs não codificadores

Todo este aumento na atenção voltada para os ncRNAs, desde a última década, tem levado diversos grupos a estabelecerem novas metodologias específicas para a identificação de ncRNAs em larga-escala. Estas metodologias combinam dados experimentais, oriundos de projetos de transcriptômica e genômica, com abordagens computacionais.

Estudos pioneiros desenvolvidos no início da última década utilizando plataformas de *microarrays*, cobrindo regiões genômicas dos cromossomos 21 e 22 humanos, demonstraram um número cerca de 10 vezes maior de regiões transcricionalmente ativas (Kampa et al, 2004; Kapranov et al, 2002; Rinn et al, 2003) do que o predito anteriormente pelo mapeamento de genes codificadores de proteínas humanos conhecidos. Nestes trabalhos, os autores utilizaram os chamados *tiling arrays*, plataformas compostas por sondas cobrindo trechos contíguos de uma determinada região de interesse do genoma (Figura 3A). Este trabalho motivou outros grupos para focarem seus estudos utilizando sondas não enviesadas para regiões codificadoras de proteínas, como as sondas intrônicas e antisenso (Figura 3B) utilizadas pioneiramente pelo nosso grupo em 2007 (Nakaya et al, 2007).

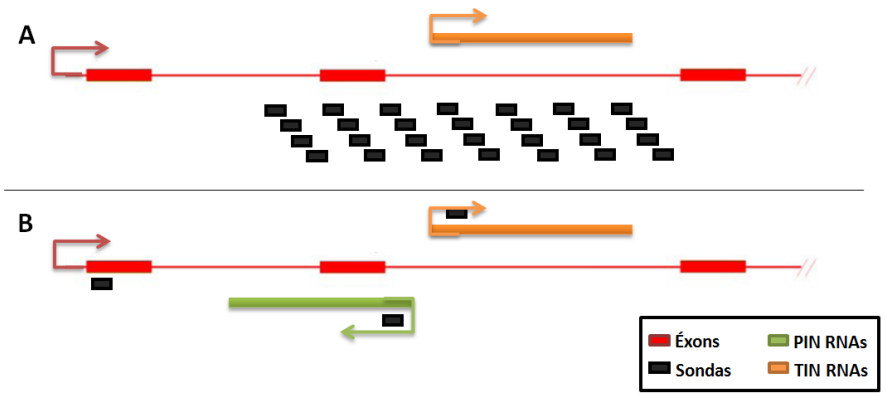


Figura 3 – Representação gráfica de duas plataformas distintas de *microarray*. Em (A), exemplo de uma plataforma de *tiling array*, nas quais as sondas são desenhadas sobrepostas, de maneira a medir a expressão gênica de toda uma região pré-selecionada. Em (B), exemplo de uma plataforma intrônico-exônica, contendo sondas desenhadas exclusivamente para transcritos de interesse, no caso mRNAs codificadores de proteínas e transcritos intrônicos previamente selecionados.

As evidências de transcrição disseminada surgidas na década passada levaram alguns grupos a reanalisarem o catalogo de ESTs disponíveis para diferentes organismos (Nakaya et al, 2007; Numata et al, 2003; Seemann et al, 2007), como também a desenvolverem suas análises abrangendo o sequenciamento de RNAs (RNA-seq) a partir de bibliotecas de cDNA que compreendem uma gama maior da atividade transcricional genômica. Exemplos são as bibliotecas não poliadeniladas (Yang et al, 2011b), e as bibliotecas orientadas (Galante et al, 2007; Mercer et al, 2012; Yassour et al, 2010), que permitem a obtenção de transcritos antes inexplorados, localizados em regiões intergênicas, intrônicas e antisenso no genoma, como também aqueles transcritos poliadenilados ou não.

Após a identificação de novos transcritos obtidos a partir de experimentos de RNA-seq em larga escala ou a partir da reanalise de ESTs públicas contra os genes de referência de um organismo em particular, é necessário verificar se estes transcritos são de fato RNAs não codificadores de proteínas, ou se são um novo gene não codificador, ou mesmo uma nova isoforma variante a partir do processamento de *splicing* de um gene já conhecido. Isto pode ser feito a partir de uma comparação contra bancos de sequências ou motivos de proteínas, como o UniProt (Magrane & Consortium, 2011) e PFAM (Punta et al, 2012). Além disso, diversos programas para predição do potencial codificador de proteínas de um transcrito veem sendo publicados (Iseli et al, 1999; Kong et al, 2007; Lin et al, 2011), que poderiam ser utilizados para analisar os transcritos que não tiverem similaridade com nenhuma sequência ou motivo proteico, descartando, desta forma, a hipótese deste transcrito ser um novo gene codificador para uma proteína ainda desconhecida.

Outra maneira de explorar todo o “RNoma” de um organismo, é buscar por RNAs estruturados, identificados a partir do cálculo da probabilidade de formar estrutura secundária, utilizando-se sua sequência primária. Na formação de uma estrutura secundária, uma região da molécula de RNA forma um pareamento de bases a partir de pontes de hidrogênio com outra região mais distante nesta mesma molécula. Para a formação dessas regiões dupla fita, é necessária a complementaridade entre os nucleotídeos ali presentes. Os métodos atualmente mais confiáveis para identificação de RNAs estruturados baseiam-se na ideia de que a evolução frequentemente preserva as estruturas secundárias dos RNAs, em preferência à sua sequencia primária. Esta abordagem faz uso de métodos estatísticos que definem a estrutura secundária de uma

molécula de RNA, a partir de uma análise da probabilidade de ocorrência de qualquer estrutura secundária em um genoma em estudo, baseando-se na covariação entre os pares de bases pareados observados em sequencias ortólogas (Figura 4) presentes em um alinhamento de sequências entre diferentes organismos, juntamente com o cálculo da energia mínima livre mais confiável dentre todas as possíveis estruturas identificadas (Clote et al, 2005; Washietl & Hofacker, 2004). Utilizando esta abordagem, a partir do alinhamento múltiplo entre sequências genômicas ou ESTs, diversos ncRNAs foram anotados em organismos dos mais variados grupos biológicos (Copeland et al, 2009; Rose et al, 2007; Seemann et al, 2007; Washietl et al, 2005a).

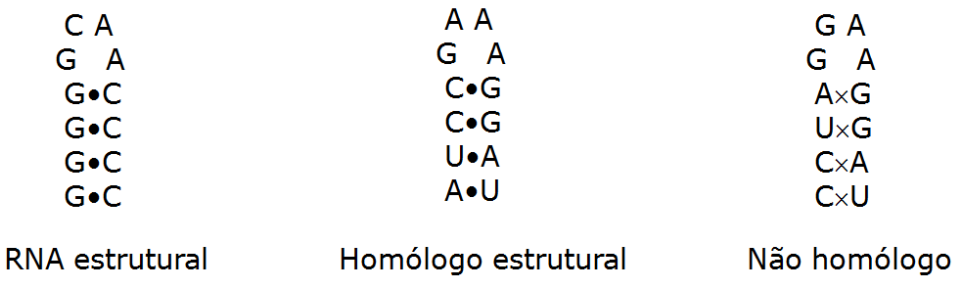


Figura 4 – Representação gráfica de uma estrutura homóloga em organismos ortólogos de acordo com a covariação das bases. Mesmo com a mudança nos nucleotídeos que constituem a sequência linear de um homólogo estrutural, as bases distintas ainda permitem com que a sequência forme um pareamento de bases na mesma região em seus grupos ortólogos, gerando o mesmo *loop* estrutural. Desta forma, uma sequência que seria incapaz de alinhar linearmente com outra, apresenta uma conservação relacionada apenas a sua estrutura. Por outro lado, quando estas variações nos nucleotídeos não permitem o pareamento de bases nestas regiões, a sequência torna-se diferente tanto linearmente, quanto estruturalmente, como no exemplo não homólogo na figura.

Capítulo 2

Objetivos

2. OBJETIVOS

Nos dias de hoje está claro que os RNAs não codificadores de proteínas apresentam-se como um tipo molecular de importância ímpar para a regulação celular. No entanto, apesar dos avanços obtidos desde o início da última década, quando os íntrons deixaram de figurar como meros “descartes” da evolução, ainda há uma grande escassez de informações acerca de toda a sua biologia. Pouco se sabe sobre como estes transcritos são regulados, ou mesmo como eles atuam regulando os mais diversos processos biológicos, sendo necessários estudos mais aprofundados envolvendo a caracterização da biologia destes transcritos. Diante do contexto apresentado, este trabalho tem como objetivos principais:

- Construir uma ferramenta *web* para catalogação das bases de dados disponíveis para as mais variadas classes de ncRNAs.

- Desenvolver um *pipeline* local para identificação e anotação de RNAs intrônicos não codificadores de proteínas.

- Garimpar os dados públicos e catalogar lncRNAs intrônicos em diferentes organismos eucariotos.

- Construir um banco de dados biológico público, contendo informações detalhadas pertinentes aos transcritos intrônicos não codificadores expressos de diversos organismos eucariotos.

- Caracterizar a possível região promotora de transcritos intrônicos não codificadores de proteínas de interesse.

- Caracterizar lncRNAs pertencentes a possíveis novas classes ou a classes já conhecidas de RNAs, bem como estudar sua conservação dentre os organismos.

- Identificar o repertório de ncRNAs presentes no parasita protostomado *Schistosoma mansoni* a partir de ESTs disponíveis em bancos públicos.

Capítulo 3

NRDR (Non-coding RNA Databases Resource): um guia e plataforma web para bancos de dados de RNAs não codificadores

3. NRDR (*Non-coding RNA Databases Resource*): um guia e plataforma web para bancos de dados de RNAs não codificadores

Enquanto realizávamos as mais diversas análises de anotação e caracterização dos ncRNAs ao longo do nosso trabalho, fizemos um extensivo levantamento de todos os bancos de dados referentes a RNAs não codificadores existentes na literatura. Atualmente, existem mais de 100 bases de dados públicas com informações relacionadas a ncRNAs, cobrindo uma vasta gama de famílias de RNAs. Estes, apresentam informações sobre estruturas, tecidos onde são encontrados, condições de expressão, diferentes espécies em que são transcritos, anotações funcionais, filogenia, taxonomia e alinhamentos, potenciais genes alvo, doenças relacionadas, dentre outras. A busca nestes bancos pode ser realizada utilizando um variado conjunto de métodos de busca, como similaridade de sequências, palavras-chave, navegação manual em tabelas, coordenadas de alinhamento genômico e localização genômica de acordo com um *locus* particular contendo grupos de transcritos agrupados.

Dada esta multiplicidade de bases disponíveis, fica claro a necessidade de algum tipo de meta-ferramenta que proporcione um ponto de entrada para a pesquisa e classificação destas informações. A importância de um portal à disposição do público para pesquisar informações referentes a RNAs não codificadores foi salientada em uma recente publicação (Bateman et al, 2011). Neste trabalho, Baterman e colaboradores defendem a construção de um repositório central contendo informações de ncRNAs, e solicitam a formação de um consórcio com este objetivo.

Desta forma, desenvolvemos o NRDR, do inglês *Non-coding RNA Databases Resource*, um portal web com o objetivo de indexar em uma interface amigável todas as bases de dados públicas disponíveis com informações relacionadas a RNAs não codificadores de proteínas. O NRDR permite aos usuários encontrar rapidamente bancos de dados utilizando critérios baseados na *família dos RNAs*, *fonte da informação*, *conteúdo da informação* e quanto aos *mecanismos de busca* disponíveis. Salientamos que o NRDR não é um repositório central, visto que ele não armazena nenhuma informação relacionada a ncRNAs, nem armazena bancos relacionados a tRNAs e rRNAs.

3.1. Levantamento da literatura, construção do banco de dados e implementação da interface web

Para o desenvolvimento do NRDR, realizamos um levantamento dos bancos de dados disponíveis atualmente de maneira manual, a partir de uma extensa varredura da literatura. Todos os 102 sites foram visitados e curados.

O banco de dados do NRDR foi desenhado manualmente utilizando a linguagem SQL (*Structured Query Language*), e implementado no sistema de gerenciamento de banco de dados MySQL (http://www.mysql.com/). O povoamento dos dados no banco foi realizado utilizando a linguagem Perl. Toda a interface web foi desenhada utilizando a linguagem PHP (*Hypertext Preprocessor*), a qual também foi utilizada para integração entre a interface e o banco.

3.2. Visão geral e interface de busca do NRDR

O portal NRDR pode ser acessado em: www.ncrnadatabases.org. Atualmente, a ferramenta armazena 102 bancos referentes a ncRNAs, sendo atualizada a cada seis meses a partir de uma curagem manual da literatura. O sítio disponibiliza uma completa descrição de cada banco de dados indexado e um link para o respectivo website.

O NRDR fornece uma interface de busca amigável, onde pode-se recuperar uma lista de bancos de dados filtrados a partir de critérios definidos pelo usuário. Estão disponíveis caixas de texto para uma busca conjunta, onde várias palavras-chave podem ser aplicadas. Como um exemplo, o usuário pode filtrar a busca utilizando o nome de uma classe de RNA, o nome de um organismo ou um PubMed ID (i.e. microRNA, *Homo sapiens*, número do PubMed ID). Além disso, o usuário pode selecionar apenas bancos que possuem *datasets* em diferentes formatos (i.e. BED, PSL, FASTA) disponíveis para download, ou aqueles bancos que possuem alguma visão gráfica do genoma (*Graphic Genome View*, no portal) disponível (i.e. Genome Browser). Para uma pesquisa mais abrangente, é possível realizar a busca diretamente por palavras-chave que aparecem na descrição do texto de cada banco de dados armazenado no NRDR (i.e. *intergenic*, *transcription*, *networks*). Por fim, também é permitido explorar o NRDR a partir de uma navegação de acordo com o organismo de interesse. As Figuras 5 e 6 demonstram capturas de tela da página de busca e de um

exemplo da descrição de um banco de dados obtido a partir de uma busca em nosso portal.

Os bancos de dados indexados pelo NRDR foram manualmente classificados utilizando quarto critérios:

1. Família dos RNAs: quais classes de ncRNAs estão presentes nos bancos de dados?

2. Fonte da informação: qual a proveniência das informações dos ncRNAs (evidências experimentais; análises computacionais; curagem manual de informação obtida experimentalmente e/ou computacionalmente; ou literatura)?

3. Conteúdo da informação: quais são os vários tipos de informações armazenadas numa base de dados (i.e. sequencia, anotação, expressão, doença)?

4. Mecanismos de busca: quais mecanismos de busca estão disponíveis em uma base de dados?

O NRDR gera automaticamente estatísticas acerca de como os bancos de dados estão distribuídos de acordo com cada critério de classificação. Estas estatísticas estão disponíveis na aba *Statistics* do portal, e foram utilizadas para descrevermos uma visão geral sobre o atual estado da pesquisa com ncRNAs.

Figura 5 – Visão geral do portal NRDR (*Non-coding RNA Databases Resource*). Captura de tela evidenciando os diferentes tipos de buscas que o usuário pode realizar.

3.3. Famílias de RNAs: diferentes classes de transcritos estão presentes em bancos de dados

Um desafio na anotação de novos ncRNAs é categorizá-los em classes bem definidas. Atualmente, o enovelamento estrutural molecular é a característica mais utilizada na classificação de ncRNAs como pertencentes a uma família particular (Gardner et al, 2009; Will et al, 2007), assumindo que transcritos que apresentam conformações estruturais em comum atuam de maneira funcional semelhante nos processos celulares. A similaridade de sequências é uma ferramenta muito importante na classificação de novas classes de ncRNAs, no entanto, apresenta uma eficácia limitada para caracterização de sequências mais divergentes.

Figura 6 – Visão geral do portal NRDR (*Non-coding RNA Databases Resource*). Captura de tela da descrição de um banco de dados (IntmiR) a partir dos que foram listados como resultado de uma busca realizada, preenchendo os critérios de interesse do usuário. A saída inicial da busca realizada na Figura 5 é uma listagem dos bancos (não mostrada na Tese) com links para suas respectivas descrições aqui descritas.

Refletindo a quantidade limitada de dados especificamente funcionais, os conjuntos complexos de ncRNAs curtos e longos que vêm sendo detectados com as tecnologias de expressão em larga escala, estão sendo anotados nos bancos dedados simplesmente como “RNAs curtos” (Johnson et al, 2007; Yang & Qu, 2012), incluindo todas as sequências menores que 200 nt, ou “RNAs longos” (Dinger et al, 2009b) (>200 nt), com pouca informação relacionada a qual classe funcional ou família estes RNAs de fato pertencem.

Recentemente, estas discrepâncias levaram a uma padronização tentativa, organizada pelo *The HUGO Gene Nomenclature Committee* (HGNC), a única organização autorizada a atribuir uma nomenclatura para transcritos humanos (Wright & Bruford, 2011). Eles classificaram estes transcritos longos em uma grande classe chamada *long non-coding RNAs* (lncRNAs), que é subdividida em dois outros grupos principais: (1) transcritos com funções conhecidas (i.e. *XIST*, *HOTAIR*), e (2) aqueles com anotação funcional desconhecida, que são sub-classificados de acordo com sua localização genômica, como: intrônicos (IT), antisenso (AS), intergênicos (LINC) ou transcritos hospedeiros de outros ncRNAs curtos (Wright & Bruford, 2011). De acordo com a classificação atual, denominada ainda como um projeto em andamento pelo HGNC, estes lncRNAs foram descritos como RNAs processados (*spliced*), possuidores de *cap* em suas extremidades 5' e poliadenilados (Wright & Bruford, 2011), o que claramente não abrange todos os diferentes tipos de lncRNAs que podem não ser processados (*unspliced*) e/ou não poliadenilados.

Devido a todas as dificuldades em categorizar RNAs, no NRDR optamos por agrupa-los de acordo com as diferentes classes de RNAs utilizadas pelos próprios bancos de dados, combinando sempre que possível com a nomenclatura proposta pelo HGNC, como exemplificado abaixo. Bases de dados que não são específicas para uma única classe de RNA, como o RFAM (Gardner et al, 2011), ncRNAdb (Szymanski et al, 2007) e NONCODE (Bu et al, 2012) foram classificados como contendo *múltiplas classes* (Tabela 2). Além dos ncRNAs categorizados pelo HGNC, outras classes de ncRNAs estão presentes nos bancos de dados, como: elementos não codificadores conservados, do inglês *Conserved Non-coding Elements* (CNEs), alguns dos quais estão associados com transcritos não codificadores bem conservados (Lee et al, 2007; Woolfe et al, 2007); RNAs estruturados, que são geralmente preditos a partir de algoritmos de enovelamento molecular estrutural em sequências genômicas ou transcriptômicas (Washietl & Hofacker, 2007); mRNA-like RNAs, que são transcritos longos não codificadores sem evidência de *splicing* (Sone et al, 2007; Zhang et al, 2010); e *Natural Antisense Transcripts* (NATs), transcritos longos que são parcialmente complementares a outros RNAs endógenos (Li et al, 2008; Yin et al, 2007; Zhang et al, 2007).

Tabela 2 – Diferentes classes de RNAs listados nos bancos de dados. Sempre que possível tentamos associar os ncRNAs descritos nas bases de dados com a classificação do HGCN (Wright & Bruford, 2011).

	Classe do RNA	Descrição
RNAs curtos		
1	snoRNAs	*small nucleolar RNAs* (i.e. CD and H/ACA box)
2	small RNAs	RNAs curtos (< 200 nucleotídeos), que podem incluir classes ainda não
3	siRNAs	*small interfering RNAs*
4	miRNAs	*microRNAs*
5	SRP RNAs	*Signal Recognition Particle RNA*
6	piRNAs	*piwi-interacting RNAs*
7	Riboenzimas	Rnase P e RNAs intrônicos grupos I, II e III removidos no processamento de
8	TERC	*Telomerase RNA Component*
RNAs longos		
9	long ncRNAs	RNAs longos(> 200 nucleotídeos), que podem incluir classes ainda não
10	NATs	*Natural Antisense Transcripts*, complementares a outros RNAs
Outros		
11	Classes múltiplas	Não específico para uma classe única de RNA, o que significa que existe uma
12	CNE	*Conserved Non-coding Elements*
13	RNAs estruturados	Específico para RNAs estruturados (i.e. estrurura secundária ou 3D)

A Figura 7 apresenta o número de bases de dados publicadas relacionadas com ncRNAs ao longo dos anos. Ela mostra um aumento substancial no número de repositórios disponíveis depois de 2005. As 102 bases de dados levantadas e armazenadas no NRDR foram categorizadas de acordo com as classes de ncRNAs nelas contidas (Tabela 6). A distribuição das bases de dados por classe de ncRNAs é apresentada na Figura 8. Algumas particularidades foram observadas entre os 102 bancos. A maioria deles (73 de 102) armazena RNAs curtos, com 68% deles (50 de 73) relacionados especificamente com microRNAs. Em 2010, Kozomora e Griffitgs-Jones destacaram que nos últimos três anos o número de sequências de miRNAs no banco miRBase tinha quase que triplicado (Kozomara & Griffiths-Jones, 2011), o que reflete o grande interesse da comunidade científica nos microRNAs. Este interesse não é novo, e é devido ao importante papel que estas moléculas desempenham na regulação da tradução de genes e, provavelmente, devido a sua correlação com diferentes patologias e processos de desenvolvimento (Davis-Dusenbery & Hata, 2010; Nicolas & Lopez-Martinez, 2010; Silahtaroglu & Stenvang, 2010).

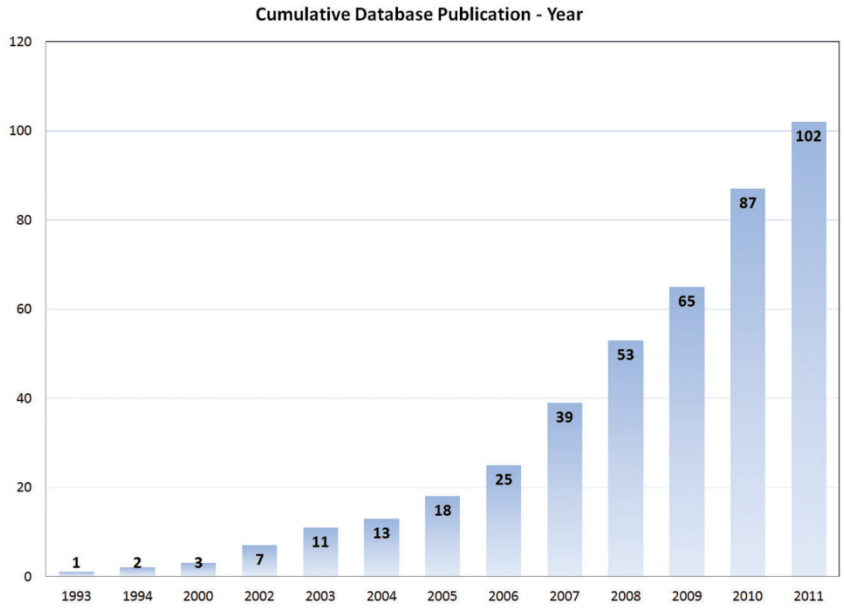


Figura 7 – Publicações referentes a bancos de dados de ncRNAs ao longo dos anos.Os valores de 2011 foram computados até novembro daquele ano.

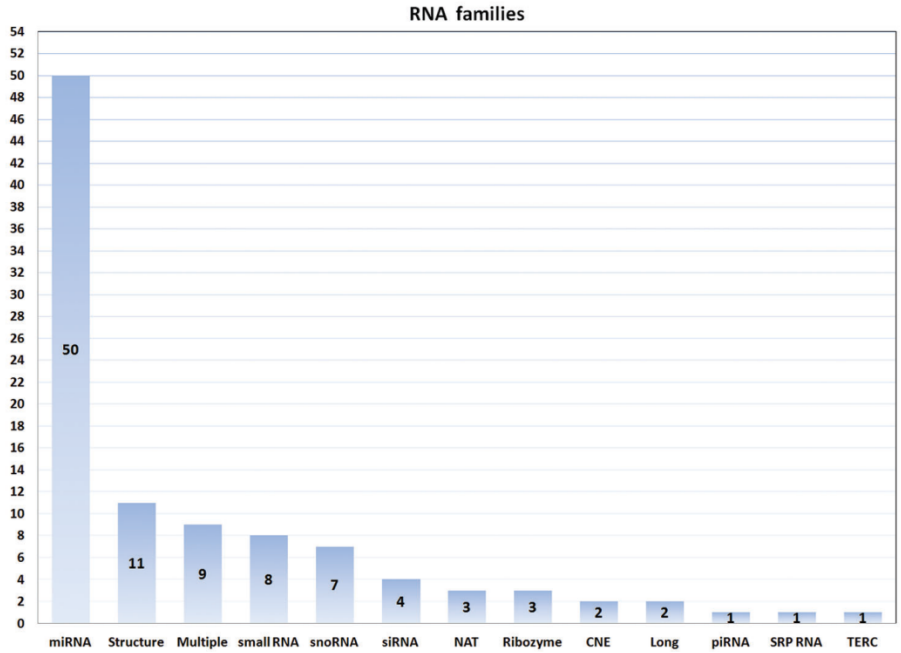


Figura 8 – Distribuição dos bancos de dados de acordo com as classes de ncRNAs.

Embora exista na literatura alguns ncRNAs longos conhecidos e funcionalmente bem anotados, como *XIST* (Plath et al, 2002), *HOTAIR* (Gupta et al, 2010) e *MALAT* (Ji et al, 2003), sua presença em bancos de dados ainda é escassa quando comparada

com RNAs curtos, de maneira que existem apenas cinco bancos específicos para ncRNAs longos (Amaral et al, 2011; Dinger et al, 2009b; Li et al, 2008; Yin et al, 2007; Zhang et al, 2007). Atualmente, existem oito bancos armazenando dados oriundos de tecnologias de sequenciamento de última geração (Backman et al, 2008; Johnson et al, 2007; Molnar et al, 2007; Yang et al, 2011a; Yang et al, 2010; Zhang et al, 2010), no entanto, apenas um deles (ncRNAimprint (Zhang et al, 2010)) é relacionado a RNAs longos (i.e. RNAs longos antissenso e mRNA-like RNAs).

É crescente a atenção que vem sendo dada nos últimos cinco anos aos ncRNAs longos, o que reflete o intensivo uso de tecnologias de larga escala para medições de expressão (Guttman et al, 2009; Kapranov et al, 2007; Nakaya et al, 2007), e no reconhecimento de que alterações na expressão de RNAs longos está presente em diversas doenças, tais como diferentes tipos de câncer humano (Gibb et al, 2011; Huarte & Rinn, 2010; Prensner & Chinnaiyan, 2011). No entanto, apenas uma fração muito pequena de tais dados funcionais estão presentes nos bancos de dados de ncRNAs. Por exemplo, a transcrição inteira do *loci* gênico *HOX* foi analisada utilizando *tiling arrays* ultra densos, identificando o *HOTAIR,* juntamente com outros 170 ncRNAs longos, como sendo estatisticamente diferencialmente expressos entre epitélio humano normal, carcinoma primário de mama, e metástases (Gupta et al, 2010). Com exceção do *HOTAIR*, nenhum das dezenas de ncRNAs longos diferencialmente expressos nos *loci HOX*, correlacionados com o câncer de mama, estão presentes nos bancos lncRNAdb (Amaral et al, 2011) ou NONCODE (Bu et al, 2012), por exemplo. Este último, inclusive, foi recentemente atualizado com o intuito de incluir uma anotação integrativa de RNAs não codificadores longos. Do mesmo modo, estudos com *microarrays* revelaram perfis de expressão contendo outras dezenas de ncRNAs longos correlacionados com um número de diferentes tipos de câncer (Brito et al, 2008; Perez et al, 2008; Reis et al, 2004; Reis et al, 2005; Tahira et al, 2011), dos quais nenhum deles encontra-se depositado em qualquer uma das bases de dados. Isto reflete a ausência atual de uma anotação sistemática e uniforme para a quantidade massiva e diversificada de ncRNAs longos expressos em diferentes tecidos humanos e de outros organismos. Outros trabalhos, utilizando sequenciamento de RNAs direcionados para uma análise profunda do transcriptoma humano (Mercer et al, 2012), ou estudos utilizando *tiling arrays* (Kapranov et al, 2005), relevaram diversas novas isoformas para genes codificadores e não codificadores de proteínas. Novamente, nenhuns destes dados encontra-se

presente nos bancos de dados atuais. Coletivamente, estes resultados apontam que a profundidade e complexidade do transcriptoma humano ainda está longe de ser totalmente caracterizada e esclarecida (Mercer et al, 2012), ao mesmo tempo em que evidenciam a atual limitação e escassez de informação existente em relação a qualquer banco especializado na catalogação de ncRNAs longos.

3.4. Fonte da informação: a confiabilidade dos dados armazenados

As informações armazenadas nestes bancos de dados foram originadas a partir de diferentes fontes, associadas a diferentes graus de confiabilidade. Classificamos as bases de dados em quatro tipos de fontes de informação: (1) *anotação in silico*, na qual os ncRNAs são caracterizados localmente a partir de análises computacionais, (2) *literatura*, na qual os dados são extraídos a partir de artigos científicos publicados, (3) *curagem manual*, onde a informação é validada localmente por um especialista humano, e (4) informação *experimental*, derivada diretamente de ensaios biológicos. Destes, a informação experimental é a que apresenta um maior grau de confiabilidade, visto que foi verificada e testada experimentalmente em laboratório. A Figura 9 mostra a distribuição das bases de dados de acordo com a fonte da informação. Toda essa informação pode ser facilmente recuperada a partir da ferramenta *online*.

Dados experimentais, por exemplo, podem ser observados em 67 dos 102 bancos de dados presentes no NRDR. Esta informação pode ser utilizada como um catálogo para validação de experimentos laboratoriais, como também pode servir de base para testes no desenvolvimento de novas metodologias em biologia computacional. Um problema, no entanto, é o fato que apenas 40 dos 102 bancos disponibilizam seus dados para *download*, dos quais apenas 24 oferecem suas sequencias para *download*, o que é um empecilho para quem deseja realizar buscas comparativas, como também para o desenvolvimento de novos algoritmos para caracterização de ncRNAs. O NRDR pode guiar rapidamente o pesquisador para os bancos que apresentam seus dados disponíveis para *download*, nos mais diversos formatos de interesse (i.e. BED, GFF, PSL, FASTA).

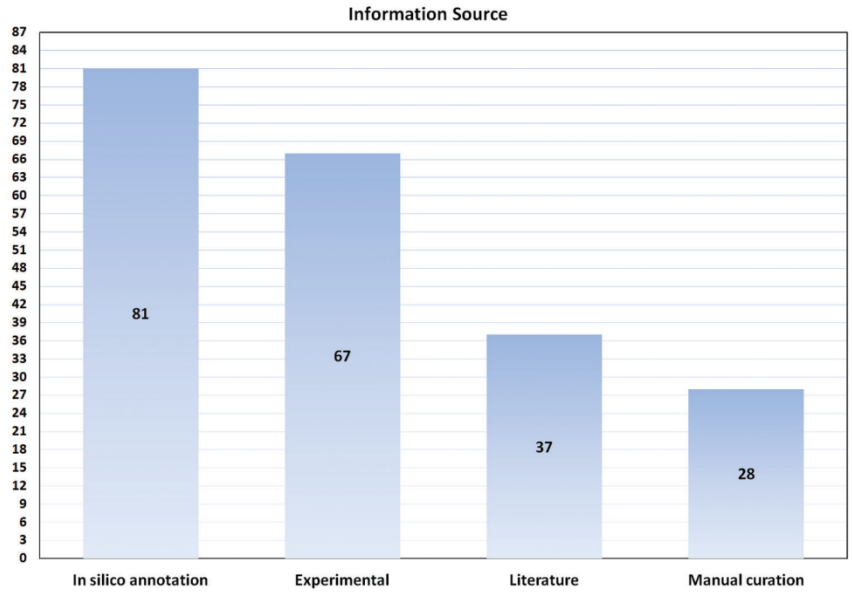


Figura 9 – Distribuição das bases de dados de acordo com cada uma das fontes da informação utilizadas na catalogação dos bancos para o NRDR.

3.5. Conteúdo da informação: a informação armazenada em bases de dados públicas

Esta seção descreve os diferentes tipos de informações armazenadas nos bancos de dados (i.e. sequência, ontologia, expressão). Isso nem sempre está claramente expresso em seu nome, e às vezes nem mesmo em sua *home page*. Por essa razão, cada banco foi examinado manualmente e um número de termos (geralmente 9 a 19 termos) foi associado a cada um deles. Esta categorização foi gerada de uma lista composta por um total de 257 termos controlados (ontologia), originados a partir da soma de todos os termos presentes nos bancos de dados levantados.

A aba *Statistics* no NRDR contém a lista de todos os termos gerados, juntamente com o número de bases de dados que foram associadas a cada um deles. Além disso, acessando a descrição de cada banco no NRDR, o usuário poderá verificar a lista de todos os termos que foram associados a aquele banco. Acessando a aba *Search* no portal, o usuário poderá realizar buscas de acordo com o conteúdo da informação, a partir de uma combinação dos termos de interesse. Na Figura 10, agrupamos e ordenamos os 10 termos mais frequentes associados às bases de dados.

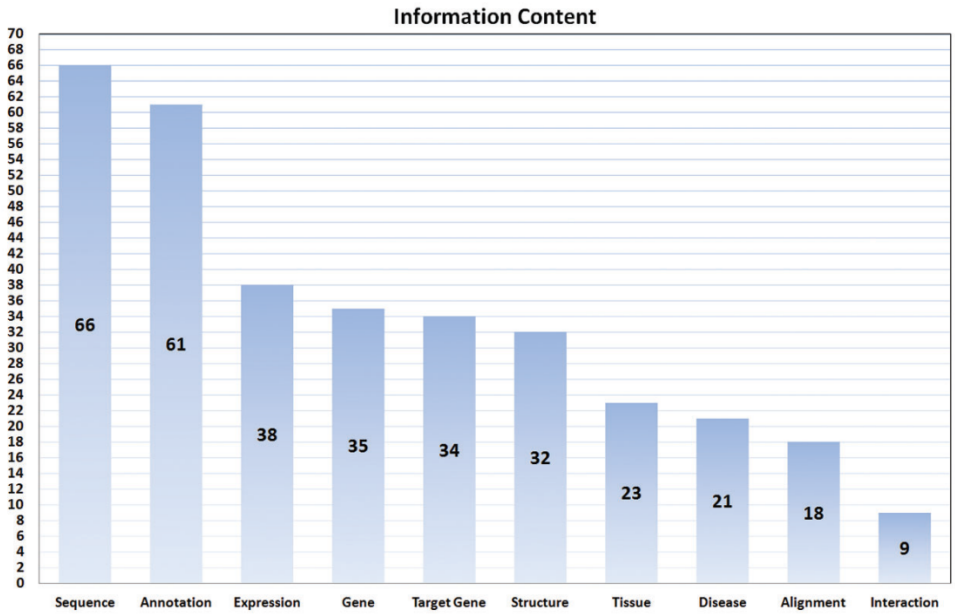


Figura 10 – Representação dos dez termos de conteúdo da informação mais presentes nas bases de dados, de acordo com as informações armazenadas nos bancos.

Explorando as informações que estão armazenadas em bancos de dados, podemos ter uma ideia geral de como os resultados das pesquisas com ncRNAs estão sendo alocadas nas bases de dados públicas. Atualmente, existe um grande número de bancos contendo informações relacionadas com expressão (38 de 102), em comparação com um número escasso de informações relacionadas com a interação destes transcritos não codificadores com outras biomoléculas (DNA, RNAs ou proteínas). Temos apenas nove bancos cobrindo informações relacionadas com interações moleculares (Cao et al, 2010; Nagaswamy et al, 2002; Piekna-Przybylska et al, 2007; Tacutu et al, 2010; Tamura et al, 2004; Turner & Mathews, 2010; Wu et al, 2006b; Xiao et al, 2009; Xin & Olson, 2009). Estes bancos, em sua grande maioria, apresentam apenas anotações relacionadas à interação entre o ncRNA e os seus transcritos alvos, em um relacionamento RNA:RNA; pouca informação existe disponível cobrindo relacionamentos do tipo RNA:proteína ou RNA:DNA.

A descoberta de novos mecanismos funcionais para este tipo molecular é um grande desafio, que é crucial para a compreensão da funcionalidade destes transcritos onipresentes no genoma de organismos eucariotos. Apesar de evidências que apontam

para funcionalidade de ncRNAs já serem do conhecimento atual, tais como: (1) especificidade de expressão por tecidos (Birney et al, 2007a); envolvimento (2) em doenças (Gibb et al, 2011; Reis et al, 2004); (3) em desenvolvimento (Amaral & Mattick, 2008); (4) em *imprinting* genômico (Seidl et al, 2006); e (5) em regulação do *splicing* alternativo (Tripathi et al, 2010); ainda é limitada a evidência dos seus mecanismos de ação e, com exceção dos miRNAs, apenas alguns poucos outros ncRNAs estão anotados nos bancos públicos como funcionalmente caracterizados.

Exemplos de transcritos com mecanismos funcionais de ação conhecidos, que estão presentes em bancos como o lncRNAdb (Amaral et al, 2011) e NONCODE (Bu et al, 2012), incluem: (1) *HOTAIR*, um ncRNA intergênico longo localizado nos clusters gênicos *HOX*, que atua regulando o estado da cromatina do *locus HOXD,* após a ligação com o complexo repressor PRC2 Polycomb e LSD1, levando ao silenciamento gênico do gene *HOXD* (Tsai et al, 2010); (2) *p21*, um transcrito longo intergênico responsivo ao p53, que se liga à proteína hnRNP-K e media uma repressão da expressão gênica global e apoptose na via do p53 (Huarte et al, 2010); e (3) *BC1*, um ncRNA curto neuronal que interage com o fator de iniciação eucariótico 4 (eIF4), resultando em uma subsequente repressão da tradução (Wang et al, 2002). Estes e outros poucos transcritos servem como paradigmas de mecanismos moleculares de atuação de ncRNAs na regulação fina dos processos celulares (Wang & Chang, 2011). O conhecimento biológico adquirido com a caracterização destes ncRNAs poderá orientar o desenvolvimento de novas abordagens computacionais e experimentais, visando a caracterização do imenso repertório de ncRNAs curtos e longos sem mecanismos biológicos de ação bem definidos, os quais estarão alimentando bancos de ncRNAs nos próximos anos.

Recentemente, algumas bases de dados envolvendo abordagens de biologia de sistemas vêm sendo descritas (Bandyopadhyay & Bhattacharyya, 2010; Chiromatzo et al, 2007; Cho et al, 2011; Friard et al, 2010; Hsu et al, 2011; Huang et al, 2009; Schmeier et al, 2011; Wang, 2008). Estes bancos, que estão indexados no NRDR, integram *datasets* de regulação (i.e. lugares de ligação de fatores de transcrição, ou marcadores epigenéticos), expressão, genes alvo e potenciais redes e vias metabólicas reguladas. No entanto, existe ainda uma ausência de dados que integrem estes transcritos com proteínas.

3.6. Mecanismos de busca: extraindo a informação dos bancos de dados

As informações de interesse em bancos de ncRNAs podem ser localizadas a partir de diferentes mecanismos de busca. Agrupamos estes mecanismos em seis categorias: *similaridade*, *palavras-chave*, *tag*, *localização genômica*, *tabular* e *densidade de ncRNAs* (Figura 11). Em alguns bancos, estas opções podem ser utilizadas em conjunto. De acordo com o conhecimento prévio que o pesquisador possui em relação a um ncRNA de interesse (i.e. sequência de nucleotídeo, nome do *locus* gênico no qual esta inserido, ou doença associada), é importante escolher as bases de dados que apresentam os mecanismos de busca mais adequados para que ele possa recuperar informações úteis referentes ao seu transcrito.

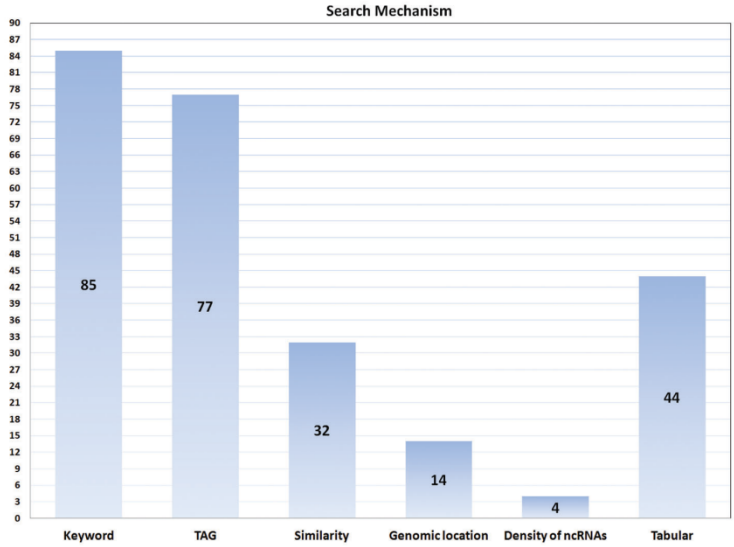


Figura 11 – Distribuição das bases de dados de acordo com os mecanismos de busca disponíveis.

Dentre todos os mecanismos de busca disponíveis, as opções de *palavras-chave* e *tag* são as mais abordagens mais comuns para buscar informações nos bancos de ncRNAs (Figura 10), os quais também são também os mais simples. A opção por *palavras-chave* consiste em caixas de texto que aceitam qualquer texto livre (i.e. código de acesso no banco, nome do gene, identificador do Ensembl, identificador GO). Já a opção *tag*, consiste de um menu simples, contendo uma lista pré-definida de termos (i.e. nome do tecido ou linhagens celulares, organismo, clado).

A busca por *similaridade* é o terceiro mecanismo mais comum, estando disponível em 32 bancos (Figura 11), onde o programa BLAST (Altschul et al, 1990) é utilizado na grande maioria dos bancos. A busca por similaridade de sequências lineares ainda é um desafio na bioinformática de RNAs, visto que famílias de ncRNAs podem divergir em sua sequência primária, conservando apenas sua estrutura secundária (Torarinsson et al, 2006). Apenas o Rfam (Gardner et al, 2011) considera a informação estrutural na busca, e ainda assim ela é associada a um pré-processamento de busca de similaridade de baixa estringência.

Apenas 14 bancos de dados oferecem a busca por coordenadas genômicas ou *localização genômica* (Figura 11), onde ncRNAs são recuperados a partir de intervalos de posições no genoma. Esta busca auxilia a exploração de regiões de interesse específico em um genoma, sendo extremamente útil para analisar diferentes *datasets* de interesse, pré-alinhados no genoma, localizados nas vizinhanças de um ncRNA particular. Um exemplo de seu uso é a investigação de potenciais sinais regulatórios (i.e. ilhas CpGs, locais de ligação de fatores de transcrição, marcadores de modificação epigenética) mapeados no genoma em regiões próximas ao ncRNA em estudo.

A opção por busca a partir da *densidade de ncRNAs* é uma variação da busca por *localização genômica*, com o objetivo de determinar a região genômica baseando-se na densidade de RNAs que estão localizados em *clusters* em uma dada janela do genoma. Esta busca é útil para identificação de transcritos em *clusters*, especialmente miRNAs, que poderiam ser originados a partir de múltiplas unidades transcricionais em um dado *locus* genômico.

Por fim, o último método de busca é o *tabular*, apresentado nos bancos como uma tabela ou uma lista. Neste caso, os *links* disponíveis nas páginas dos bancos levam a tabelas que listam as informações. Toda a busca é realizada manualmente a partir da navegação nestas tabelas.

3.7. Perspectivas futuras do NRDR

A explosão de dados de sequências, decorrente da utilização das novas plataformas de sequenciamento em larga escala, tem estimulado a criação de novas bases de dados públicas referentes a ncRNAs. O NRDR é uma tentativa de facilitar a

descoberta e exploração das informações que estão disponíveis nestes bancos de dados. Com o contínuo aumento do número de projetos de sequenciamento de genomas e transcriptomas, espera-se que os bancos disponíveis irão aumentar consideravelmente o seu conteúdo, em paralelo com o surgimento de outras novas bases de dados especializadas. Desta forma, o NRDR foi projetado de maneira que novos bancos possam ser facilmente adicionados, permitindo que o portal mantenha sua funcionalidade mesmo com o surgimento de novos repositórios integrativos. Como dito anteriormente, a atualização das bases disponíveis no NRDR é realizada automaticamente a cada seis meses, após uma curagem manual da literatura.

3.8. Contribuições do autor para o trabalho

Todas as atividades referentes ao NRDR foram desenvolvidas em conjunto com o aluno de doutorado em Bioinformática Alexandre Rossi Paschoal, orientando dos professores Dr. Alan Mitchel Durham (IME/USP) e Dra. Zilá Simões (FMRP/USP).

Ambos os alunos realizaram todo o levantamento dos dados em conjunto, como também idealizaram o banco de dados e desenvolveram toda a interface web. O aluno Alexandre Paschoal ficou responsável também pelo desenvolvimento do código MySQL do banco e de seu povoamento.

3.9. Publicação

Os resultados apresentados neste capítulo foram publicados no artigo: "*Paschoal, AR; ***Maracaja-Coutinho, V**; Setubal, JC; Simões, ZLP; Verjovski-Almeida, S; Durham, AM. *Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNAs databases.* **RNA Biology**, 9(3):274-82, 2012." (Paschoal et al, 2012), ANEXO 1 desta Tese.

* Ambos autores contribuíram igualmente para este trabalho.

Capítulo 4

IntromeDB: uma base de dados para expressão de RNAs não codificadores intrônicos de organismos eucariotos

4. IntromeDB: uma base dados para expressão de RNAs não codificadores intrônicos de organismos eucariotos

Após a investigação de todas as bases de dados públicas referentes a ncRNAs, e descritas no capítulo anterior, verificamos uma ausência de bancos que armazenassem informações relacionadas a transcritos intrônicos, principal classe de ncRNA estudada pelo nosso grupo. Como mostraremos ao longo de toda esta tese de Doutorado, já existem evidências claras de que estes ncRNAs intrônicos não são transcritos espúrios, meros descartes de um processamento do mRNA no mecanismo de *splicing*. Uma base de dados pública contendo informações acerca da expressão destes ncRNAs é de suma importância para facilitar o acesso a esses dados, permitindo sua exploração em busca de um maior entendimento do seu comportamento em diferentes linhagens celulares, tecidos e organismos ao longo da evolução. Isso nos motivou a desenvolver uma ferramenta web, chamada IntromeDB, na qual armazenamos dados referentes a atividade transcricional intrônica de 21 espécies de eucariotos, com dados públicos obtidos a partir de ESTs disponíveis no dbEST (Boguski et al, 1993), *datasets* públicos de *oligoarrays* intrônicos humanos orientados disponíveis no GEO (Edgar et al, 2002) e *datasets* de bibliotecas de RNA-seq orientadas obtidas a partir de células humanas do projeto ENCODE (Rosenbloom et al, 2012) e disponíveis no *UCSC Genome Browse*r (Kuhn et al, 2012).

4.1. Visão geral do IntromeDB

A ferramenta web IntromeDB pode ser acessada através do endereço: www.intromedb.org. O portal consiste de três bancos de dados principais: (a) *Eukaryotes dbEST*, (b) *Human oligoarrays* e (c) *Human ENCODE* (Figura 12). Atualmente, ele é constituído por *datasets* gerados a partir de duas plataformas experimentais: ESTs sequenciadas a partir de diferentes gerações de plataformas de sequenciamento e *oligoarrays* customizados intrônicos-exônicos (plataforma descrita originalmente em (Nakaya et al, 2007)). Os dados de expressão foram obtidos a partir de bases de dados públicas, e processados utilizando *pipelines* desenvolvidos localmente (descritos abaixo), visando a identificação dos lncRNAs intrônicos expressos em diferentes tecidos, codições e organismos, armazenados em nossos bancos. O IntromeDB é atualizado

continuamente, a medida que dados relacionados a lncRNAs intrônicos vão sendo publicados, a partir de rotinas automatizadas.

O portal fornece uma interface de busca amigável, onde os usuários podem recuperar informações relacionadas a expressão de lncRNAs intrônicos nos três bancos de dados (*Eukaryotes dbEST*, *Human Oligoarrays* e *Human ENCODE*), a partir de critérios definidos pelo próprio usuário (Figura 12). Os resultados das buscas são fornecidos com uma série de informações relacionadas ao lncRNA intrônico, e anotações do seu gene codificador hospedeiro, em determinada espécie, célula ou tecido de interesse. Adicionalmente, seu resultado é integrado com links para acessar suas informações em bases de dados públicas consolidadas (i.e. *UCSC Genome Browser* (Dreszer et al, 2012), *Gene Ontology database* (Ashburner et al, 2000) e diferentes bancos do NCBI (Jenuth, 2000)) (ver Figura 12 e tópico 4.4. para uma descrição mais detalhada).

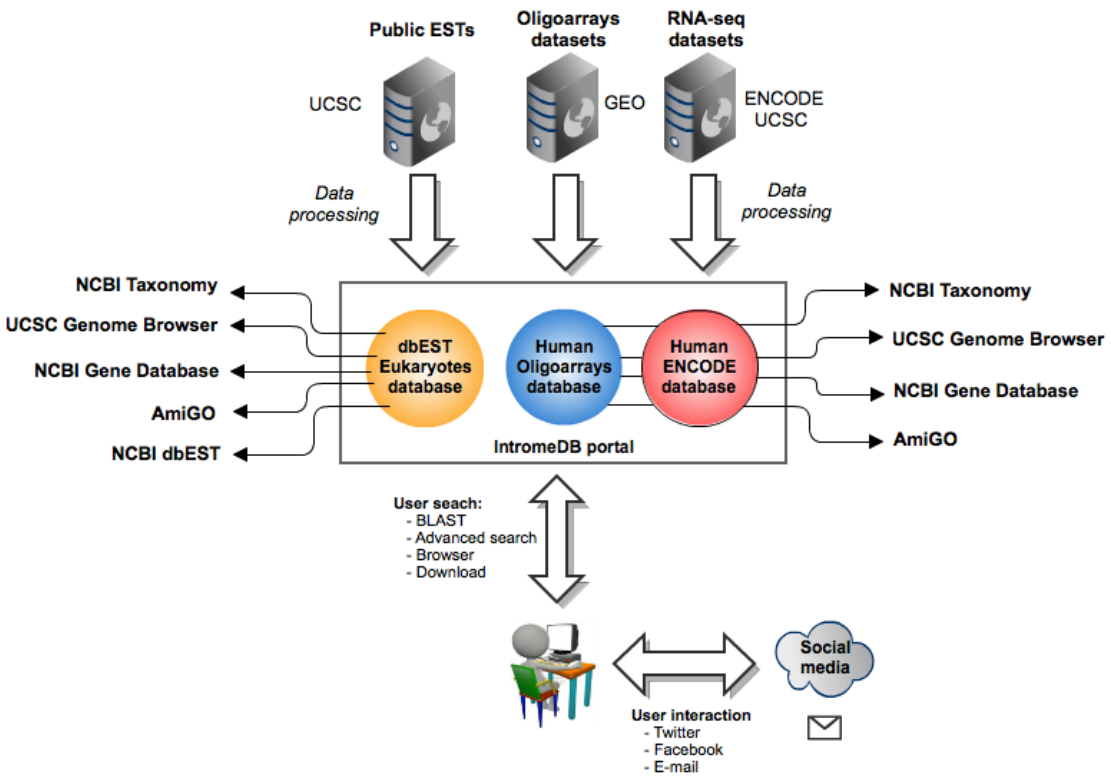


Figura 12 – Visão geral do IntromeDB. Representação gráfica do portal, evidenciando toda sua estrutura, desde o processamento dos dados para gerar as informações armazenadas em nossos bancos, passando pelas formas de extração da informação nele contida, conexões com outros bancos e maneiras de interação com o usuário. As esferas coloridas representam as três bases de dados existentes no portal.

Com o IntromeDB estamos propondo uma nova maneira de interação com a comunidade científica através da Web 2.0. O portal oferece recursos de aplicações web que facilitam o compartilhamento de informações e a colaboração participativa dos usuários. O usuário não só utilizará o e-mail para interagir conosco. Disponibilizamos recursos para os usuários utilizarem as redes sociais (Facebook e Twitter) para colaborar com o banco e interagir com outros usuários, diminuindo as distâncias entre pesquisadores e comunidade em geral (Figura 12 e tópico 4.5 para uma descrição mais completa).

4.2. Construção do banco de dados e implementação da interface web

Os três bancos que constituem o IntromeDB foram modelados e tiveram seus códigos SQL gerados utilizando a ferramenta DBDesigner4 (http://www.fabforce.net/dbdesigner4/), sendo posteriormente implementado em MySQL. A importação dos dados para os bancos foi realizada utilizando a linguagem Perl.

Toda a interface do IntromeDB foi desenvolvida utilizando a plataforma de gerenciamento de conteúdo WordPress (www.wordpress.com), instalada localmente. Utilizamos como *layout* o tema Dzonia Lite 1.4 (http://wordpress.org/extend/themes/dzonia-lite), o qual foi modificado utilizando a linguagem HTML (*Hyper Text Markup Language*). Toda a programação em HTML do portal está de acordo com as normas CSS (*Cascading Style Sheets*). A integração entre a interface e o banco de dados MySQL foi desenvolvida utilizando as linguagens CGI (*Common Gateway Interface*) e Perl.

4.3. Bases de dados disponíveis

***Eukaryotic dbEST intronic lncRNAs*: elucidando a atividade transcricional intrônica sem evidência de *splicing* em diversos organismos eucariotos**

Para a identificação de lncRNAs intrônicos nos mais diversos organismos eucariotos, implementamos um *pipeline* genérico localmente que considera todas as ESTs presentes no banco de dados dbEST, e todos os datasets de mRNAs e genes

RefSeq que continham *tracks* de coordenadas de alinhamento genômico disponível no *UCSC Genome Browser* (Dreszer et al, 2012) (tabelas *all_est*, *all_mrna* e *refSeqAli*). Os arquivos de coordenadas genômicas para eucariotos estão disponíveis para *download* no formato PSL, que foram convertidas para o formato BED. Apenas sequências que mapeiam uma única vez, e que apresentavam uma cobertura maior que 70% e identidade maior que 95%, em relação à sequência do seu respectivo genoma, são utilizadas nas etapas posteriores. Regiões intrônicas com menos que 30 nt são descartadas, através da junção dos dois éxons adjacentes.

Primeiramente, criamos um *dataset* de genes codificadores de proteínas de referência, utilizando dados de mRNAs e RefSeq (excluindo todos os NC_RefSeq, que são constituídos por ncRNAs), compreendendo, desta forma, todas as potenciais isoformas de *splicing* para cada gene. mRNAs que alinham a éxons de dois ou mais RefSeq sem sobreposição de coordenadas presentes na mesma fita genômica são descartados pelo fato de serem possíveis fusões gênicas, ou mesmo erros de sequenciamento, e complicarem o processamento dos dados e a anotação errônea de potenciais transcritos intergênicos como intrônicos. Em paralelo, o *dataset* de ESTs é processado com o intuito de identificar o grupo de ESTs sem evidência nenhuma de *splicing*, que apresentava sobreposições de coordenadas (*contigs*), composto por pelo menos duas ESTs. Então, os *contigs* têm suas coordenadas cruzadas contra o grupo de genes de referência, visando à identificação dos *contigs* exônicos, intergênicos, totalmente intrônicos (TINs) ou parcialmente intrônicos (PINs, i.e. *contigs* que mapearam em um éxon de um gene de referência, cobrindo as regiões flanqueadoras intrônicas em cada um dos lados em pelo menos 30 bases contiguas) (Nakaya et al, 2007).

Apenas os contigs intrônicos sem potencial para codificar uma proteína, de acordo com a ferramenta *Coding Potential Calculator* – CPC (Kong et al, 2007), foram considerados. A Figura 13 evidencia graficamente como foi realizada a “clusterização” das ESTs públicas. Os processamentos para busca da sobreposição de coordenadas foram realizados utilizando a ferramenta BEDtools (Quinlan & Hall, 2010) e *scripts* desenvolvidos localmente em linguagem Perl. Todos os passos do *pipeline* estão descritos na Figura 14.

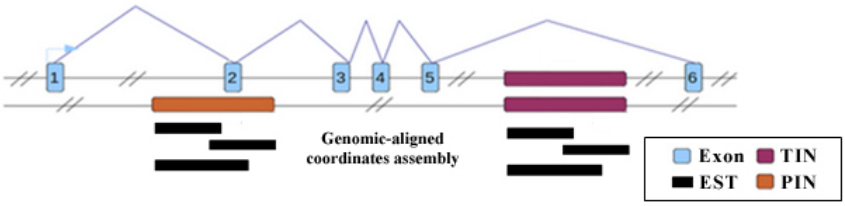


Figura 13 – Representação gráfica da distribuição genômica dos genes hospedeiros codificadores de proteínas e seus respectivos de RNAs intrônicos TINs e PINs. Na figura, representamos como é realizada a "clusterização" de coordenadas das ESTs públicas que dão origem aos *datasets* de RNAs TINs e PINs.

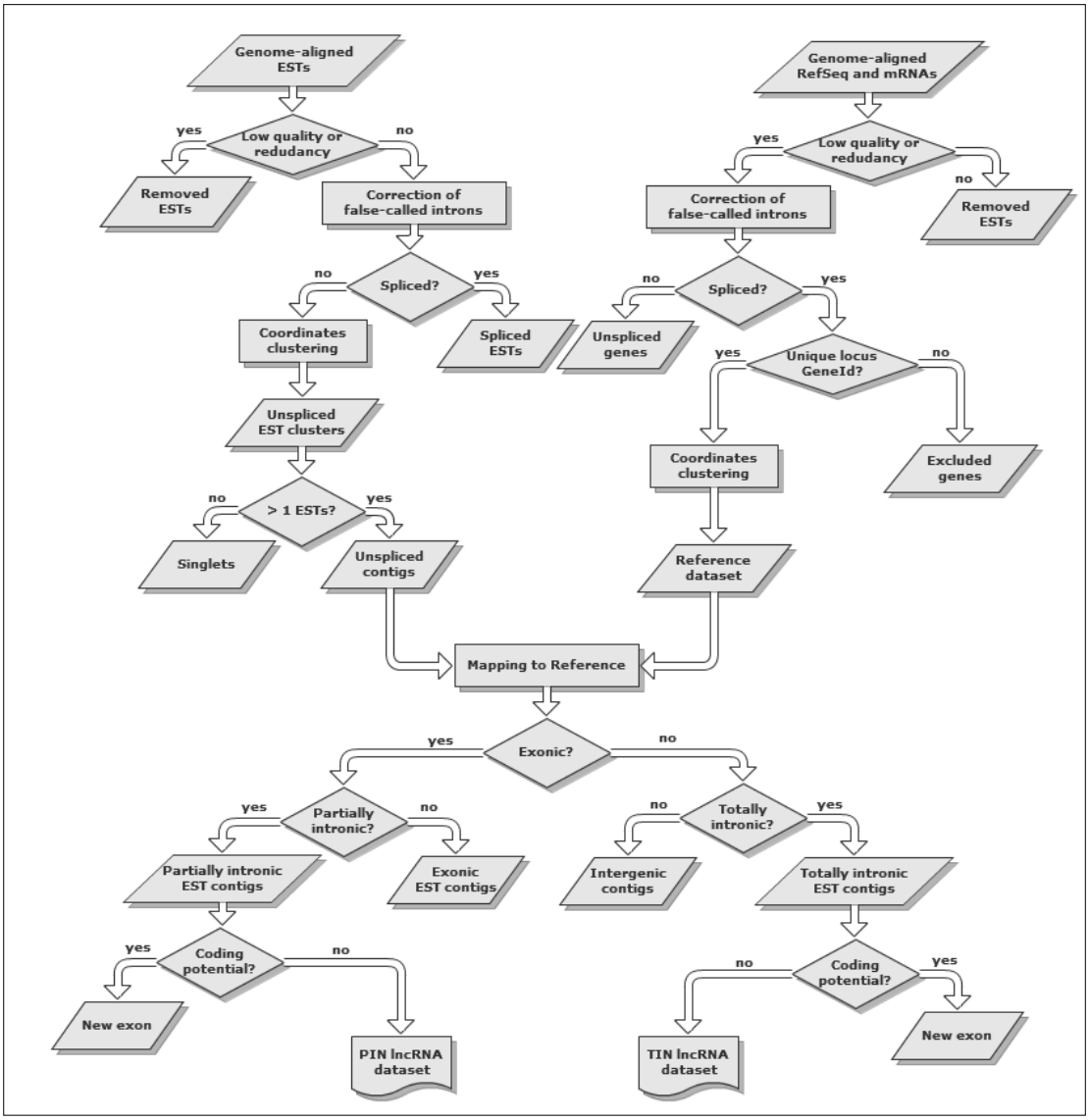


Figura 14 – *Pipeline* desenvolvido para a identificação de potenciais ncRNAs intrônicos em organismos eucariotos com alinhamento de coordenadas disponíveis no *UCSC Genome Browser* (Dreszer et al, 2012). Todas as etapas de "clusterização" de coordenadas genômicas são realizadas utilizando a ferramenta BEDtools (Quinlan & Hall, 2010), em conjunto com *scripts* em linguagem Perl. Utilizamos a ferramenta CPC (Kong et al, 2007) para predição do potencial codificador de proteínas dos transcritos.

Aplicando o nosso *pipeline* para identificação de lncRNAs intrônicos em ESTs públicas de 21 organismos eucariotos com coordenadas de alinhamento genômico disponíveis no *UCSC Genome Browser*, fomos capazes de elucidar parte da atividade transcricional intrônica de espécies pertencentes a diferentes ramos da evolução (Tabela 3).

Tabela 3 – Lista dos 21 organismos armazenados na base de dados *Eukaryotic dbEST* do portal IntromeDB. Na tabela, descrevemos para cada espécie: a sequencia de referência do genoma utilizada, o número de ESTs do dbEST analisados, o número de genes RefSeq não-redudantes e RNAs TIN e PIN identificados pelo nosso *pipelin*e.

Organismo	Montagem do genoma	No. de ESTs disponíveis	RefSeq não-redundantes disponíveis	RNAs TIN identificados	RNAs PIN identificados
Vertebrados					
Homo sapiens (humano), dbEST	hg19	8.260.885	18.302	72.844	10.865
Gallus gallus (galinha)	galGal3	618.543	4.264	1.454	465
Bos taurus (boi)	bosTau6	1.566.611	12.286	7.710	2.168
Cavia porcellus (porquinho-da-índia)	cavPor3	24.425	303	0	2
Canis familiaris (cachorro)	canFam2	399.910	1.154	105	23
Pan troglodytes (chimpanzé)	panTro3	17.639	7.344	2.155	849
Equus caballus (cavalo)	equCab2	37.539	581	8	1
Oryzias latipes (medaka)	oryLat2	703.843	499	15	21
Pongo abelii (orangotango)	ponAbe2	46.027	3.167	60	5
Mus musculus (camundongo)	mm9	4.367.838	17.963	36.539	5.262
Rattus norvegicus (rato)	rn4	1.121.086	13.610	10.007	1.939
Macaca mulatta (macaco-rhesus)	rheMac2	68.320	1.837	12	1
Oryctolagus cuniculus (coelho-europeu)	oryCun2	36.332	938	20	0
Ovis aries (ovelha)	oviAri1	297.638	151	7	3
Taeniopygia guttata (mandarim)	taeGut1	92.700	808	48	11
Xenopus tropicalis (rã)	xenTro3	1.369.022	7.344	2.155	849
Danio rerio (peixe-zebra)	danRer7	1.512.254	12.937	3.372	716

Deuterostomados basais					
Ciona intestinalis	ci2	1.206.102	709	131	46
Strongylocentrus purpuratus (ouriço-do-	strPur2	201.064	298	7	2
Insetos					
Drosophila melanogaster	dm3	1.029.541	11.408	2.617	818
Nematóides					
Caenorhabditis elegans	ce10	377.606	19.322	72	110

Este estudo demonstra que há de fato uma intensa atividade transcricional intrônica nos mais variados organismos eucariotos; pertencentes a diferentes graus de complexidade, como insetos, nematoides, deusterostomados basais e vertebrados (Tabela 3). Isso é mais um indício que a aquisição de sequencias íntronicas e sua transcrição pode realmente ter sido um marco importante para a evolução dos organismos mais complexos.

É importante notar, que o pequeno número de ESTs disponível para algumas espécies (Tabela 3), concomitante com o fato de que a maioria dos projetos de sequenciamento de ESTs são realizados utilizando apenas transcritos poliadenilados, limitando a identificação de um número maior de lncRNAs intrônicos. Outro fator importante, que limita a identificação de lncRNAs intrônicos é o fato de que o sequenciamento de ESTs em larga escala por RNAseq gera *datasets* que atualmente não são mais armazenados no dbEST e sim arquivados no NCBI, no banco *Short Reads Archive* (SRA), ou em bancos próprios mantidos pelos desenvolvedores dos projetos. Apenas fazendo *download* para banco local pode-se fazer análises detalhadas e integradas. Além disso, a maioria dos trabalhos atuais que usam RNAseq e buscam lncRNAs exclui os ESTs que mapeiam em introns, olhando apenas para os lncRNAs intergênicos (Halvardson et al, 2013; Liu et al, 2012; Mercer et al, 2012). Nossa ferramenta inclui um *pipeline* que identifica os lncRNAs que mapeiam em introns, e este será um facilitador importante para aumentar o interesse e o acesso aos dados de lncRNAs intrônicos. Espera-se que a facilidade ao acesso às novas tecnologias de sequenciamento irá permitir uma disponibilidade maior de ESTs para os organismos da

Tabela 3 e outros organismos, levando a uma catalogação mais completa da atividade transcricional intrônica ao longo da evolução dos eucariotos. A disponibilidade de uma base de dados contendo estes transcritos em um portal eletrônico como o IntromeDB poderá auxiliar futuras caracterizações e identificação de lncRNAs intrônicos em outros estudos genômicos e de transcriptoma nos mais diversos organismos.

***Human intronic-exonic oligoarrays*: perfil transcricional intrônico em diversos tecidos humanos**

Em 2007, nosso grupo desenhou uma plataforma "customizada" de *oligoarray* 44k intrônico-exônico da Agilent, com sondas fita-específicas, composta de 13.757 sondas cobrindo genes codificadores de proteínas, e 20.017 sondas cobrindo lncRNAs totalmente (TIN) ou parcialmente (PIN) intrônicos senso e antisenso (Figura 15). Esta plataforma vem sendo utilizada pelo nosso grupo ao longo dos anos, o que tem gerado informações biológicas importantes acerca da biologia destes transcritos. Estes trabalhos forneceram diversos perfis transcricionais intrônicos em diferentes células e tecidos humanos.

No IntromeDB, reanalizamos a expressão das sondas representadas no *oligoarray* para cada um dos lncRNAs intrônicos e genes codificadores de proteínas em 28 linhagens ou tecidos humanos (Tabela 4). Posteriormente, catalogamos os níveis de expressão de cada um dos tecidos em uma base de dados relacional disponível no IntromeDB.

Tabela 4 – Lista dos *datasets* de células e tecidos utilizados e armazenados no banco *Human Intronic-Exonic Oligoarrays* do portal IntromeDB. Notar que mantemos pacientes e doadores como tecidos diferentes, uma vez que o perfil transcricional é particular para cada indivíduo.

Número GEO	Tecido/Células	Referência
GSE5452	Próstata	(Nakaya et al, 2007)
GSE5452	Fígado	(Nakaya et al, 2007)
GSE5452	Rim	(Nakaya et al, 2007)
GSE5453	Células LNCaP	(Nakaya et al, 2007)
GSE18134	Células MSC de cordão umbilical, doador 1	(Secco et al, 2009)
GSE18134	Células MSC de cordão umbilical, doador 2	(Secco et al, 2009)
GSE18134	Células MSC de cordão umbilical, doador 3	(Secco et al, 2009)
GSE18134	Células MSC de cordão umbilical, doador 4	(Secco et al, 2009)
GSE18134	Células MSC de sangue de cordão umbilical, doador 1	(Secco et al, 2009)
GSE18134	Células MSC de sangue de cordão umbilical, doador 2	(Secco et al, 2009)
GSE18134	Células MSC de sangue de cordão umbilical, doador 3	(Secco et al, 2009)
GSE18134	Células MSC de sangue de cordão umbilical, doador 4	(Secco et al, 2009)
GSE18911	Células CD34+ normais, doador 1	(Baratti et al, 2010)
GSE18911	Células CD34+ normais, doador 2	(Baratti et al, 2010)
GSE18911	Células CD34+ normais, doador 3	(Baratti et al, 2010)
GSE18911	Células CD34+ normais, doador 4	(Baratti et al, 2010)
GSE18911	Células CD34+ com MDS, doador 1	(Baratti et al, 2010)
GSE18911	Células CD34+ com MDS, doador 2	(Baratti et al, 2010)
GSE18911	Células CD34+ com MDS, doador 3	(Baratti et al, 2010)
GSE18911	Células CD34+ com MDS, doador 4	(Baratti et al, 2010)
GSE18911	Células de estroma normais, doador 1	(Baratti et al, 2010)
GSE18911	Células de estroma normais, doador 2	(Baratti et al, 2010)
GSE18911	Células de estroma normais, doador 3	(Baratti et al, 2010)
GSE18911	Células de estroma normais, doador 4	(Baratti et al, 2010)
GSE18911	Células de estroma com MDS, doador 1	(Baratti et al, 2010)
GSE18911	Células de estroma com MDS, doador 2	(Baratti et al, 2010)
GSE18911	Células de estroma com MDS, doador 3	(Baratti et al, 2010)
GSE18911	Células de estroma com MDS, doador 4	(Baratti et al, 2010)

MSC = *Mesenchymal stem cells*, MDS = *Myelodysplastic syndrome*.

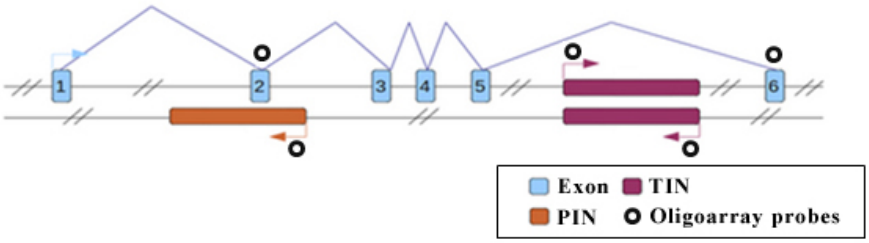


Figura 15 – Representação gráfica da distribuição genômica dos transcritos intrônicos e exônicos, e suas respectivas sondas, no *oligoarray* 44k intrônico-exônico.

Para tanto, baixamos do banco de dados GEO (registros GPL4051 e GPL9193) e reanotamos as duas plataformas de *oligoarrays* intrônicos-exônicos desenhadas por nosso grupo. Uma descrição gráfica da distribuição das sondas está representada na Figura 15. As coordenadas genômicas de cada sonda presente no *array* foram cruzadas contra o conjunto de genes de referência para o genoma humano identificados no tópico anterior, e as sondas foram anotadas de acordo com cada classe de transcrito: exônico, lncRNA TIN senso/antisenso e lncRNA PIN antisenso (Figuras 15 e 16).

Após isso, baixamos todos os dados brutos de expressão de experimentos públicos que utilizaram estas plataformas (Tabela 4), e realizamos uma nova análise de expressão, utilizando critérios definidos abaixo, a fim de identificar aqueles transcritos expressos em cada tecido/célula com dados disponíveis (Figura 16). Notar que mantemos pacientes e doadores distintos do mesmo tecido como um *dataset* diferente, visto que o perfil transcricional é particular para cada doador.

Os dados foram normalizados utilizando o método de LOWESS (*Locally Weighted Linear Regression Algorithm*), implementado utilizando o pacote R, com o intuito de corrigir possíveis vieses sistemáticos da incorporação dos *dyes*. As sondas que apresentaram sinal de intensidade acima da média mais dois desvios-padrão dos controles negativos presentes na plataforma foram utilizadas nas análises posteriores. Uma sonda foi considerada expressa se pelo menos três dos quatro sinais em um tecido estivesse acima do ponto de corte de detecção (todos os experimentos apresentam quatro replicatas). Finalmente, os dados foram normalizados entre tecidos ou células por quantil, utilizando a ferramenta *Spotfire DecisionSite® for Microarray Analysis* (*TIBCO*

Software Inc, Somerville, MA, USA). Todos os passos do *pipeline* estão descritos na Figura 15.

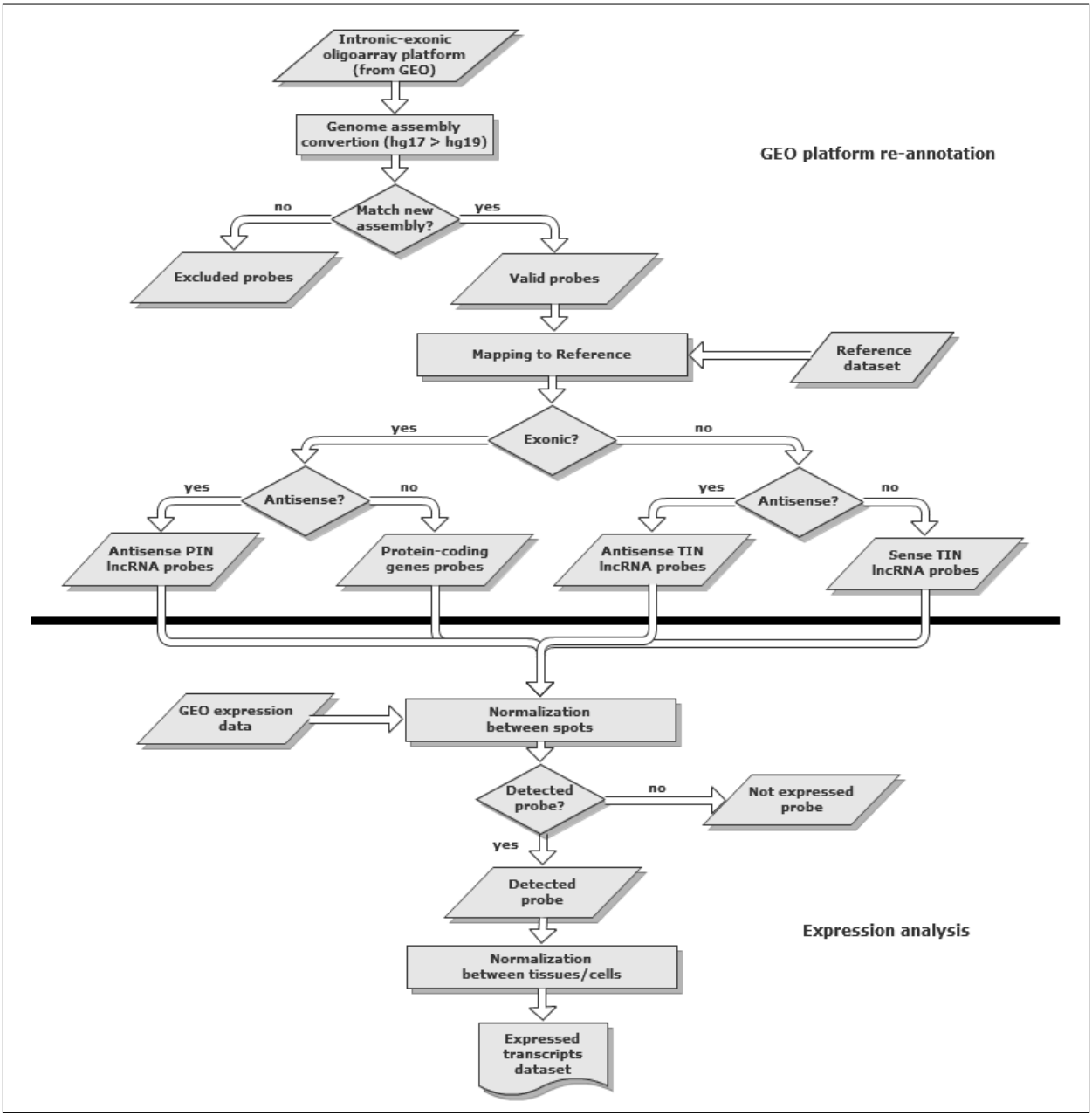


Figura 16 – *Pipeline* utilizado para re-anotação da plataforma de *oligoarray* previamente desenhada pelo nosso grupo, e todos os passos percorridos na re-análise da expressão dos mRNAs codificadores de proteínas e lncRNAs intrônicos presentes no *oligoarray*.

Esta base de dados disponível na internet servirá como um catálogo público para análise do perfil transcricional de lncRNAs intrônicos humanos. Com ela, o pesquisador poderá facilmente explorar como está se comportando o seu lncRNA de interesse nos mais diversos tecidos e condições de interesse, auxiliando na caracterização e formulação de novas hipóteses funcionais referentes ao transcrito.

***Human ENCODE*: perfil transcricional intrônico fita-específico em diversas linhagens de células humanas do projeto ENCODE**

Foi utilizado um total de 98 *datasets* de RNA-seq fita-específico do projeto ENCODE (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=299892339&c=chr19&g=wgEncodeCshlLongRnaSeq), referentes a 39 linhagens celulares humanas (Tabela 5), que foram obtidos a partir de sequenciamento fita-específico de fragmentos de RNAs humanos contendo tamanho maior que 200 nucleotídeos, utilizando a plataforma de sequenciamento Illumina GAIIx. Os dados possuíam diversas replicatas, contendo um perfil transcricional de RNAs poliadenilados ou não, bem como RNAs obtidos a partir da célula inteira ou de diferentes compartimentos celulares como núcleo, nucléolo, nucleoplasma ou citosol.

Para identificar o conjunto de unidades transcricionais intrônicas de referência expressas nas bibliotecas fita-específicas das linhagens celulares humanas do projeto ENCODE, utilizamos todos os *contigs* (tabelas *bedRnaElements*) em formato BED disponíveis para cada linhagem que apresentavam um BKPM (do inglês, *Bases per Kilobases per Million mapped bases*) maior ou igual a um. Estes *contigs* tiveram então suas coordenadas genômicas agrupadas entre si, utilizando uma versão modificada do *pipeline* descrito na Figura 14, mais acima neste mesmo tópico. Esta modificação leva em consideração a fita na qual os *contigs* são originados. Desta forma, também levamos em consideração aqueles transcritos com sobreposição de coordenadas com éxons, que são originados a partir da fita oposta ao gene codificador de proteínas, e que não possuem 30 bases contiguas extendidas para as regiões flanqueadoras intrônicas a estes éxons. Com isso, preferimos defini-los apenas como RNAs antisenso sobrepondo a éxons. Este agrupamento levou em conta todos os *contigs* presentes em todas as bibliotecas como um único *dataset*.

Tabela 5 – Lista das linhagens de células disponíveis na página do ENCODE no *UCSC Genome Browser* que foram analisadas e apresentam dados armazenados no banco *Human ENCODE* do portal IntromeDB.

Linhagem Celular	Descrição
GM12878	Linfoblastóides
H1-hESC	Células-tronco embrionárias
K562	Leucemia
A549	Células epiteliais de carcinoma de pulmão
Células B CD20+	Células B
HeLa-S3	Carcinoma cervical
HepG2	Carcinoma hepatocelular
HUVEC	Células endoteliais da veia de cordão umbilical
IMR90	Fibroblastos de pulmão fetal
MCF-7	Adenocarcinoma de glândulas mamárias
Monócitos CD14+	Monócitos CD14+
SK-N-SH	Neuroblastoma
AG04450	Fibroblastos de pulmão fetal
BJ	Fibroblastos de pele
CD34+ mobilizadas	Células CD34+ mobilizadas
HAoAF	Fibroblastos da aorta adventícia de dois indivíduos
HAoEC	Células endoteliais aórticas (torácicas) de dois indivíduos
HCH	Condrócitos indiferenciados de dois indivíduos
HFDPC	Células das papilas dérmicas do folículo
HMEC	Células do epitélio mamário
HMEpC	Células do epitélio mamário
hMNC-PB	Células mononucleares de dois indivíduos
hMSC-AT	Células-tronco mesenquimais indiferenciadas de dois indivíduos
hMSC-BM	Células-tronco mesenquimais indiferenciadas de dois indivíduos
hMSC-UC	Células-tronco mesenquimais indiferenciadas de dois indivíduos
HOB	Osteoblastos indiferenciados de dois indivíduos
HPC-PL	Pericitos indiferenciados
HPIEpC	Células epiteliais placentárias da membrana amniótica
HSaVEC	Células endoteliais da veia safena de dois indivíduos
HSMM	Mioblastos do músculo esquelético
HVMF	Fibroblastos mesenquimais placentarios
HWP	Pré-adipócitos brancos indiferenciados de dois indivíduos
NHDF	Fibroblastos dérmicos da mama
NHEK	Queratinócitos epidérmicos
NHEM.f M2	Melanócitos epidérmicos do prepúcio de dois indivíduos
NHEM M2	Melanócitos epidérmicos da bochecha de dois indivíduos
NHLF	Fibroblastos de pulmão
SkMC	Células do músculo estriado esquelético
SK-N-SH RA	Neuroblastoma diferenciado com ácido retinóico

Desta forma, fomos capazes de identificar um total de 71.311 unidades transcricionais antisenso humanas que apresentavam sobreposição de coordenadas com um éxon mas eram originados a partir de sua fita oposta (antisenso), um total de 25.972 unidades transcricionais antisenso totalmente inseridas em regiões intrônicas (TINs antisenso), e um total de 67.941 unidades transcricionais senso totalmente inseridas em regiões intrônicas (TINs senso).

Para traçar o perfil transcricional deste conjunto de referência de RNAs intrônicos e antisenso em relação com cada uma das 39 linhagens de células diferentes (Tabela 5), os re-mapeamos contra os *contigs* de cada linhagem, determinando, dessa forma, em quais células cada transcrito de referência estava sendo expresso. Ainda hoje as análises de RNA-seq excluem os RNAs intronicos, de maneira que este conjunto de referência de RNAs intrônicos é o maior *dataset* já mapeado de transcritos intrônicos a partir de dados de RNA-sesq.

4.4. Recuperando informações no IntromeDB e sua integração com bases de dados consolidadas

O IntromeDB disponibiliza uma interface web amigável, na qual o usuário é capaz de recuperar as informações referentes à expressão de lncRNAs em diferentes tecidos, células ou espécies a partir de buscas em um dos dois bancos que fazem parte de sua estrutura, utilizando critérios escolhidos pelo próprio usuário. Primeiramente, é necessário escolher no menu principal um dos bancos a ser explorado: *Eukaryotes dbEST*, *Human Oligoarrays* ou *Human ENCODE* (Figura 17). Daí, o usuário terá quatro opções de busca: *Search*, *BLAST*, *Browser* e *Download*. A Figura 16 demonstra uma captura de tela da página de busca do IntromeDB.

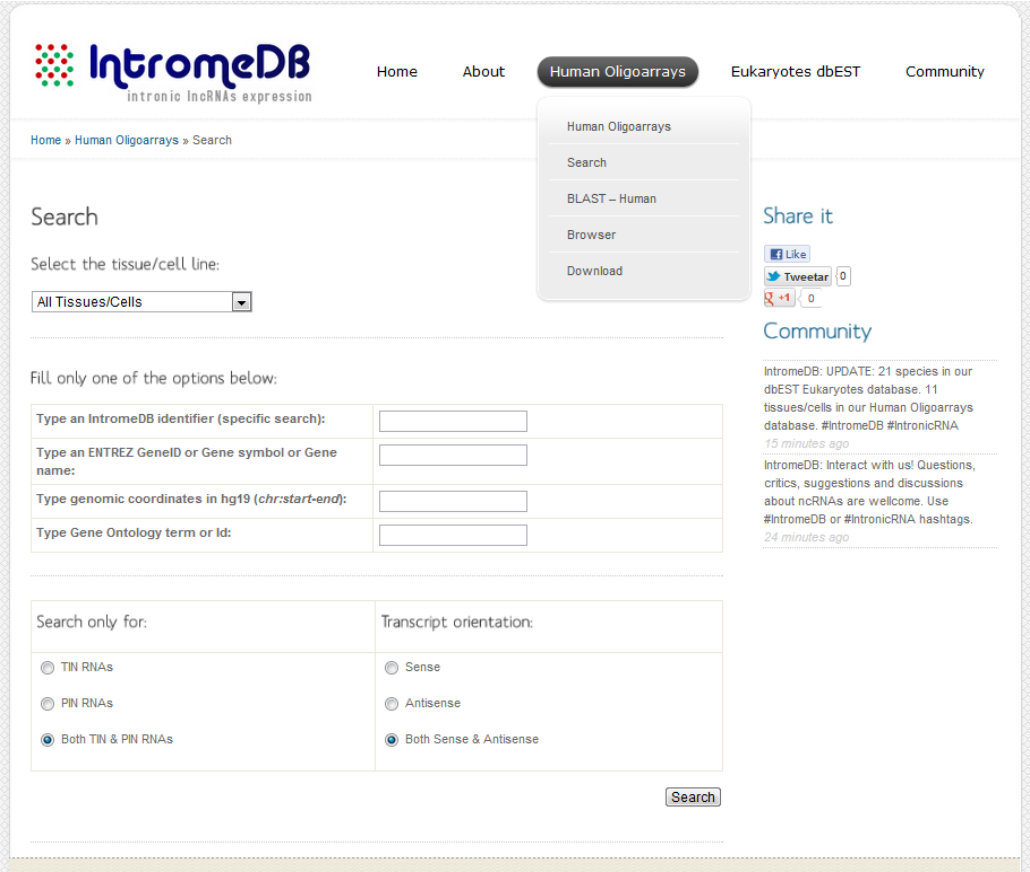


Figura 17 – Captura de tela da interface de busca do IntromeDB. Nesta figura, tomamos como base uma busca no banco *Human Oligoarrays*. No topo superior direito está representado o campo *Share it*, onde os usuários são capazes de difundir o portal nas redes sociais (Facebook, Twitter e Google+). No lado direito inferior, representado pelo campo *Community*, os usuários podem acompanhar as discussões que estão sendo realizadas no Twitter por outros usuários do portal. Estas discussões podem ser feitas apenas pela menção do nosso id (*@IntromeDB*), ou através das *hashtags #IntromeDB* e *#intronicRNA* em suas publicações no Twitter.

Na opção *Search*, o usuário primeiramente seleciona o tecido/célula ou organismo de interesse em um menu simples. Caixas de texto livre estão disponíveis para o usuário escolher uma das seguintes opções de busca: (1) identificador do IntromeDB (i.e. hg19_TIN145, dm3_PIN_367); (2) símbolo do gene, GeneID do ENTREZ ou nome do gene hospedeiro de interesse (i.e. *NCAM2*, *4685*, *neural adhesion molecule 2*); (3) intervalos de coordenadas genômicas (i.e. cromossomo:início-fim); e (4) termo ou identificador de *gene ontology* do *locus* de um gene hospedeiro de interesse (i.e. *translation*, GO:00001666). Finalmente, o usuário pode combinar a busca a partir da seleção da classe de lncRNA intrônico (TIN, PIN ou ambos) e da orientação do transcrito

(senso, antisenso ou ambos). Esta última opção não é disponível para o banco *Eukaryotes dbEST*, visto que os lncRNAs ali armazenados são constituídos de um grupo de dados de ESTs que não possuem informação acerca da fita na qual o transcrito está sendo originado.

A opção *BLAST*, permite buscas por similaridade contra *datasets* armazenados em nos três bancos que constituem o portal. Para o banco *Eukaryotes dbEST*, é possível buscar contra todos os organismos em conjunto, ou contra uma espécie particular de interesse. Após a busca por similaridade, o usuário poderá recuperar o identificador do IntromeDB retornado no resultado da busca e utilizá-lo para buscar informações referentes ao lncRNA no portal.

Finalmente, também é possível explorar os dados armazenados no IntromeDB a partir de uma navegação na aba *Browser*, acessando as informações de acordo com os organismos/tecidos de interesse e, subsequente, de acordo com o *locus* gênico de interesse. Ainda disponibilizamos diferentes *datasets* das informações armazenadas no IntromeDB em formatos BED, FASTA ou *tab-delimited* para *Download*, permitindo que o usuário utilize estes dados em outros aplicativos (i.e. Genome Browser (Dreszer et al, 2012), Galaxy (Giardine et al, 2005)).

Cada registro de lncRNA intrônico, resultado de uma busca no banco, é mostrado em páginas contendo diversas informações relacionadas ao transcrito. Exemplos de informações são: o organismo ao qual o transcrito pertence, a classe de lncRNA intrônico (TIN ou PIN), suas coordenadas genômicas, as ESTs que constituem o *contig* completo daquele transcrito, sua sequência de nucleotídeos. Adicionalmente, fornecemos também anotações relacionadas ao *locus* gênico hospedeiro destes RNAs intrônicos, como uma anotação do gene e seus termos de ontologias gênicas (GO). Para os lncRNAs presentes nos bancos *Human Oligoarrays* e *Human ENCODE*, o registro de um transcrito também fornece informações relacionadas à orientação do transcrito, às coordenadas genômicas das sondas do *oligoarray* (para o banco *Human Oligoarrays*), e dados de expressão referentes tanto aos lncRNAs intrônicos como também aos genes hospedeiros, em todos os tecidos ou linhagens celulares com dados experimentais disponíveis. Estes dados de expressão estão representados em histogramas, para os lncRNAs e para os genes hospedeiros, contendo as intensidades das sondas normalizadas para cada tecido/linhagem armazenados nos bancos.

Todas essas informações estão integradas com *links* diretos para bancos de dados consolidados, como o *NCBI Taxonomy* (Federhen, 2012), *NCBI Genes* (Maglott et al, 2011), *dbEST* (Boguski et al, 1993), *UCSC Genome Browser* (Dreszer et al, 2012) e o *Gene Ontology database* (Ashburner et al, 2000).

Finalmente, com o intuito de estreitar o relacionamento com os usuários do portal, como também uma forma de divulgação e popularização do seu uso, foram criadas contas nas principais redes sociais da atualidade: Facebook (http://www.facebook.com/intromedb) e Twitter (@*IntromeDB*), para que os usuários colaborem conosco e interajam entre eles. Também criamos duas *hashtags* no Twitter (*#intronicRNA* e *#IntromeDB*), que podem ser utilizadas pelos usuários para participação em discussões simplesmente mencionando uma delas, ou o nome do nosso perfil do Twitter (@*IntromeDB*) em seus *posts*.

Na Figura 16, lado direito, apresentamos uma *timeline* intitulada *Community*, onde estas discussões no Twitter podem ser acompanhadas diretamente pelo nosso portal. As discussões podem ser acompanhas também pelo site do Twitter, buscando por qualquer uma das *hashtags*.

Acreditamos que esta estratégia será importante para a popularização do IntromeDB, como também para manter nosso banco de dados constantemente mais confiável e atualizado, visto que as discussões nestas redes sociais serão acompanhadas e moderadas pelos próprios usuários do banco utilizando o conceito de *crowdsourcing*, onde a própria comunidade tem interesse em zelar pela informação ali portada. No entanto, todo o conteúdo é gerenciado e moderado de maneira automática pela nossa equipe, de maneira que aqueles tópicos de maior relevância serão levados em consideração para melhoramentos e atualizações no portal. Esta prática já é utilizada há algum tempo na área de tecnologia da informação, sendo amplamente aplicada tanto em plataformas para difusão do conhecimento, tal qual a Wikipedia, como também para o melhoramento comunicabilidade entre usuários e colaboradores de *softwares* e sistemas operacionais disponibilizados com código aberto como Linux, MySQL, etc.

4.5. Casos de uso do IntromeDB

Com o intuito de apresentar a importância de um banco exclusivo para ncRNAs intrônicos, demostramos aqui dois casos de uso do banco: (i) identificação de potencial ncRNA com expressão constitutiva, para uso como controle para experimentos de expressão de transcritos não codificadores; e (ii) avaliação do perfil transcricional não codificador ao longo do *locus* gênico *GAS6*.

Identificação de potencial ncRNA com expressão constitutiva, para uso como controle para experimentos de expressão de transcritos não codificadores

Como apresentado ao longo deste trabalho, o estudo de transcritos não codificadores nos mais diversos organismos está deixando de ser algo raro e começa a fazer parte dos estudos de expressão gênica em larga-escala (*microarrays* ou RNA-seq). Com isso, novas hipóteses acerca de mecanismos funcionais em relação aos mais diversos ncRNAs tenderão a surgir cada vez mais, sendo necessário o uso de novos controles para comparação do nível de expressão destes ncRNAs. Assim, como forma de demonstrar um uso efetivo da ferramenta, apresentamos a identificação de um potencial transcrito antisenso não codificador de proteínas com expressão constitutiva em distintas linhagens celulares examinadas.

Como forma de contextualização em um ambiente real, vamos hipotetizar que dois ncRNAs com sondas disponíveis em uma plataforma de oligoarrays da *Agilent* ou *Affymetrix* apresentaram-se como ativos em cinco estudos distintos de expressão gênica de um determinado laboratório, utilizando diferentes linhagens celulares e tecidos humanos. Isso chamou a atenção do grupo, que decidiu por uma exploração mais profunda destes ncRNAs utilizando PCR quantitativo para medir sua expressão gênica e verificar o quão expressos são em outras células e tecidos. Desta forma, poderiam ser utilizados como controles em nossos experimentos. No entanto, para evitar gastos desnecessários de tempo e de reagentes, utilizaramos o IntromeDB para verificar se este ncRNA está presente no banco e, posteriormente, avaliar o comportamento da expressão destes ncRNAs em 39 linhagens celulares distintas.

Para tanto, tomaramos como base o banco de dados *Human ENCODE*. Primeiramente, como possuíamos as sequências base dos transcritos que foram

utilizados para desenho de suas sondas nos arrays, realizaramos uma busca por similaridades contra o nosso banco utilizando a versão online do BLAST que está disponível no portal (Figura 18A). Como o resultado do BLAST, verificaramos que as sequências tiveram similaridades com os transcritos com identificadores “contig_AS_Intron_158518” e “contig_AS_Intron_49356”. Assim, por meio de uma busca simples por palavras-chave utilizando estes identificadores, somos capazes de explorar o perfil de expressão destes transcritos nas 39 linhagens celulares presentes no banco (Figura 18B). Com isso, verificaramos que o primeiro transcrito estava expresso em 35 das 39 linhagens celulares, enquanto que o segundo estava apenas em 5 delas (Figura 18C). Com estes resultados, decidiramos por continuar experimentalmente sua busca por um ncRNA controle com o transcrito identificador “contig_AS_Intron_158518”. Vale lembrar, que estes dados estão armazenados no banco apenas com informação de “presença” ou “ausência” de expressão. Assim, para confirmar que de fato são ncRNAs constitutivos, é necessária a confirmação experimental de que estes transcritos tambe estão apresentam o mesmo nível de expressão nos mais variados tecidos, células e condições examinadas.

***Avaliação do perfil transcricional não codificador ao longo do* locus *gênico* GAS6**

Como segundo caso de uso, imaginemos que nosso grupo tenha interesse pelo gene *GAS6*, o qual, de acordo com as sequências de referência disponíveis no *UCSC Genome Browser*, apresenta três isorformas de *splicing*. Afim de identificar a população de ncRNAs dentro do *loci* gênico do *GAS6*, utilizamos uma busca simples por palavras-chave pelo nome do gene no banco *Human dbEST* do IntromeDB (Figura 19A). Como resultado, verificaramos que haviam 21 ncRNAs ao longo de todo o *loci* (Figura 19B). Ao acessar a página referente a cada transcrito, fomos capazes de recuperar as coordenadas genômicas, como também as sequências referentes a cada um dos ncRNAs, e utilizamos essas informações como entrada no *UCSC Genome Browser*, verificando assim a distribuição genômica dos 21 ncRNAs presentes ao longo do loci (Figura 19C). Assim, somos capazes de elaborar novas hipóteses a serem testadas experimentalmente sobre potenciais ncRNAs ao longo do *loci* que poderiam atuar como reguladores de alguma das diferentes isoformas do gene *GAS6.*

Estes dois casos são pequenos exemplos do quão importante uma base de dados como essa pode ser para a comunidade científica. A organização de informações em um local com fácil acesso de maneira gratuita só tende a auxiliar a pesquisa científica, principalmente em uma área com mudanças constantes como a biologia molecular

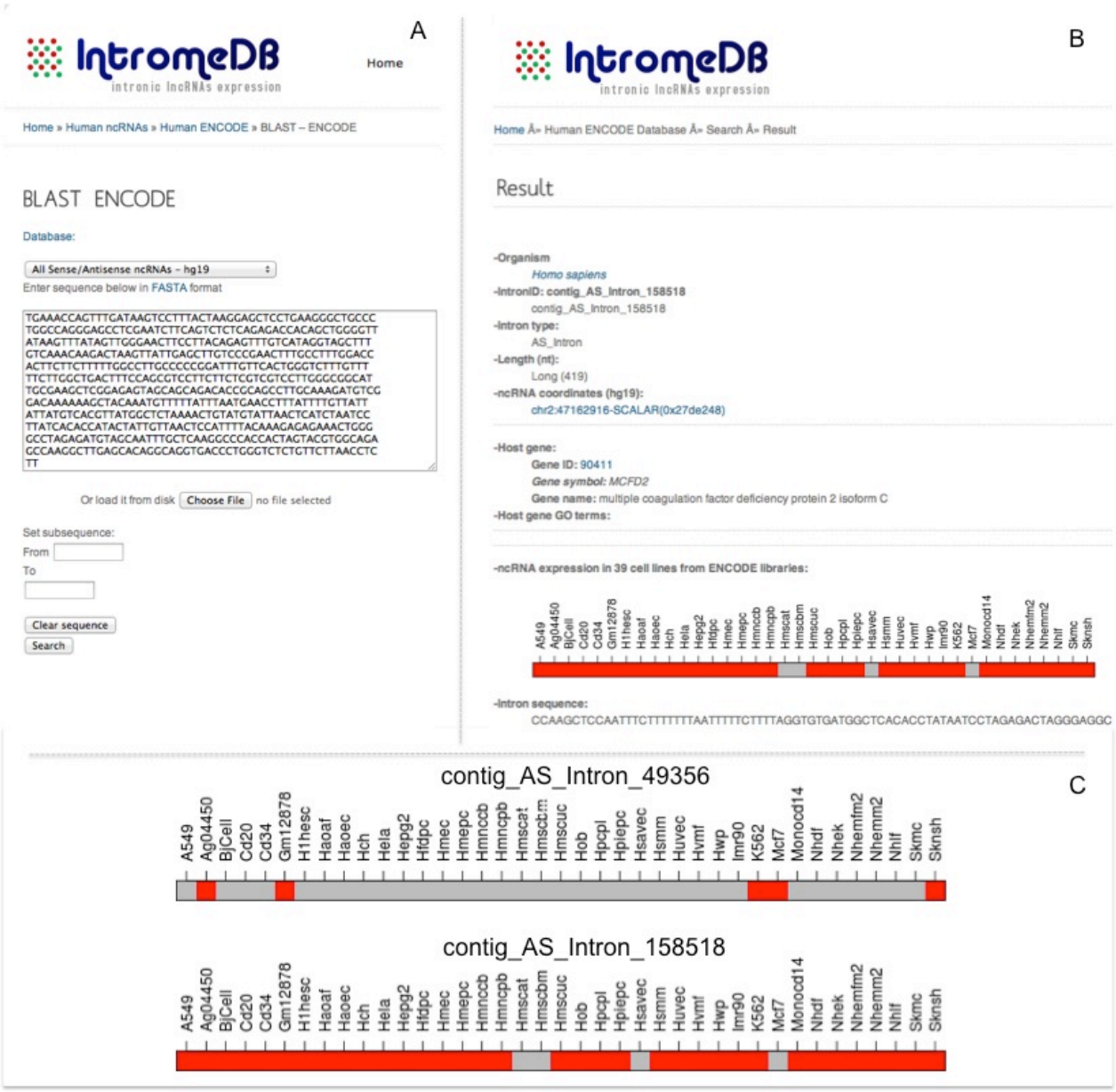


Figura 18 – Exemplo de uso do IntromeDB em que buscamos por um ncRNA controle. Em (A) podemos verificar uma busca por similaridade no banco, utilizando a ferramenta BLAST. Dessa maneira, os usuários poderão verificar se transcritos de interesse possuem informações armazenadas no banco. Existindo o ncRNA procurado, poderão fazer uma simples busca no portal utilizando o identificador deste transcrito. Como como resultado (B), terão um conjunto distinto de informações em relação ao RNA de interesse. Em (C), podemos verificar o resultado final do estudo de caso em questão,

com o “contig_AS_Intron_158518” apresentando-se expresso em 35 das linhagens estudadas, enquanto que o “contig_AS_Intron_49356” apresenta-se expresso em apenas 4. Em vermelho estão destacadas as linhagens celulares em que o ncRNA está expresso. A cor cinza representa ausência de expressão.

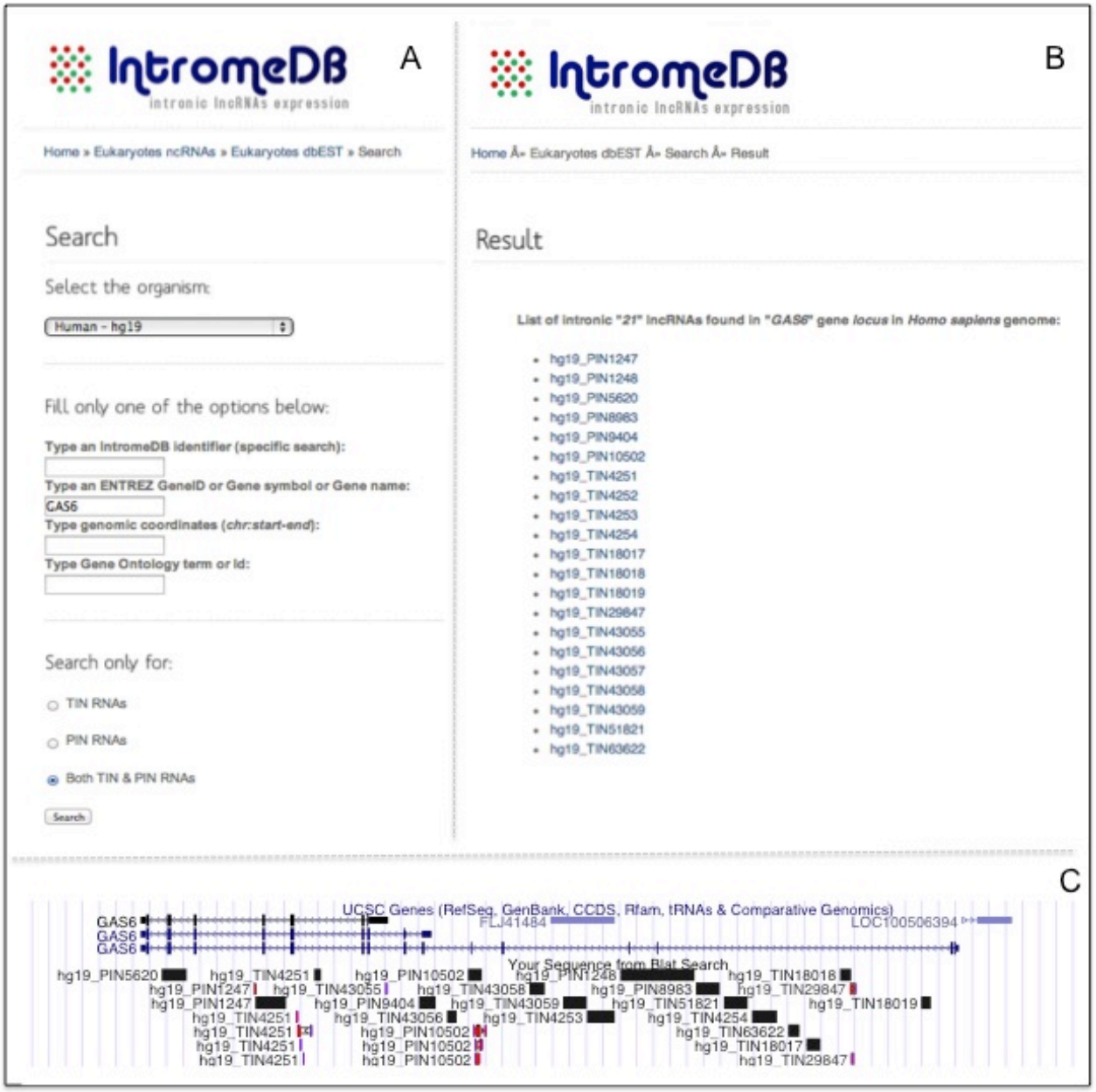


Figura 19 – Exemplo de uso do IntromeDB em que estudamos o perfil transcricional intrônico no *locus* gênico *GAS6*. Em (A) podemos verificar uma busca simples por palavra-chave pelo gene *GAS6*. Como como resultado (B), observa-se a lista de ncRNAs que estão presentes ao longo de todo o *locus*. Por fim, os pesquisadores são capazes de utilizar as informações presentes em cada entrada do banco e utiliza-las em outras ferramentas e bancos. Neste exemplo, a sequência de cada um dos ncRNAs foi mapeada contra o genoma humano utilizando a ferramenta BLAT presente no UCSC Genome Browser (C). Desta forma, pode-se observar visualmente a distribuição

genômica dos ncRNAs em relação ao gene codificador e, assim, elaborar novas hipóteses e experimentos a serem testados.

4.6. Perspectivas futuras do IntromeDB

Visando o melhoramento e aumento das informações armazenadas no banco *Eukaryotes dbEST* do IntromeDB, implementaremos o nosso pipeline para identificação de *datasets* de transcritos intrônicos não codificadores em organismos de interesse biológico, que apresentem um número considerável de ESTs disponíveis no dbEST e/ou no SRA, mas que não possuam arquivos de coordenadas pré-mapeadas em formato PSL disponível no site do UCSC. Isso será realizado a partir do mapeamento de suas ESTs em seus respectivos genomas, de maneira que as coordenadas genômicas sejam posteriormente recuperadas em formato PSL.

Adicionalmente, projetos desenvolvidos pelo nosso laboratório e colaboradores, utilizando a plataforma de *oligoarray* 44k intrônica-exônica estão em constante desenvolvimento. Além disso, uma nova versão contendo 244 mil sondas para transcritos intrônicos, intergênicos e exônicos foi construída e também vem sendo utilizada pelo nosso grupo e colaboradores. Atualizaremos o portal IntromeDB conforme novos resultados que utilizem estas duas plataformas de *oligoarrays* venham a ser publicados. Finalmente, desejamos atualizar o registro de cada um dos lncRNAs armazenados em nossos bancos, conforme informações relacionadas a estudos experimentais ou computacionais venham a ser desenvolvidos, visando a caracterização funcional dos mesmos.

4.6. Publicação

Neste momento estamos finalizando a escrita do artigo científico: "**Maracaja-Coutinho, V**; Setubal, JC; Verjovski-Almeida, S. *IntromeDB: a database for intronic long non-coding RNAs expressed in eukaryotes".* Pretendemos concluir a escrita do manuscrito nos próximos 3 meses.

Capítulo 5

Anotação e níveis de expressão de RNAs intrônicos no fígado humano

5. Anotação e níveis de expressão de RNAs intrônicos no fígado humano

O fígado é o maior dos órgãos internos humanos, composto por uma infinidade de células responsáveis por diversos processos bioquímicos vitais para manter a homeostase do corpo. Ele é o principal órgão na modulação do metabolismo, participando na desintoxicação celular, síntese protéica, dentre outros inúmeros mecanismos celulares. Apesar de sua importância, ainda hoje não se tem uma caracterização dos mecanismos moleculares referentes à atuação de RNAs não codificadores de proteínas no fígado.

Com isso, neste trabalho, descrevemos a avaliação do perfil de expressão de transcritos não-codificadores intrônicos ao longo de todo o genoma humano em hepatócitos humanos, o tipo celular mais proeminente em tecido de fígado, obtido por MPSS (Huang et al, 2007), em conjunto com uma reanálise da porção intrônica dos dados de expressão para tecido de fígado obtido anteriormente em nosso oligoarray 44k intrônico-exônico (Nakaya et al, 2007). Além disso, desenvolvemos e aplicamos um *pipeline* para anotação estrutural, por homologia e por "clusterização" de coordenadas de alinhamentos genômicos destes lncRNAs intrônicos, identificando possíveis ncRNAs intrônicos longos precursores de novas ou já conhecidas classes de RNAs.

5.1. Metodologia utilizada

Mapeamento e expressão das assinaturas de MPSS em lncRNAs intrônicos de hepatócitos humanos

Para identificação dos lncRNAs intrônicos expressos em hepatócitos humanos, reanotamos as assinaturas de MPSS de Huang e colaboradores (Huang et al, 2007), focando apenas nas mensagens intrônicas. Para isso, baixamos os dados brutos das assinaturas de expressão do site: http://202.127.18.238/hepatocytes/, e cruzamos contra os RNAs TINs e PINs identificados no dbEST (Capítulo 4 desta Tese), utilizando a ferramenta megablast (Zhang et al, 2000) com um tamanho de palavra de 17 bases, permitindo apenas alinhamentos perfeitos entre a assinatura e a possível mensagem intrônica e excluindo todas as assinaturas que mapeavam mais de uma vez no genoma ou transcriptoma humano (Figura 20). Como descrito por Huang (Huang et al, 2007), as assinaturas tiveram a frequência de ocorrência normalizada por Transcritos Por Milhão

(TPM). Quando mais de uma assinatura mapeava no mesmo transcrito intrônico, suas frequências foram somadas.

A orientação dos transcritos intrônicos expressos foi definida a partir da observação da fita do *locus* gênico ao qual ele estava inserido. Quando uma assinatura de MPSS mapeava em um fragmento intrônico em sua orientação reversa, a assinatura e o transcrito eram definidos como pertencentes à fita menos (-) do genoma. Quando o mapeamento apresentava a mesma orientação, a assinatura e o transcrito eram definidos como pertencentes à fita mais (+). Desta forma, o lncRNA intrônico era definido como senso, se o transcrito orientado apresentava-se na mesma fita de expressão que o seu gene hospedeiro, por outro lado, quando apresentavam diferentes fitas, o transcrito era definido como uma mensagem antisenso.

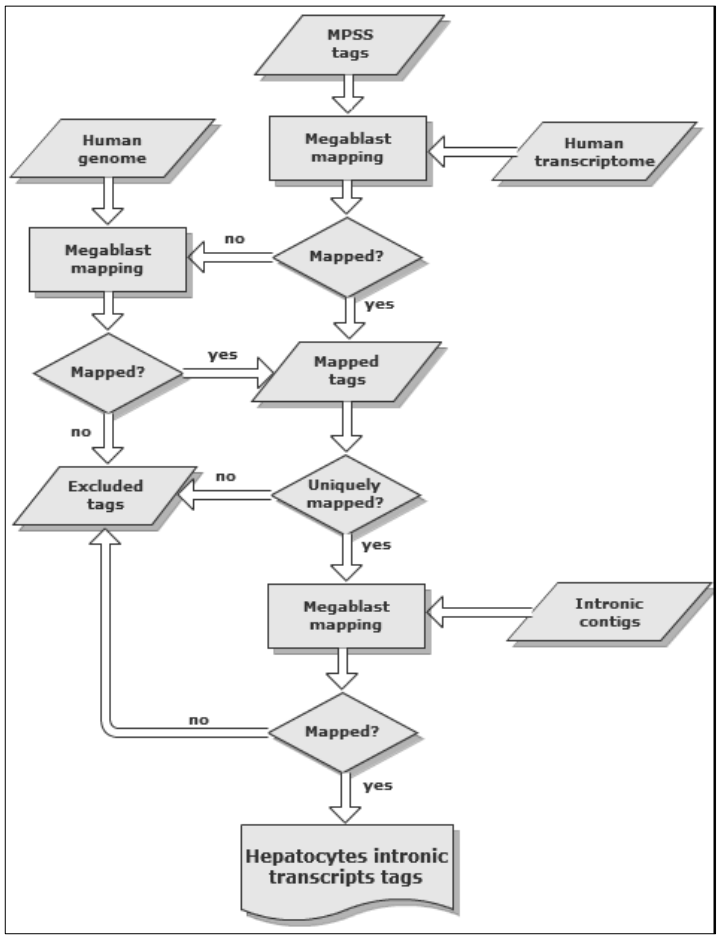


Figura 20 - Fluxograma do *pipeline* desenvolvido para o mapeamento das assinaturas de expressão de MPSS contra os transcritos intrônicos humanos identificados no dbEST.

Identificação dos lncRNAs intrônicos expressos no fígado humano utilizando dados públicos de *microarrays*

Como análise adicional ao perfil de expressão medido por assinaturas de MPSS de lncRNAs intrônicos em hepatócitos, reanalisamos os dados de expressão obtidos para fígado humano em um estudo anterior do nosso grupo (Nakaya et al, 2007) utilizando um *oligoarray* 44k intrônico-exônico customizado contendo 11.574 sondas intrônicas selecionadas a partir de aproximadamente 68 mil messagens TINs e PINs, juntamente com 7.464 sondas correspondentes a éxons dos genes codificadores de proteínas, selecionadas a partir de sondas desenhadas pela Agilent em seu *dataset Whole Human Genome Oligo Microarray*.

Os dados de expressão originais de quatro replicas para tecido de fígado humano foram baixados do banco *Gene Expression Omnibus* – GEO (GSE5452). Para cada uma das quatro replicatas a média da intensidade do controle negativo mais dois desvios-padrão foi utilizada como ponto de corte para determinar se o transcrito intrônico estava expresso ou não. Os valores finais de expressão para cada transcrito intrônico foram determinados pela média da intensidade das sondas nas quatro replicas.

Desenvolvimento de *pipeline* para anotação dos lncRNAs intrônicos expressos no fígado humano

Para anotação dos lncRNAs intrônicos, desenvolvemos um pipeline genérico para categorização de transcritos em classes de ncRNAs já conhecidas a partir de buscas por homologia, estrutura ou sobreposição de coordenadas de alinhamentos genômicos de *datasets* de interesses, quando disponíveis (Figura 21).

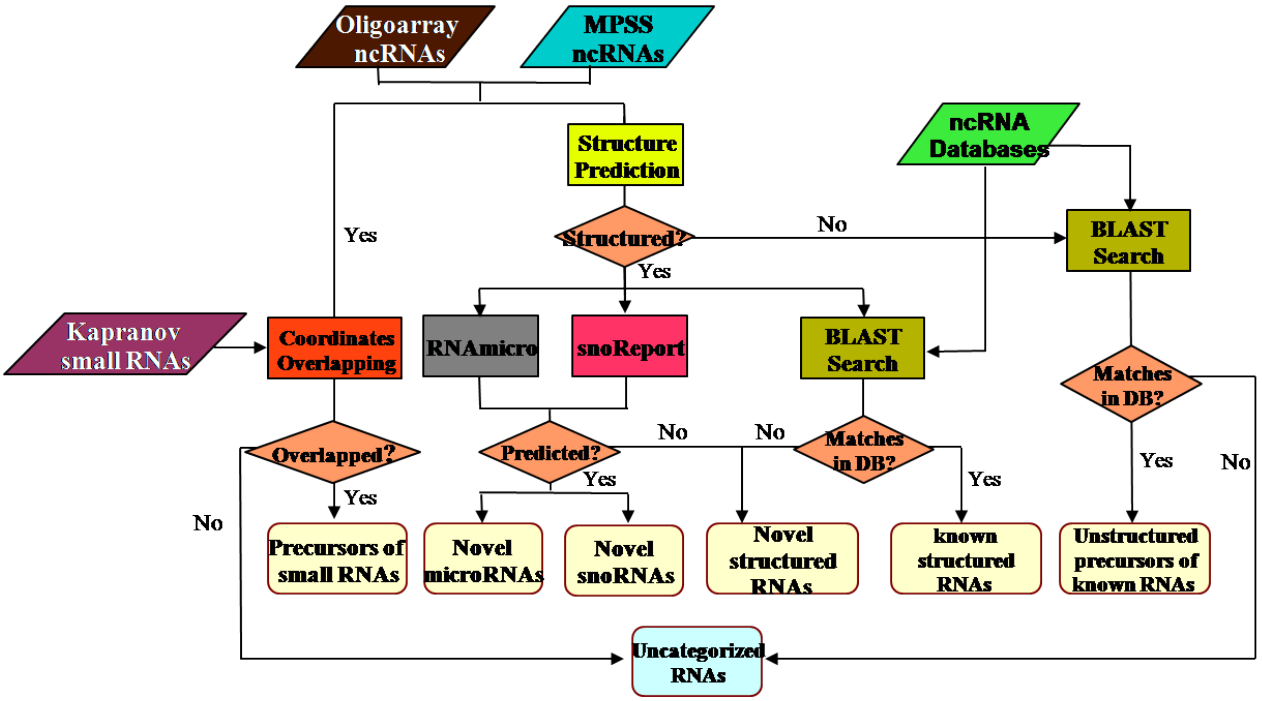


Figura 21 – *Pipeline* desenvolvido para anotação de lncRNAs a partir de diferentes tipos de bases de dados disponíveis. O pipeline baseia-se basicamente na busca por RNAs estruturais a partir da predição de estruturas secundárias, como também na busca por homologia em bancos de RNAs disponíveis, ou por busca de sobreposição de coordenadas de alinhamento genômico de datasets de interesse disponíveis na literatura.

Predição de estruturas secundárias conservadas e termodinâmicamente estáveis

As buscas por estruturas secundárias conservadas e termodinâmicamente estáveis foram realizadas utilizando a ferramenta RNAz (Washietl et al, 2005b). Utilizamos como entrada para o RNAz o alinhamento múltiplo de 14 espécies de vertebrados, que foram gerados com a ferramenta Multiz (Blanchette et al, 2004), e baixados do portal Galaxy (Giardine et al, 2005). Para a composição das espécies no alinhamento selecionamos arbitrariamente dois organismos representativos para cada clado evolutivo. As espécies selecionadas foram: humano, macaco-rhesus, lêmur, camundongo, rato, coelho, vaca, cão, elefante, marsupial, galinha, sapo, peixe e lampréia. Estruturas preditas com uma classificação positiva de probabilidade de $P > 0,5$ foram consideradas. A fita das estruturas preditas foi identificada utilizando a ferramenta RNAstrand (Reiche & Stadler, 2007).

Anotação baseada em homologias

Primeiramente, os lncRNAs intrônicos estruturados identificados em nossa análise foram anotados utilizando uma busca por BLAST (Altschul et al, 1990), através do script rnazBlast.pl do pacote RNAz (Washietl et al, 2005b), contra os bancos mais populares de RNAs não codificadores: Noncode (Bu et al, 2012), ncRNAdb (Szymanski et al, 2007), Rfam (Gardner et al, 2011) e miRBase (Kozomara & Griffiths-Jones, 2011).

Posteriormente, para aqueles RNAs estruturados que não apresentaram homologia contra nenhum ncRNA presente em bancos públicos, utilizamos as ferramentas snoReport (Hertel et al, 2008) e RNAmicro (Hertel & Stadler, 2006), com o intuito de identificarmos, respectivamente, novos snoRNAs e miRNAs ainda não descritos. Para o restante dos lncRNAs intrônicos não estruturados, realizamos as buscas com BLAST contra os mesmos bancos mencionados acima.

lncRNAs intrônicos precursores de ncRNAs curtos

Para identificar os lncRNAs intrônicos precursores de RNAs curtos no fígado humano, cruzamos os transcritos intrônicos determinados como expressos em nossas análises de MPSS e *microarrays*, contra o conjunto de RNAs curtos identificados por Kapranov (Kapranov et al, 2007). Primeiramente, convertemos as coordenadas genômicas dos RNAs curtos da montagem hg17 para a montagem hg19 do genoma humano utilizando a ferramenta liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver/). Após isso, cruzamos suas coordenadas genômicas contra os transcritos intrônicos. Apenas aqueles com pelo menos 70% de sobreposição foram considerados.

Análise de Ontologia Gênica (GO)

As análises de anotação e enriquecimento funcional a partir de ontologias gênicas foram realizadas utilizando a ferramenta DAVID (Dennis et al, 2003). Utilizamos como base para o cálculo do enriquecimento de ontologias os GeneIDs dos respectivos genes hospedeiros para cada um dos transcritos intrônicos; esta opção assume que o lncRNA intrônico age em *cis* sobre o gene codificador de proteína do locus genômico onde ele é transcrito, e que portanto a função afetada pelo lncRNA será a função do

gene codificador. O DAVID não inclui em suas análises eventuais duplicatas contendo o mesmo GeneID em um dado grupo de dados, de maneira que apenas um deles é contado. O teste estatístico de Fisher foi utilizado para avaliar a significância estatística com um ponto de corte de 0,05.

5.2. Resultados obtidos e discussão

Atualização da atividade transcricional humana intrônica sem evidência de *splicing*

Primeiramente, realizamos uma atualização da lista da atividade transcricional intrônica humana identificada pelo nosso grupo em estudos anteriores (Louro et al, 2008; Nakaya et al, 2007). Nesta atualização, baseada no banco de dados dbEST de setembro de 2009, identificamos um total de 82.800 unidades transcricionais intrônicas no genoma humano. Deste total, 72.055 eram mensagens de RNAs não codificadores de proteínas totalmente intrônicos (TINs) e 10.745 eram mensagens não codificadoras parcialmente intrônicas (PINs) (Tabela 6). Nosso estudo identificou que 71% dos genes humanos bem anotados (RefSeq) (setembro de 2009) possuem alguma evidência de transcrição intrônica, com 19% das mensagens totalmente intrônicas sendo originadas a partir do primeiro íntron do gene “hospedeiro”. O aumento observado no número de mensagens intrônicas nesta atualização em comparação com nossos estudos anteriores reflete o número crescente de novas ESTs depositadas nos bancos públicos. Utilizamos o termo contig ao nos referir a cada um dos 82.800 fragmentos de cDNA parciais reconstruídos a partir da sobreposição de coordenadas genômicas de pelo menos duas ESTs montadas em nosso estudo.

As diferenças entre os RNAs PIN de 2006 e os demais *datasets* são devido a diferentes critérios utilizados naquele trabalho, mais precisamente, a exigência de que a mensagem na fita oposta a um éxon estendesse mais de 30 bp dentro da região intrônica em apenas um dos lados da vizinhança do éxon. Posteriormente, exigimos que a mensagem estendesse para a região intrônica em ambos os lados do éxon.

Tabela 6 - Levantamento da transcrição intrônica humana em ESTs disponíveis no dbEST de 2006 a 2009. O *dataset* de 2006 foi extraído de (Nakaya et al, 2007). O *dataset* de 2007 foi extraído de (Louro et al, 2008).

Dataset	2006	2007	2009
	Nakaya et al	Louro et al	Este trabalho
ESTs públicas	5,340,464	7,946,717	8,260,885
RNAs TIN	55,139	67,915	72,055
RNAs PIN	12,592*	10,23	10,745
Total de RNAs intrônicos	67,731	78,147	82,800
Tamanho mediano dos RNAs TIN (bp)	573	*	595
Tamanho mediano dos RNAs PIN (bp)	719	*	973

* Naquela ocasião, não foi calculado o tamanho dos *contigs*.

Expressão de ncRNAs intrônicos em hepatócitos e tecido de fígado humano utilizando bases de dados públicas

Mapeamos os 82.800 contigs de RNAs TIN e PIN, identificados em nossa análise a partir de dados do dbEST utilizando a ferramenta megablast contra um dataset público de 60.635 assinaturas de MPSS (*Massively Parallel Signature Sequencing*) obtidas a partir de hepatócitos microdissecados de fígado humano (Huang et al, 2007) utilizando a ferramenta megablast, com uma palavra com 17nt de tamanho (*match* perfeito da assinatura).

O fluxograma do pipeline desenvolvido para esta análise esta descrito na Figura 20. Neste mapeamento, observamos que 2.723 assinaturas únicas orientadas de MPSS expressas em hepatócitos tinham sobreposição com os transcritos intrônicos identificados no dbEST, representando assim um transcrito intrônico expresso. Utilizamos estas assinaturas para identificar a fita na qual os fragmentos intrônicos estava sendo transcritos. Os resultados podem ser vistos na Tabela 4.

Os níveis de expressão relativos para cada ncRNA intrônico no perfil transcricional pode ser inferido a partir da contagem das assinaturas de MPSS que apresentam sobreposição com qualquer um dos contigs de RNAs intrônicos. Para aqueles contigs que apresentavam duas ou mais assinaturas mapeadas, a soma de todas elas foi considerada. Nossa análise mostra que 19.6% das mensagens intrônicas em hepatócitos de fígado humano tem uma expressão acima de 3 Transcritos Por

Milhão (TPM). Deste total, 22.5% eram pinRNAs e 77.5% eram tinRNAs. No trabalho a partir do qual as assinaturas de MPSS foram obtidas (Huang et al, 2007) os autores demonstraram que tags mesmo possuindo baixa contagem, com valor igual ou mesmo menor que 3 TPM, eram seguramente transcritos expressos, como confirmado a partir de experimentos de RT-PCR.

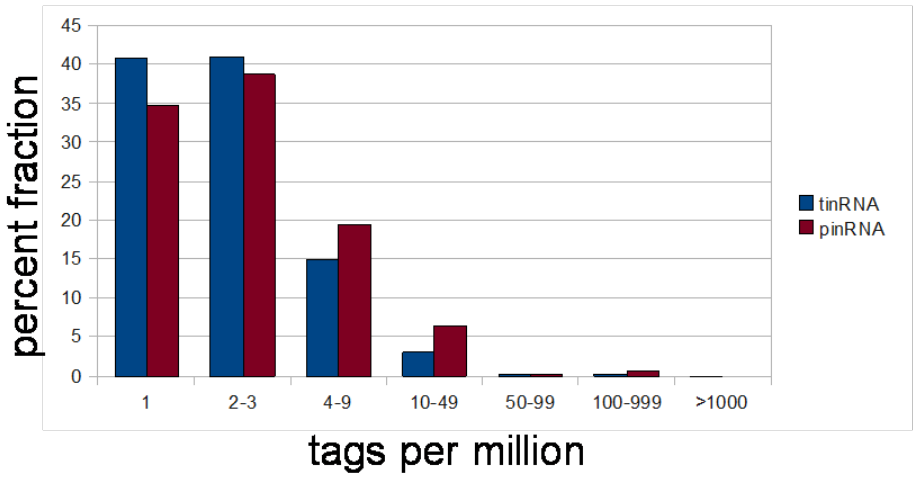


Figura 22 - Distribuição das assinaturas de expressão dos transcritos totalmente (RNAs TIN) e parcialmente (RNAs PIN) intrônicos em hepatócitos humanos. Os *contigs* de tinRNAs e pinRNAs estão representados em azul e vermelho, respectivamente.

Em estudo anterior (Nakaya et al, 2007), nosso grupo gerou dados de expressão fita-específica a partir de uma plataforma "customizada" de *oligoarray* intrônico-exônico 44k para três tecidos humanos (fígado, rim e próstata). Re-analisamos a porção intrônica do *oligoarray* para as amostras de fígado (Nakaya et al, 2007) usando os dados brutos originais disponíveis no banco de dados GEO (*Gene Expression Omnibus*). A partir dos 82.200 *contigs* intrônicos previamente mapeados, identificamos 2.483 tinRNAs e 1.000 pinRNAs expressos no fígado, de acordo com o que foi detectado pelas sondas do *oligoarray* (Tabela 7).

Tabela 7 - Orientação dos transcritos intrônicos expressos em fígado, identificados a partir das abordagens de MPSS e da plataforma de *oligoarray* 44k intrônico-exônica.

	MPSS		Oligoarray	
Tipo de RNA	**Antisenso**	**Senso**	**Antisenso**	**Senso**
RNAs TIN	694 (30,54%)	1.579 (69,46%)	1.120 (45,11%)	1.363 (54,89%)
RNAs PIN	450 (100%)	-	1.000 (100%)	-
Total	1.144 (42,02%)	1.579 (57,98%)	2.120 (60,87%)	1.363 (39,13%)

Os dados de expressão combinados (MPSS e *oligoarray*) evidenciaram que 4.694 tinRNAs e 1.405 pinRNAs estavam expressos no fígado humano (total de 6.099). Mais da metade deles (52,5%) eram transcritos na fita antisenso do gene hospedeiro.

Entre as 2.723 assinaturas únicas de MPSS que tinham sobreposição com transcritos intrônicos do dbEST, apenas 408 (15%) delas estavam representadas por um *contig* com sonda presente no *oligoarray*. Destes, 23,5% (96 assinaturas únicas) foram identificadas como expressas em ambas as técnicas de MPSS e *oligoarray*. Isto é, a assinatura de MPSS detectada nos hepatócitos de fígado tinham sobreposição com o *contig* de ncRNA detectado como expresso no *array*. Pelo fato da técnica de MPSS teoricamente cobrir todo o transcriptoma, sem a necessidade do desenho de sondas pré-conhecidas como no microarranjo, esperávamos que as assinaturas de MPSS fossem cobrir a maior parte dos RNAs intrônicos detectados nos experimentos de *oligoarray*.

Esta baixa sobreposição entre ambos os estudos pode ser explicada pelas diferenças nas amostras utilizadas nas duas abordagens, células de hepatócitos isoladas a partir de *Laser Capture Microdissection* no ensaio de MPSS e tecido de fígado nos experimentos de *oligoarray*. Outros fatores que afetam a chance de sobreposição entre a assinatura de MPSS e a sequência exata de um *contig* intrônico são o tamanho curto das assinaturas (17 bp), o possível viés de termos esta assinatura perto de locais de restrição de endonucleases utilizadas nos protocolos de sequenciamento do MPSS, e o possível uso alternativo de sítio de poliadenilação para a adição de caudas poliadeniladas. Tais fatores podem levar à perda de sobreposição entre uma assinatura de MPSS e um ncRNA intrônico que, apesar de não sobrepostos,

podem eventualmente mapear em uma região próxima nas imediações de uma mesma região intrônica.

Para verificar essa possível ocorrência, computamos o número de vezes em que uma assinatura de MPSS mapeava com uma região intrônica no mesmo íntron de um transcrito com sonda no *oligoarray*, sem exigir que a assinatura estivesse mapeando exatamente sobre a sequência do transcrito intrônico. Esta análise resultou em um total de 2.898 regiões intrônicas com assinaturas de MPSS na vizinhança de transcritos intrônicos com sondas no *array*, das quais 355 (12% dos *loci* intrônicos com sondas no *array*) foram detectados como expressos nas duas técnicas na vizinhança de RNAs TIN senso; e 256 (9%) na vizinhança de RNAs TIN antisenso. Por outro lado, identificamos 77 (3%) regiões intrônicas mapeadas por assinaturas de MPSS na vizinhança da porção intrônica de RNAs PIN. Estas assinaturas podem representar outros 688 novos transcritos não identificados em nossa análise anterior.

Anotação dos transcritos intrônicos expressos no fígado humano

Nesta seção descrevemos nossas análises dos 6.099 transcritos intrônicos expressos no fígado, discutidos na seção anterior. Primeiramente, identificamos transcritos com sequências com estruturas secundárias conservadas filogeneticamente entre organismos vertebrados e termodinamicamente estáveis, considerando que ambas são evidências adicionais de uma possível atividade biológica. Posteriormente, realizamos buscas de similaridade contra bancos de RNAs não codificadores de proteínas selecionados e verificamos aqueles que seriam possíveis precursores de RNAs curtos.

ncRNAs estruturais termodinamicamente estáveis e conservados em fígado humano

Sequências de RNA tendem a enovelar-se em estruturas secundárias e terciárias para desenvolverem suas funções (i.e. interações RNA-RNA, ou interações RNA-proteínas). Estas estruturas podem ser preditas computacionalmente através de abordagens conservacionais, em conjunto com termodinâmica do enovelamento das

moléculas (Washietl & Hofacker, 2004). Estruturas secundárias conservadas em sequências de RNA entre diferentes organismos podem ser interpretadas como um sinal de pressão seletiva natural, apontando para papéis funcionais para estas sequências (Washietl & Hofacker, 2004).

Entre os transcritos intrônicos identificados na técnica de MPSS, predissemos 676 estruturas secundárias com uma taxa FDR de 23,4%. Já para os transcritos intrônicos do *oligoarray*, predissemos 325 estruturas, com um FDR de 21,5%. No total 921 ncRNAs intrônicos possuem predições de estrutura secundária, com 903 na mesma fita dos ncRNAs expressos. Apesar de uma taxa de falso positivo de algo em torno de 20%, este valor é considerado bom tendo em vista os atuais métodos disponíveis para predição de RNAs estruturais (Gruber et al, 2007). As estruturas e sua anotação podem ser visualizadas nos links: http://www.bioinf.uni-leipzig.de/data/TIN_PIN/mpss/ e http://www.bioinf.uni-leipzig.de/data/TIN_PIN/array/.

ncRNAs intrônicos como precursores de classes novas ou previamente já conhecidas de ncRNAs

Dentre os RNAs identificados como expressos no *oligoarray* e MPSS em fígado humano, apenas cerca de 20% e 30%, respectivamente, apresentaram similaridade com RNAs de classes conhecidas (piwiRNAs, microRNAs, snRNAs e scALU RNAs) (23). Imaginamos que grande parte destes RNAs seria de membros ainda não descritos de classes conhecidas ou mesmo classes completamente novas de RNAs.

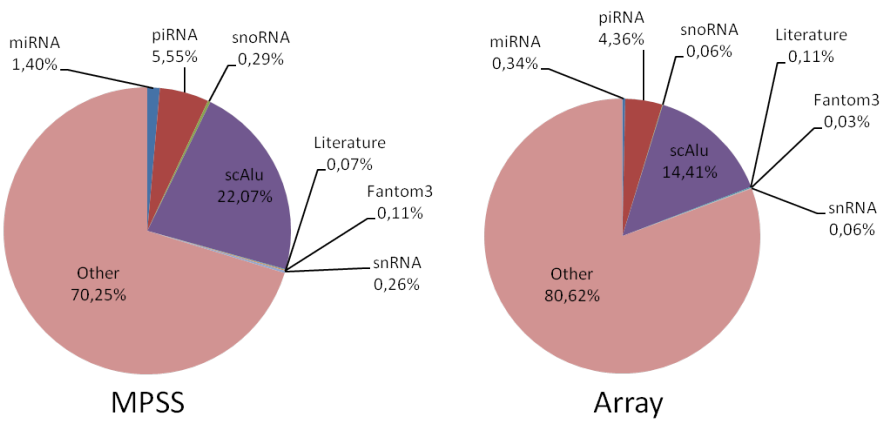
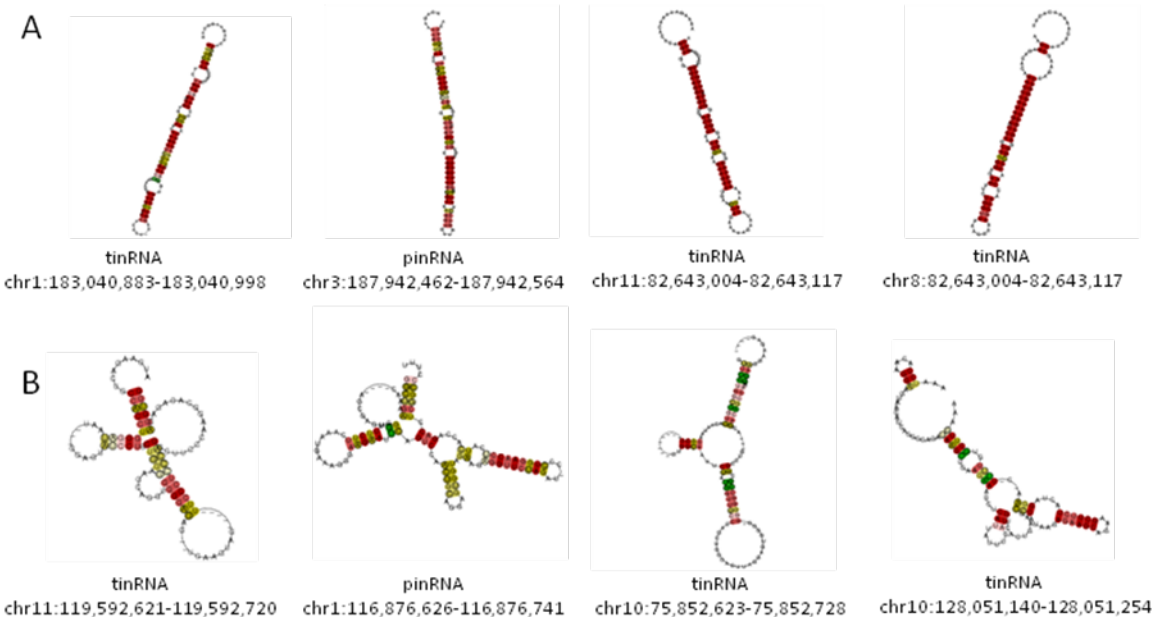


Figura 23 – Distribuição da anotação dos RNAs intrônicos identificados nas duas abordagens de expressão gênica.



24 – Quatro exemplos de possíveis novos microRNAs e snoRNAs processados a partir dos lncRNAs intrônicos preditos em nosso estudo de acordo com as ferrametas RNAmicro e snoReport, respectivamente. (A) microRNAs e (B) snoRNAs, com suas respectivas coordenadas genômicas (hg19).

O fato interessante desta anotação é o fato de que entre os RNAs intrônicos anotados, uma percentagem considerável (22% no MPSS e 14% no *array*) continha um trecho similar a scALU RNAs (small cytoplasmatic Alu RNAs), contendo geralmente em torno de 120 nt de tamanho. Esta classe de RNA é um tipo de sequência repetitiva Alu derivada dos 7 SL RNA (Kriegs et al, 2007). Além disso, todos os 1.103 (601 do MPSS e 502 do *array*) RNAs intrônicos que apresentaram scAlu RNAs inseridos são derivados de 11 scAlu distintos. Fizemos uma busca por RNAs intrônicos com scAlu RNAs inseridos em sua sequência no *dataset* original de transcritos intrônicos identificados no banco de dados dbEST. Verificamos que um total de 16% dos RNAs intrônicos do dbEST apresentou os mesmos scAlu inseridos em sua sequência. Já se sabe na literatura que a inserção de sequências transponíveis no genoma pode gerar novos ncRNAs funcionais. Um exemplo disso é o ncRNA neuronal BC1, originado a partir da retrotransposição de $tRNA^{Ala}$ em roedores (DeChiara & Brosius, 1987; Muslimov et al, 2002). Como ocorrido com o ncRNA BC1, uma hipótese é que estes RNAs foram inseridos no genoma ao longo da evolução e com o tempo sua maquinaria de

transcrição acabou por transcrever um RNA maior (acima de 2000k) que o scAlu original, gerando uma nova classe de RNA. Isto implicaria também na hipótese de que parte dos ncRNAs teria evoluído a partir da retrotransposição de genes já conhecidos. No entanto, experimentos ainda precisam ser realizados a fim de validar esta hipótese.

Especulamos que uma fração considerável, entre 70% e 80% dos RNAs sem similaridade com os bancos, representa outros membros de novas classes de RNAs. Tais classes poderiam estar participando da regulação de uma série de processos celulares fundamentais da célula. Esta hipótese pode ser apoiada pela grande evidência recentemente acumulada na literatura de funções biológicas importantes que ncRNAs estariam desempenhando na célula (Amaral et al, 2011; Cabianca et al, 2012; Feng et al, 2006; Louro et al, 2009; Mohammad et al, 2012; Nakaya et al, 2007; Qu & Adelson, 2012; Secco et al, 2009).

ncRNAs intrônicos longos processados em RNAs curtos

Outra hipótese para explicar a ausência de muitos destes transcritos intrônicos nos bancos de dados pode ser pelo fato de que vários deles podem ser processados em RNAs curtos ainda não descritos, para daí desenvolverem seu papel funcional na célula. Testamos esta hipótese pelo mapeamento de nossas sequências contra aquelas descritas por (Kapranov et al, 2007). Estes pesquisadores realizaram um mapeamento ao longo de todo o genoma humano de transcritos não codificadores de proteínas que apresentavam evidência de processamento em RNAs curtos em linhagens celulares de carcinoma hepatocelular de fígado (HepG2). Cruzamos estes RNAs curtos identificados por Kapranov et al. contra nosso grupo de transcritos intrônicos expressos em fígado. Na análise utilizando os dados do MPSS, identificamos 493 (18%) RNAs intrônicos possíveis precursores de RNAs curtos, enquanto que nos transcritos obtidos no *oligoarray*, este número foi de 520 (15%). Por outro lado, a sobreposição destes RNAs curtos com RNAs estruturais é de apenas 5% para os dados de MPSS e de 3% para os dados de *oligoarray*, sugerindo que estes transcritos devem provavelmente pertencer a classes de RNAs distintas daquelas dos RNAs estruturados aqui identificados.

Concluímos que 17% destes transcritos são provavelmente processados em diferentes classes de RNAs curtos para desenvolver suas funções regulatórias na célula. Algumas possíveis funções desenvolvidas por estes RNAs poderiam ser a regulação da

arquitetura da cromatina (Yu et al, 2008); e através de interação *duplex* com promotores, ativando ou reprimindo a expressão de genes codificadores de proteínas (Janowski et al, 2007; Schwartz et al, 2008).

Análise de Ontologia Gênica (GO)

Os *loci* gênicos com transcrição intrônica, levando-se em conta todos os 82.800 *contigs*, são enriquecidos em genes pertencentes a subcategorias GO relacionadas com ligação de proteína (*protein binding*), atividade catalítica (*catalytic activity*) e atividade de transporte (*transporter activity*). Os RNAs intrônicos expressos em hepatócitos com três ou mais TPM de acordo com a análise de MPSS são transcritos a partir de regiões intrônicas de genes codificadores de proteínas enriquecidos nestas mesmas categorias GO. Os transcritos intrônicos com estruturas preditas são transcritos a partir de regiões de genes codificadores de proteínas significativamente enriquecidos nas categorias de ligação de proteína (*protein binding*) e regulador da atividade enzimática (*enzyme regulator activity*). Estes resultados podem indicar que muitos destes transcritos apresentam um papel funcional na regulação celular.

Mais especificamente, uma hipótese para a função de algumas destas longas mensagens intrônicas é que eles poderiam esta atuando como co-reguladores da modulação de importantes processos biológicos, como já descrito para o transcrito longo intergênico HOTAIR, o qual é transcrito a partir do cluster *HOXC* e liga-se a proteínas do grupo Polycomb (PRC2), levando a um silenciamento do cluster *HOXD* que está em um locus genômico vizinho, remodelando a cromatina para um estado repressivo (Rinn et al, 2007). Outras hipóteses funcionais são a ligação destas mensagens intrônicas (ou produtos de algum processamento delas) diretamente a outras sequências de DNA ou RNA, realizando uma regulação transcricional (Beltran et al, 2008) ou pós-transcricional (Martianov et al, 2007).

5.3. Conclusões

Neste trabalho, apresentamos uma atualização da atividade transcricional intrônica humana em fígado, baseando-se em ESTs e mRNAs disponíveis em bancos

públicos. Focamos apenas em mensagens totalmente (RNAs TIN) ou parcialmente (RNAs PIN) intrônicas sem nenhuma evidência de *splicing*. Realizamos uma análise de expressão destes transcritos intrônicos no fígado humano, a partir de uma reanálise de dados públicos referentes a duas abordagens de expressão gênica: MPSS e *microarrays*.

Além disso, desenvolvemos um *pipeline* para anotação de transcritos não codificadores de proteínas baseado em evidências estruturais, sobreposição de coordenadas de alinhamentos genômicos e homologia de sequências, o qual aplicamos no conjunto de ncRNAs intrônicos identificados como expressos no fígado humano. Diferente dos demais *pipelines* apresentados nesta Tese, este poderá ser utilizado para a anotação de transcritos não codificadores em distintas classes de ncRNAs (microRNAs, piwiRNAs, snoRNAs, etc). Esta análise demonstrou que grande parte destes transcritos pertence a classes ainda não descritas de RNAs.

Todos os resultados aqui apresentados servem para demonstrar o quanto ainda se tem para desvendar em relação à participação dos ncRNAs intrônicos nos processos celulares. O conjunto de transcritos aqui identificados no fígado humano serve como um catálogo para futuras caracterizações funcionais destes RNAs em tecidos hepáticos. Finalmente, o *pipeline* desenvolvido pode ser aplicado em qualquer projeto de transcriptoma humano, auxiliando na anotação dos transcritos identificados.

5.4. Contribuições do autor para o trabalho

Este trabalho foi realizado em colaboração com o doutorando da Universität Leipzig (Alemanha), Dominic Rose, orientado pelo professor Dr. Peter Stadler.

As etapas de identificação da atividade transcricional intrônica humana a partir das ESTs públicas, as análises de expressão no fígado em MPSS e *microarrays*, e as análises de ontologias gênicas foram desenvolvidos pelo autor desta tese. Por outro lado, as análises de anotação dos ncRNAs foram desenvolvidas em conjunto por Rose e pelo autor.

Capítulo 6

Caracterização da região promotora e conservação de lncRNAs intrônicos e intergênicos expressos em tecidos pancreáticos não tumorais e neoplásicos humanos

6. Caracterização da região promotora e conservação de lncRNAs intrônicos e intergênicos expressos em tecidos pancreáticos não tumorais e neoplásicos humanos

O adenocarcinoma ductal pancreático (PDAC, do inglês *Pancreatic Ductal Adenocarcinoma*) é a mais comum neoplasia de pâncreas humano e é responsável por mais de 85% dos casos de tumores pancreáticos (Hezel et al, 2006). O PDAC é uma doença devastadora com prognóstico bastante ruim, para a qual o único tratamento é a cirurgia de ressecção (Yokoyama et al, 2009). No entanto, apenas 15 a 20% dos pacientes retiram o tumor, com apenas 20% destes apresentando uma sobrevida de cinco anos, resultando numa taxa média de sobrevivência de cinco anos de 3 a 5% dos casos totais (Hezel et al, 2006). A agressividade do PDAC está associada principalmente à falta de ferramentas de diagnóstico precoce e à resposta limitada aos tratamentos atualmente disponíveis (Yokoyama et al, 2009).

Neste trabalho, utilizamos um microarranjo de cDNA "customizado" contendo aproximadamente 4.000 elementos, previamente descrito pelo nosso grupo (Brito et al, 2008; Reis et al, 2004), para investigar os padrões de expressão de uma coleção de transcritos codificadores de proteínas e potenciais RNAs não codificadores em amostras clínicas de tumor primário e metastático, pancreatite crônica e tecido pancreático histologicamente normal. Esta plataforma contém sondas que interrogam mRNAs RefSeq de genes associados com câncer, de acordo com a literatura, como também transcritos que mapeiam regiões intrônicas e intergênicas do genoma, além de alguns ncRNAs já conhecidos. Neste projeto, buscamos identificar e caracterizar diferentes transcritos humanos não codificadores de proteínas envolvidos com o câncer de pâncreas. Identificamos indícios de que estes transcritos apresentam uma certa conservação com outros organismos, possivelmente refletindo o fato que sofreram uma pressão evolutiva, como também apresentam evidências que apoiam a teoria de que são unidades transcricionais independentes ativas no genoma. Desta forma, o trabalho fornece um catálogo base de lncRNAs intrônicos e intergênicos para futuros ensaios funcionais referentes à atuação e biologia deste tipo molecular no câncer de pâncreas.

4.1. Metodologia utilizada

Microarray, extração do RNA e análises de expressão

Primeiramente, reanotamos todas as sondas presentes no *microarray* 4k para a montagem atual do genoma (hg19), e genes de referência RefSeq e UCSC de outubro de 2010. Cada lamina contém 722 *spots* referentes a ESTs que mapeiam em regiões intrônicas de genes codificadores de proteínas, 74 referentes a RefSeqs de lncRNAs conhecidos, 188 a regiões intergênicas do genoma e 2.371 correspondentes a éxons de genes codificadores de proteínas.

Alvos fluorescentes de cRNAs gerados a partir de 38 amostras de tecido pancreático (15 adenocarcinomas primários, 9 tecidos adjacentes histologicamente normais, 6 amostras metastáticas e 9 amostras de pancreatite crônica) foram hibridizados individualmente com o microarranjo em replicatas. Todas as amostras de tecido pancreático utilizadas e as etapas de extração de RNAs, hibridização e análise de expressão dos *microarrays* foram processadas pela aluna de doutorado Ana Tahira, orientanda do professor Dr. Eduardo Reis, e estão descritas em artigo publicado na revista Molecular Cancer (Tahira et al, 2011).

Análises das possíveis regiões regulatórias de lncRNAs intrônicos e intergênicos expressos no pâncreas

Primeiramente, o potencial codificador de proteína dos lncRNAs intrônicos e intergênicos expressos foi determinado, a partir da ferramenta *Coding Potential Calculator* (CPC) (Kong et al, 2007). Utilizamos o pacote de ferramentas BEDTools (Quinlan & Hall, 2010), em conjunto com *scripts* em Perl desenhados localmente, para cruzarmos as coordenadas de mapeamento genômico de diversos *datasets* utilizados no trabalho, e disponíveis no *UCSC Genome Browser* (Dreszer et al, 2012) e outros bancos públicos: i) microRNAs (Kozomara & Griffiths-Jones, 2011) e snoRNAs (Lestrade & Weber, 2006); ii) assinaturas de CAGE derivadas a partir de bibliotecas de RNAs poliadenilados de seis linhagens celulares do projeto RINKEN (Kodzius et al, 2006); iii) dados de ChIP-seq de locais de ligação no DNA de H3K4me3 (Mikkelsen et al, 2007) e iv) ilhas CpGs preditas no genoma humano (Gardiner-Garden & Frommer, 1987). A Figura 25 descreve graficamente como foram feitos os mapeamentos.

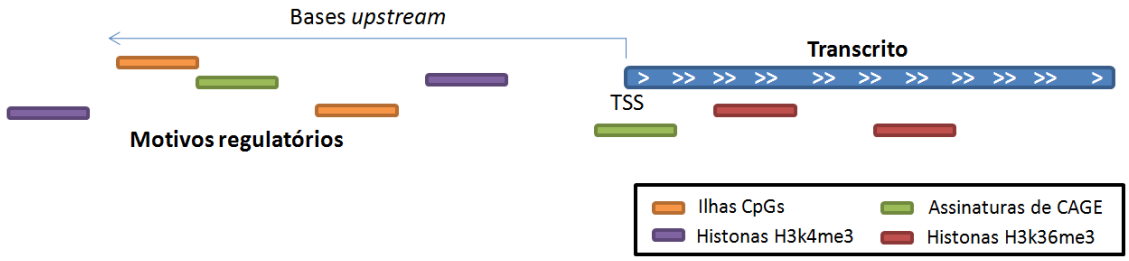


Figura 25 – Representação gráfica do mapeamento de coordenadas genômicas de transcritos de interesse (i.e. lncRNAs, mRNAs codificadores de proteínas ou sequências randômicas), em relação a motivos regulatórios de interesse. A barra maior em azul, representa o RNA em estudo. As barras menores representam os motivos regulatórios de interesse: a) ilhas CpGs, em laranja; b) histonas H3K4me3, em roxo; c) assinaturas de CAGE, em verde; e d) histonas H3K36me3, em vermelho. Os mapeamentos foram realizados utilizando a ferramenta BEDtools e *scripts* em linguagem Perl. Para os motivos de ilhas CpGs, histonas H3k4me3, as marcas que se apresentaram em até 10 kb *upstream* foram computadas. Para as histonas H3K4me3, apenas aquelas que tinha sobreposição com os transcritos foram computadas.

Para testar a significância estatística da sobreposição entre o nosso grupo de lncRNAs e os *datasets* de motivos regulatórios (H3K4me3, ilhas CpGs e assinaturas de CAGE), geramos 100 grupos randômicos controle a partir de regiões intrônicas e intergênicas do genoma humano, contendo o mesmo tamanho e conteúdo GC dos transcritos não codificadores expressos. Os dados de assinaturas de CAGE foram pré-processados, de maneira que as assinaturas contendo sobreposição de coordenadas foram “clusterizadas”, e apenas aqueles *clusters* contendo pelo menos cinco assinaturas foram considerados para as análises posteriores.

Primeiramente, computamos a distância das marcas de histona H3K4me3, ilhas CpGs preditas e assinaturas de CAGE mais próximas do nosso conjunto de lncRNAs, mRNAs codificadores de proteínas expressos no *array* e para os 100 grupos de sequências randômicas. Elementos regulatórios de transcrição que mapeavam com regiões 5’ UTR de transcritos conhecidos (genes RefSeq e UCSC) foram removidos com o intuito de evitar a contribuição do mapeamento com sinais de inicio de transcrição de genes conhecidos com elementos regulatórios de início de transcrição de lncRNAs localizados próximo ao início do gene codificador.

Apenas elementos regulatórios presentes em uma distância até 10 kb na vizinhança dos lncRNAs foram considerados. Após isso, o teste estatístico não paramétrico *Kolmogorov-Smirnov* (KS), implementado a partir do pacote *Deducer* em linguagem R (www.deducer.org), foi utilizado para comparar a distribuição das distâncias das histonas H3K4me3, ilhas CpGs preditas e assinaturas de CAGE encontradas na vizinhança dos lncRNAs intrônicos e intergênicos, em relação a cada um dos 100 grupos do controle randômico. A distribuição dos motivos regulatórios nos lncRNAs intrônicos ou intergênicos e mRNAs codificadores de proteínas foi considerada significativamente diferente, em comparação com os controles, se todos os *p-values* do teste KS calculados utilizando cada um dos grupos randômicos fossem menores que 0,05.

Análises de conservação genômica e de estrutura secundária de lncRNAs intrônicos e intergênicos expressos no pâncreas humano

Da mesma forma que o tópico anterior, utilizamos o pacote BEDtools para cruzar as coordenadas genômicas dos lncRNAs expressos contra regiões genômicas conservadas em vertebrados (*phastCons 46way vertebrate*), mamíferos (*phastCons 46way placental*) e primatas (*phastCons 46way primates*); calculadas anteriormente utilizando o programa PhastCons (Siepel et al, 2005). Os dados foram baixados a partir de *datasets* disponíveis no *UCSC Genome Browser* (Dreszer et al, 2012). Utilizamos o teste de Fischer para verificar a significância estatística (ponto de corte $p < 0,05$) do enriquecimento das regiões genômicas conservadas, sobrepostas aos *loci* gênicos dos lncRNAs intrônicos e intergênicos, em relação ao enriquecimento observado com sobreposição com as regiões genômicas dos 100 grupos de sequências randômicas.

Finalmente, utilizamos a ferramenta RNAz (Washietl et al, 2005b) para predizermos estruturas secundárias conservadas e termodinamicamente estáveis. Da mesma forma que no Capítulo 3 desta Tese, utilizamos como entrada para o RNAz o alinhamento múltiplo de 17 espécies de vertebrados pertencentes a diferentes ordens da biologia, os quais foram baixados do portal Galaxy (Giardine et al, 2005). Apenas estruturas preditas com $P > 0,5$ foram consideradas.

4.2. Resultados obtidos e discussão

Tecido pancreático humano não tumoral e neoplásico apresentam longos ncRNAs intrônicos e intergênicos

Após a filtragem dos dados extraídos da lâmina de *microarray* 4k, 1.607 transcritos foram detectados como expressos em pelo menos um tipo de tecido histológico, dos quais 1.267 eram mRNAs codificadores de proteínas e 340 potenciais RNAs não codificadores, incluindo transcritos com nenhuma sobreposição a éxons de genes RefSeq, i.e., que mapeiam inteiramente dentro de um íntron ou em regiões intergênicas. Deste total de potenciais lncRNAs, consideramos para o restante do trabalho apenas aquelas sequências expressas que apresentaram um mapeamento no genoma com pelo menos 90% de identidade e cobertura, resultando em 335 transcritos (22 ncRNAs NC_RefSeq conhecidos, 240 potenciais ncRNAs intrônicos e 73 intergênicos).

Frações comparáveis de mRNAs codificadores de proteínas e potenciais transcritos não codificadores foram detectadas em todos os tipos histológicos de tecidos (Tabela 8). Este resultado demonstra que grande parte dos RNAs intrônicos e intergênicos investigados no *microarray* está de fato expressa nos tecidos pancreáticos analisados (Tabela 8). A fração de lncRNAs intrônicos detectados como expressos no microarranjo (240/722 = 0,33) é comparável àquela dos lncRNAs NR_RefSeq conhecidos (22/74 = 0,30) e lncRNAs intergênicos (73/188 = 0,39), e menor que a fração dos mRNAs codificadores de proteínas expressos (1.267/2.371 = 0,53). Esta menor fração detectada para lncRNAs nos tecidos pancreáticos (0,30 a 0,39), em comparação com os mRNAs codificadores de proteínas (0,53), reflete a observação de outros trabalhos que apontam que os transcritos não codificadores são geralmente menos abundantes e mais tecido-específicos que os mRNAs codificadores (Birney et al, 2007a; Nakaya et al, 2007).

Tabela 8 – Sondas detectadas no microarranjo de acordo com o tipo de sonda e a histologia do tecido pancreático.

Tipo	Nº sondas no *array*	NT (n=9)	T (n=15)	M (n=6)	PC (n=8)	Nº de *probes* expressas *
mRNA codificador	2371	1106	1167	1198	1230	1267
ncRNA conhecido (RefSeq)	74	20	19	22	18	22
RNA intrônico	722	206	202	238	235	240
RNA intergênico	188	68	68	74	77	78
Total	3355	1400	1456	1532	1560	1607

*Para ser considerada expressa, a sonda deveria ser detectada acima da media de intensidades do *background* da lâmina em pelo menos 75% das amostras em pelo menos um tipo histológico

Dentre os 240 *loci* gênicos que abrigam lncRNAs intrônicos detectados em tecidos pancreáticos, apenas 62 possuem sondas interrogando éxons de mRNAs neste mesmo *locus*. Destes, 31 (50%) apresentaram apenas sondas cobrindo transcritos intrônicos detectadas como expressos nos microarranjos, apontando para um subconjunto de lncRNAs originados a partir de uma transcrição intrônica totalmente independente, ao invés de um potencial produto de um *splicing* do pré-mRNA. Os demais 31 *loci* foram detectados como expressos tanto por sondas exônicas, quanto intrônicas. Para cada um destes *loci*, calculamos a correlação de Pearson entre a expressão do mRNA codificador e lncRNA intrônico correspondente em todas as amostras de tecido de pâncreas. Foram detectadas, em geral, correlações baixas ($-0,5 < r < 0,5$ para 27 dos 31 *loci*), com 11 *loci* apresentando uma correlação negativa entre a expressão do lncRNA intrônico e o mRNA, e 20 apresentando uma correlação positiva. Esta baixa correlação observada entre a expressão de lncRNAs intrônicos e os éxons do mesmo *loci* sugere que estas moléculas são transcritas, processadas e acumuladas na célula em níveis distintos em relação aos mRNAs produzidos no mesmo *loci*, dando suporte mais uma vez à hipótese de que estes transcritos intrônicos não são meros descartes do processamento de um pré-mRNA.

Finalmente, o potencial codificador das 335 sequencias que mapearam em regiões intrônicas e intergênicas foi investigado utilizando a ferramenta CPC (Kong et al, 2007). Esta análise demonstrou que a maioria das sequências (322/335, 96%) apresentou pouco ou nenhum potencial codificador. Desta forma, sugerimos que a maioria das transcrições intrônicas e intergênicas detectadas em tecidos pancreáticos é realmente constituida de RNAs não codificadores de proteínas.

Embora seja claro que os lncRNAs exercem diversas funções celulares através de vários mecanismos moleculares (Louro et al, 2009; Mattick, 2009; Wilusz et al, 2009), alguns pesquisadores sugerem que uma fração desta porção não codificadora do transcriptoma corresponda a um ruído transcricional resultante da atividade da RNA polimerase em regiões de cromatina aberta ou segmentos intrônicos processados de mRNAs (van Bakel et al, 2010). Nossas medidas de expressão para os lncRNAs intrônicos detectados no pâncreas não permitem distinguirmos entre (a) resquícios intrônicos resultantes do processamento de *splicing* de pré-mRNAs, ou (b) unidades transcricionais intrônicas independentes localizadas dentro de regiões intrônicas humanas. Concentramos o nosso trabalho em frações de RNAs selecionados por conter poliadenilação, seguido por transcrição reversa utilizando *oligo-dTs*, visando minimizar a chance de hibridizarmos alvos não-poliadenilados resultantes do *splicing* de mRNAs. Adicionalmente, realizamos diversas análises visando a obtenção de evidências que apoiam a hipótese de que estes lncRNAs intrônicos e intergênicos detectados em tecidos pancreáticos sejam realmente transcritos não codificadores funcionais.

lncRNAs intrônicos e intergênicos expressos no pâncreas humano são originados a partir de regiões conservadas do genoma

A conservação de sequencias entre espécies é geralmente vista como um indicador de funcionalidade de uma dada característica genômica. Buscamos por evidências de conservação entre o grupo de ncRNAs intrônicos e intergênicos expressos em tecidos pancreáticos, a partir do mapeamento de suas coordenadas genômicas levando em conta os elementos de DNA conservados entre vertebrados, mamíferos e primatas, obtidos a partir do UCSC Genome Browser (Dreszer et al, 2012). Como esperado, após normalização pelo número de elementos conservados em cada grupo, a sobreposição com elementos conservados foi maior em relação aos primatas, mamíferos e vertebrados, respectivamente (Figura 26).

A sobreposição dos ncRNAs com elementos evolutivamente conservados foi maior que o esperado ao acaso, como visto a partir de uma análise similar utilizando sequências intrônicas e intergênicas randomicamente selecionadas com o mesmo tamanho e conteúdo GC (teste de Fisher $p < 0,05$, Figura 26). Este resultado sugere que pelo menos uma fração destes lncRNAs está sob seleção evolutiva dentre os

vertebrados e provavelmente estes lncRNAs são biologicamente funcionais. Para os transcritos não codificadores restantes, a ausência de conservação em sua sequência linear não deve ser tomada como uma ausência de funcionalidade, visto que já é sabido que ncRNAs funcionalmente caracterizados também apresentam pouca conservação em relação a sua sequência linear, mantendo uma conservação estrutural quando leva-se em conta a covariação das bases que formam a sua estrutura (Pang et al, 2006; Torarinsson et al, 2006). Como proposto por Washielt e colaboradores, o mapeamento de regiões contendo estruturas secundárias de RNAs pode conduzir à descoberta de novos lncRNAs funcionais (Washietl et al, 2005a).

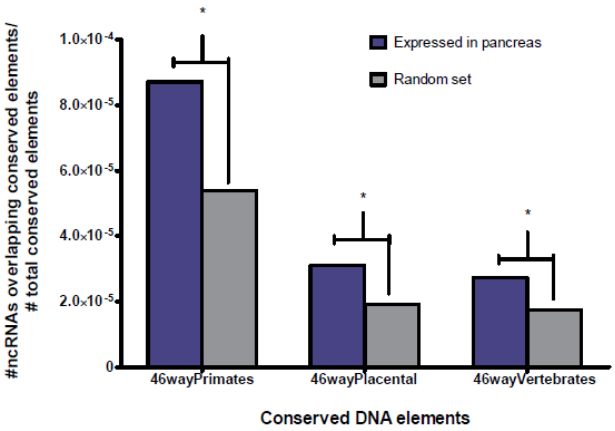


Figura 26 – Os lncRNAs intrônicos e intergênicos estão enriquecidamente localizados em regiões contendo elementos conservados de DNA. Coordenadas genômicas de elementos de DNA conservados identificados em vertebrados, placentários e primatas (ver Materiais e Métodos para maiores detalhes) foram cruzadas com as coordenadas dos lncRNAs expressos em tecidos pancreáticos (barras azuis). Estes transcritos apresentaram maior sobreposição com elementos evolutivamente conservados de DNA do que observado com um conjunto controle de sequências genômicas aleatórias com mesmo tamanho e conteúdo GC (barras cinzas) (teste de Fisher, $p < 0,05$). A altura das barras indica o número de lncRNAs que tiveram sobreposição com elementos conservados de DNA em cada grupo taxonômico dividido pelo número total de elementos conservados de DNA presentes em cada grupo.

Desta forma, utilizando a ferramenta RNAz (Washietl et al, 2005b) encontramos que uma fração destes ncRNAs expressos no pâncreas humano (49 das 335 sequências analisadas, 15%) tende a enovelar-se em domínios estruturais estáveis e conservados evolutivamente ($P > 0.5$), e pode ter papéis fundamentais no seu processamento ou função biológica. Em conjunto, todas estas observações fornecem indícios adicionais

para a suposição de que parte destes transcritos que mapeiam regiões intrônicas e intergênicas do genoma humano deve exercer papéis funcionais nas células pancreáticas. As estruturas identificadas estão disponíveis em: http://verjo102.iq.usp.br/sites/tahira/structures.html.

Já está bem documentado na literatura que pequenos RNAs regulatórios podem ser gerados a partir do processamento de RNAs longos precursores, transcritos de regiões intrônicas e intergênicas do genoma (Wilusz et al, 2009). Para investigar a fração do nosso grupo de lncRNAs expressos em tecidos que poderia originar pequenos RNAs, comparamos suas coordenadas genômicas contra as coordenadas de snoRNAs e microRNAs conhecidos (Kozomara & Griffiths-Jones, 2011; Lestrade & Weber, 2006). Encontramos apenas uma discreta sobreposição com um único snoRNA presente no *locus SNOR89*. Pelo fato dos precursores de pequenos RNAs regulatórios possuírem um tamanho médio de mil bases, estendemos as coordenadas dos nossos lncRNAs até mil bases e refizemos esta análise. Apenas um conjunto adicional de sete RNAs curtos apresentou sobreposição com lncRNAs expressos em pâncreas: *hsa-mir-1259*, *hsa-mir-326*, *hsa-mir-4269*, *hsa-mir-675*, *SNORD12*, *SNORD12B* e *SNORD12C*. Recentemente, uma nova classe de snoRNAs associados a longos ncRNAs foi descrita, os sno-lncRNAs, os quais apresentam um papel funcional distinto dos snoRNAs comuns, estando associados à síndrome de Prader-Willi (Yin et al, 2012). Estes transcritos podem estar associados a esta nova classe, que talvez também tenham um papel associado com o câncer de pâncreas. Estes resultados deixam abertas as hipóteses de que estes transcritos devem representar outras classes ainda desconhecidas e não caracterizadas de novos RNAs curtos.

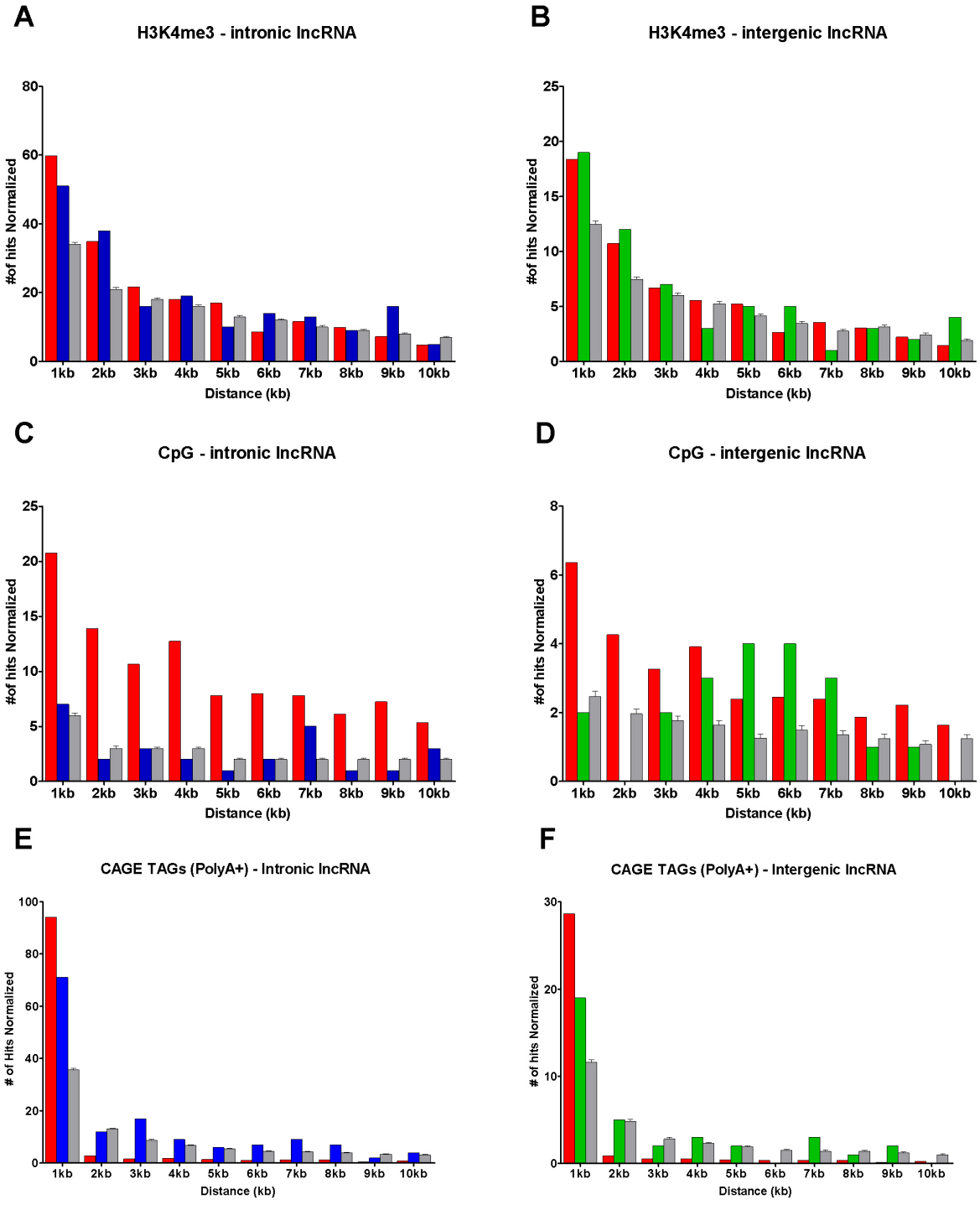
Loci genômicos que originam lncRNAs intrônicos de pâncreas apresentam enriquecimento de marcadores de cromatina associados a regiões promotoras e locais de início de transcrição de transcritos contendo 5'-cap

Dada a escassez de informações acerca da biogênese de ncRNAs originados a partir de regiões intrônicas e intergênicas, buscamos por elementos regulatórios no genoma que poderiam estar associados ao seu controle transcricional. Primeiramente, investigamos a distribuição da trimetilação da lisina 4 em histona 3 (H3K4me3), uma modificação na cromatina associada com regiões de ativação da transcrição (Mikkelsen

et al, 2007), nas proximidades dos lncRNAs identificados em pâncreas. As coordenadas genômicas para marcas de H3K4me3 medidas em 13 linhagens celulares (Mikkelsen et al, 2007) foram obtidas a partir do *UCSC Genome Browser* (Dreszer et al, 2012). Apenas aquelas marcas com p < 10e-5 foram utilizadas como forma de eliminar algum ruído experimental. A marca para H3K4me3 mais próxima em relação aos limites conhecidos das sequências de cada ncRNA expresso em tecidos pancreáticos (baseado no mapeamento das sequências das ESTs no genoma) foram selecionadas e a distância entre elas anotada. Como controle, realizamos a mesma análise utilizando um conjunto de 100 sequências de DNA intrônicas ou intergênicas aleatórias no genoma com mesmo tamanho e conteúdo GC.

Um enriquecimento de marcas H3K4me3 foi observado nas proximidades dos limites conhecidos para o grupo de transcritos intrônicos expressos nas amostras de tecido pancreático (27, painel A, barras azuis). A distribuição relativa das marcas de histonas em relação às vizinhanças conhecidas dos transcritos foi significativamente diferente (teste KS, maior p-value < 0,05) das distribuições levando em conta os conjuntos randômicos, indicando que não é uma distribuição ao acaso. A mesma análise foi realizada com um grupo de lncRNAs expressos a partir de regiões intergênicas. Apesar de observarmos uma maior frequência de marcas de H3K4me3 nas proximidades de transcritos intergênicos, não encontramos diferença estatística na sua distribuição em relação ao grupo controle randômico (Figura 27, painel B, barras verdes).

Como esperado, também verificamos uma maior frequência de marcas H3K4me3 nas proximidades do início transcricional de mRNAs codificadores de proteínas (27, painel A, barras vermelhas). Esta distribuição é estatisticamente diferente daquela obtida com o conjunto controle, composto por sequências randômicas (teste KS, maior p-value < 0,01). Nenhuma diferença estatística significante foi observada entre os lncRNAs intrônicos/intergênicos e os mRNAs expressos no pâncreas, indicando que as distribuições das marcas de histonas associadas a regiões promotoras são similares tanto no grupo codificador, como no não codificador de proteínas. Estes níveis comparáveis de enriquecimento de marcas de histonas H3K4me3 na vizinhança de lncRNAs e mRNAs, sugerem que a transcrição de mRNAs e de lncRNAs intrônicos são iniciadas por regiões promotoras similares. Além disso, fornece evidências adicionais de que estes lncRNAs intrônicos são de fato unidades transcricionais independentes.



27 – *Loci* genômicos que transcrevem lncRNAs intrônicos são enriquecidos com marcadores de histonas associados a regiões promotoras (H3K4me3) e locais de início de transcrição de transcritos contendo o 5'-cap. Distribuição da distância (eixo X) dos locais de ligação de marcadores de cromatina H3K4me3 (painéis A e B), ilhas CpG (painéis C e D) e assinaturas de expressão de CAGE (painéis E e F), em relação às coordenadas genômicas de transcritos intrônicos (barras azuis) e intergênicos (barras verdes) expressos em tecido pancreático (eixo Y). Para comparação, utilizamos o mesmo número de sequências de mRNAs codificadores de proteínas (barras vermelhas) e sequências randômicas intrônicas e intergênicas com mesmo tamanho e conteúdo GC dos lncRNAs expressos no pâncreas (barras em cinza claro).

Também verificamos a distribuição de ilhas CpG anotadas em relação às ESTs (codificadoras e não codificadoras de proteínas) presentes no nosso microarranjo e expressas em tecidos pancreáticos. Nesta análise, utilizamos as coordenadas genômicas de ilhas CpGs disponíveis no *UCSC Genome Browser* (Dreszer et al, 2012). Primeiramente, cruzamos as coordenadas anotadas para ilhas, contra as coordenadas das ESTs presentes no *array* que representavam mRNAs codificadores expressos em tecidos pancreáticos, os quais demonstraram enriquecimento nas proximidades destes transcritos (Figura 27, painel C, barras vermelhas). Este mapeamento apresentou-se significativamente diferente em relação ao observado em um grupo com sequências randômicas (teste KS, $p < 0,0001$). No entanto, nenhuma associação estatisticamente significante, levando em conta motivos de ilhas CpGs, foi observado para lncRNAs intrônicos e intergênicos em relação a conjuntos randômicos com mesmo tamanho e conteúdo GC (teste KS, $p > 0,05$) (Figura 27, painéis C e D, barras azuis e verdes). Isto sugere que a metilação de ilhas CpGs não está envolvida na regulação da transcrição da maioria dos lncRNAs intrônicos e intergênicos expressos em tecidos pancreáticos.

Finalmente, realizamos a comparação do local de início de transcrição dos lncRNAs intrônicos e intergênicos com o local de mapeamento de CAGE *tags* obtidos a partir do sequenciamento da ponta 5' do RNA que contem o cap de metil-guanosina (Takahashi et al, 2012), de RNAs poliadenilados de seis diferentes linhagens celulares do projeto RIKEN. É importante saber que este *dataset* não inclui bibliotecas de CAGE derivadas de tecidos pancreáticos. Primeiramente, as coordenadas das *tags* foram agrupadas de maneira que apenas *clusters* contendo pelo menos cinco *tags* foram considerados para as análises posteriores. Após isso, calculamos a distância do *cluster* de CAGE *tags* mais próximo de lncRNAs intrônicos ou intergênicos, mRNAs codificadores de proteínas e sequências genômicas escolhidas randomicamente. Foi observado um enriquecimento significativo (KS teste, $p < 0,05$) de *tags* localizadas até 1kb do potencial local de início de transcrição de lncRNAs intrônicos expressos em tecidos pancreáticos (Figura 27, painel E, barras azuis). Apesar da alta frequência observada de CAGE *tags* próximas ao local de início de transcrição de lncRNAs intergênicos, nenhum enriquecimento estatístico foi observado quando comparado com sequências randômicas controle (Figura 27, painel F, barras verdes e cinzas). Estes resultados, juntamente com todos os outros demonstrados em relação aos mapeamentos de assinaturas de cromatina, e evidências conservacionais, corroboram

mais uma vez com a hipótese de que estes lncRNAs intrônicos são unidades transcricionais independentes.

6.3. Conclusões

Neste trabalho, descrevemos que RNAs não codificadores de proteínas originados a partir de regiões intrônicas e intergênicas do genoma estão expressas em tecidos pancreáticos neoplásicos e não tumorais humanos. Enriquecimento de marcas de cromatina associadas a regiões promotoras, em conjunto com a observação que muitos destes ncRNAs intrônicos apresentam níveis de expressão diferentes em comparação com os mRNAs codificadores de proteína presentes nos *loci* gênicos nos quais são transcritos, fornecem evidências adicionais de que estas mensagens não são erros transcricionais, nem mesmo produtos do processamento de *splicing*, e sim unidades transcricionalmente ativas independentes. Adicionalmente, evidências conservacionais estruturais e de sequência do *locus* genômico de onde são originados sugerem que estes transcritos sofreram uma certa pressão seletiva ao longo da evolução. Estas pressões evolutivas normalmente estão correlacionadas com funcionalidade.

Em resumo, nosso trabalho fornece um conjunto de candidados não codificadores de proteínas que aparenta apresentar uma relevância biológica em tecidos pancreáticos tumorais e não tumorais. Os resultados destacam a importância de investigar a participação destes transcritos no entendimento das bases moleculares da doença.

6.4. Contribuições do autor para o trabalho

Este trabalho foi realizado em colaboração com a doutoranda do Laboratório de Expressão Gênica em Câncer do Instituto de Química da USP, Ana Tahira, orientanda do professor Dr. Eduardo Reis.

Os dados aqui apresentados fizeram parte de um trabalho mais amplo, contendo outros resultados não mostrados nesta tese, mas que podem ser observados no artigo científico publicado na Molecular Cancer, referenciado no próximo tópico. As etapas de extração de RNAs, hibridização e análise de expressão dos *microarrays* foram

desenvolvids pela aluna Ana Tahira. A re-anotação da lâmina e o estudo do potencial codificador foram realizados pelo autor da tese, enquanto que as análises de conservação e da região promotora dos ncRNAs foram desenvolvidas em conjunto por Tahira e pelo autor.

6.5. Publicação

Os resultados apresentados neste capítulo foram publicados no artigo: “Tahira, AC; Kubrusly, MS; Faria, MF; Dazzani, B; Fonseca, RS; **Maracaja-Coutinho, V**; Verjovski-Almeida, S; Machado, MCC; Reis, RM. *Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer.* **Molecular Cancer**, 10:141, 2011.” (Tahira et al, 2011), ANEXO 2 desta Tese.

Capítulo 7

Análise das ESTs públicas do protostomado <u>Schistosoma mansoni</u> **revelam um grande repertório de ncRNAs**

7. Análise das ESTs públicas do protostomado *Schistosoma mansoni* revelam um grande repertório de ncRNAs

O *Schistosoma mansoni* é um parasita humano prostostomado causador da esquistossomose. Esta doença ocorre em 76 países, sendo estimado que 779 milhões de pessoas vivem em situação de risco para adquiri-la, e que em torno de 207 milhões de pessoas estão atualmente infectadas (Steinmann et al, 2006). O genoma do *Schistosoma ssp* é organizado em oito cromossomos, sete deles autossômicos e um sexual. Em 2003, nosso grupo publicou o transcriptoma do *S. mansoni* na *Nature Genetics* (Verjovski-Almeida et al, 2003), fornecendo novas perspectivas para a genômica funcional deste organismo (Verjovski-Almeida et al, 2004). Seis anos mais tarde, em 2009, as sequências genômicas do *S. mansoni* (Berriman et al, 2009) e *S. japonicum* (Schistosoma japonicum Genome & Functional Analysis, 2009) foram publicadas. No entanto, nenhuma análise comparativa entre o transcriptoma e o genoma destes organismos foi realizada.

Desta forma, pelo fato do *S. mansoni* ser um organismo extensamente estudado pelo nosso grupo, e o grande interesse em entendermos os mecanismos envolvidos na evolução dos RNAs não codificadores, especialmente aqueles originados a partir de regiões intrônicas do genoma, resolvemos re-analisar todo o repertório de ESTs públicas do *S. mansoni*, comparando-as com o genoma recém-sequenciado, e identificar os potenciais ncRNAs presentes neste organismo mais basal na evolução dos eucariotos.

7.1. Metodologia utilizada

Bases de dados e pré-processamentos das ESTs

Primeiramente, baixamos todas as 205.892 sequencias de ESTs e mRNAs de *Schistosoma mansoni* disponíveis no GenBank até dezembro de 2010. Aquelas sequencias menores que 100 bases foram eliminadas. Após isso, filtramos as sequências contra possíveis vetores, utilizando a ferramenta cross_match, do pacote Phred/Phrap/Consed (Ewing et al, 1998), contra o banco UniVec, do NCBI.

Posteriormente, utilizamos a ferramenta BLAST (Altschul et al, 1990) para cruzar todas as sequências restantes contra as predições gênicas de *Schistosoma mansoni*

publicadas juntamente com o seu genoma (Berriman et al, 2009), e disponíveis no site do Sanger Institute http://www.sanger.ac.uk/resources/downloads/helminths/schistosoma-mansoni.html). Desta forma, teríamos em mãos todo o repertório de ESTs públicas pertencentes às predições gênicas conhecidas, como também aquelas predições sem evidência de ESTs.

Montagem e anotação das ESTs ainda não anotadas

O conjunto de ESTs que não pertenciam a genes conhecidos poderiam ser novos genes codificadores de proteínas, novos RNAs não codificadores, ou até mesmo contaminantes pertencentes a outros organismos que não foram filtrados nos *pipelines* utilizados em seus projetos de sequenciamento. Para eliminar essa última possibilidade, utilizamos a ferramenta BLAST (Altschul et al, 1990) contra o genoma ou DNA de todos os potenciais organismos que poderiam atuar como contaminantes: a) elementos transponíveis, b) DNA mitocontrial, c) DNA ribossomal, d) *Biomphalaria ssp*., e) humano, f) boi, g) hamster, h) camundongo, e i) bactérias.

Com as sequencias finais em mãos, realizamos o processo de montagem das ESTs utilizando a ferramenta CAP3 (Huang & Madan, 1999), com limite mínimo de 30 bases de sobreposição e 90% de identidade. As sequências *contigs* e *singlets* resultantes do processo de montagem foram posteriormante mapeadas contra o genoma do *Schistosoma mansoni* utilizando a ferramenta BLAST (Altschul et al, 1990). Então, os conjuntos de sequências *contigs* e *singlets* que mapearam ou não no genoma foram comparados contra todas as proteínas do banco de dados Uniprot (Wu et al, 2006a), utilizando a ferramenta BLASTX (Altschul et al, 1990). As sequências que não tiveram nenhuma similaridade com as proteínas disponíveis no Uniprot tiveram seu potencial de codificar uma proteína avaliado utilizando a ferramenta CPC (Kong et al, 2007). Com estas últimas análises, chegamos ao conjunto de ESTs que potencialmente pertencem a novos genes codificadores de proteínas ainda não descritos no *S. mansoni*, como também no repertório de potenciais RNAs codificadores do organismo. A Figura 27 dos resultados deste capítulo apresenta um fluxograma onde é possível visualizar todo o *pipeline* utilizado para mapeamento e anotação das ESTs.

7.2. Resultados obtidos e discussão

ESTs públicas *versus* predições gênicas: quantos genes codificadores e não codificadores de proteínas existem em *S. mansoni*?

Primeiramente, realizamos uma comparação entre o transcriptoma e o genoma de S. *mansoni*, visando à identificação da percentagem de transcritos que poderiam estar relacionados a potenciais ncRNAs (Figura 28). Utilizamos todas as 205.892 ESTs e mRNAs públicas de *S. mansoni* disponíveis no GenBank em dezembro de 2010. As ESTs que tiveram alta similaridade com sequências vetores foram eliminadas (133 ESTs). Dentre as 205.759 ESTs restantes, verificamos que 154.707 (75,1%) mapearam com genes anotados de *S. mansoni* (i.e. 13.215 predições gênicas *Smp* (Berriman et al, 2009), mais outros genes de ncRNAs (RNAs ribossomais, transportadores, microRNAs e *small nucleolar* RNAs), descritos e disponíveis no website do Sanger Institute, em: http://www.sanger.ac.uk/resources/downloads/helminths/schistosoma-mansoni.html).

A partir das 51.052 ESTs que não mapearam com genes *Smp* (i.e. predições gênicas feitas pelo Sanger para *S. mansoni*) ou outros genes descritos, 10.942 foram filtradas e eliminadas por terem alto grau de similaridade com elementos repetitivos e genes mitocondriais. As 40.110 ESTs (19,5%) restantes foram então montadas utilizando o programa CAP3, gerando um total de 5.166 *contigs* (compostos de 22.553 ESTs, 11%) e 17.557 EST *singlets* (8,5%) (Figura 28).

As ESTs montadas foram divididas em dois grupos. Um contendo sequências que mapearam no genoma fora de qualquer região codificadora de genes *Smp* preditos (15.536 ESTs (7,5% do total) que montaram em 3.311 *contigs,* mais 9.080 *singlets* de ESTs (4,4% do total) restantes (total de 24.616 ESTs, 11,9%); o outro grupo contém sequencias que não mapearam no genoma (7.017 ESTs montaram em 1.855 *contigs*, mais 8.477 *singlets*; um total de 15.494 ESTs, 7,5%) (Figura 28). No total, nossa análise demonstrou que 87% das ESTs públicas de *S. mansoni* mapeiam no genoma, evidenciando o fato de que uma porção significante (11,9%) apresenta evidência de transcrição em regiões nas quais nenhuma predição de genes *Smp* foi feita anteriormente (Berriman et al, 2009). Além disso, 4.076 genes *Smp* e 2.717 outros

genes (Sanger Institute) foram preditos no genoma de *S. mansoni*, mas não apresentaram nenhuma evidência de EST.

Figura 28 – Fluxograma do mapeamento genômico e toda anotação das ESTs públicas de *S. mansoni* disponíveis no GenBank.

Por outro lado, as 26.616 ESTs que não tiveram *match* com nenhuma predição gênica *Smp*, mas que mapearam no genoma, foram montadas em 3.311 contigs (15.536 ESTs, 7,5%) (Figura 28). Dentre eles, 522 *contigs* (2.547 ESTs, 1,2%) tiveram alta similaridade com 154 proteínas conhecidas do Uniprot, pertencentes a outros organismos e que ainda não haviam sido preditas em *S. mansoni*. Além disso, um *contig* adicional (composto de 2 ESTs, 0,00003%), que não tinha similaridade com proteínas UNIPROT, foi predito como tendo um alto potencial codificador pela ferramenta CPC (Figura 28). Os 2.788 *contigs* restantes (12.987 ESTs, 6,3%) são potenciais RNAs não codificadores de proteínas, visto que eles não apresentaram similaridades com nenhuma proteína presente no Uniprot e não apresentaram potencial codificador de acordo com o CPC. Levando em conta as 9.080 ESTs *singlets* (4,4%) que mapearam no genoma, mas não mapearam com predições gênicas, encontramos 960 ESTs (0,5%) mapeando proteínas conhecidas do Uniprot. As 8.120 ESTs *singlets* (3,9%) restantes também são potenciais RNAs não codificadores, visto que não mapeiam com nenhuma proteína conhecida, como também não tiveram nenhum possível potencial codificador predito pelo CPC (Kong et al, 2007).

Desta forma, concluímos que, no total, as 21.107 ESTs (10,3%) que mapearam no genoma de *S. mansoni* e não apresentaram nenhum potencial codificador de proteínas são fortíssimos candidatos a ncRNAs. Estas ESTs estão localizadas em 10.908 regiões distintas do genoma (2.788 *contigs*, mais 8.120 *singlets*) com evidências de transcrição de ncRNAs. Nossos resultados também apontam para 356 proteínas conhecidas presentes no Uniprot (2.547 ESTs montadas em 522 *contigs*, e 960 singlets) que também são expressas em *S. mansoni*, mapeiam em sua sequencia genômica, mas não haviam sido preditas anteriormente pelo projeto genoma original (Berriman et al, 2009).

Diversas ESTs públicas de *S. mansoni* não mapeiam em seu genoma

Um total de 15.585 ESTs (7,7%) não mapeou no genoma de *S. mansoni*, e são potencialmente transcritas a partir de uma região ainda não sequenciada do seu genoma, ou representariam potenciais contaminantes adicionais presentes no banco de dados de seu transcriptoma, especialmente os ESTs *singlets*. Dentre os 1.855 *contigs* (7.017 ESTs, 3,4%), encontramos que 329 (1.857 ESTs, 0,1%) deles mapeavam com

sequencias contaminantes (i.e. *M. musculus*, *H. sapiens*, *B. glabrata*, *B. taurus*, *R. norvegicus* e bactérias) (Figura 28). Dentre os 1.516 contigs (5.160 ESTs, 2,5%), 488 *contigs* (2.186 ESTs) mapearam com 434 proteínas UNIPROT. Por outro lado, 1.068 *contigs* (2.974 ESTs) não mapearam com proteínas UNIPROT. Deste grupo, 2 *contigs* (2 ESTs, 1,4% de todos os transcritos) apresentaram potencial codificador de acordo com o CPC, enquanto que 1.066 *contigs* não apresentaram nenhum potencial.

Dentre as 8.477 ESTs *singlets* que não mapeiam no genoma, encontramos que 956 mapearam com potenciais contaminantes. Dentre as 7.521 *singlets* restantes, 1.276 mapearam com 764 proteínas do UNIPROT, enquanto que 6.245 ESTs (3%) não mapearam com proteínas e não apresentaram potencial codificador (Figura 28).

Concluímos que 9.213 ESTs (4,4%), dentre *contigs* e *singlets* que não mapeiam no genoma, são potenciais ncRNAs. Estes dados também apontam que 1.443 proteínas únicas do UNIPROT são potencialmente expressas em *S. mansoni*, mas ainda não tiveram sua região genômica sequenciada.

Resumindo, dentre as 205.892 ESTs públicas de *S. mansoni*, um total de 30.320 ESTS (dentre *contigs* e *singlets*, 14,7% do total) não apresentou nenhum potencial codificador de proteínas. Entre essas, 21.107 ESTs (10,2%) mapeiam no genoma, enquanto que 9.213 ESTs (4,5%) não mapeiam. A fração total de transcrição em *S. mansoni* compreendida por ncRNAs provavelmente revela um menor nível transcricional desta classe de transcritos quando comparada à atividade transcricional de genes codificadores de proteínas. Já se sabe que ncRNAs longos em humanos são transcritos em uma taxa bem menor que os genes codificadores de proteínas (Kapranov et al, 2007), como também que os ncRNAs representam entre 10 a 20% dos repositórios humanos de ESTs públicas (Nakaya et al, 2007).

Cobertura das ESTs em relação ao genoma e predições gênicas: uma surpreendente atividade transcricional intrônica

Após todo o levantamento dos possíveis ncRNAs presentes no transcriptoma de *S. mansoni*, calculamos a percentagem de bases no genoma que é coberta por predições gênicas, e o número de bases do genoma que são cobertas por ESTs públicas presentes no dbEST. O genoma de *S. mansoni* possui 362.876.148 bases,

distribuídas em 5.754 *scaffolds* genômicos maiores que 2 kb (Berriman et al, 2009). Deste total de bases, 165.206.376 bp (45,5%) são *loci* de genes preditos. Dentro destes *loci*, 15.852.242 bp são éxons de predições gênicas (4,3% do total de bases no genoma, 9,6% dos *loci*), enquanto que 149.354.134 bp (41% do total de bases no genoma, 90,4% dos *loci*) são íntrons.

Baseados no levantamento das ESTs públicas de *S. mansoni* que vêm sendo acumuladas e que mapeiam com as regiões sequenciadas do genoma, verificamos que um total de 16.516.608 bp está coberto por pelo menos uma EST, o que significa que pelo menos 4,6% do genoma de *S. mansoni* é transcrito. Deste total de 16.516.608 bp transcritas, 12.717.085 bases (77% das bases transcritas) estão localizadas em *loci* de predições gênicas (3,5% das bases genômicas). Um total de 7,7% das bases dos *loci* gênicos está coberto por ESTs, incluindo regiões exônicas e intrônicas.

Analisando mais profundamente os éxons do genoma (composto de 15.852.242 bp), verificamos que 8.652.015 bp estão cobertas por ESTs públicas (42% das bases transcritas), representando uma cobertura de 55% das bases dos éxons de genes preditos (2,4% das bases do genoma); além disso, outras 1.557.580 bp estão localizadas em regiões UTRs (Figura 29). Levando-se em conta os íntrons de genes preditos (149.354.134 bp), verificamos que 4.065.070 bp foram cobertos por ESTs públicas (34% do total de bases transcritas). Detectamos que 3.799.523 das bases transcritas (1% do total de bases do genoma, e 23% das bases transcritas) estão localizadas em regiões intergênicas. A Figura 29 resume todos estes números.

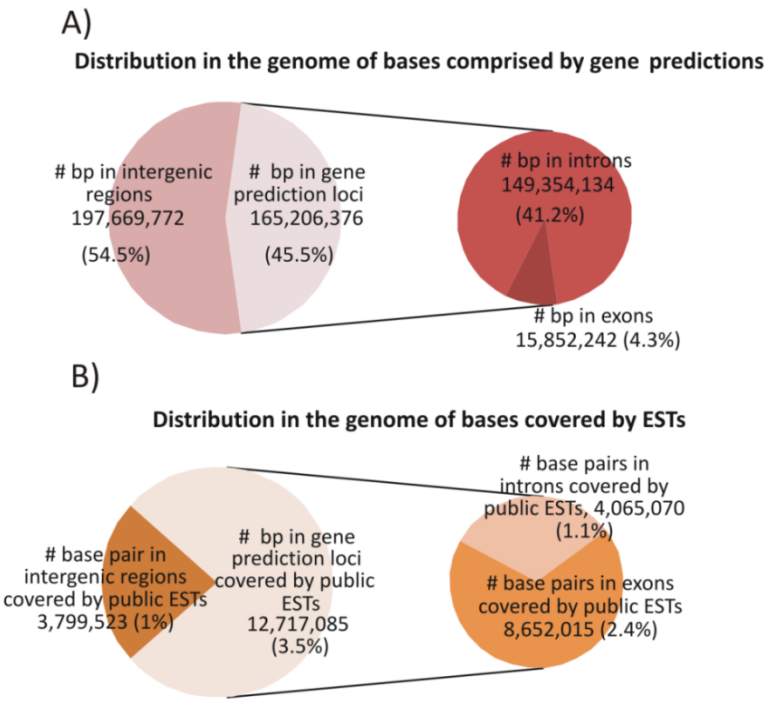


Figura 29 – Análises das bases genômicas de *S. mansoni*. (A) Distribuição das bases que compreendem predições gênicas. (B) Distribuição das bases cobertas por ESTs públicas. As percentagens da direita de cada figura estão relacionadas à subcategoria correspondente da parte da esquerda da mesma.

Em 2004, Mattick (Mattick, 2004) levantou a hipótese de que a complexidade de um organismo é provavelmente derivada da expansão das regiões não codificadoras de proteínas no genoma, especialmente pelo fato que não há um aumento considerável de genes codificadores de proteínas ao longo da evolução. Por outro lado, é evidente o aumento de regiões não codificadoras nos genomas de organismos mais complexos. Tal expansão se deu especialmente nas regiões intrônicas de genes codificadores. Mattick calculou a razão entre o DNA não codificador e o DNA total (ncDNA/DNAtotal) em uma grande variedade de organismos pertencentes aos diferentes reinos e classes, e verificou que organismos complexos como *M. musculus* e *H. sapiens*, têm uma razão superior a 0,9 (Mattick, 2004).

Baseando-se no genoma e nas predições gênicas (Berriman et al, 2009), calculamos a razão ncDNA/DNAtotal em *S. mansoni*, e obtivemos um resultado consideravelmente alto de 0,96, quando levamos em conta a complexidade do parasita, e comparamos tal resultado com o genoma humano. Vale salientar que outro

platelminto, *Schmidtea mediterranea*, apresenta um genoma de 480 MB (http://genome.wustl.edu/genomes/view/schmidtea_mediterranea/), em torno de 100 MB maior que *S. mansoni*, mas aparenta não apresentar um número maior de genes codificadores de proteínas, resultando numa taxa ncDNA/DNAtotal maior que *S. mansoni*. Tais observações sugerem que a expansão do ncDNA pode realmente ser um dos mecanismos utilizados ao longo da evolução a fim de adicionar complexidade aos platelmintos, como também melhorar suas maneiras de interação com o meio. *S. mansoni* pode ter sofrido uma redução genômica (Keeling & Slamovits, 2005) em relação a *S. mediterranea* devido a sua forma de parasitismo.

No entanto, ainda há uma limitada evidencia da atividade transcricional do genoma de *S. mansoni*. Apenas 4,6% de todas as bases estão cobertas por ESTs públicas, como descrito acima. As regiões intrônicas preditas no genoma de *S. mansoni* são bastante longas, compreendendo 41% das bases genômicas totais e 90% dos *loci* genômicos das predições gênicas. A transcrição detectada em nestas regiões corresponde a apenas 1,1% do total de bases em seu genoma, apesar de 34% das bases transcritas estarem nos íntrons, onde uma transcrição disseminada vem sendo detectada (Birney et al, 2007a; Kapranov et al, 2007). Em humanos, uma análise de 5,3 milhões de ESTs públicas apontaram para a presença de pelo menos uma EST mapeando em íntrons de 74% de todos os genes RefSeq (Nakaya et al, 2007). Refizemos esta análise para as atuais oito milhões de ESTs humanas disponíveis no dbEST, e encontramos aproximadamente 70.000 *loci* intrônicos únicos, cobrindo em torno de 42 milhões de bases (1,7% do genoma humano) com evidências de transcrição.

Na etapa piloto do projeto ENCODE (Birney et al, 2007a) foi realizado um grande esforço para estudar a transcricional intrônica referente a um segmento correspondente a 1% de todo o genoma humano, utilizando diferentes metodologias como tiling-arrays, RNA-seq e sequenciamento *paired-end*. Naquela época, o projeto ENCODE identificou que 93% deste 1% de segmentos estudados apresentaram transcrição. Tanto regiões intrônicas, quanto intergênicas demonstraram indícios transcricionais (Birney et al, 2007a), sugerindo que esta transcrição pode ser extrapolada para todo o genoma humano. Na etapa final do projeto, onde todo o genoma humano foi explorado, verificou-se que mais de 75% do genoma apresentava uma transcrição disseminada (Djebali et al, 2012). Em outros organismos eucariotos complexos, como *C. elegans*, 70% do genoma é transcrito, este organismo vem sendo analisado em um projeto ENCODE

próprio (Gerstein et al, 2010). Em *D. melanogaster*, 85% do seu genoma é transcrito (Graveley et al, 2011).

Observando todos estes números, não é de surpreender que pelo menos 34% das bases transcritas no genoma de *S. mansoni* venha de regiões intrônicas. Além disso, a limitada cobertura da transcrição intrônica em relação ao genoma total (1,1% das bases totais) pode ser explicada pelo pouco número de projetos de sequenciamento de ESTs envolvendo S. *mansoni* até o momento (Almeida et al, 2012; Franco et al, 1995; Franco et al, 2000; Merrick et al, 2003; Verjovski-Almeida et al, 2003). Outro fator limitante é que tais projetos foram realizados utilizando apenas transcritos poliadenilados, e já se sabe que muitos transcritos, especialmente ncRNAs, não apresentam Poli(A) (Kiyosawa et al, 2005). Isso sugere que um grande esforço envolvendo sequenciamento de alta estringência e abordagens de *tiling-arrays* ainda são necessários para podermos obter uma maior cobertura de toda a atividade transcricional do genoma de *S. mansoni*.

7.3. Conclusões

Neste trabalho, reanalisamos todo o repertório de ESTs públicas do protostomado parasita humano *Schistosoma mansoni*, onde foi possível identificar um grande repertório de potenciais novos genes e novos RNAs não codificadores para este organismo em regiões intrônicas e intergênicas do seu genoma. Também evidenciamos parte de predições gênicas que não haviam nenhum sinal de atividade transcricional nos bancos públicos, como também um conjunto de ESTs com evidências para codificar proteínas mas que não haviam sido preditas para o genoma do organismo.

A análise demonstra o quanto há a ser explorado no genoma e ESTs de organismos não modelo. Atualmente, inúmeras espécies possuem parte de suas ESTs ou parte do seu genoma sequenciados e disponíveis em bancos públicos, e muitas delas sequer tiveram sua atividade transcricional não codificadora de proteínas explorada. O *pipeline* aqui desenvolvido servirá como um guia para a exploração do transcriptoma não codificador em organismos não modelos. Além disso, o repertório de transcritos não codificadores identificados no genoma de *S. mansoni* servirá como base para futuras caracterizações deste tipo molecular no organismo.

7.4. Contribuições do autor para o trabalho

Este trabalho foi realizado em colaboração com a pós-doutoranda do Laboratório de Expressão Gênica em Eucariotos do Instituto de Química da USP, grupo do qual faço parte, a Dra. Katia Oliveira. Os dados aqui apresentados, foram todos desenvolvidos pelo autor da tese e fizeram parte de um trabalho mais amplo, contendo outros resultados não mostrados, mas que podem ser observados no artigo científico publicado nos Anais da Academia Brasileira de Ciências, referenciado no próximo tópico. Parte das análises da distribuição das bases do genoma de *Schistosoma manson*i foi realizada em conjunto com o Dr. João Paulo Kitajima.

7.5. Publicação

Os resultados apresentados neste capítulo foram publicados no artigo: “Oliveira, KC; Carvalho, MLP; **Maracaja-Coutinho, V**; Kitajima, JP; Verjovski-Almeida, S. *Non-coding RNAs in schistosomes: an unexplored world.* **An. Acad. Bras. Ciên.**, 83(2):673-94, 2011.” (Oliveira et al, 2011), ANEXO 3 desta Tese.

Capítulo 8

Construção de plataformas "customizadas" de oligoarrays 244k intrônico-exônico-intergênico humanas

8. Construção de plataformas "customizadas" de *oligoarray* 244k intrônico-intergênico-exônico humano

O nosso laboratório possui uma vasta experiência com *oligoarrays*. Como descrito previamente, já trabalhamos com plataformas de microarranjos "customizadas" de 4k e 44k. No entanto, apesar de traçarmos o perfil de expressão de milhares de transcritos, há uma limitação de espaço nestas duas plataformas, de maneira que muitos dos transcritos intrônicos identificados a partir das análises de ESTs disponíveis nos bancos públicos não puderam ter sondas desenhadas. A fim de criarmos uma plataforma capaz de medir o perfil transcricional de todos os *contigs* de transcritos intrônicos com pelo menos duas ESTs com coordenadas sobrepostas disponíveis no dbEST, optamos por desenhar um novo *chip* de microarranjo Agilent 244k.

Para a construção das plataformas de *oligoarray* 244k intrônico-exônico-intergênico, aplicamos uma versão modificada do *pipeline* descrito no tópico 4.3 desta tese em todas as ESTs humanas disponíveis no dbEST. A modificação se dá no aproveitamento e “clusterização” das coordenadas genômicas das ESTs caracterizadas como intergênicas. Adicionalmente, utilizamos a base de dados para transcritos exônicos fornecida pela própria Agilent para incluir sondas para todo o repertório de transcritos codificadores de proteínas no nosso *array*.

Como a plataforma 244k Agilent permite o desenho de aproximadamente 244 mil sondas, tínhamos a possibilidade de desenhar sondas para todos os RNAs TINs e PINs identificados pelo nosso pipeline intrônico, como também para todos os transcritos presentes nos nossos antigos arrays 4k e 44k; e podiamos incluir a plataforma comercial com sondas exônicas da Agilent. A Figura 30 e a Tabela 8 descrevem graficamente o nosso novo *chip* 244k. Todos os RNAs TINs tiveram sondas desenhadas para ambas as fitas senso e antisenso em relação ao gene codificador do *locus* ao qual está inserido. Já para os RNAs PIN, os transcritos exônicos aos quais eles faziam par senso/antisenso também tiveram suas sondas desenhadas. Como ainda havia espaço para sondas no *array*, selecionamos transcritos intergênicos sem evidência de *splicing* e desenhamos sondas para ambas as fitas (+) e (-) destes transcritos.

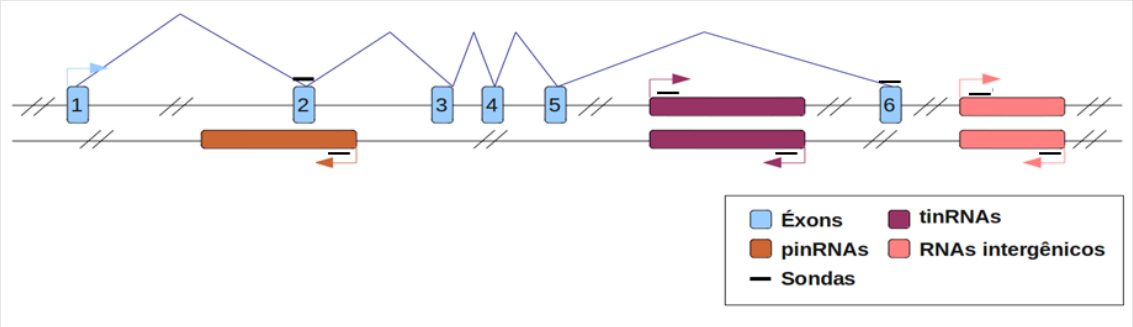


Figura 30 – Desenho esquemático da distribuição das sondas desenhadas em nosso novo *oligoarray* 244k intrônico-exônico-intergênico. O gene representado apresenta seis éxons. Além destas sondas, foram incluídas no *chip* todas as sondas comerciais exônicas da Agilent, que representam um trecho de um éxon dos genes codificadores de proteínas humanos conhecidos, como também as sondas intrônicas presentes no nosso *oligoarray* 44k e uma sonda representando um trecho de cada uma das duas fitas dos cDNAs presentes no *array* 4k usado anteriormente no grupo.

Tabela 9 – Distribuição das sondas desenhadas no novo *microarray* 244k intrônico-exônico-intergênico.

Tipo	Sondas
RNA PIN	10721
Exon sobreposto ao RNA PIN	11461
RNA TIN senso	67251
RNA TIN antisenso	67254
Intergênicos	3838 fita (+) / 3838 fita (-)
Exônicos da Agilent	43376
Exonicos *chip* 4k	2328 (exônicas) / 2328 (antisenso)
Não exônicos *chip* 4k fita (senso/antisenso)	1024 fita (+) / 1024 fita (-)
Intrônicos *chip* 44k (senso/antisenso)	30869
Exons do 44k não presentes no *dataset* Agilent	59

Dada a crescente busca para uma caracterização funcional de todos os grupos de lncRNAs (intrônicos e intergênicos), decidimos desenhar uma segunda plataforma 244k compreendendo não somente todos os lncRNAs intrônicos e mRNAs codificadores de proteínas, mas também todos aqueles lncRNAs intergênicos identificados a partir de ESTs presentes no dbEST. Desta forma, somos capazes de estimar quase toda a atividade transcricional humana baseando-se em ESTs públicas, sejam elas intrônicas, intergênicas ou exônicas.

Neste novo chip selecionamos apenas uma sonda representativa de cada um dos genes codificadores presentes na plataforma comercial da Agilent, resultando em um total de 18.698 sondas para mRNAs. Além disso, foi tentado acrescentar sondas nas duas fitas dos 32.657 RNAs intergênicos identificados a partir da comparação das coordenadas genômicas de todas as ESTs humanas em relação ao conjunto de genes de referência, resultando em 31.299 sondas para a fita mais (+) e 31.294 sondas para a fita menos (-). Desta forma, somos capazes de cobrir praticamente todos os *contigs* intergênicos do genoma, e não apenas os cerca de quatro mil selecionados randomicamente na construção anterior.

As duas plataformas contendo 244 mil sondas vêm sendo utilizadas em diferentes projetos em desenvolvimento no nosso grupo, que gerarão novas publicações científicas e contribuirão para uma maior caracterização funcional destes lncRNAs intrônicos, antisenso e intergênicos.

Capítulo 9

Considerações finais

9. CONSIDERAÇÕES FINAIS

Como pôde ser observado, esta tese de Doutorado constou de uma coleção de diversos estudos referentes a RNAs não codificadores de proteínas em organismos eucariotos. Os estudos envolveram o desenvolvimento de ferramentas, *pipelines*, identificação e caracterizações distintas de lncRNAs presentes em diferentes espécies. Esta variedade de trabalhos, envolvendo organismos distintos, foi importante para estabelecer alguns métodos para projetos iniciais de caracterização funcional, e também fornecer indícios sobre o comportamento destes transcritos nos mais variados organismos.

No trabalho, cumprindo o primeiro objetivo desta Tese, de **construir uma ferramenta web para catalogação das bases de dados disponíveis para as mais variadas classes de ncRNAs,** apresentamos o banco NRDR – *Non-coding RNA Databases Resource* –, um repositório *online* para bancos de dados de ncRNAs, que serve como um guia para pesquisadores encontrarem os principais *datasets* para utilizar em sua pesquisa. Também apresentamos o banco de dados IntromeDB, referente ao quarto objetivo proposto na Tese, relacionado à construção de uma **base dados pública para dados de expressão e caracterização de lncRNAs intrônicos, expressos em diferentes organismos eucariotos**, obtidos a partir dos mais variados tecidos e condições biológicas. Estas ferramentas são importantes catálogos com informações úteis para qualquer projeto genoma ou transcriptoma, disponibilizando bases confiáveis que axiliarão pesquisadores nos processos de anotação dos transcritos gerados em seus projetos.

Cumprindo com o segundo objetivo da tese, referente ao **desenvolvimento de *pipelines* para a identificação e anotação de RNAs intrônicos não codificadores de proteínas**, apresentamos diferentes abordagens para identificação de ncRNAs intrônicos, tanto a partir da "clusterização" de coordenadas de alinhamentos genômicos, como também a partir do uso de ferramentas de comparação de sequências, mais úteis para organismos não modelos, nos quais o genoma não apresenta-se totalmente completo, ou possua ainda alguns *gaps* na sua montagem final. Combinando os *pipelines* para identificação de ncRNAs com o *pipeline* apresentado para anotação de transcritos não codificadores, bem como os bancos de dados desenvolvidos (NRDR e

IntromeDB), esta Tese fornece ferramentas computacionais úteis para análises de bioinformática de RNAs para qualquer organismo.

Com a aplicação deste *pipelines*, fomos capazes de cumprir com o terceiro objetivo da Tese, de realizar uma **catalogação e atualização da atividade transcricional intrônica humana e de outros 21 organismos eucariotos**, levando em conta as ESTs públicas que possuiam alinhamento de coordenadas genômicas disponíveis em bancos de dados. Como também cumprimos com o sétimo objetivo, que consistiu de uma re-análise de todo o repertório de ESTs públicas do parasita humano protostomado *Schistosoma mansoni*, a partir de sequencias de ESTs disponíveis no NCBI. Desta forma, foi possível **identificar um grande repertório de potenciais novos genes e novos RNAs não codificadores intrônicos e intergênicos para este organismo não modelo**. Além disso, identificamos todo o repertório de ncRNAs intrônicos presentes nos *datasets* públicos referentes ao RNA-seq de 39 linhagens celulares humanas do projeto ENCODE.

A aplicação destes *pipelines* também nos permitiu analisar o perfil transcricional intrônico de alguns destes ncRNAs identificados no fígado humano, a partir da reanálise de dados públicos de expressão obtidos por técnicas como MPSS e *oligoarrays*. Este estudou foi capaz de identificar um catálogo completo daqueles transcritos intrônicos longos e sem evidência de *splicing*, que seriam precursores de transcritos pertencentes a classes já conhecidas de RNAs. No entanto, verificamos que a sua grande maioria pertencia a classes ainda não caracterizadas de RNAs, evidenciando o quanto ainda há para ser explorado e caracterizado no mundo de RNAs intrônicos de eucariotos.

Realizamos uma caracterização computacional de um conjunto de lncRNAs intrônicos e intergênicos identificados como expressos em tecidos pancreáticos normais ou neoplásicos humanos. Estas análises cumpriram com o quinto objetivo da Tese, relacionado ao **estudo da região promotora de ncRNAs intrônicos**, e parte do sexto objetivo, no que se refere ao **estudo da conservação de ncRNAs intrônicos**. Estes lncRNAs demonstraram estar sendo transcritos a partir de regiões genômicas enriquecidas com marcas de cromatina associadas a regiões promotoras e transcritos contendo cap de metil-guanosina em sua cauda 5'. Além disso, também apresentaram conservação de sequência entre espécies no *locus* onde são transcritos, como também conservação estrutural, sendo capazes de gerar estruturas secundárias conservadas

entre espécies e termodinamicamente estáveis. Evidências deste tipo nos fornecem indícios funcionais iniciais acerca da biologia destes transcritos. No entanto, com a exploração futura destes dados em conjunto com novos *datasets* funcionais, como os gerados em setembro de 2012 com as publicações finais do projeto ENCODE humano, e novos ensaios experimentais, seremos capazes de caracterizar diversos destes transcritos funcionalmente, elucidando todos os mecanismos moleculares envolvidos na sua regulação e atuação nos processos celulares.

Por fim, desenhamos dois novos *oligoarrays* 244k exônico-intrônico-intergênico humanos, contendo todos os *contigs* intrônicos e intergênicos identificados no genoma humano a partir das ESTs presentes no dbEST. Estas novas plataformas permitirão explorar simultaneamente o perfil transcricional de todos os genes codificadores de proteínas e lncRNAs intrônicos e intergênicos sem evidência de *splicing* humanos em qualquer que seja o tecido ou condição de interesse, auxiliando na caracterização funcional destes lncRNAs.

Todos estes estudos foram importantes para realizarmos uma maior caracterização acerca da biologia destes transcritos, obtendo informações fundamentais referentes à conservação e evolução, que seriam indícios funcionais a partir de uma pressão seletiva ao longo do tempo, como também obtivemos informações importantes ligadas a modulação de sua transcrição via mecanismos genéticos já conhecidos. A reunião destas informações em bancos públicos, como os aqui apresentados, juntamente com os *pipelines* desesenvolvidos, são de fundamental importância para podermos formular novas hipóteses a serem testadas referentes à biologia deste tipo molecular nas células eucarióticas.

REFERÊNCIAS

Almeida GT, Amaral MS, Beckedorff FC, Kitajima JP, DeMarco R, Verjovski-Almeida S (2012) Exploring the Schistosoma mansoni adult male transcriptome using RNA-seq. *Experimental parasitology* 132: 22-31

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* 39: D146-151

Amaral PP, Mattick JS (2008) Noncoding RNA in development. *Mamm Genome* 19: 454-492

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29

Backman TW, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, Givan SA, Carrington JC, Kasschau KD (2008) Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Res* 36: D982-985

Bandyopadhyay S, Bhattacharyya M (2010) PuTmiR: a database for extracting neighboring transcription factors of human microRNAs. *BMC Bioinformatics* 11: 190

Baratti MO, Moreira YB, Traina F, Costa FF, Verjovski-Almeida S, Olalla-Saad ST (2010) Identification of protein-coding and non-coding RNA expression profiles in CD34+ and in stromal cells in refractory anemia with ringed sideroblasts. *BMC medical genomics* 3: 30

Bateman A, Agrawal S, Birney E, Bruford EA, Bujnicki JM, Cochrane G, Cole JR, Dinger ME, Enright AJ, Gardner PP, Gautheret D, Griffiths-Jones S, Harrow J, Herrero J, Holmes IH, Huang HD, Kelly KA, Kersey P, Kozomara A, Lowe TM, Marz M, Moxon S, Pruitt KD, Samuelsson T, Stadler PF, Vilella AJ, Vogel JH, Williams KP, Wright MW, Zwieb C (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17: 1941-1946

Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* 22: 756-769

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* 74: 3171-3175

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM (2009) The genome of the blood fluke Schistosoma mansoni. *Nature* 460: 352-358

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007a) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S,

Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Program NCS, Baylor College of Medicine Human Genome Sequencing C, Washington University Genome Sequencing C, Broad I, Children's Hospital Oakland Research I, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007b) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14: 708-715

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". *Nature genetics* 4: 332-333

Brito GC, Fachel AA, Vettore AL, Vignal GM, Gimba ER, Campos FS, Barcinski MA, Verjovski-Almeida S, Reis EM (2008) Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. *Mol Carcinog* 47: 757-767

Brosnan CA, Voinnet O (2009) The long and the short of noncoding RNAs. *Curr Opin Cell Biol* 21: 416-425

Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, Zhao H, Liu Z, Liu C, Chen R, Zhao Y (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res* 40: D210-215

Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y, Gabellini D (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* 149: 819-831

Cao Y, Wu J, Liu Q, Zhao Y, Ying X, Cha L, Wang L, Li W (2010) sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA* 16: 2051-2057

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116: 499-509

Chen D, Farwell MA, Zhang B (2010) MicroRNA as a new player in the cell cycle. *J Cell Physiol* 225: 296-301

Chen LL, Carmichael GG (2010) Long noncoding RNAs in mammalian cells: what, where, and why? *Wiley Interdiscip Rev RNA* 1: 2-21

Chiromatzo AO, Oliveira TY, Pereira G, Costa AY, Montesco CA, Gras DE, Yosetake F, Vilar JB, Cervato M, Prado PR, Cardenas RG, Cerri R, Borges RL, Lemos RN, Alvarenga SM, Perallis VR, Pinheiro DG, Silva IT, Brandao RM, Cunha MA, Giuliatti S, Silva WA, Jr. (2007) miRNApath: a database of miRNAs, target genes and metabolic pathways. *Genet Mol Res* 6: 859-865

Cho S, Jun Y, Lee S, Choi HS, Jung S, Jang Y, Park C, Kim S, Kim W (2011) miRGator v2.0: an integrated system for functional investigation of microRNAs. *Nucleic Acids Res* 39: D158-162

Clote P, Ferre F, Kranakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* 11: 578-591

Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74

Copeland CS, Marz M, Rose D, Hertel J, Brindley PJ, Santana CB, Kehr S, Attolini CS, Stadler PF (2009) Homology-based annotation of non-coding RNAs in the genomes of Schistosoma mansoni and Schistosoma japonicum. *BMC genomics* 10: 464

Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322: 1845-1848

Crick F (1970) Central dogma of molecular biology. *Nature* 227: 561-563

Davis-Dusenbery BN, Hata A (2010) Mechanisms of control of microRNA biogenesis. *J Biochem* 148: 381-392

De Lucia F, Dean C (2011) Long non-coding RNAs and chromatin regulation. *Curr Opin Plant Biol* 14: 168-173

DeChiara TM, Brosius J (1987) Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content. *Proc Natl Acad Sci U S A* 84: 2624-2628

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4: P3

Dindot SV, Person R, Strivens M, Garcia R, Beaudet AL (2009) Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. *Genome Res* 19: 1374-1383

Dinger ME, Amaral PP, Mercer TR, Mattick JS (2009a) Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct Genomic Proteomic* 8: 407-423

Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS (2009b) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res* 37: D122-126

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR (2012) Landscape of transcription in human cells. *Nature* 489: 101-108

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP,

Haussler D, James Kent W (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 40: D918-923

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30: 207-210

Eswaran J, Cyanam D, Mudvari P, Reddy SD, Pakala SB, Nair SS, Florea L, Fuqua SA, Godbole S, Kumar R (2012) Transcriptomic landscape of breast cancers through mRNA sequencing. *Scientific reports* 2: 264

Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M (2004) CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* 24: 3804-3814

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175-185

Federhen S (2012) The NCBI Taxonomy database. *Nucleic acids research* 40: D136-143

Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* 20: 1470-1484

Francia S, Michelini F, Saxena A, Tang D, de Hoon M, Anelli V, Mione M, Carninci P, di Fagagna FD (2012) Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature*

Franco GR, Adams MD, Soares MB, Simpson AJ, Venter JC, Pena SD (1995) Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library. *Gene* 152: 141-147

Franco GR, Valadao AF, Azevedo V, Rabelo EM (2000) The Schistosoma gene discovery program: state of the art. *Int J Parasitol* 30: 453-463

Friard O, Re A, Taverna D, De Bortoli M, Cora D (2010) CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics* 11: 435

Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome biology* 8: R40

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196: 261-282

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res* 39: D141-145

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res* 37: D136-140

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17: 669-681

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330: 1775-1787

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15: 1451-1455

Gibb EA, Brown CJ, Lam WL (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer* 10: 38

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE (2011) The developmental transcriptome of Drosophila melanogaster. *Nature* 471: 473-479

Gruber AR, Neubock R, Hofacker IL, Washietl S (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res* 35: W335-338

Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G, Croft LJ, Taft RJ, Rizzi E, Askarian-Amiri M, Bonnal RJ, Callari M, Mignone F, Pesole G, Bertalot G, Bernardi LR, Albertini A, Lee C, Mattick JS, Zucchi I, De Bellis G (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* 10: 163

Guil S, Soler M, Portela A, Carrere J, Fonalleras E, Gomez A, Villanueva A, Esteller M (2012) Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nature structural & molecular biology* 19: 664-670

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464: 1071-1076

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458: 223-227

Halvardson J, Zaghlool A, Feuk L (2013) Exome RNA sequencing reveals rare and novel alternative transcripts. *Nucleic Acids Res* 41: e6

Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* 24: 158-164

Hertel J, Stadler PF (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* 22: e197-202

Hezel AF, Kimmelman AC, Stanger BZ, Bardeesy N, Depinho RA (2006) Genetics and biology of pancreatic ductal adenocarcinoma. *Genes Dev* 20: 1218-1249

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res* 39: D163-169

Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, Ju YW, Huang HD (2009) sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res* 37: D150-154

Huang J, Hao P, Zhang YL, Deng FX, Deng Q, Hong Y, Wang XW, Wang Y, Li TT, Zhang XG, Li YX, Yang PY, Wang HY, Han ZG (2007) Discovering multiple transcripts of human hepatocytes using massively parallel signature sequencing (MPSS). *BMC Genomics* 8: 207

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9: 868-877

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142: 409-419

Huarte M, Rinn JL (2010) Large non-coding RNAs: missing links in cancer? *Hum Mol Genet* 19: R152-161

Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*: 138-148

Janowski BA, Younger ST, Hardy DB, Ram R, Huffman KE, Corey DR (2007) Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat Chem Biol* 3: 166-173

Jenuth JP (2000) The NCBI. Publicly available tools and resources on the Web. *Methods in molecular biology* 132: 301-312

Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Muller-Tidow C (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22: 8031-8041

Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V (2007) CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res* 35: D829-833

Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 21: 93-102

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 14: 331-342

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296: 916-919

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316: 1484-1488

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 15: 987-997

Keeling PJ, Slamovits CH (2005) Causes and effects of nuclear genome reduction. *Curr Opin Genet Dev* 15: 601-608

Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K (2005) Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res* 15: 463-474

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P (2006) CAGE: cap analysis of gene expression. *Nat Methods* 3: 211-222

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* 35: W345-349

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39: D152-157

Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* 23: 158-161

Kuhn RM, Haussler D, Kent WJ (2012) The UCSC genome browser and associated tools. *Briefings in bioinformatics*

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett

N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921

Lee AP, Yang Y, Brenner S, Venkatesh B (2007) TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. *BMC Genomics* 8: 441

Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 34: D158-162

Li JT, Zhang Y, Kong L, Liu QR, Wei L (2008) Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. *Nucleic Acids Res* 36: 4833-4844

Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27: i275-282

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH (2012) Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis. *The Plant cell* 24: 4333-4345

Louro R, El-Jundi T, Nakaya HI, Reis EM, Verjovski-Almeida S (2008) Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics* 92: 18-25

Louro R, Nakaya HI, Amaral PP, Festa F, Sogayar MC, da Silva AM, Verjovski-Almeida S, Reis EM (2007) Androgen responsive intronic non-coding RNAs. *BMC biology* 5: 4

Louro R, Smirnova AS, Verjovski-Almeida S (2009) Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* 93: 291-298

Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 39: D52-57

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011: bar009

Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 445: 666-670

Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet* 5: 316-323

Mattick JS (2009) The genetic signatures of noncoding RNAs. *PLoS Genet* 5: e1000459

Mattick JS, Makunin IV (2006) Non-coding RNA. *Hum Mol Genet* 15 Spec No 1: R17-29

Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet* 10: 155-159

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* 30: 99-104

Merrick JM, Osman A, Tsai J, Quackenbush J, LoVerde PT, Lee NH (2003) The Schistosoma mansoni gene index: gene discovery and biology by reconstruction and analysis of expressed gene sequences. *J Parasitol* 89: 261-269

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553-560

Mohammad F, Pandey GK, Mondal T, Enroth S, Redrup L, Gyllensten U, Kanduri C (2012) Long noncoding RNA-mediated maintenance of DNA methylation and transcriptional gene silencing. *Development* 139: 2792-2803

Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature* 447: 1126-1129

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621-628

Muslimov IA, Lin Y, Heller M, Brosius J, Zakeri Z, Tiedge H (2002) A small RNA in testis and brain: implications for male germ cell development. *J Cell Sci* 115: 1243-1250

Nag A, Jack T (2010) Sculpting the flower; the role of microRNAs in flower development. *Curr Top Dev Biol* 91: 349-378

Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE (2002) NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* 30: 395-397

Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reis EM, Verjovski-Almeida S (2007) Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome biology* 8: R43

Nicolas FE, Lopez-Martinez AF (2010) MicroRNAs in human diseases. *Recent Pat DNA Gene Seq* 4: 142-154

Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, Hayashizaki Y, Tomita M (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res* 13: 1301-1306

Oliveira KC, Carvalho ML, Maracaja-Coutinho V, Kitajima JP, Verjovski-Almeida S (2011) Non-coding RNAs in schistosomes: an unexplored world. *Anais da Academia Brasileira de Ciencias* 83: 673-694

Oliver HF, Orsi RH, Ponnala L, Keich U, Wang W, Sun Q, Cartinhour SW, Filiatrault MJ, Wiedmann M, Boor KJ (2009) Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics* 10: 641

Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* 22: 1-5

Paschoal AR, Maracaja-Coutinho V, Setubal JC, Simoes ZL, Verjovski-Almeida S, Durham AM (2012) Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA biology* 9: 274-282

Perez DS, Hoage TR, Pritchett JR, Ducharme-Smith AL, Halling ML, Ganapathiraju SC, Streng PS, Smith DI (2008) Long, abundantly expressed non-coding transcripts are altered in cancer. *Hum Mol Genet* 17: 642-655

Piekna-Przybylska D, Decatur WA, Fournier MJ (2007) New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* 13: 305-312

Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B (2002) Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* 36: 233-278

Prensner JR, Chinnaiyan AM (2011) The emergence of lncRNAs in cancer biology. *Cancer Discov* 1: 391-407

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 29: 742-749

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. *Nucleic Acids Res* 40: D290-301

Qu Z, Adelson DL (2012) Evolutionary conservation and functional roles of ncRNA. *Frontiers in genetics* 3: 205

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841-842

Reiche K, Stadler PF (2007) RNAstrand: reading direction of structured RNAs in multiple sequence alignments. *Algorithms Mol Biol* 2: 6

Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, Egidio CM, Paquola AC, Machado AA, Festa F, Yamamoto D, Alvarenga R, da Silva CC, Brito GC, Simon SD, Moreira-Filho CA, Leite KR, Camara-Lopes LH, Campos FS, Gimba E, Vignal GM, El-Dorry H, Sogayar MC, Barcinski MA, da Silva AM, Verjovski-Almeida S (2004) Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23: 6684-6692

Reis EM, Ojopi EP, Alberto FL, Rahal P, Tsukumo F, Mancini UM, Guimaraes GS, Thompson GM, Camacho C, Miracca E, Carvalho AL, Machado AA, Paquola AC, Cerutti JM, da Silva AM, Pereira GG, Valentini SR, Nagai MA, Kowalski LP, Verjovski-Almeida S, Tajara EH, Dias-Neto E, Bengtson MH, Canevari RA, Carazzolle MF, Colin C, Costa FF, Costa MC, Estecio MR, Esteves LI, Federico MH, Guimaraes PE, Hackel C, Kimura ET, Leoni SG, Maciel RM, Maistro S, Mangone FR, Massirer KB, Matsuo SE, Nobrega FG, Nobrega MP, Nunes DN, Nunes F, Pandolfi JR, Pardini MI, Pasini FS, Peres T, Rainho CA, dos Reis PP, Rodrigus-Lisoni FC, Rogatto SR, dos Santos A, dos Santos PC, Sogayar MC, Zanelli CF (2005) Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. *Cancer Res* 65: 1693-1699

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M (2003) The transcriptional activity of human Chromosome 22. *Genes Dev* 17: 529-540

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129: 1311-1323

Rose D, Hackermuller J, Washietl S, Reiche K, Hertel J, Findeiss S, Stadler PF, Prohaska SJ (2007) Computational RNomics of drosophilids. *BMC Genomics* 8: 406

Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, Diekhans M, Fujita PA, Goldman M, Gravell RC, Harte RA, Hinrichs AS, Kirkup VM, Kuhn RM, Learned K, Maddren M, Meyer LR, Pohl A, Rhead B, Wong MC, Zweig AS, Haussler D, Kent WJ (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res* 40: D912-917

Schistosoma japonicum Genome S, Functional Analysis C (2009) The Schistosoma japonicum genome reveals features of host-parasite interplay. *Nature* 460: 345-351

Schmeier S, Schaefer U, MacPherson CR, Bajic VB (2011) dPORE-miRNA: polymorphic regulation of microRNA genes. *PLoS One* 6: e16657

Schwartz JC, Younger ST, Nguyen NB, Hardy DB, Monia BP, Corey DR, Janowski BA (2008) Antisense transcripts are targets for activating small RNAs. *Nat Struct Mol Biol* 15: 842-848

Secco M, Moreira YB, Zucconi E, Vieira NM, Jazedje T, Muotri AR, Okamoto OK, Verjovski-Almeida S, Zatz M (2009) Gene expression profile of mesenchymal stem cells from paired umbilical cord units: cord is different from blood. *Stem cell reviews* 5: 387-401

Seemann SE, Gilchrist MJ, Hofacker IL, Stadler PF, Gorodkin J (2007) Detection of RNA structures in porcine EST data and related mammals. *BMC Genomics* 8: 316

Seidl CI, Stricker SH, Barlow DP (2006) The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J* 25: 3565-3575

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034-1050

Silahtaroglu A, Stenvang J (2010) MicroRNAs, epigenetics and disease. *Essays Biochem* 48: 165-185

Sone M, Hayashi T, Tarui H, Agata K, Takeichi M, Nakagawa S (2007) The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J Cell Sci* 120: 2498-2506

St Laurent G, 3rd, Faghihi MA, Wahlestedt C (2009) Non-coding RNA transcripts: sensors of neuronal stress, modulators of synaptic plasticity, and agents of change in the onset of Alzheimer's disease. *Neurosci Lett* 466: 81-88

Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J (2006) Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *Lancet Infect Dis* 6: 411-425

Szymanski M, Erdmann VA, Barciszewski J (2007) Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res* 35: D162-164

Tacutu R, Budovsky A, Fraifeld VE (2010) The NetAge database: a compendium of networks for longevity, age-related diseases and associated processes. *Biogerontology* 11: 513-522

Tahira AC, Kubrusly MS, Faria MF, Dazzani B, Fonseca RS, Maracaja-Coutinho V, Verjovski-Almeida S, Machado MC, Reis EM (2011) Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Molecular cancer* 10: 141

Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature protocols* 7: 542-561

Tamura M, Hendrix DK, Klosterman PS, Schimmelman NR, Brenner SE, Holbrook SR (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res* 32: D182-184

Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* 16: 885-889

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* 39: 925-938

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329: 689-693

Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* 38: D280-282

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. *PLoS Biol* 8: e1000371

Verjovski-Almeida S, DeMarco R, Martins EA, Guimaraes PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY, Jr., Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sa RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Wilson RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E (2003) Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. *Nature genetics* 35: 148-157

Verjovski-Almeida S, Leite LC, Dias-Neto E, Menck CF, Wilson RA (2004) Schistosome transcriptome: insights and perspectives for functional genomics. *Trends Parasitol* 20: 304-308

Wang H, Iacoangeli A, Popp S, Muslimov IA, Imataka H, Sonenberg N, Lomakin IB, Tiedge H (2002) Dendritic BC1 RNA: functional role in regulation of translation initiation. *J Neurosci* 22: 10232-10241

Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Mol Cell* 43: 904-914

Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14: 1012-1017

Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. *Trends Cell Biol* 21: 354-361

Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* 342: 19-30

Washietl S, Hofacker IL (2007) Identifying structural noncoding RNAs using RNAz. *Curr Protoc Bioinformatics* Chapter 12: Unit 12 17

Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005a) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 23: 1383-1390

Washietl S, Hofacker IL, Stadler PF (2005b) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 102: 2454-2459

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239-1243

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* 3: e65

Williamson (1977) DNA insertions and gene structure. *Nature* 270: 295-297

Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* 23: 1494-1504

Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, McEwen GK, Elgar G (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev Biol* 7: 100

Wright MW, Bruford EA (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum Genomics* 5: 90-98

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B (2006a) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34: D187-191

Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbo G, Chen L, Lu H, Zhao Y, Chen R (2006b) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res* 34: D150-152

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 37: D105-110

Xin Y, Olson WK (2009) BPS: a database of RNA base-pair structures. *Nucleic Acids Res* 37: D83-88

Yan MD, Hong CC, Lai GM, Cheng AL, Lin YW, Chuang SE (2005) Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum Mol Genet* 14: 1465-1474

Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH (2011a) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res* 39: D202-209

Yang JH, Qu LH (2012) DeepBase: annotation and discovery of microRNAs and other noncoding RNAs from deep-sequencing data. *Methods Mol Biol* 822: 233-248

Yang JH, Shao P, Zhou H, Chen YQ, Qu LH (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res* 38: D123-130

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL (2011b) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* 12: R16

Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol* 11: R87

Yin QF, Yang L, Zhang Y, Xiang JF, Wu YW, Carmichael GG, Chen LL (2012) Long noncoding RNAs with snoRNA ends. *Mol Cell* 48: 219-230

Yin Y, Zhao Y, Wang J, Liu C, Chen S, Chen R, Zhao H (2007) antiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics* 8: 319

Yokoyama Y, Nimura Y, Nagino M (2009) Advances in the treatment of pancreatic cancer: limitations of surgery and evaluation of new therapeutic strategies. *Surg Today* 39: 466-475

Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, Gong Y, Cheng H, Laxman B, Vellaichamy A, Shankar S, Li Y, Dhanasekaran SM, Morey R, Barrette T, Lonigro RJ, Tomlins SA, Varambally S, Qin ZS, Chinnaiyan AM (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* 17: 443-454

Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui H (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451: 202-206

Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH (2010) ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. *RNA* 16: 1889-1901

Zhang Y, Li J, Kong L, Gao G, Liu QR, Wei L (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res* 35: D156-161

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7: 203-214

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* 40: 939-953

ANEXOS

Anexo 1: *Paschoal A.R.; ***Maracaja-Coutinho V.**; Setubal J.C.; Simões Z.L.; Verjovski-Almeida S.; Durham A.M. (2012). *Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases*. **RNA Biology** 9(3). ***Ambos autores contribuíram igualmente para o trabalho**.

Anexo 2: Tahira A.C.; Kubrusly M.S.; Faria M.F.; Dazzani B.; Fonseca R.S.; **Maracaja-Coutinho V.**; Verjovski-Almeida S.; Machado M.C.C.; Reis E.M. (2011). *Long intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer*. **Molecular Cancer** 10:141.

Anexo 3: Oliveira K.C.; Carvalho M.L.P.; **Maracaja-Coutinho V.**; Kitajima J.P.; Verjovski-Almeida S. (2010). *Non-coding RNAs in Schistosomes: an unexplored world*. **An. Acad. Bras. Cienc.** 83(2) 673-694.

Anexo 4: Histórico acadêmico.

Anexo 6: Súmula curricular.

RNA Biology 9:3, 1–9; March 2012; © 2012 Landes Bioscience

Non-coding transcription characterization and annotation

A guide and web resource for non-coding RNA databases

Alexandre Rossi Paschoal,[3,5,†] Vinicius Maracaja-Coutinho,[2,5,†] João Carlos Setubal,[2] Zilá Luz Paulino Simões,[4] Sergio Verjovski-Almeida[2] and Alan Mitchell Durham[1,*]

[1]Departamento de Ciência da Computação; Instituto de Matemática e Estatística; Universidade de São Paulo; São Paulo, Brazil; [2]Departamento de Bioquímica; Instituto de Química; Universidade de São Paulo; São Paulo, Brazil; [3]Engenharia da Computação; Universidade Tecnológica Federal do Paraná – Campus Cornélio Procópio; Cornélio Procópio, Brazil; [4]Departamento de Biologia; Faculdade de Filosofia Ciências e Letras de Ribeirão Preto; Universidade de São Paulo; São Paulo, Brazil; [5]Programa Interunidades em Bioinformática; Instituto de Matemática e Estatística; Universidade de São Paulo; São Paulo, Brazil

[†]Both authors contributed equally to this work.

Keywords: ncRNA, non-coding RNA, database, transcription, bioinformatics, web tool

Abbreviations: ncRNA, non-coding RNA; lincRNA, long intergenic non-coding RNA; lncRNA, long non-coding RNA; nt, nucleotide; HGNC, The HUGO Gene Nomenclature Committee; CNE, conserved non-coding elements; NAT, natural antisense transcript; tRNA, transfer RNA; rRNA, ribosomal RNA; mRNA, messenger RNA; miRNA, microRNA; snoRNA, small nucleolar RNA; siRNA, small interfering RNA; SRP RNA, signal recognition particle RNA; piRNA, piwi-interacting RNA; TERC, telomerase RNA component

Large-scale transcriptome projects have shown that the number of RNA transcripts not coding for proteins (non-coding RNAs) is much larger than previously recognized. High-throughput technologies, coupled with bioinformatics approaches, have produced increasing amounts of data, highlighting the role of non-coding RNAs (ncRNAs) in biological processes. Data generated by these studies include diverse non-coding RNA classes from organisms of different kingdoms, which were obtained using different experimental and computational assays. This has led to a rapid increase of specialized RNA databases. The fast growth in the number of available databases makes integration of stored information a difficult task. We present here NRDR, a Non-coding RNA Databases Resource for information retrieval on ncRNA databases (www.ncrnadatabases.org). We performed a survey of 102 public databases on ncRNAs and we have introduced four categorizations to classify these databases and to help researchers quickly search and find the information they need: RNA family, information source, information content and available search mechanisms. NRDR is a useful databases searching tool that will facilitate research on ncRNAs.

Introduction

High-throughput technologies such as tiling arrays or deep sequencing,[1-4] combined with a large number of bioinformatics approaches,[5-8] have produced increasing amounts of expression data and functional information related to the characterization of non-coding RNAs (ncRNAs). Presently, there are over 100 public databases with ncRNA information, which cover a wide range of ncRNA families. These databases have information on secondary structure of RNAs, on tissues where they are expressed, on expression conditions, on specific species, on functional annotation, phylogeny, taxonomy and alignment analysis, on putative target genes, on related diseases, and other information. These databases can be searched using a wide range of different methods such as sequence similarity search, keyword or tag search, manual browsing of tables, genomic alignment coordinates and genomic location in relation to a given locus with clustered transcripts.

Given this multitude of databases it is clear that some sort of metaresource, providing a single point-of-entry for search and classification of the information, is needed. In this work we present the *Non-coding RNA Databases Resource* (NRDR), which is a web portal that indexes currently available public databases with non-coding RNA information in a user-friendly way. NRDR allows users to quickly find databases using criteria based on RNA family, information source, information content and other available search mechanisms.

The importance of a publicly available portal for searching information on ncRNAs has been highlighted by a recent publication.[9] It must be noted that Bateman et al.[9] argue for a central repository of ncRNA information and calls for the formation of a consortium with this goal. NRDR is not a central repository, since it does not store any ncRNA information. In addition, NRDR does not index databases that are specific for tRNAs and rRNAs. Nevertheless, by currently indexing 102 ncRNA databases, we believe that NRDR is a useful tool that will facilitate research on ncRNAs.

NRDR overview. NRDR can be accessed at www.ncrnadatabases.org. The website includes a description of each

*Correspondence to: Alan Mitchell Durham; Email: aland@usp.br
Submitted: 11/22/11; Revised: 01/11/12; Accepted: 01/13/12
http://dx.doi.org/10.4161/rna.9.3.19352

indexed database and a link to the respective website. The complete list and full description for each database can also be found on the **Supplementary Material**. NRDR is updated every six months, driven by manual curation of literature.

NRDR provides a search interface where users can retrieve a list of databases filtered by user-defined criteria. Free-text boxes are available for a combined search where multiple keywords can be applied. As an example, the user can filter the data using an RNA class name, database name, organism name or PubMed ID (e.g., microRNA, *Homo sapiens*, ID number). Additionally, the user can select only databases that have data sets available for downloading, or those with a Graphic Genome View available (e.g., Genome Browser). For a more comprehensive search, there is a provision for the user to search directly by the keywords that appear in the text description of each database (e.g., "intergenic," "transcription," "networks"). Finally, it is also possible to explore the NRDR data by browsing it according to the organism of interest. **Figure 1** shows the NRDR search page and an example of a database description retrieved from our server.

Databases indexed by NRDR were manually classified using four criteria:

1. RNA families: what classes of ncRNAs are present in a database?
2. Information source: what is the provenance of the ncRNA information? (from experimental evidence; from computational analysis only; from manual curation of information obtained experimentally and/or computationally; and from literature).
3. Information content: what are the various types of information stored in a database (e.g., sequence, annotation, expression)?
4. Search mechanisms: What search mechanisms are available in a database?

NRDR automatically generates statistics on how the databases are distributed according to each classification criterion. These statistics are available on the NRDR web resource (see *Statistics* link) and was used below to provide an overview of current work on ncRNA research.

RNA families: different classes of transcripts are present in databases. A challenge in the annotation of novel ncRNAs is to categorize them into well-defined classes. Currently, most of the specific ncRNA families are defined by the structural molecular folding,[10,11] assuming that transcripts with a common structural conformation act in a similar functional manner in cellular processes. Sequence similarity is a very important tool to classify new ncRNAs, but limits the effectiveness of characterizing more divergent sequences.[11]

Reflecting the limited amount of specific functional data, the complex sets of short and long ncRNAs being detected by high-throughput expression technologies have been simply annotated in databases as "small RNAs"[12,13], including all RNA sequences shorter than 200 nt, or "long RNAs"[14] (> 200 nt), with little information about which functional class or family they belong.

Recently, these difficulties prompted a standardization proposed by the HUGO Gene Nomenclature Committee (HGNC), the sole organization authorized to assign nomenclature to human transcripts.[15] They classified these long transcripts as a large class named long non-coding RNAs (lncRNAs), which is subdivided into two main classes: (1) transcripts with known function (e.g., XIST or HOTAIR), and (2) those with unknown functional annotation, which are sub-classified according to their genomic location, such as: intronic (IT), antisense (AS), intergenic (LINC) or host transcripts of other small ncRNAs.[15] Under the present

Figure 1. Overview of our Non-coding RNA Databases Resource (NRDR) web tool. (A) Initial screen showing the different types of search that a user can perform. (B) Example of a database description, from one of the databases that were listed in the output results, which satisfied the search criteria shown in (A). The actual output results screen from the search (not shown) is a list of databases with links to their respective descriptions, available in NRDR.

categorization, termed as an ongoing project by HGNC, these lncRNAs were described as spliced, capped and polyadenylated RNAs,[15] which clearly does not encompass all different lncRNAs that may be unspliced and/or non-polyadenylated.[16-18]

Due to all the difficulties in categorizing RNAs, in NRDR we chose to group them according to the different classes of RNAs used in the databases themselves, combined whenever possible with HGNC nomenclature, as exemplified below. Databases that are not specific for a single RNA class, such as RFAM,[10] ncRNAdb,[19] and NONCODE[20] were classified as containing *multiple classes*. Apart from those ncRNAs categorized by HGNC, other classes of ncRNAs are present in databases, such as: Conserved Non-coding Elements (CNEs), some of which are associated with well-conserved non-coding transcripts[21,22]; structured RNAs that were usually predicted by molecular structural folding algorithms applied to transcriptomic or genomic sequences[23]; *mRNA-like RNAs*, long spliced non-coding transcripts[24,25]; and Natural Antisense Transcripts (NATs), long transcripts that are partially complementary to other endogenous RNAs.[26-28]

Figure 2 presents the number of published databases related to ncRNAs over the years. It shows a substantial increase in the number of repositories available after 2005. The 102 surveyed databases were categorized by the classes of ncRNAs they contain (**Table 1**, see the Statistics page of the web resource). The distribution of databases per ncRNA class is presented in **Figure 3**. Some particularities were observed among the 102 databases. The majority of databases (73 of 102) host small RNAs, with 68% of them (50 of 73) related specifically to microRNAs. In 2010, Kozomara and Griffiths-Jones highlighted that in the previous three years the number of miRNA sequences in the miRBase database had almost tripled.[29] This reflects the high interest of the research community in microRNAs. This interest is not new, and is due to the role these molecules play in regulating gene translation, and also probably because microRNAs have been correlated with different pathologies, diseases and with the development process.[30-32]

Although there are in the literature some well-known functionally annotated long ncRNAs, such as *XIST*,[33] *HOTAIR*[34] and *MALAT1*,[35] their presence in databases is still scarce when compared with small RNAs (there are only 5 databases specific to long ncRNAs[14,26-28,36]). Currently there are eight databases with data originated from next generation sequencing technologies,[12,13,24,37-40] however only one of them (ncRNAimprint[24]) is related to long RNAs (e.g., antisense long RNAs and mRNA-like RNAs). Increasing attention has been given in the past five years to long ncRNAs, which reflects the intensive use of large-scale technologies for expression measurements[1,16,41] and the recognition that altered expression of long ncRNAs is present in many diseases such as different types of human cancer.[42-44] Nevertheless, only a very small fraction of such functional data are present in ncRNA databases. For example, transcription has been analyzed by ultra-dense tiling arrays covering the entire *HOX* gene loci, and *HOTAIR* has been identified along with another 170 long ncRNAs as being differentially expressed among normal human breast epithelia, primary breast carcinomas, and distant metastases.[34] Except for *HOTAIR* itself, none of the dozens of long ncRNAs differentially expressed in the *HOX* loci and related to breast cancer[34] are present for example in the lncRNAdb[36] or in the NONCODE database, a large repository of ncRNAs which has been recently updated to include an integrative annotation of long non-coding RNAs.[20] Similarly, microarray analysis revealed expression profiles comprised of dozens of long ncRNAs correlated with a number of different types of cancer.[45-49] Again,

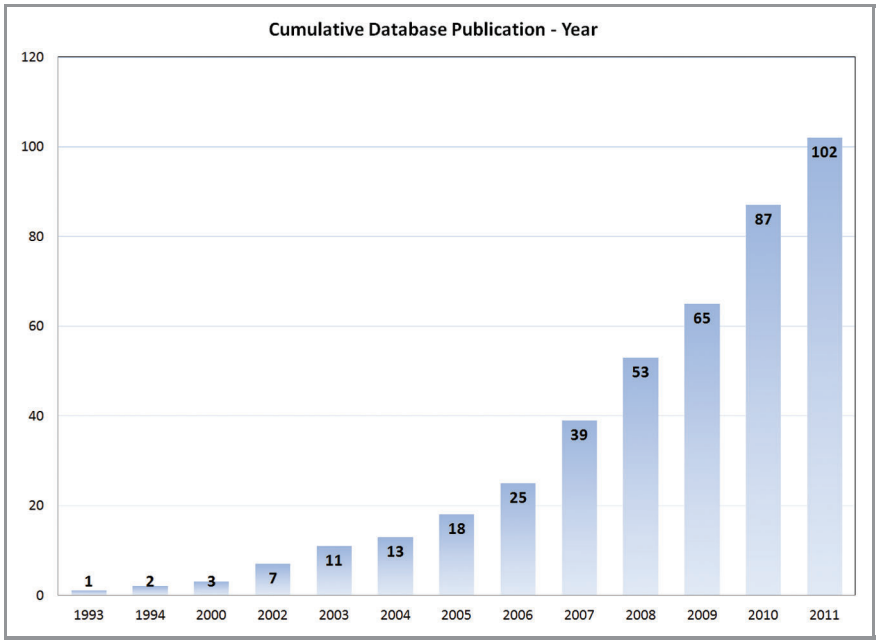


Figure 2. Publication of ncRNA databases over the years. Values for 2011 were computed until November of that year.

Table 1. Different RNA classes listed in the databases. Whenever possible we tried to associate the ncRNAs described on databases with HGCN classification

RNA class	Description
Small RNAs	
snoRNAs	small nucleolar RNAs (e.g., CD and H/ACA box)
small RNAs	small RNAs (< 200 nucleotides) that can include classes not yet classified by HGNC
siRNAs	small interfering RNAs
miRNAs	microRNAs
SRP RNAs	Signal Recognition Particle RNA
piRNAs	piwi-interacting RNAs
Ribozymes	Rnase P and group I, II and III intronic RNAs removed in the splicing process or involved in RNA catalyzes
TERC	Telomerase RNA Component
Long RNAs	
long ncRNAs	Long RNAs (> 200 nucleotides), that can include classes not yet classified by HGNC
NATs	Natural Antisense Transcripts, complementary to other RNAs
Other	
Multiple classes	Not specific to a unique class of RNA, which means there is a diversity of RNA families
CNE	Conserved Non-coding Elements
Structured RNAs	Specific to Structured RNAs (e.g., Secondary Structure or 3D Structure)

none of these ncRNAs are in the databases, possibly reflecting the lack of uniform and systematic annotation nomenclature for the diverse and extensive amount of long ncRNAs expressed in human tissues. In another example, a recent paper using targeted RNA sequencing for an in-depth analysis of the human transcriptome reveals multiple additional isoforms of coding and non-coding genes;[50] similar results had been previously shown by tiling arrays.[51] Also, very long ncRNAs abundantly transcribed in intergenic genomic regions have been identified recently, and expression of these regions was shown to be associated with neoplastic transformation.[52] Again, none of these data are present in any database. Collectively, these results reveal that the range, depth and complexity of the human transcriptome is far from fully characterized,[50] and at the same time they point to a

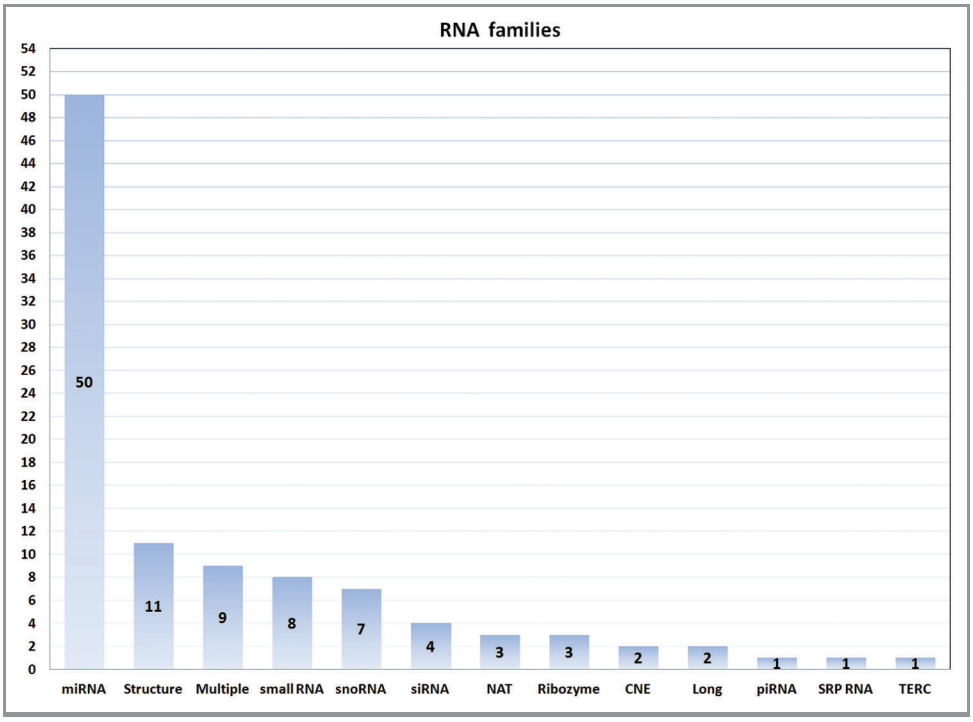


Figure 3. Database distribution per RNA class.

considerable limitation regarding the completeness of any database related to long ncRNAs.

Information source: the reliability of stored data. Information stored in these databases comes from different sources. We associate databases to four types of information source: (1) in silico annotation, where ncRNAs are characterized locally by computational analyses, (2) literature, where data are extracted from published articles, (3) manual curation, where information is locally validated by a human expert, and (4) experimental information, derived directly from biological assays. Of these, experimental information is generally the most reliable source. **Figure 4** shows the distribution of databases per information source. This information can be easily retrieved on our web tool.

Experimental data on ncRNAs is present in 67 of the 102 databases. This information could be used as a validation catalog for laboratory experiments, as well as for testing and developing new approaches in computational biology. However, just 40 of these make the data available for downloading (see Supplementary Tables for a complete list, or the Statistics page of the web resource), of which only 24 offer the data directly in FASTA format or similar sequence file. NRDR can quickly guide the investigator to the databases providing a given downloadable format of interest (e.g., BED, GFF or PSL).

Information content: information stored in public domain databases. Information content describes the various types of information stored in the databases, such as sequence, annotation, expression. This is not always clearly stated in the database name and sometimes not even in the homepage; for this reason, each database was manually examined and a number of terms (usually from 9 to 19 terms) were associated to that database, out of a list with 257 terms that we have generated from all terms present in the databases. The Statistics link at the NRDR portal contains the list of all generated terms along with the number of databases that were associated to each given term; in addition, upon accessing the NRDR description of each database, one can see the list of all terms that have been associated to that database under the information content category. In the Search link of the portal the databases can be searched by information content using multiple combinations of terms; in addition, a search for sequence under information content can be combined with a filter that discriminates the databases according to the sequence formats available for download (FASTA, GFF, BED, PSL, etc.). In **Figure 5**, we grouped and ranked the top 10 most frequent terms associated to the databases.

Browsing the information stored in databases, one can get a general idea of the results of ncRNA research that are finding their way into public databases. There is a large number of databases containing information related to expression (38 out of 102), in comparison to the scarce information related to interaction of these non-coding transcripts with other biomolecules (DNA, RNA, or protein). There are only nine databases with information on molecular interactions.[53-61] Most of these annotations are related to interaction between ncRNA and their target transcripts in an RNA:RNA relationship; little information is available covering RNA:protein or RNA:DNA relationships. Discovering new functional mechanisms of these molecules is a big challenge, which is crucial for understanding the functionality of these ubiquitous transcripts in eukaryotic genomes. Despite evidence pointing to the functionality of ncRNAs, such as: (1) tissue specificity;[62] (2) involvement in disease;[43,45] (3) development;[63] (4) genomic imprinting;[64] (5) alternative splicing regulation;[65] there is still limited evidence of their mechanisms of action, and

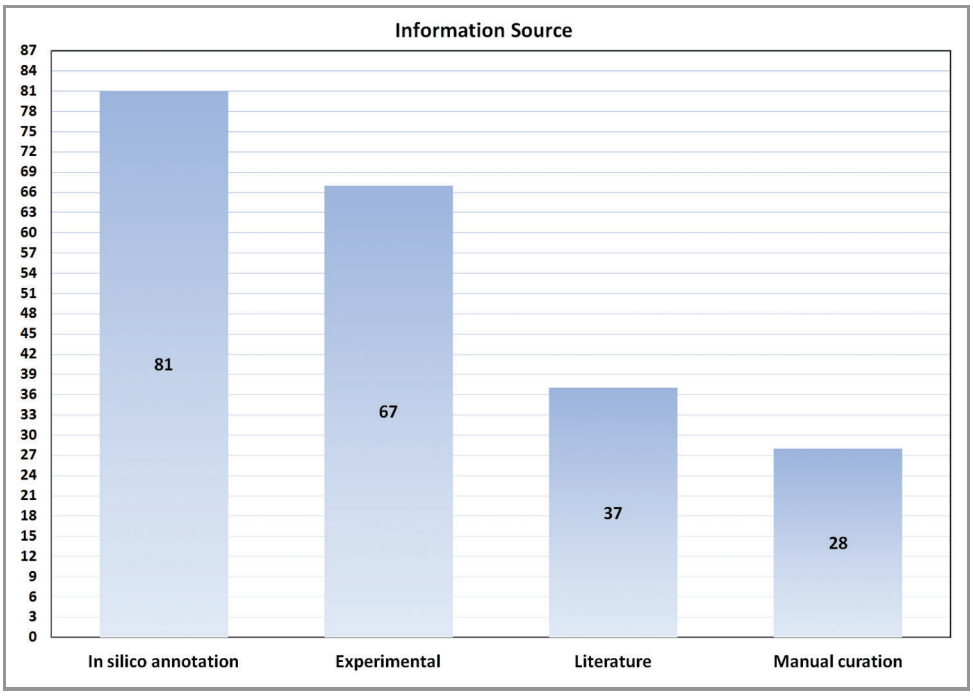


Figure 4. Database distribution per information source.

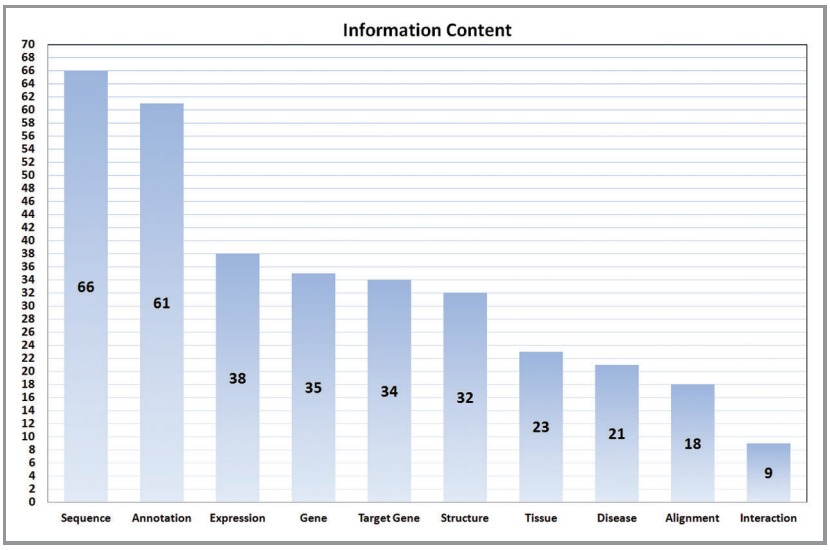


Figure 5. Top ten ranking of information content terms based on the information stored in the databases.

besides the miRNAs, only a few other ncRNAs are annotated in the databases as functionally characterized.

Examples of transcripts with known mechanisms of action that are present in databases such as lncRNAdb[36] and NONCODE[20] include: (1) *HOTAIR*, a long intergenic ncRNA located within the *HOXC* gene cluster that regulates the chromatin state in the *HOXD* locus after binding to PRC2 Polycomb repressor complex[66] and to LSD1, a histone lysine demethylase,[67] leading to gene-silencing of the *HOXD* gene cluster; (2) *p21*, a p53-responsive long intergenic transcript, that binds to hnRNP-K protein and mediates global gene repression and apoptosis in the p53 pathway;[68] and (3) *BC1* (Brain Cytoplasmic RNA 1), a neuronal small ncRNA that interacts with eukaryotic initiation factor 4 (eIF4), resulting in a subsequent translational repression.[69] These transcripts and a few others serve as paradigms of molecular mechanisms of ncRNAs[70] acting in the cellular processes fine-tuning regulation. Biological knowledge acquired with the characterization of these ncRNAs could guide the development of new experimental and computational approaches to characterize the immense repertoire of small and long ncRNAs without well-defined biological mechanisms of action, which should feed the ncRNA databases in the coming years.

Recently, a number of databases involving systems biology approaches have been described.[71-78] These databases, which are indexed by NRDR, integrate data sets of regulation (e.g., transcription factor binding sites, or epigenetic markers), expression, target genes and putative regulated networks and pathways. However, there is still a lack of data integrating these transcripts with proteins.

Search mechanisms: extracting information from databases. Information of interest from ncRNA databases can be located using different search mechanisms. We have grouped the search mechanisms into six categories: similarity, keyword, tag, genomic location, tabular and density of ncRNAs (**Fig. 6**). In some databases these options can be used together. Depending on the prior knowledge one has about the RNA of interest (e.g., nucleotide sequence, gene locus name in which it is inserted or associated disease), it is important to choose the databases with appropriate search mechanisms that can provide the necessary information retrieval capabilities.

Of all the search mechanisms, the keyword and tag options are the most common ways to search information in ncRNA databases (**Fig. 6**), being also the simplest ones. The keyword option consists of a text box that accepts free text (e.g., accession code, gene name, Ensembl accession, GO accession etc.). The tag option consists of a pre-defined list of terms (e.g., tissue or cell line names, organism, clade) usually in a pull-down menu.

Similarity search is the third most common mechanism, available in 32 databases (**Fig. 6**). The BLAST program[79] is used for this purpose in the majority of the databases. Similarity search alone has the drawback that ncRNA families can be very divergent in primary sequence, conserving only their secondary structure.[80] Only the Rfam database[10] considers structural information in the search, and even so it is still associated with a pre-processing by a low stringency similarity search.

Only 14 databases offer search by genomic coordinates (**Fig. 6**), where ncRNAs are retrieved based on intervals of positions in the genome. This search helps the examination of specific regions of interest in a genome, and is extremely useful to explore different pre-aligned data sets of interest in the vicinities of a given ncRNA of interest. An example of the use of such a query is the investigation of putative regulatory signals (e.g., CpG islands, transcription factor binding sites, epigenetic modification markers) mapping to the genome near the ncRNA sequence of interest.

The density of ncRNAs search option is a variation of the search by genomic location, with the goal of determining a genomic region by the density of RNAs that are clustered in that region. This search can be useful to identify clusters of transcripts (especially of miRNAs) that could be originated from multiple transcriptional units in a given genomic locus.

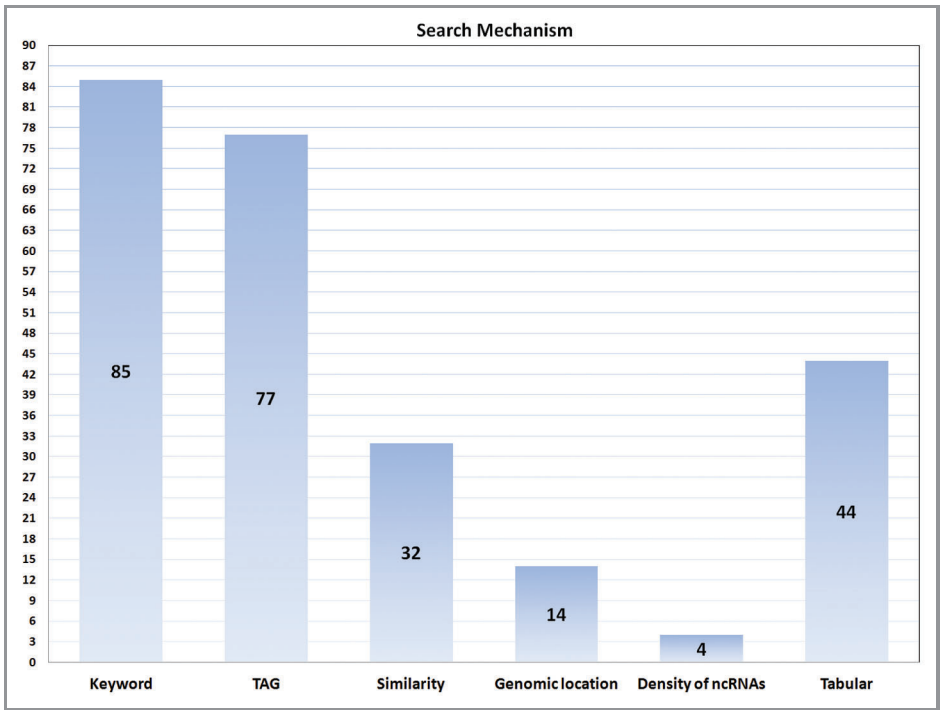


Figure 6. Database distribution per search mechanism.

The last search mechanism is the tabular form (as a table or list). Here, the links available in the databases' webpages lead to tables listing the information. Search is performed by manually browsing these tables.

Conclusion and Perspectives

The explosion in sequence information spurred by new sequencing technologies has created in turn an abundance of ncRNA online databases. NRDR is an attempt to facilitate information discovery as available in these databases. With the rate of sequencing projects continuing to increase we can expect that some of the existing databases will vastly increase their contents in terms of sequencing, while new specialized databases will also continue to be created. NRDR was designed in such a way that new online resources can easily be added, thus allowing NRDR to keep pace with new developments.

Acknowledgments

This work was funded in part by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to AMD; and from CNPq, Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), Financiadora de Estudos e Projetos (FINEP) and the FP-7 European Community SEtTReND grant agreement number 241865 to SVA. ARP received a fellowship from CAPES; VMC received a fellowship from CAPES and subsequently from FAPESP. SVA received an established investigator fellowship award from CNPq.

Supplemental Materials

Supplemental materials can be found at:
www.landesbioscience.com/journals/rnabiology/article/19352

References

1. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 2007; 316:1484-8; PMID:17510325; http://dx.doi.org/10.1126/science.1138341
2. Guffanti A, Iacono M, Pelucchi P, Kim N, Soldà G, Croft LJ, et al. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 2009; 10:163; PMID:19379481; http://dx.doi.org/10.1186/1471-2164-10-163
3. Vockenhuber MP, Sharma CM, Statt MG, Schmidt D, Xu Z, Dietrich S, et al. Deep sequencing-based identification of small non-coding RNAs in Streptomyces coelicolor. RNA Biol 2011; 8:468-77; PMID:21521948; http://dx.doi.org/10.4161/rna.8.3.14421
4. Ling KH, Brautigan PJ, Hahn CN, Daish T, Rayner JR, Cheah PS, et al. Deep sequencing analysis of the developing mouse brain reveals a novel microRNA. BMC Genomics 2011; 12:176; PMID:21466694; http://dx.doi.org/10.1186/1471-2164-12-176
5. Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, et al & Athanasius F Bompfünewerer Consortium. RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zool B Mol Dev Evol 2007; 308:1-25; PMID:17171697
6. Jossinet F, Ludwig TE, Westhof E. RNA structure: bioinformatic analysis. Curr Opin Microbiol 2007; 10:279-85; PMID:17548241; http://dx.doi.org/10.1016/j.mib.2007.05.010
7. Machado-Lima A, del Portillo HA, Durham AM. Computational methods in noncoding RNA research. J Math Biol 2008; 56:15-49; PMID:17786447; http://dx.doi.org/10.1007/s00285-007-0122-6
8. Soldà G, Makunin IV, Sezerman OU, Corradin A, Corti G, Guffanti A. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Brief Bioinform 2009; 10:475-89; PMID:19383843; http://dx.doi.org/10.1093/bib/bbp022
9. Bateman A, Agrawal S, Birney E, Bruford EA, Bujnicki JM, Cochrane G, et al. RNAcentral: A vision for an international database of RNA sequences. RNA 2011; 17:1941-6; PMID:21940779; http://dx.doi.org/10.1261/rna.2750811

10. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res 2009; 37(Database issue): D136-40; PMID:18953034; http://dx.doi.org/10.1093/nar/gkn766
11. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 2007; 3:e65; PMID:17432929; http://dx.doi.org/10.1371/journal.pcbi.0030065
12. Yang JH, Shao P, Zhou H, Chen YQ, Qu LH. deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res 2010; 38(Database issue):D123-30; PMID:19966272; http://dx.doi.org/10.1093/nar/gkp943
13. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. CSRDB: a small RNA integrated database and browser resource for cereals. Nucleic Acids Res 2007; 35(Database issue):D829-33; PMID:17169981; http://dx.doi.org/10.1093/nar/gkl991
14. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res 2009; 37(Database issue):D122-6; PMID:18829717; http://dx.doi.org/10.1093/nar/gkn617
15. Wright MW, Bruford EA. Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. Hum Genomics 2011; 5:90-8; PMID:21296742
16. Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 2007; 8: R43; PMID:17386095; http://dx.doi.org/10.1186/gb-2007-8-3-r43
17. Louro R, El-Jundi T, Nakaya HI, Reis EM, Verjovski-Almeida S. Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. Genomics 2008; 92:18-25; PMID:18495418; http://dx.doi.org/10.1016/j.ygeno.2008.03.013
18. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. Genome Biol 2011; 12:R16; PMID:21324177; http://dx.doi.org/10.1186/gb-2011-12-2-r16
19. Szymanski M, Erdmann VA, Barciszewski J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Res 2007; 35(Database issue):D162-4; PMID: 17169980; http://dx.doi.org/10.1093/nar/gkl994
20. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, et al. NONCODE v3.0: integrative annotation of long non-coding RNAs. Nucleic Acids Res 2012; 40(Database issue): D210-5; PMID:22135294; http://dx.doi.org/10.1093/nar/gkr1175
21. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, et al. CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 2007; 7:100; PMID:17760977; http://dx.doi.org/10.1186/1471-213X-7-100
22. Lee AP, Yang Y, Brenner S, Venkatesh B. TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. BMC Genomics 2007; 8:441; PMID:18045502; http://dx.doi.org/10.1186/1471-2164-8-441
23. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, et al. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 2007; 17:852-64; PMID:17568003; http://dx.doi.org/10.1101/gr.5650707
24. Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH. ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA 2010; 16:1889-901; PMID:20801769; http://dx.doi.org/10.1261/rna.2226910
25. Sone M, Hayashi T, Tarui H, Agata K, Takeichi M, Nakagawa S. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J Cell Sci 2007; 120:2498-506; PMID: 17623775; http://dx.doi.org/10.1242/jcs.009357
26. Yin Y, Zhao Y, Wang J, Liu C, Chen S, Chen R, et al. antiCODE: a natural sense-antisense transcripts database. BMC Bioinformatics 2007; 8:319; PMID:17760969; http://dx.doi.org/10.1186/1471-2105-8-319
27. Zhang Y, Li J, Kong L, Gao G, Liu QR, Wei L. NATsDB: Natural Antisense Transcripts DataBase. Nucleic Acids Res 2007; 35(Database issue):D156-61; PMID:17082204; http://dx.doi.org/10.1093/nar/gkl782
28. Li JT, Zhang Y, Kong L, Liu QR, Wei L. Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. Nucleic Acids Res 2008; 36:4833-44; PMID:18653530; http://dx.doi.org/10.1093/nar/gkn470
29. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011; 39(Database issue):D152-7; PMID: 21037258; http://dx.doi.org/10.1093/nar/gkq1027
30. Davis-Dusenbery BN, Hata A. Mechanisms of control of microRNA biogenesis. J Biochem 2010; 148:381-92; PMID:20833630
31. Nicolas FE, Lopez-Martinez AF. MicroRNAs in human diseases. Recent Pat DNA Gene Seq 2010; 4:142-54; PMID:21288192; http://dx.doi.org/10.2174/187221510794751659
32. Silahtaroglu A, Stenvang J. MicroRNAs, epigenetics and disease. Essays Biochem 2010; 48:165-85; PMID: 20822493; http://dx.doi.org/10.1042/bse0480165
33. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 2002; 36:233-78; PMID:12429693; http://dx.doi.org/10.1146/annurev.genet.36.042902.092433
34. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010; 464:1071-6; PMID:20393566; http://dx.doi.org/10.1038/nature08975
35. Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 2003; 22:8031-41; PMID:12970751; http://dx.doi.org/10.1038/sj.onc.1206928
36. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 2011; 39(Database issue):D146-51; PMID:21112873; http://dx.doi.org/10.1093/nar/gkq1138
37. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res 2011; 39(Database issue):D202-9; PMID:21037263; http://dx.doi.org/10.1093/nar/gkq1056
38. Backman TW, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, et al. Update of ASRP: the Arabidopsis Small RNA Project database. Nucleic Acids Res 2008; 36(Database issue):D982-5; PMID:17999994; http://dx.doi.org/10.1093/nar/gkm997
39. Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. Nature 2007; 447:1126-9; PMID:17538623; http://dx.doi.org/10.1038/nature05903
40. Fei Z, Joung JG, Tang X, Zheng Y, Huang M, Lee JM, et al. Tomato Functional Genomics Database: a comprehensive resource and analysis package for tomato functional genomics. Nucleic Acids Res 2011; 39 (Database issue):D1156-63; PMID:20965973; http://dx.doi.org/10.1093/nar/gkq991
41. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009; 458:223-7; PMID:19182780; http://dx.doi.org/10.1038/nature07672
42. Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer? Hum Mol Genet 2010; 19(R2):R152-61; PMID:20729297; http://dx.doi.org/10.1093/hmg/ddq353
43. Gibb EA, Brown CJ, Lam WL. The functional role of long non-coding RNA in human carcinomas. Mol Cancer 2011; 10:38; PMID:21489289; http://dx.doi.org/10.1186/1476-4598-10-38
44. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov 2011; 1:391-407; PMID:22096659; http://dx.doi.org/10.1158/2159-8290.CD-11-0209
45. Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene 2004; 23:6684-92; PMID: 15221013; http://dx.doi.org/10.1038/sj.onc.1207880
46. Reis EM, Ojopi EP, Alberto FL, Rahal P, Tsukumo F, Mancini UM, et al & Head and Neck Annotation Consortium. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. Cancer Res 2005; 65:1693-9; PMID:15753364; http://dx.doi.org/10.1158/0008-5472.CAN-04-3506
47. Brito GC, Fachel AA, Vettore AL, Vignal GM, Gimba ER, Campos FS, et al. Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. Mol Carcinog 2008; 47:757-67; PMID:18348187; http://dx.doi.org/10.1002/mc.20433
48. Perez DS, Hoage TR, Pritchett JR, Ducharme-Smith AL, Halling ML, Ganapathiraju SC, et al. Long, abundantly expressed non-coding transcripts are altered in cancer. Hum Mol Genet 2008; 17:642-55; PMID: 18006640; http://dx.doi.org/10.1093/hmg/ddm336
49. Tahira AC, Kubrusly MS, Faria MF, Dazzani B, Fonseca RS, Maracaja-Coutinho V, et al. Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. Mol Cancer 2011; 10:141; PMID:22078386; http://dx.doi.org/10.1186/1476-4598-10-141
50. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 2012; 30:99-104; PMID:22081020; http://dx.doi.org/10.1038/nbt.2024
51. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 2005; 15:987-97; PMID: 15998911; http://dx.doi.org/10.1101/gr.3455305
52. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol 2010; 8:149; PMID:21176148; http://dx.doi.org/10.1186/1741-7007-8-149
53. Cao Y, Wu J, Liu Q, Zhao Y, Ying X, Cha L, et al. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. RNA 2010; 16: 2051-7; PMID:20843985; http://dx.doi.org/10.1261/rna.2193110
54. Tamura M, Hendrix DK, Klosterman PS, Schimmelman NR, Brenner SE, Holbrook SR, Sr. SCOR: Structural Classification of RNA, version 2.0. Nucleic Acids Res 2004; 32(Database issue):D182-4; PMID:14681389; http://dx.doi.org/10.1093/nar/gkh080
55. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 2010; 38(Database issue):D280-2; PMID:19880381; http://dx.doi.org/10.1093/nar/gkp892

56. Tacutu R, Budovsky A, Fraifeld VE. The NetAge database: a compendium of networks for longevity, age-related diseases and associated processes. Biogerontology 2010; 11:513-22; PMID:20186480; http://dx.doi.org/10.1007/s10522-010-9265-8
57. Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, et al. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 2006; 34(Database issue):D150-2; PMID:16381834; http://dx.doi.org/10.1093/nar/gkj025
58. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009; 37(Database issue): D105-10; PMID:18996891; http://dx.doi.org/10.1093/nar/gkn851
59. Piekna-Przybylska D, Decatur WA, Fournier MJ. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. RNA 2007; 13:305-12; PMID:17283215; http://dx.doi.org/10.1261/rna.373107
60. Xin Y, Olson WK. BPS: a database of RNA base-pair structures. Nucleic Acids Res 2009; 37(Database issue): D83-8; PMID:18845572; http://dx.doi.org/10.1093/nar/gkn676
61. Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, et al. NCIR: a database of non-canonical interactions in known RNA structures. Nucleic Acids Res 2002; 30:395-7; PMID:11752347; http://dx.doi.org/10.1093/nar/30.1.395
62. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al, ENCODE Project Consortium, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute & Children's Hospital Oakland Research Institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799-816; PMID:17571346; http://dx.doi.org/10.1038/nature05874
63. Amaral PP, Mattick JS. Noncoding RNA in development. Mamm Genome 2008; 19:454-92; PMID:18839252; http://dx.doi.org/10.1007/s00335-008-9136-7
64. Seidl CI, Stricker SH, Barlow DP. The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. EMBO J 2006; 25:3565-75; PMID:16874305; http://dx.doi.org/10.1038/sj.emboj.7601245
65. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 2010; 39:925-38; PMID:20797886; http://dx.doi.org/10.1016/j.molcel.2010.08.011
66. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 2007; 129:1311-23; PMID:17604720; http://dx.doi.org/10.1016/j.cell.2007.05.022
67. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 2010; 329:689-93; PMID:20616235; http://dx.doi.org/10.1126/science.1192002
68. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 2010; 142:409-19; PMID:20673990; http://dx.doi.org/10.1016/j.cell.2010.06.040
69. Wang H, Iacoangeli A, Popp S, Muslimov IA, Imataka H, Sonenberg N, et al. Dendritic BC1 RNA: functional role in regulation of translation initiation. J Neurosci 2002; 22:10232-41; PMID:12451124
70. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell 2011; 43:904-14; PMID:21925379; http://dx.doi.org/10.1016/j.molcel.2011.08.018
71. Cho S, Jun Y, Lee S, Choi HS, Jung S, Jang Y, et al. miRGator v2.0: an integrated system for functional investigation of microRNAs. Nucleic Acids Res 2011; 39(Database issue):D158-62; PMID:21062822; http://dx.doi.org/10.1093/nar/gkq1094
72. Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, et al. sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. Nucleic Acids Res 2009; 37 (Database issue):D150-4; PMID:19015153; http://dx.doi.org/10.1093/nar/gkn852
73. Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 2008; 14:1012-7; PMID:18426918; http://dx.doi.org/10.1261/rna.965408
74. Bandyopadhyay S, Bhattacharyya M. PuTmiR: a database for extracting neighboring transcription factors of human microRNAs. BMC Bioinformatics 2010; 11:190; PMID:20398296; http://dx.doi.org/10.1186/1471-2105-11-190
75. Friard O, Re A, Taverna D, De Bortoli M, Corá D. CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC Bioinformatics 2010; 11:435; PMID:20731828; http://dx.doi.org/10.1186/1471-2105-11-435
76. Chiromatzo AO, Oliveira TY, Pereira G, Costa AY, Montesco CA, Gras DE, et al. miRNApath: a database of miRNAs, target genes and metabolic pathways. Genet Mol Res 2007; 6:859-65; PMID:18058708
77. Schmeier S, Schaefer U, MacPherson CR, Bajic VB. dPORE-miRNA: polymorphic regulation of microRNA genes. PLoS One 2011; 6:e16657; PMID:21326606; http://dx.doi.org/10.1371/journal.pone.0016657
78. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011; 39(Database issue):D163-9; PMID:21071411; http://dx.doi.org/10.1093/nar/gkq1107
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10; PMID:2231712
80. 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 Res 2006; 16:885-9; PMID:16751343; http://dx.doi.org/10.1101/gr.5226606

MOLECULAR CANCER

RESEARCH Open Access

Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer

Ana C Tahira[1], Márcia S Kubrusly[2], Michele F Faria[1], Bianca Dazzani[1], Rogério S Fonseca[1], Vinicius Maracaja-Coutinho[1], Sergio Verjovski-Almeida[1], Marcel CC Machado[2] and Eduardo M Reis[1*]

Abstract

Background: Pancreatic ductal adenocarcinoma (PDAC) is known by its aggressiveness and lack of effective therapeutic options. Thus, improvement in current knowledge of molecular changes associated with pancreatic cancer is urgently needed to explore novel venues of diagnostics and treatment of this dismal disease. While there is mounting evidence that long noncoding RNAs (lncRNAs) transcribed from intronic and intergenic regions of the human genome may play different roles in the regulation of gene expression in normal and cancer cells, their expression pattern and biological relevance in pancreatic cancer is currently unknown. In the present work we investigated the relative abundance of a collection of lncRNAs in patients' pancreatic tissue samples aiming at identifying gene expression profiles correlated to pancreatic cancer and metastasis.

Methods: Custom 3,355-element spotted cDNA microarray interrogating protein-coding genes and putative lncRNA were used to obtain expression profiles from 38 clinical samples of tumor and non-tumor pancreatic tissues. Bioinformatics analyses were performed to characterize structure and conservation of lncRNAs expressed in pancreatic tissues, as well as to identify expression signatures correlated to tissue histology. Strand-specific reverse transcription followed by PCR and qRT-PCR were employed to determine strandedness of lncRNAs and to validate microarray results, respectively.

Results: We show that subsets of intronic/intergenic lncRNAs are expressed across tumor and non-tumor pancreatic tissue samples. Enrichment of promoter-associated chromatin marks and over-representation of conserved DNA elements and stable secondary structure predictions suggest that these transcripts are generated from independent transcriptional units and that at least a fraction is under evolutionary selection, and thus potentially functional.
Statistically significant expression signatures comprising protein-coding mRNAs and lncRNAs that correlate to PDAC or to pancreatic cancer metastasis were identified. Interestingly, *loci* harboring intronic lncRNAs differentially expressed in PDAC metastases were enriched in genes associated to the MAPK pathway. Orientation-specific RT-PCR documented that intronic transcripts are expressed in sense, antisense or both orientations relative to protein-coding mRNAs. Differential expression of a subset of intronic lncRNAs (*PPP3CB*, *MAP3K14* and *DAPK1 loci*) in metastatic samples was confirmed by Real-Time PCR.

Conclusion: Our findings reveal sets of intronic lncRNAs expressed in pancreatic tissues whose abundance is correlated to PDAC or metastasis, thus pointing to the potential relevance of this class of transcripts in biological processes related to malignant transformation and metastasis in pancreatic cancer.

Keywords: pancreatic cancer, molecular markers, noncoding RNAs, intronic transcription, metastasis, *MAPK* , pathway, cDNA microarrays

* Correspondence: emreis@iq.usp.br
[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brasil
Full list of author information is available at the end of the article

Background

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic neoplasm and accounts for > 85% of pancreatic tumor cases [1]. PDAC is a devastating disease with very poor prognosis for which the only curative treatment is resection surgery [2]. However, only 15-20% of patients have resectable pancreatic tumor, and from these only 20% presents a 5-year survival, which results in an average 5-year survival rate of 3-5% [1]. PDAC aggressiveness is mainly associated to the lack of early diagnosis tools and the limited response to available treatments [2].

Large-scale gene expression studies of tumor samples have been extensively employed to delineate the molecular pathways and cellular processes involved in tumorigenesis and progression of PDAC [3] and to search for novel biomarkers for diagnosis and molecular targets for therapeutic intervention in pancreatic cancer [4]. In spite of the wealth of information generated in recent years on the most frequent molecular alterations found in PDAC [5], there are still important open question in pancreatic cancer biology such as the profound resistance of primary and metastatic PDAC to chemo- and radiotherapy [6]. Regarding the identification of molecular markers for pancreatic cancer diagnostic/prognostic, while some promising candidate genes have been proposed [4], none have been proven effective to significantly improve early detection and to reduce mortality/morbidity of the disease. Thus, a better understanding of the molecular basis of pancreatic cancer is required for the identification of more effective diagnostic markers and therapeutic targets.

Over the last decade, advances in genome-wide analyses of the eukaryotic transcriptome have revealed that the majority of the human genome is transcribed, producing large numbers of long (> 200 nt) noncoding RNAs (lncRNAs) mapping to intronic and intergenic regions [7-10]. These include subsets of polyadenylated and non-adenylated transcripts that accumulate differently in the nucleus and cytoplasm of cells [10,11]. While only a small fraction of lncRNAs have been characterized in detail, it is clear that these transcripts may act through diverse molecular mechanisms and play regulatory and structural roles in important biological processes, such as in genomic imprinting, chromosome inactivation, cell differentiation and development, cell proliferation, protein nuclear import, organization of nuclear domains and apoptosis (see [12] for a review).

Altered expression of lncRNAs has been documented in different types of human cancer [13-15] prompting an increasing interest in their use as biomarkers for diagnosis and prognosis as well as potential therapeutic targets [14,16-19]. Increased expression of the lncRNA *MALAT-1* has been observed in several types of tumors, including metastatic non-small cell lung cancer [19]. Recently, augmented levels of *HOTAIR* in primary breast tumors were shown to correlate with breast cancer invasiveness and metastasis [18]. Measurement of lncRNA *PCA3* in patient urine samples has been shown to allow more sensitive and specific diagnosis of prostate cancer than the widely used marker prostate-specific antigen (PSA) [16]. The lncRNA *HULC* is highly expressed in hepatocarcinoma patients and detected in the blood by conventional PCR methods [20].

There are several reports of aberrant expression of microRNAs in PDAC [21,22], and there is potential in their use as biomarkers for disease diagnosis [23,24]. However, there is a paucity of information regarding the expression of lncRNAs in pancreatic cancer. In an interesting study performed by Ting et al. it was observed the aberrant overexpression of satellite repeat RNAs (HSATII) ranging from 100 to 5000 nt in patients with PDAC [25]. Interestingly, detection of HSATII by RNA in situ hybridization was able to correctly diagnose PDAC in tumor biopsies, including cases in which the histopathology was non-diagnostic [25].

Our group has previously shown that most (at least 74%) annotated protein-coding gene *loci* generate intragenic lncRNAs that map to intronic regions [26]. Possible relevance of intronic lncRNAs to neoplastic processes was proposed following the observation that subsets of these transcripts are present in gene expression signatures correlated to the degree of malignancy in prostate cancer [17] or to tissue histology in head and neck tumors [27] and renal cell carcinoma [28]. In addition, a number of intronic lncRNAs were found to be regulated by androgen stimulation of cultured prostate cancer cells [29], indicating that these transcripts are expressed in a regulated manner and thus, corroborating the idea that intronic lncRNAs are biologically relevant.

In this study, we used a custom cDNA microarray platform with probes for lncRNAs expressed from intronic and intergenic regions of the human genome, as well as for a selected set of cancer-related protein-coding genes to generate expression profiles from a collection of tumor and non-tumor pancreatic tissue samples. Expression of intronic/intergenic lncRNAs subsets was detected across all samples tested. Enrichment of promoter-associated chromatin marks indicate that these transcripts originate from independent transcriptional units. Over-representation of conserved DNA elements and stable secondary structure predictions suggest that at least a fraction of these transcripts are under evolutionary selection and thus potentially functional.

Importantly, we identified expression signatures comprising long noncoding RNAs that are significantly correlated with primary and metastatic ductal pancreatic adenocarcinoma. This suggests that lncRNAs are modulated during tumorigenesis and tumor progression and therefore may participate in molecular processes relevant to malignant transformation and metastasis in pancreatic cancer.

Results

Long noncoding RNAs from intronic and intergenic regions are expressed in neoplastic and non-tumor pancreatic tissues

In this work, a custom spotted cDNA microarray with approximately 4,000 elements was used to investigate the expression patterns of a collection of protein-coding transcripts and putative noncoding RNAs in clinical samples of primary and metastatic tumor, chronic pancreatitis and histologically normal pancreatic tissue. This array platform has been described previously [17,28] and contains probes that interrogate 2,371 RefSeq mRNAs from genes associated with cancer in the literature, as well as 984 transcripts mapping to intronic or intergenic regions of the genome and to known lncRNAs. Fluorescent cRNA targets generated from 38 pancreatic tissue samples (15 primary adenocarcinoma, 9 histologically normal adjacent tissue, 6 metastatic samples and 8 chronic pancreatitis) were individually hybridized to microarrays in replicate. After data filtering (see Methods for details), 1,607 transcripts were detected as expressed in at least one histological type, being 1,267 protein-coding mRNAs and 340 putative noncoding RNAs, including transcripts with no overlap to RefSeq exons, i.e, mapping to intronic and intergenic regions. Only candidate lncRNAs sequences that showed genomic alignments with at least 90% identity and coverage were further analyzed, resulting in 335 transcripts (22 known lncRNAs, 240 putative lncRNAs mapped to intronic regions and 73 to intergenic regions).

Expression of comparable fractions of protein-coding mRNAs and putative long noncoding transcripts mapping to intronic and intergenic regions was detected in all histological tissue types (Table 1). The fraction of intronic lncRNAs detected as expressed in the microarray (240/722 = 0.33) is comparable to that of known RefSeq lncRNAs (22/74 = 0.30) and intergenic lncRNAs (73/188 = 0.39), and lower than the fraction of expressed protein-coding mRNAs (1267/2371 = 0.53). The smaller fraction of lncRNAs detected in pancreatic tissues (0.30-0.39) compared to protein-coding mRNAs (0.53) reflects the observation from other studies that noncoding RNAs are generally less abundant and more tissue-specific than protein coding mRNAs [9,26]. In fact, we observed that the lncRNAs detected in pancreatic tissue samples by array hybridization were on average less abundant than protein-coding transcripts (average intensities 24.9 and 31.6, respectively).

Of the 240 gene loci harboring intronic lncRNAs that were detected in pancreatic tissues, only 62 had array probes interrogating exons of mRNAs from the same *loci*. From these, 31 (50%) were detected only in intronic regions, pointing to a subset of lncRNAs that conceivably are generated by independent intronic transcription rather than pre-mRNA splicing. Thirty one *loci* were detected by both exonic and intronic probes (50%). For each of these *loci*, Pearson correlation between the expression of the lncRNA and mRNA across all pancreatic tissue samples was calculated. Correlations were generally low ($-0.5 < r < 0.5$ for 27 out 31 *loci*), with 11 *loci* displaying a negative correlation between expression of the intronic lncRNA and the mRNA, and 20 showing a positive correlation.

To obtain further information regarding the correlation between the 240 intronic lncRNAs expressed in pancreatic tissues and the adjacent exons from the same *loci* we analyzed their expression in a set of nine RNA-seq libraries [30]. For each *locus* in each library, the number of tags was normalized by RPKM (Reads Per Kilobase of exon model per Million mapped reads).

Table 1 Gene expression detected in the microarrays according to probe type and pancreatic tissue histology

Type	# probes in the array	Detected as expressed in				# expressed probes *
		NT (n = 9)	T (n = 15)	M (n = 6)	CP (n = 8)	
protein-coding mRNA	2371	1106	1167	1198	1230	1267
Known lncRNA (RefSeq)	74	20	19	22	18	22
Intronic lncRNA	722	206	202	238	235	240
Intergenic lncRNA	188	68	68	74	77	78
Total	3355	1400	1456	1532	1560	1607

*To be considered as expressed, a probe signal should be detected above the median array intensity value in at least 75% of samples from at least one histological type.

NT, non-tumor pancreatic tissue; T, primary pancreatic adenocarcinoma; M, metastases from primary pancreatic tumors; CP; chronic pancreatitis.

Pearson correlations between each intronic lncRNA and the adjacent upstream/downstream exons were calculated if all elements (intronic lncRNA, upstream and downstream exons) were detected in at least 4 out 9 RNAseq libraries. Seventy five loci (75/240, 31%) satisfied these criteria and were further analyzed. As expected, we found a high correlation between the expression of exons flanking intronic lncRNAs (66/75 with $r > 0.5$). We found that about a third of exon/intron pairs showed positive correlation of expression (24/75, $r > 0.5$). The expression of more than half of exon/intron pairs were poorly correlated (46/75, $-0.5 < r < 0.5$), and a small fraction was negatively correlated (5/75, $r < -0.5$). The overall low correlation observed between the expression of intronic lncRNAs and adjacent exons suggest that for the most part intronic lncRNAs are processed and accumulate in the cell at rates distinct from mRNAs produced in the same *loci*, arguing against them being simply remainings of splicing lariats.

To gain further insight into the expression pattern of the 335 putative lncRNAs, we investigated their expression in other tissue types using publicly available RNA-seq datasets generated from nine different tissue histologies [30]. By cross-referencing the genome mapping coordinates of the pancreatic-expressed lncRNAs with coordinates of the RNAseq reads we found that approximately 80% of the former (267/335) are detected in at least one other human tissue (Figure 1).

Coding potential of the 335 sequences mapping to intronic and intergenic regions was investigated using the Coding Potential Calculator (CPC) software [31]. This analysis showed that most sequences (322/335, 96%) have little or no protein coding potential. Thus, we suggest that most of the intronic and intergenic transcripts detected in pancreatic tissues are indeed noncoding RNAs.

To document the length of the intronic transcripts expressed in pancreatic samples we compared the set of intronic RNA sequences (n = 240) with sequences resulting from the assembly of ESTs and mRNAs deposited in GenBank that map to intronic regions of the genome, previously generated in our group and which is available as a UCSC Genome Browser custom annotation track [26]. We found that 190 out 240 intronic transcripts (79%) are represented by an assembled sequence contig. The mean length of the intronic contigs is 779 bp, whereas the individual ESTs have a mean length of 428 bp, suggesting that the ESTs spotted on the microarray are partial sequences of longer noncoding RNA transcripts.

We also investigated the proximity of 73 intergenic lncRNAs to UTRs of annotated genes to evaluate if these transcripts could represent untranslated regions of incomplete mRNAs. We found that 14 intergenic transcripts (14/73, i.e. 19%) map within 1 kb from a known 3'or 5' UTR and could potentially extend the 3' or 5' untranslated region of a known protein-coding mRNA. The remaining 59 transcripts map at least 1 kb away from a known mRNA and possibly constitute yet unannotated intergenic noncoding RNAs.

To investigate if the intronic and intergenic RNAs detected in pancreatic tissue could be precursors of small regulatory ncRNAs, we compared the set of 335 lncRNAs expressed in pancreas to microRNA and

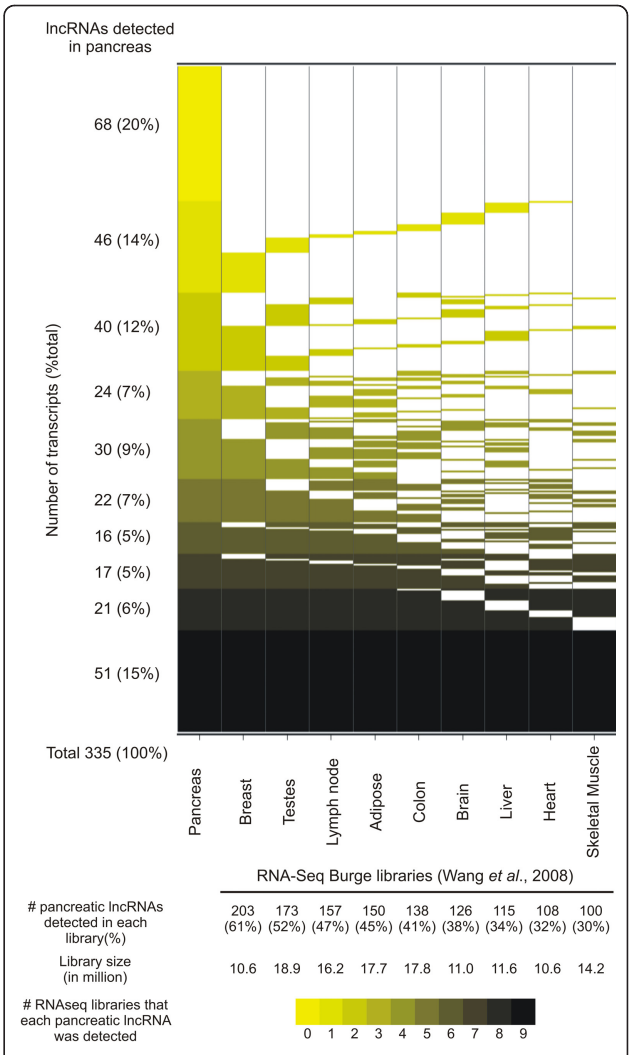


Figure 1 Intronic and intergenic lncRNAs detected in pancreatic tissues comprise both tissue-specific and broadly expressed transcripts. Genomic coordinates of 335 intronic/intergenic transcripts detected as expressed in pancreatic tissues were cross-referenced with genomic coordinates of transcripts detected in RNAseq libraries from nine different tissues [30]. LncRNAs were grouped and colored by the number of tissue libraries where their expression was detected (Y axis). RNAseq libraries are ordered in the × axis according to the number of pancreatic lncRNAs detected in each library.

snRNA sequence databases [32,33]. No significant similarity to known small RNA was found, except for one sequence that mapped to the *SNOR89 locus*. Considering that the average length of micro RNA precursors (> 1,000 nt) is greater than the average EST length (468 nt) we extended the genomic coordinates of probe sequences by 1 kb at both ends and repeated the sequence comparison with known small RNA datasets. Using this approach, we found four putative extended EST sequences that show high similarity to seven additional small RNAs: *hsa-mir-1259, hsa-mir-326, hsa-mir-4269, hsa-mir-675, SNORD12, SNORD12B* and *SNORD12C*.

Sequence conservation among species is generally viewed as an indication of functional significance of a given genomic feature. We searched for evidence of sequence conservation within the set of intronic/intergenic lncRNAs expressed in pancreatic tissues by comparing their mapping coordinates with those from conserved DNA elements in vertebrates (phastCons 46way vertebrates), mammals (phastCons 46way placental) and primates (phastCons 46way primates) obtained from the UCSC genome browser. After normalization by the number of conserved elements present in each group, relatively greater overlap with conserved DNA elements was observed within primate, mammalian and vertebrate sequences, in this order (Additional File 1, Figure S1). The overlap of intronic/intergenic RNAs with evolutionarily conserved DNA elements was greater than the expected by chance alone, as judge by the overlap attained with a set of randomly selected intronic/intergenic DNA sequences with same length and CG% content (Fisher's exact test $p < 0.05$, Additional File 1, Figure S1). In addition, a fraction of the lncRNAs mapping to intronic/intergenic regions (49 sequences out 335 analyzed, 15%) appear to fold into stable RNA secondary structures (P > 0.5) (http://verjo102.iq.usp.br/sites/tahira/structures.html), as predicted by the RNAz program [34]. Altogether, these observations provide additional support to the notion that at least a fraction of the noncoding transcripts mapping to intronic/intergenic regions may exert functional roles in pancreatic cells.

Enrichment of promoter-associated chromatin marks (H3K4me3) and start sites of capped transcripts suggest that intronic lncRNAs are independent transcriptional units

Given the paucity of information about the biogenesis of lncRNAs originated from intronic and intergenic regions, we searched for regulatory elements in the genome that could be associated to their transcriptional control. First, we investigated the distribution of trimethylation of lysine 4 in histone 3 (H3K4me3), a chromatin modification associated with regions of transcription initiation [35], in the vicinity of intronic/intergenic lncRNAs. Genomic coordinates of H3K4me3 marks measured in 13 cell lineages [35] were obtained from the UCSC Genome Browser. Only H3K4me3 marks with a $p < 10^{-5}$ were used to limit the experimental noise. The nearest H3K4me4 mark relative to the known boundaries of intronic/intergenic lncRNAs (based on sequenced ESTs) expressed in pancreatic tissues was selected and the distance annotated. As a control, the same analysis was performed using 100 random sets of intronic or intergenic DNA sequences with same length and GC content.

An enrichment of H3K4me3 marks was observed closer to the known boundaries for the set of intronic transcripts expressed in pancreatic tissue samples (Figure 2, panel A, blue bars). The distance distribution of H3K4me3 marks relative to known boundaries of intronic transcripts was significantly different (KS test, highest $p < 0.05$) from that observed for a random set of sequences (same length and %CG), indicating that it is not explained by chance alone. The same analysis was performed with the set of expressed intergenic regions. Although we observed a higher frequency of H3K4me3 marks closer to the known boundaries of the intergenic transcripts, we found no statistically significant difference in their distribution relative to the random control set (Figure 2, panel B, green bars).

As expected, we observed a higher frequency of H3K4me3 marks closer to the known boundaries of protein-coding mRNAs (Figure 2, panel A, red bars). This distribution is statistically different from the one obtained with a control comprising a random sequence set (KS test, highest $p < 0.01$). No statistically significant difference was observed between pancreas-expressed intronic/intergenic lncRNAs and mRNAs regarding the distributions of promoter-associated H3K4me3 marks, indicating that these distributions are similar. The enrichment of promoter-associated H3K4me3 at the vicinity of intronic/intergenic pancreatic-expressed transcripts argues that these transcripts are independent transcriptional units.

We also investigated the distribution of annotated CpG islands relative to EST probes representing protein-coding mRNAs and noncoding intronic/intergenic RNAs expressed in pancreatic tissues. To pursue this analysis we used the genomic coordinates of CpG islands available as a UCSC genome browser track. First, we cross-referenced the coordinates of annotated CpG islands with those of EST probes representing mRNAs expressed in pancreatic tissues, which showed an enrichment towards EST boundary coordinates (Figure 2, panel C, red bars), significantly different from the distribution observed by a random sequence set (KS test, p <

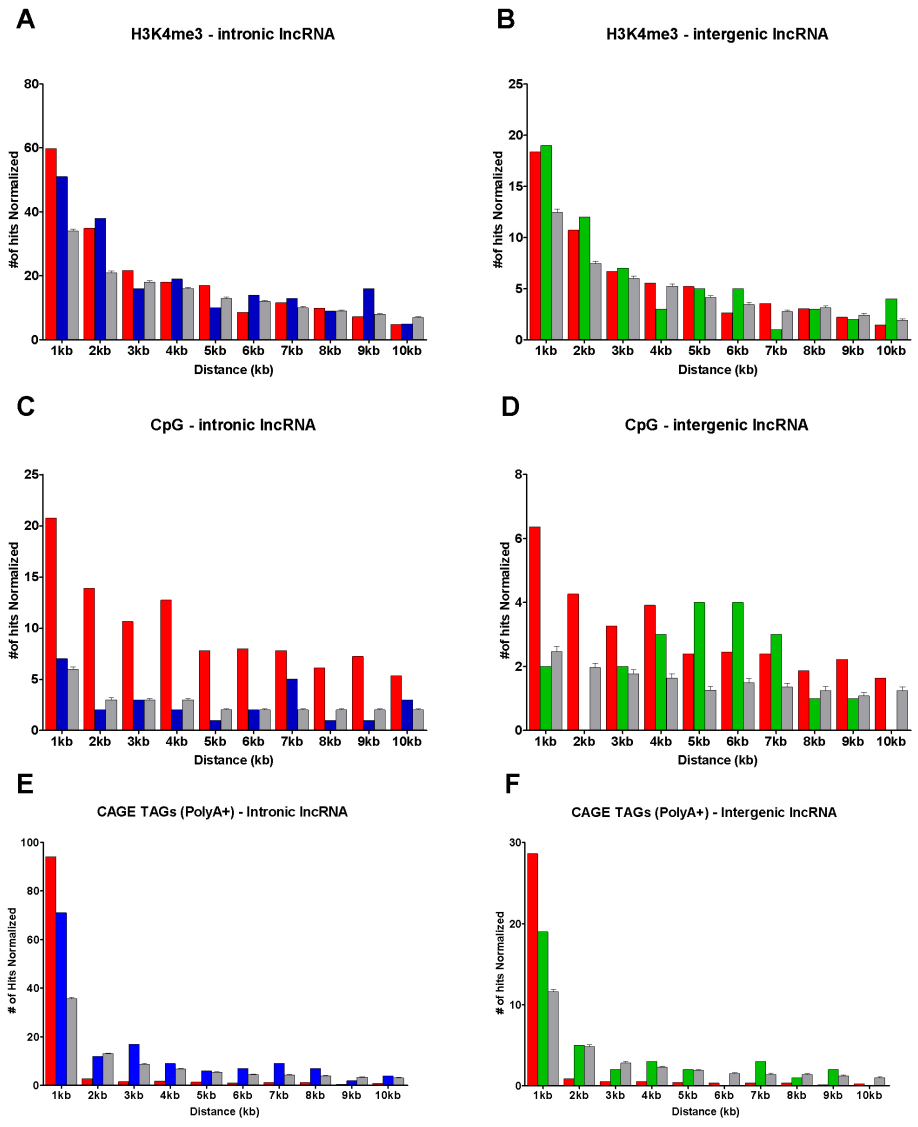


Figure 2 Genomic *loci* encoding intronic lncRNAs are enriched in promoter-associated histone marks and start sites of capped transcripts. Distance distribution (X axis) of promoter-associated chromatin marks H3K4me3 binding sites (panels A, B), CpG islands (panels C, D) and CAGE Tags (panels E, F) relative to genomic coordinates of intronic (blue bars) and intergenic (green bars) transcripts expressed in pancreatic tissues (Y axis) were calculated. For comparison, distribution distances were calculated for an equal number of protein-coding mRNAs (red bars) and for randomly selected intronic or intergenic genomic sequences with the same length and % GC of pancreas expressed lncRNAs (light gray bars).

0.001). No statistically significant association with CpG islands was observed for intronic or intergenic sequence sets relative to random sets with same length and CG% content (KS test, *p-value* > 0.05) (Figure 2, panels C and D, blue and green bars).

We also compared the known start sites of intronic/intergenic lncRNAs with CAGE tags generated from poly(A+) RNA from 6 different cell lineages (RIKEN). We note that this set does not include CAGE libraries derived from pancreatic tissues. As pre-processing, coordinates of overlapped tags were clustered and only clusters containing at least 5 tags were considered for further analysis. Next, we calculated the distance of the closest CAGE tag cluster to intronic/intergenic lncRNAs, protein-coding mRNAs, and to random genomic sequences. A significant enrichment (KS test, $p < 0.05$) of CAGE tags within 1kb of the known start of intronic lncRNAs expressed in pancreatic tissues was observed (Figure 2, panel E, blue bars). Although a higher frequency of CAGE tags closer to the known start site of intergenic lncRNAs was observed, the enrichment was not statistically significant when

compared to the random control set (Figure 2, panel F, green and gray bars).

Identification of a gene expression signature correlated to ductal pancreatic cancer comprising protein-coding and lncRNAs

To gain further insights on the putative biological relevance of intronic/intergenic noncoding RNAs in pancreatic cancer we investigated their relative expression in tumor and non-tumor pancreatic tissues. Identification of genes specifically deregulated in malignant pancreatic epithelial cells is frequently confounded by an augmented stromal component in the latter due to the presence of proliferating stromal cells and infiltrating inflammatory cells [36,37]. A similar desmoplastic reaction is observed in chronic pancreatitis [38,39]. To favor the identification of genes specifically altered in neoplastic pancreatic cells, we performed a two-class analysis comparing the expression profiles of 15 primary adenocarcinoma samples with nine histologically normal tissue fragments adjacent to tumors combined to eight samples of chronic pancreatitis. Using this approach, we found 147 transcripts differentially expressed in pancreatic tumor samples relative to non-tumor tissues (FDR ≤ 10%). This expression signature comprised 104 protein-coding mRNAs and 43 lncRNAs, being 34 intronic and 9 intergenic transcripts (See Additional file 2, Table S1 for a complete list). As shown in Figure 3, except by one sample (210, primary tumor), the 147-gene signature efficiently discriminated tumor and non-tumor tissues. Conceivably, the prevalence of an inflammatory component in the 210 sample could explain this sample showing an expression profile more similar to chronic pancreatitis samples.

Next, we performed a meta-analysis to compare the list of protein-coding mRNAs presented in the pancreatic

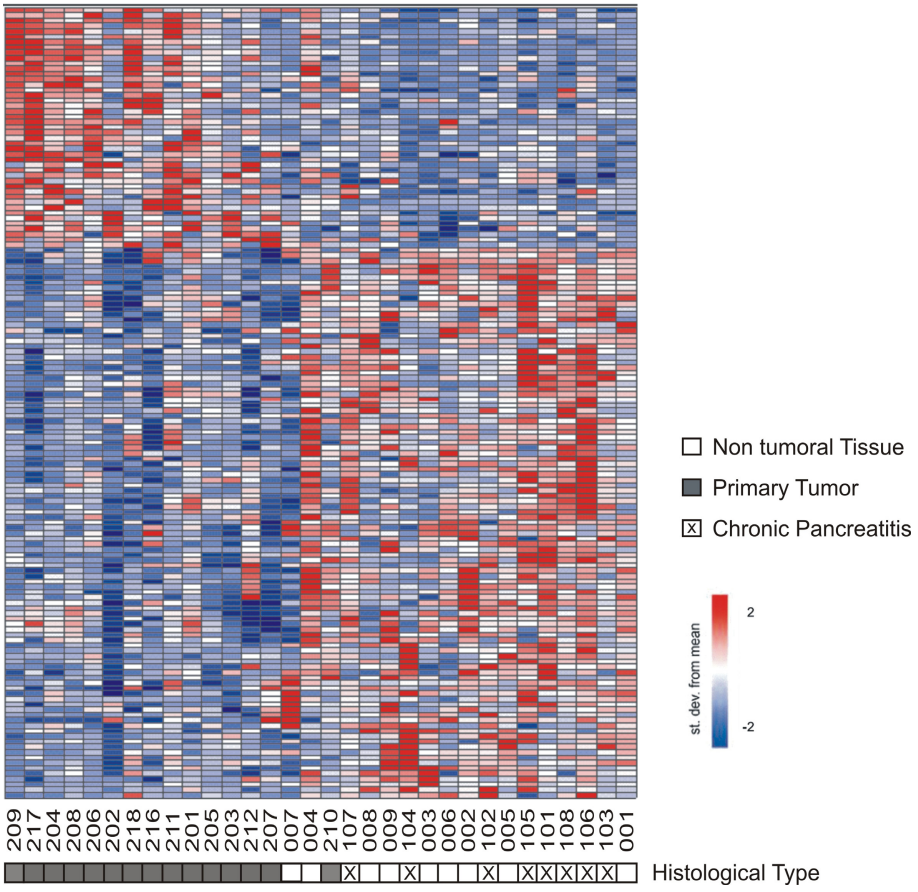


Figure 3 A gene expression signature of pancreatic adenocarcinoma. A two-class statistical analysis (see ***Methods***) identified 147 transcripts (rows) differentially expressed (FDR ≤10%) between primary adenocarcinoma samples and histologically normal and chronic pancreatitis samples combined (columns). Patient ID numbers are shown below the columns. Forty three transcripts mapping to intronic or intergenic regions were identified (43/147, i.e. 29%). Expression level of each gene is represented by the number of standard deviations above (red) or below (blue) the average value for that gene across all samples. Samples are ordered according to their individual correlation to the average profile of the primary tumor samples. Sample tissue histology is shown below each patient ID.

tumor expression signature with those identified in other gene expression studies with clinical samples of pancreatic cancer, retrieved from the Pancreatic Expression Database [40] (see Additional file 3, Table S2). Twenty four out 104 protein-coding transcripts detected in our analysis (24/104, i.e. 23%) were reported in at least one of the 12 studies, comprising 15 different analyses, deposited in the Pancreatic Expression Database. From these, expression changes of 17 genes were confirmed by other studies, whereas 3 showed partial agreements and 4 showed an inverted pattern of expression. Confirmed genes included genes already reported in the literature and proposed as biomarkers of pancreatic cancer such as *S100A6*, *TIMP1*, *NF-κB*, *VCL* and *S100P* [5,41-47]. It is worth mentioning that overexpression of *S100P* was detected in 11 different studies (Additional file 3, Table S2). We also detected upregulation of *MBOAT* in pancreatic tumors. Increased expression of MBOAT in ductal pancreatic adenocarcinoma has been already reported and was shown to inversely correlate to patient survival in pancreatic cancer [39].

A gene enrichment analysis using the DAVID analysis suite [48] using as input either the list of protein-coding mRNAs or of intronic transcripts differentially expressed in pancreatic tumors was performed to investigate the over-representation of specific molecular functions, biological processes and cellular components of the Gene Ontology annotation [49]. For this analysis, intronic transcripts were annotated according to the gene *locus* where they map on the genome. Only categories having corrected EASE score < 0.05 [48] were considered as overrepresented.

Among the set of 104 protein-coding mRNAs differentially expressed in pancreatic tumors we found an enrichment of gene categories encoding proteins involved in "focal adhesion" ($p < 0.03$; *TRIP6*, *TRIM25*, *VCL*, *SDC1*, *ARPC2*, *DLC1*, *CDH1*, *ITGB5*), "RNA transport and localization " ($p < 0.01$; *THOC7*, *THOC2*, *RAN*, *NUP85*, *THOC3*) and localizing to "basolateral plasma membrane" ($p < 0.05$; *NOTCH4*, *ARPC1B*, *CDH1*, *TRIP6*, *VCL*, *TRIM25*, *SDC1*, *DLC1*). Deregulation in pancreatic cancer of genes encoding proteins involved in focal adhesion has already been reported in the literature [39]. Noteworthy, the gene category "RNA transport and localization" comprises genes associated to the TREX-complex (*THOC2*, *THOC3*). Increased expression of this complex (Thoc1) in breast cancer correlates with tumor size and the metastatic state of the tumor progression [50], thus suggesting that modulation of the TREX-complex could also have a role in pancreatic cancer. No enriched gene category was found amongst gene *loci* that harbor differentially expressed intronic lncRNAs.

Ingenuity Pathway Analysis (IPA) [51] was used to identify pathways and gene networks represented amongst the sets of protein-coding mRNAs identified in the pancreatic tumor gene expression signature. The most enriched network, "cellular movement, cell-to-cell signaling interactions and endocrine system" ($p < 10^{-43}$) comprised 23 differentially expressed transcripts and included most genes represented in the enriched gene categories identified using DAVID (see Additional file 4, Figure S2). Gene networks associated with "cellular movement, skeletal and muscular system development and function and inflammatory response" (17 genes, $p < 10^{-29}$) and "carbohydrate metabolism, small molecule biochemistry and infectious disease" (16 genes, $p < 10^{-28}$) were also identified.

Identification of genes correlated to metastasis in pancreatic cancer

A hallmark of pancreatic cancer is the high prevalence of metastatic disease, whose molecular basis is poorly understood. To search for protein-coding and long non-coding RNAs with expression levels correlated to the metastatic phenotype in pancreatic cancer, we compared expression profiles from 15 primary adenocarcinoma samples with those obtained from 6 distant metastases originated from primary pancreatic adenocarcinoma. Metastatic samples were collected from secondary tumors appearing in different target sites (1 from peritoneum, 1 from ganglion, 4 from liver), from different patients. Using a significance threshold of FDR ≤ 5%, we identified a metastasis-associated signature comprising 355 differentially expressed transcripts (Figure 4). From these, 221 are protein-coding mRNAs and 134 are non-coding RNAs (134/355, 38% of signature), from which 101 map to intronic, 27 to intergenic genomic regions and 6 are known lncRNAs (a complete list is available as Additional file 5, table S3).

Gene enrichment analysis using protein-coding mRNAs differentially expressed in metastatic samples identified the over-representation of genes involved in "nucleic acid transport" and "RNA localization " ($p < 0.03$; *THOC7*, *THOC2*, *RAN*, *NUP85*, *THOC3*, *NUP88*).

A similar analysis performed with gene *loci* harboring intronic lncRNAs differentially expressed in metastasis showed enrichment of gene categories pertaining to "MAPK signaling pathway" ($p < 0.03$; *ARRB1*, *ATF2*, *MAPK1*, *MAP2K5*, *MAP3K1*, *MAP3K14*, *PPP3CB*, *RAPGF2* and *TGFβR2*), "phosphate metabolic process" ($p < 0.05$; *ABL2*, *ENPP2*, *PTEN*, *CSNK1D*, *TYK2*, *MAPK1*, *MAP2K5*, *MAP3K1*, *MAP3K14*, *PPP3CB*, *PPP2R2A*, *PASK*, *TNK2*, *DAPK1* and *TGFβR2*), "non-membrane-bounded organelle" ($p < 0.02$; *ABL2*, *GPHN*, *ITPR1*, *SORBS1*, *TYK2*, *MAPK1*, *MAP2K5*, *MAP3K1*, *NDRG1*, *DST*, *MCPH1*, *USH1C*, *MAEA*, *BBS5*, *SLC4A7*,

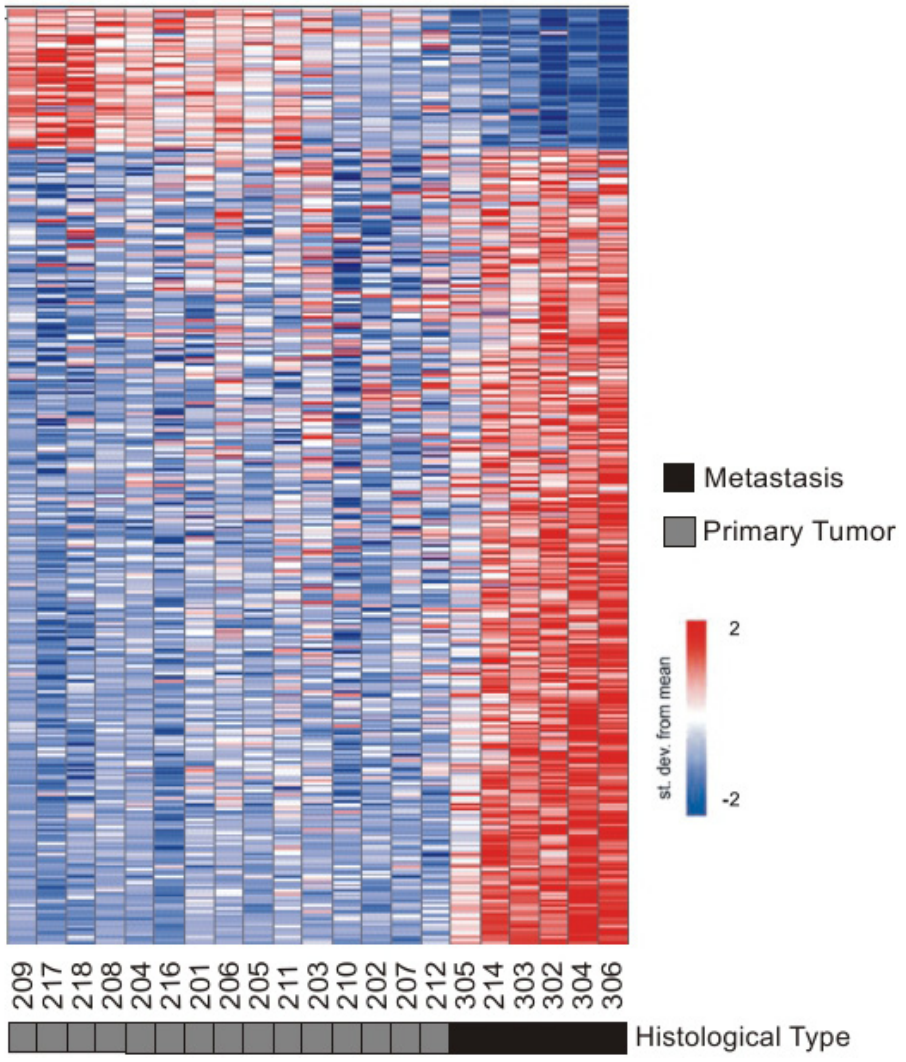


Figure 4 A gene expression signature correlated to metastasis in pancreatic adenocarcinoma. Three hundred fifty five transcripts (rows) identified as differentially expressed (FDR ≤ 5%) between metastatic (dark box) and primary tumor (dark gray) samples from 21 patients (columns). Patient ID numbers are shown below the columns. One hundred thirty four intronic and intergenic lncRNAs were identified, comprising 38% of the metastasis signature. Expression level of each transcript is represented by the number of standard deviations above (red) or below (blue) the average value across all samples. Samples are ordered according to their individual correlation to the average profile of primary tumor samples. Tissue histology is shown below each patient ID.

RAPH1, *CNN3*, *NR2C2*, *DMD*, *DAZAP1*, *PHF12*, *NOP58*, *ATF3*, *ALMS1*, *STON2*, *DAPK1* and *MYO5A*) and "actin filament-based process"($p < 0.05$; *ABL2*, *SORBS1*, *PACSIN2*, *CNN3*, *MYO5A* and *DST*).

Ingenuity Pathway Analysis identified significantly enriched gene networks amongst protein-coding genes differentially expressed in metastasis. The most enriched gene network of differentially expressed protein-coding mRNAs ($p < 10^{-41}$) included genes related to "cellular movement, gene expression and immune cell trafficking" (see Additional file 6, Figure S3). Among these we found that up regulation of *S100A4*, *NCAM1* and *LIMK1* had already been associated with metastatic behavior in pancreatic cancer [52-54]. While the remaining genes in the network had not been associated with metastasis in pancreatic cancer yet, most of them have previously shown to be involved with malignancy or metastatic behavior in other types of cancer (see Additional file 7, Table S4 for a complete list). Other gene networks enriched in protein-coding mRNAs deregulated in metastatic tumor samples were "cell cycle, genetic disorder, metabolic disease" (21 genes, $p < 10^{-33}$), "cardiovascular system development and functions, embryonic development and tissue development" (21 genes, $p < 10^{-30}$) and "cancer,

tumor morphology and genetic disorder" (19 genes, $p < 10^{-29}$). IPA analysis also highlighted the prevalence of genes related to cell death within the metastasis-signature. It comprised 42 protein-coding mRNAs related to apoptosis (19 down-regulated and 23 up-regulated), in line with the notion that perturbation of the normal programmed cell death is involved in the metastatic phenotype in pancreatic cancer [6,55]. Interestingly, we found 6 intronic lncRNAs mapped to *locus* of apoptosis-related genes among those present in the metastasis signature (*ATF2*, *TGFβR2*, *MAP2K5*, *MAP3K1*, *DAPK1* and *PTEN*).

Metastasis-associated intronic lncRNAs are expressed with antisense and/or sense orientation relative to corresponding protein-coding genes

To document in more detail the structure of lnc RNAs mapping intronic regions of gene *loci* related to the MAPK pathway or to apoptosis, which were over-represented in the metastasis-signature, we investigated their orientation relative to the corresponding protein-coding mRNA. Orientation-specific RT-PCR was employed to determine the strandedness of the eleven intronic transcripts mapping to introns of *MAPK* pathway or/and apoptosis-related genes, namely *ARRB1*, *ATF2*, *MAPK1*, *MAP2K5*, *MAP3K1*, *MAP3K14*, *PPP3CB*, *RAPGF2*, *TGFβR2*, *DAPK1* and *PTEN*. These experiments were performed using total RNA isolated from pancreatic tumor tissue samples or from cultured MIA PaCa-2 cells. Ten transcripts showed evidence of being transcribed with the same (sense) orientation of the corresponding protein-coding mRNA in both pancreatic tissue samples and MIA PaCa-2 cells (Figure 5). Interestingly, a transcript with antisense orientation relative to the protein-coding mRNA was detected in an intronic region of *PPP3CB* in MIA PaCa-2 cells. When RNA from pancreatic tumors was used, antisense intronic transcription was detected in three additional *loci* (*ATF2*, *TGFBR2* and *MAP3K1*), which produced both sense and antisense messages (Figure 5).

The relative abundance of the eleven intronic lncRNAs identified in genes from the *MAPK* pathway or related to apoptosis was evaluated by quantitative Real-Time PCR in RNA samples isolated from primary tumors and distant metastasis. We initially measured the abundance of each of the 11 intronic transcripts in three samples of primary adenocarcinoma and three samples of metastasis. In spite of a great variability due to small sample size, 7 out 11 intronic transcripts showed a similar expression change (same direction) as measured in the microarray. Since the amount of RNA from clinical samples were limiting, we selected for further validation in additional samples three intronic lncRNAs, being one antisense (*PPP3CB*) and two with the same orientation (*MAP3K14* and *DAPK1*) relative to the protein-coding gene. As shown in Figure 6, statistically significant increased expression of all three intronic lncRNAs was observed.

We next asked if the expression changes of these intronic transcripts would reflect in the expression of

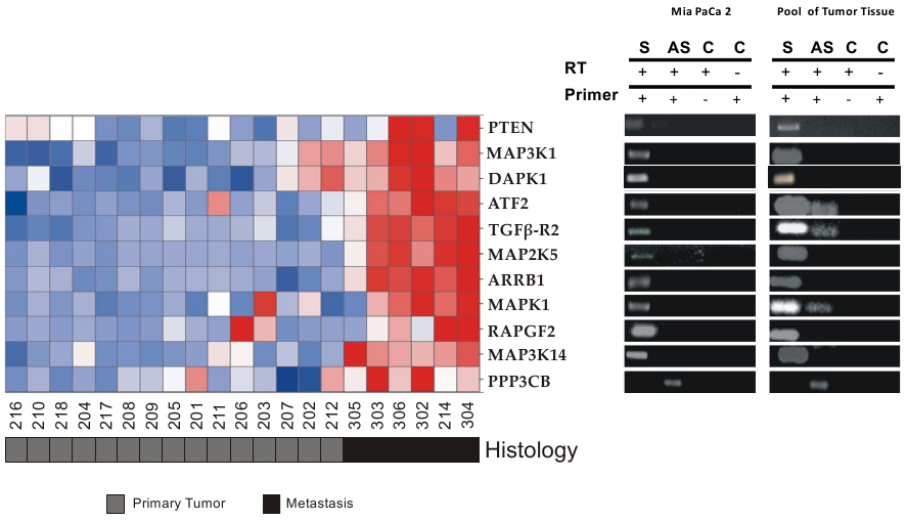


Figure 5 LncRNAs deregulated in pancreatic cancer metastasis map to intronic regions of genes associated with the MAPK pathway and/or related to apoptosis. Transcriptional orientation of eleven intronic lncRNAs mapping to MAPK pathway and/or apoptosis-related gene ***loci*** (***ARRB1***, ***ATF2***, ***MAPK1***, ***MAP2K5***, ***MAP3K1***, ***MAP3K14***, ***PPP3CB***, ***RAPGF2***, ***TGFβR2***, ***DAPK1*** and ***PTEN***) was investigated by strand-oriented reverse transcription followed by PCR. For each gene, sense (S) or antisense (AS) transcription was measured. Controls (C) for the absence of self-annealing during reverse transcription (RT) were obtained by performing RT reactions in the absence of primers (RT+, Primer -). Controls for the absence of genomic DNA contamination were obtained by omitting reverse transcriptase in the RT reaction (RT-, Primer +).

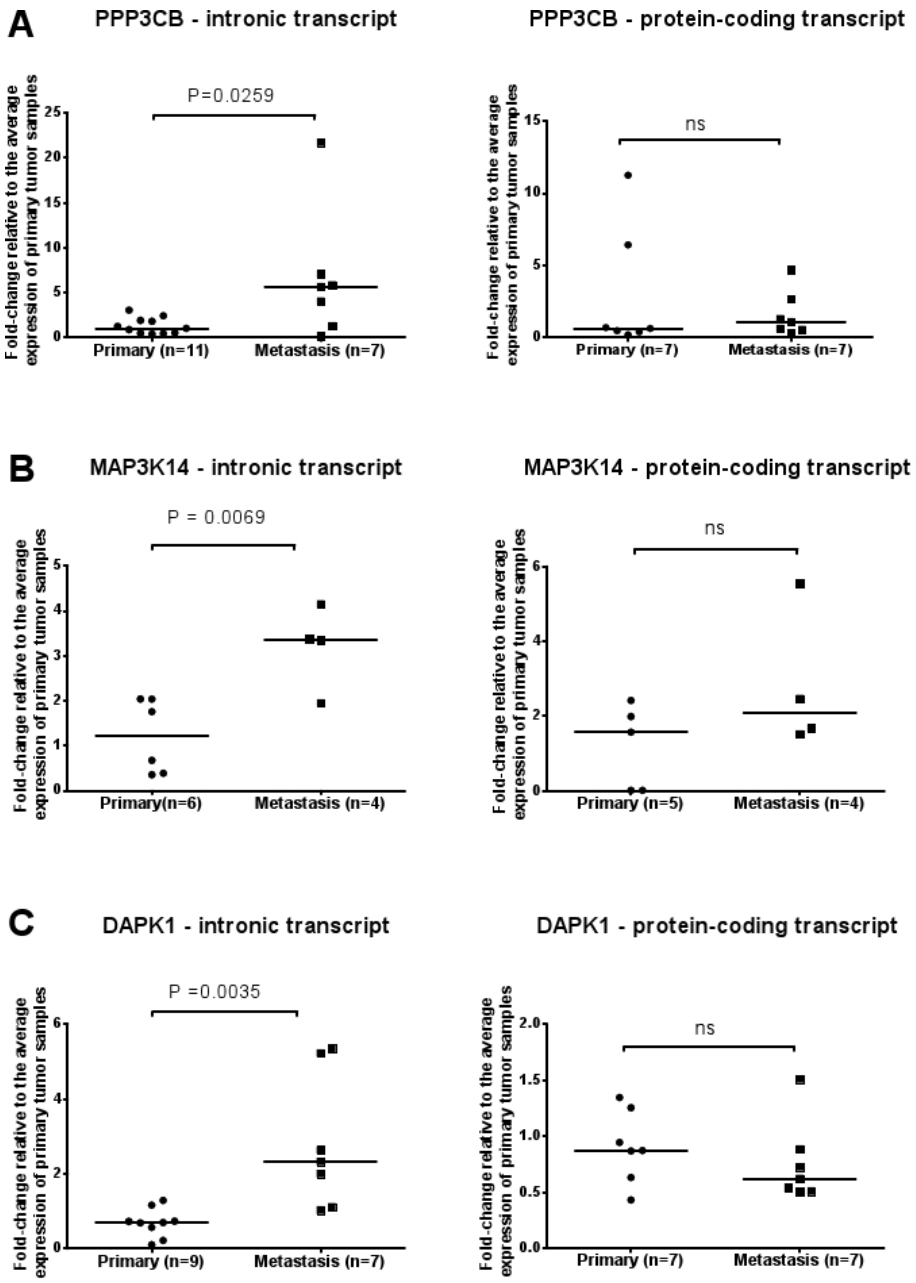


Figure 6 Expression changes of intronic lncRNAs mapping to *PPP3CB*, *MAP3K14* and *DAPK1 loci* are not accompanied by changes in the corresponding protein-coding mRNAs. Relative levels of intronic lncRNAs and protein-coding mRNAs from ***PPP3CB*** (A), ***MAP3K14*** (B) and ***DAPK1*** (C) ***loci*** were determined by Real-Time PCR in clinical samples of primary adenocarcinoma (circles) or distant metastasis (squares) from pancreatic cancer patients. For each transcript, the number of tested samples is indicated in the × axis. For each gene, results are expressed as fold-change relative to the average expression of primary tumor samples. For each sample, qPCR assays were performed in triplicate and mean values are shown. House-keeping gene HMBS was used as the endogenous control for normalization across patient samples. All intronic transcripts (left panels) were differentially expressed in metastatic samples at a significance threshold of $p < 0.05$ (***PPP3CB***, $p = 0.0259$; ***MAP3K14***, $p = 0.0069$; DAPK1, $p = 0.0035$). Protein-coding transcripts (right panels) were not significantly differentially expressed.

the corresponding protein-coding mRNA. Quantitative RT-PCR experiments with primers interrogating protein-coding mRNAs from *PPP3CB, MAP3K14* and *DAPK1 loci* showed no statistically significant expression change between primary tumors and metastasis samples. To investigate the co-expression of protein-coding mRNAs and intronic lncRNAs from the same *loci* we measured the Pearson correlation of their expression

measurements in all samples tested. A high Pearson correlation (r = 0.78, $p < 0.013$) was observed for the *MAP3K14 locus*, suggesting that the intronic sense transcript may be a by-product of pre-mRNA processing of the protein-coding transcript (see discussion for details). No significant correlation between protein-coding mRNAs and intronic lncRNAs was observed for the other two *loci* ($p > 0.05$), leaving open the possibility that intronic RNAs mapping to *PPP3CB* (antisense) and *DAPK1* (sense) *loci* are noncoding RNAs originated from independent transcriptional events.

Discussion

In this work we investigated gene expression profiles from clinical samples of pancreatic cancer using a custom cDNA microarray enriched in probes that interrogate long potentially noncoding RNAs mapping to intronic and intergenic regions of the human genome, plus a collection of protein-coding genes previously associated with cancer in the literature. By comparing expression profiles of 38 pancreatic clinical samples with four distinct tissue histologies (primary adenocarcinoma, adjacent non-tumor tissue, chronic pancreatitis, metastasis), we detected in all types of pancreatic tissues studied a proportion of intronic and intergenic transcripts comparable to the one observed for protein-coding mRNAs. There are several reports of aberrant expression of microRNAs [21-24], but this is to our knowledge the first time that the expression of lncRNAs has been studied in pancreatic cancer.

We observed that most intronic and intergenic transcripts expressed in pancreatic tissues have little or no coding potential (96% of total). Comparison with sequence contigs resulting from the assembly of EST/mRNA data produced in our group [26] showed that these transcripts have a mean size of at least 779 nt, being longer than the EST probes deposited in the microarrays, which represent indeed only parts of longer noncoding RNAs transcribed from intronic regions. Most putative intergenic transcripts (~81%) were located more than 1 kb apart from an UTR of an annotated gene, suggesting that for the most part, these are indeed intergenic transcripts rather than uncharacterized untranslated regions of incomplete mRNAs.

While it is clear that lncRNAs may exert diverse cellular functions through multiple molecular mechanisms [12,56,57], it has been suggested that a fraction of the transcriptome noncoding complement may correspond to transcriptional noise resulting from RNA polymerase activity in regions of open chromatin or intronic segments of processed mRNAs [58]. Our expression measurements of intronic lncRNAs do not permit to distinguish between i) intron lariats resulting from splicing of a pre-mRNA or ii) independent transcriptional units located within intron-annotated genomic regions. We have focused on poly(A+)-selected RNA fractions followed by oligo-dT primed reverse transcription to minimize the chance of labeling targets from non-polyadenylated spliced lariats. We argue that the identification of subsets of transcripts that map to intronic regions and whose steady-state levels allows the detection by microarrays indicate that these are not rapidly turned-over intron lariats. We have also performed a series of analysis to obtain additional evidence to support the notion that intronic/intergenic lncRNAs detected in pancreatic tissues are indeed *bona fide* cellular transcripts, as discussed below.

We first sought independent confirmation of intronic/intergenic lncRNA expression using RNAseq data generated from 9 distinct tissue libraries [30]. We found that approximately 80% of intronic/intergenic lncRNAs detected in pancreatic tissues were also detected in at least one RNAseq library (Figure 1). Most transcripts confirmed by the RNAseq data were detected i) only in a single tissue type other than pancreas, or ii) in all 9 tissue libraries plus pancreas, indicating the prevalence of subsets of noncoding transcripts with broad or specific tissue-type expression patterns, respectively (Figure 1).

While only a fraction of the intronic/intergenic lncRNAs expressed in pancreatic tissues overlapped evolutionarily conserved DNA elements in vertebrates, mammals and primates, we observed a significant enrichment ($p < 0.05$) compared to randomly selected control regions. This result suggests that at least a fraction of these lncRNAs are under purifying selection in the vertebrate lineage and therefore must be biologically functional. For the remaining transcripts, absence of sequence conservation should not be taken as evidence of no biological relevance, since it is known that well-characterized functional lncRNAs are poorly conserved across their global sequence [59].

As proposed by Washielt et al. [60], mapping conserved RNA secondary structure may lead to the discovery of novel functional lncRNAs. We found that a small fraction of lncRNAs expressed in pancreas (15%. i.e. 49/335) are predicted to form stable structural domains that could be important for their processing or biological function. It is well documented in the literature that small regulatory RNAs can be generated by processing of long RNA precursors transcribed from intronic and intergenic regions of the genome [56]. To ask what fraction of our set of lncRNAs expressed in pancreatic tissues could be precursor of small RNAs we compared their sequences to those of known microRNA and snoRNA [32,33]. Only a discrete overlap was found, indicating that long intronic/intergenic transcripts are predominantly not precursors of known microRNAs/

snoRNAs, yet leaving open the possibility that these transcripts may represent precursors of uncharacterized novel small RNAs.

We found significant enrichment of H3K4me3, a promoter-associated chromatin mark frequently found in RNA Pol II transcribed regions [35,61], in the vicinity (up to 2 kb) of intronic ($p < 0.05$) noncoding transcripts as compared to randomly selected genomic DNA sequences. A comparable H3K4me3 enrichment was observed nearby known protein-coding transcripts, suggesting that transcription of protein-coding mRNAs and intronic lncRNAs initiates at promoter regions with similar chromatin contexts. We also observed a significant enrichment of CAGE tags proximal to known start sites of intronic lncRNAs expressed in pancreatic tissues, corroborating the notion that at least a fraction of these is independent transcriptional units. Since pancreatic tissues were absent from the study that generated the CAGE tags used for cross-reference, these results possibly underestimate the co-localization of intronic/intergenic lncRNAs with *bona fide* transcription start sites of capped transcripts.

Differently from protein-coding mRNAs, we did not find significant enrichment of CpG island in the vicinity of intronic and intergenic RNA sequences expressed in pancreatic tissues. Based on this observation, we propose that methylation of CpG islands is not involved in the transcriptional regulation of most intronic/intergenic lncRNAs expressed in pancreatic tissues. Nonetheless, the full set of observations regarding the structure, conservation and genomic context argues that at least a fraction of intronic/intergenic transcripts detected in pancreatic tissues are independent transcriptional units rather than transcriptional noise originated from random Pol II firing [62], prompting us to investigate in more detail their relative expression levels in tumor and non-tumor pancreatic tissues.

Differential expression of intronic lncRNAs in prostate and renal cancer has already been documented [17,28]. Here we extend these observations to pancreatic cancer, asking whether there were sets of intronic/intergenic lncRNAs deregulated in clinical samples of pancreatic tumor. Comparing expression profiles from primary tumors with samples from histologically non-malignant pancreatic tissue and chronic pancreatitis (CP) we identified a 147-gene signature correlated with primary pancreatic tumor. This strategy was devised to favor the identification of tumor specific markers rather than transcripts associated with the stromal cell component, which is augmented in both tumor and CP samples [36,37]. We sought to validate the pancreatic cancer expression signature by performing a meta-analysis with published gene expression studies of pancreatic cancer. Only 23% of the protein-coding mRNAs present in our pancreatic cancer signature were also identified in other reports. This modest overlap can be accounted for by differences in platforms and the heterogeneity of pancreatic tumor samples. Notwithstanding, we observed a high agreement (17/24, 71%) between the expression changes measured in our signature and those retrieved from published data, which provides independent support for our result and validates our sample set and methodological approach. This set included genes already reported in the literature as differentially expressed in pancreatic cancer and that have been investigated as biomarkers for pancreatic cancer (i.e. *S100A6* [47], *S100P* [46], *TIMP1* [63] and *NF-κB* [64]). In agreement with previous findings [5], the analysis of gene enriched categories in the pancreatic cancer expression signature indicated the over-representation of genes involved in focal adhesion. Over-representation of focal adhesion genes in the pancreatic cancer signature is suggestive that deregulation of genes encoding proteins involved in the connection and signaling to the extracellular matrix plays an important role in the malignant transformation and/or maintenance of pancreatic adenocarcinomas. This set included integrin beta 5 (*ITGB5*), which we found to be upregulated in pancreatic adenocarcinoma. Itgb5 protein has been investigated as diagnostic biomarker in non-small cell lung cancer [65] and is target of the inhibitor drug EMD121974, which is under clinical trial [66]. Thus, *ITGB5* is an attractive candidate to be tested as biomarker and/or new drug target in pancreatic cancer.

Interestingly, a significant fraction (29%) of the 147-gene signature correlated with primary pancreatic tumor was comprised by lncRNAs mapping to intronic or intergenic regions, suggesting that noncoding RNAs could exert roles related to tumorigenesis of pancreatic cancer. This result prompted us to investigate the existence of subsets of lncRNAs with expression levels altered in metastatic samples.

We identified a statistically significant metastasis signature of 355 differentially expressed transcripts that includes 220 protein-coding genes, 134 intronic/intergenic transcripts and 6 known lncRNAs (Figure 4 and Additional file 5, Table S3). In addition to protein-coding genes previously shown to be deregulated in pancreatic metastasis (7 out of 19), the metastasis signature comprises known genes already associated to metastasis in other types of cancer (Additional file 7, Table S4), thus pointing to potentially interesting candidates for testing as new targets for treatment of the metastatic disease in pancreatic cancer.

The significant fraction of lncRNAs in the metastasis signature (38% of total) suggests that deregulation of these lncRNAs could also be associated with the metastatic process. Expression changes of protein-coding

mRNAs from genes of the MAPK pathway has already been described in pancreatic carcinoma [67-69]. Here we found 9 intronic lncRNAs mapped to genes correlated to the MAPK pathway in the metastasis signature. We also identified expression changes in gene *loci* related to apoptosis, including 42 protein-coding mRNAs and 6 intronic lncRNAs; this pathway was one out of 12 described by Jones *et al.* [6] as genetically altered in pancreatic cancer. Four intronic lncRNAs belong to both categories. These results prompted us to document in more detail the nature of the 11 transcripts mapping to intronic regions of gene *loci* associated with the MAPK pathway or related to apoptosis, i.e., their relative orientation to the corresponding protein-coding mRNAs.

Strand-specific RT-PCR assays using RNA aliquots from tumor tissue samples showed that 4 intronic transcripts have antisense orientation relative to the protein-coding mRNA: *PPP3CB*, *ATF2*, *TGFBR2* and *MAPK1*. Antisense transcripts originated in *PPP3CB* intronic regions were also detected in MIA PaCa-2 cells. The antisense orientation relative to the corresponding protein-coding mRNA provide strong evidence to support that these noncoding RNAs are produced from independent transcriptional units, possibly under control of a different promoter region.

Transcripts mapping to intronic regions with the same orientation of the corresponding protein-coding mRNA were detected in the *ATF2*, *TGFBR2* and *MAPK1*, as well as in the 7 other gene *loci* tested (*ARRB1*, *MAP3K1*, *MAP3K14*, *MAP2K5*, *PTEN*, *DAPK1* and *RAPGF2*), in both tissue and MIA PaCa-2 RNA samples. These sense-oriented intronic transcripts could indeed be *bona fide* RNAs originated from independent transcription, but also result from reverse transcription of unprocessed mRNA precursors or of stable RNA lariats generate during pre-mRNA splicing. Further experiments will be necessary to determine the precise nature of these sense-oriented intronic RNAs.

The relative abundance of two sense (*DAPK1, MAP3K14*) and one antisense-oriented (*PPP3CB*) intronic transcripts in samples of primary pancreatic adenocarcinoma and pancreatic metastases was independently accessed by qRT-PCR, confirming the results measured in the microarray hybridizations. Four additional intronic lncRNAs showed concordant results between qRT-PCR and the microarrays (*ARRB*, *RAPGF2*, *ATF2* and *PTEN*). Expression changes of 4 intronic lncRNAs were not concordant between qRT-PCR and microarray (*MAP3K1*, *TGFBR2*, *MAP2K5* and *MAPK1*). The amount of RNA and the number of patient tissue samples available for the qRT-PCR experiments were limiting, and the marginally significant and non-validated lncRNA candidates were tested only in few samples in an initial round of validation. It is possible that some of the intronic lncRNA candidates that failed the initial round of validation would still be validated as differentially expressed if tested in additional tissue samples. However, an alternative explanation for the non-validation of some candidates is the presence of array hybridization artifacts such as cross-hybridization or target amplification biases.

Intragenic lncRNAs have been shown to modulate in *cis* the expression of mRNAs expressed in the same *locus* [29,70,71]. We measured the relative abundance of mRNAs produced in the *PPP3CB*, *DAPK1* and *MAP3K14 loci* in the same samples and did not observe statistically significant expression differences between primary tumors and metastasis. This result indicates that intronic RNAs produced in these *loci* do not affect in *cis* the abundance of the corresponding protein-coding transcripts. This conclusion is also supported by the absence of significant correlation between expression levels of protein-coding and noncoding RNAs originating from *PPP3CB* and *DAPK1 loci*. The possibility that intronic lncRNAs differentially expressed in metastatic samples may exert regulatory functions acting in *trans* is compelling and warrants further studies.

It has been shown that a significant portion of the noncoding component of the human transcriptome is comprised of non-polyadenylated RNAs [10]. We note that our analysis was limited to the set of lncRNAs interrogated by the array platform (Table 1) and by the use of poly(A+)-enriched RNA, and therefore is not comprehensive in terms of describing the full complement of lncRNAs expressed in pancreatic tissues. Thus, additional studies using unbiased approaches such as RNAseq or tiling arrays will be required to catalog all poly(A+) and poly(A-) transcripts expressed in pancreatic tissues with distinct degrees of malignancy and for the identification of novel regulatory lncRNA candidates involved in the malignant transformation and tumor progression.

Conclusions

In this work we report that noncoding RNAs originating from intronic and intergenic genomic regions are expressed in tumor and non-tumor pancreatic tissues. Enrichment of promoter-associated chromatin marks plus the observation of antisense orientation of intronic transcripts relative to mRNAs expressed from the same *loci* provide evidence that these messages are not by-products of random transcription or pre-mRNA splicing but rather, are independent transcriptional units. Further investigation will be required to determine the biogenesis of these lncRNAs.

We identified gene expression signatures correlated to primary and metastatic stages of pancreatic cancer,

which in addition to protein-coding mRNAs comprise collections of long intronic and intergenic noncoding RNAs. Further studies will be necessary to reveal possible biological functions and molecular mechanisms exerted by these lncRNAs in tumorigenesis and/or progression of pancreatic tumors.

In summary, our work contributes with novel candidate biomarkers of pancreatic cancer and highlights the importance of investigating the biological relevance of long noncoding RNAs in order to fully understand the molecular basis of the disease.

Methods

Patient Samples and Cell Lines

A total of 38 pancreatic samples stored in freshly-frozen tissue collections were obtained with informed consent from patients seen at Hospital das Clínicas, Faculdade de Medicina da Universidade de São Paulo (HC-FMUSP). Primary tumor tissues (T) were obtained from 15 patients with no evidence of metastasis. Nine samples of histologically normal pancreatic tissue fragments (NT) were dissected from non-neoplasic tissue sections adjacent to tumor sites. Six samples of metastases originated from primary pancreatic tumors (M) were obtained from biopsies in affected organs (one from peritoneum, one from ganglion and four from liver tissues). Eight tissue samples from patients with chronic pancreatitis (CP) were also collected. All tissue sections were reviewed by a pathologist for histological confirmation and whenever necessary, macro-dissected to guarantee that 80% or more of the sections used for gene expression analysis were composed of neoplastic/pancreatitis tissue.

Pancreatic carcinoma cell lines MIA PaCa-2 were obtained from the American Type Culture Collection and maintained using DEMEM supplemented with 10% (v/v) fetal calf serum (FCS), 3 mM L-glutamine, 100 μg/ml streptomycin and 100 U/ml penicillin.

RNA extraction and microarray target preparation

Total RNA was extracted from pancreatic tissue samples (50-100 mg tissue) using Trizol (Invitrogen) according to manufacturer's recommendations. RNA cleanup including a DNase I digestion step was performed using RNeasy spin columns (Qiagen). RNA integrity was measured by the relative abundance of 28S/18S ribosomal subunits, verified through micro fluid capillary electrophoresis (Agilent Bioanalyzer 2100).

To generate cRNA targets, 1 μg of total RNA from each sample was linearly amplified in two rounds of reverse transcription followed by *in vitro* transcription according to Wang et al. [72]. Briefly, oligo dT-T7 was used to prime first-strand cDNA synthesis (SuperScript III First Strand Synthesis - Invitrogen). After second-strand cDNA synthesis (cDNA Polymerase Mix - Clontech), cRNA targets were produced by *in vitro* transcription (MegaScript T7 - Ambion). In the second round of amplification, cRNAs produced in the first-round were reverse transcribed using random hexamer primers and used as template for *in vitro* transcription in the presence of amino-allyl UTP. Prior to hybridization, cRNA targets were labeled by coupling with mono-reactive Cy5-esters (Amersham). Quantification of cDNA yield and dye incorporation was performed using a NanoDrop spectrophotometer (Thermo Scientific). Typically, 50-100 μg of cRNA were obtained following two rounds of linear amplification. Two sets of cRNA targets were generated from each RNA sample and independently hybridized to microarray slides.

Microarray design and hybridization

Construction of the spotted custom-cDNA microarray platform was described previously [17]. Probes were selected from the over 1 million EST clone collection generated during the Human Cancer Genome Project, a large-scale EST sequencing project that used cDNA libraries generated from poly(A) mRNA derived from over 20 different types of human tumors [73,74]. Transcripts from the EST dataset were annotated as protein-coding, putative intronic lncRNA or intergenic lncRNA following mapping to the human genome sequence and cross-referencing with genome mapping coordinates of annotated genes (RefSeq dataset) [17]. Intronic/intergenic lncRNAs used for microarray spotting were randomly selected from the annotated EST dataset. Transcripts annotated as "intronic lncRNAs" comprise sequences that mapped within an intronic region of a protein-coding gene. Transcripts annotated as "intergenic lncRNAs" comprise sequences that map to genomic regions devoid of any annotated gene. To be annotated as intronic or intergenic a given transcript could not overlap a genomic region spanning an exon of annotated protein-coding genes. "Known lncRNAs" refer to transcripts whose genomic coordinates overlap fully with the coordinates of noncoding RNAs from the RefSeq dataset (accession Id = NR_nnnnnn). Transcripts annotated as "protein-coding gene" overlapped with exons of protein-coding RefSeq transcripts in genomic space. To account for possible unannotated intron retention events, partial transcripts mapping to exon/intron boundaries were annotated as "exonic".

In the course of this work microarray probes were re-mapped to the latest version of the human genome (hg19) and re-annotated to reflect updated RefSeq and UCSC gene models (Oct. 2010). Each glass-slide contains 3,355 cDNA fragments spotted in duplicate, plus positive (cDNA from housekeeping genes) and negative (plant and bacterial DNA) controls. Spotted cDNAs

comprise 722 ESTs mapping to intronic regions of well-annotated (RefSeq) protein-coding genes, 74 ESTs mapping to known RefSeq lncRNAs, 188 ESTs mapping to intergenic regions of the genome. The array also contained 2,371 ESTs mapping to exons of protein-coding genes associated with cancer based on a literature search [17], comprising genes involved in apoptosis, tumorigenesis, metastasis, cancer metabolism and cancer progression.

For each sample, cRNA targets were ressuspended in a final volume of 200 μl of 1× Microarray Hybridization Solution v.2.0 (GE Healthcare) containing 25% formamide, denaturated at 92°C for 2 minutes and incubated with microarrays at 42°C for 16 hours using an automated slide processor (GE Healthcare). Following sequential washes in 1× SSC; 0.2% SDS, 0.1× SSC; 0.2% SDS and 0.1× SSC; 0.2% SDS, microarray slides were scanned immediately in a Generation III Microarray Systems Scanner (Molecular Dynamics/GE Healthcare). For each sample, two slides were hybridized with different preparations of cRNA targets. As probes are spotted in duplicate in the arrays, a total of 4 replicate measurements were collected for each cDNA for each sample.

Data processing and analysis

Cy5-intensity measurements from hybridized targets were extracted from array images using the ArrayVision software (Imaging Research Inc.). To make the expression values comparable across all samples tested, the raw data was normalized by the quantile method [75]. Next, for each slide, the fifty percent of probes with the lowest intensity values were filtered out. A mean expression value for each probe, in each patient sample, was calculated when at least 3 out of 4 replicates showed valid measurements. Only probes with valid measurements in at least 75% of the samples in any of the histological groups (NT, T, PA or M) were selected, resulting in a total of 1,607 probes for further analysis. The ComBat program was used to remove systematic variations in gene expression across experiments resulting from the use of different batches of microarrays [76]. Inter-slide Pearson correlations using normalized intensities from all probes in the array were calculated before and after filtering and normalization of data intensities. Raw data intensities showed average inter-slide correlations of 0.63, whereas normalized data showed inter-slide correlation of 0.83.

Raw and normalized microarray intensities were deposited in the Gene Expression Omnibus database (GEO - http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE30134.

Significance analysis of microarrays (SAM) approach [77] was employed to identify gene expression signatures correlated to tissue histology, with the following parameters: two or multi-class response, 1000 permutations, K-Nearest Neighbors Imputer, and false discovery rate (FDR) ≤ 10% or ≤5%. For representation of gene expression measurements in heat-maps, samples were ordered according to their individual correlation to the average profile of the primary tumor samples.

Bioinformatics analyses

We used the BEDTools software package [78] to cross-reference genome mapping coordinates (GRCh37 build, hg19) of our dataset of intronic/intergenic noncoding sequences with those from the various datasets used in this analysis and available through the UCSC Genome Browser [66]: i) RNAseq data of PolyA$^+$ RNA-derived libraries from 9 tissues [30], ii) RIKEN CAGE tag data from PolyA$^+$ RNA-derived libraries from 6 cell lineages [79]; iii) ChIP-seq data of H3K4m3 DNA binding sites [35], iv) conserved DNA elements in vertebrates, mammalian and primates calculated with PhastCons program [66], v) predicted CpG islands [80] and vi) intronic noncoding RNAs assembled from EST/mRNA GenBank data [26].

To test the statistical significance of the overlap between our dataset of intronic/intergenic lncRNAs and the datasets of conserved elements and regulatory motifs (H3K4me3, CpG islands, CAGE tags), we generated 100 groups of randomly selected sequences from intronic or intergenic regions of the human genome matching in number, length and CG content our set of expressed noncoding sequences. As pre-processing of CAGE tag data, coordinates of overlapped tags were clustered and only clusters containing at least 5 tags were considered for further analysis. Fischer's exact test was used to test the statistical significance ($p < 0.05$ threshold) of the enrichment of conserved DNA elements in intronic/intergenic lncRNAs relative to the enrichment observed for the 100 random sequence sets.

For the analysis of transcription regulatory elements, we first computed the distance of the closest H3K4me3 marks, predicted CpG islands and CAGE tags to our set of expressed intronic/intergenic lncRNAs, expressed protein-coding mRNAs, and of the set of 100 random groups. Transcription regulatory elements mapping to 5'UTRs of known transcripts (RefSeq and UCSC genes) were removed to avoid the contribution of signals at start sites of known genes to the enrichment of regulatory elements at start sites of intronic lncRNAs mapping nearby.

Only regulatory elements distant up to 10 kb of sequence boundaries were considered. Next, non-parametric Kolmogorov-Smirnov (KS) test statistics, implemented using the Deducer package under R language [81] was used to compare the distance distributions of H3K4me3 marks, predicted CpG islands and CAGE tags

observed for intronic or intergenic noncoding sequences and those calculated for each of the 100 control random sets. Distance distributions of regulatory motifs from intronic/intergenic lncRNAs/protein-coding mRNAs were considered significantly different from the obtained by chance if all KS *p-values* calculated using each random set were smaller than 0.05.

Protein-coding potential of intronic and intergenic lncRNAs was evaluated using the Coding Potential Calculator software [31] with default parameters. RNAz program [34] was used to predict structurally conserved and thermodynamically stable RNA secondary structures. Only predicted structures with P > 0.5 were considered as containing conserved secondary structures.

Orientation-specific RT-PCR

Aliquots of DNAse-treated total RNA from a pool of six pancreatic tumor tissues samples or from Mia PaCa-2 cells were used as template in orientation-specific reverse transcription reactions. Reactions were performed with 200 ng total RNA plus 2.5 μM of oligonucleotide primers designed to detect sense or antisense strand intronic transcripts, relative to the orientation of the mRNA from the same *locus*. SuperScript III™ Super Mix (Invitrogen) was used according manufacture's recommendations, with the following modification: reverse transcription reaction was increased to 57°C to limit RNA self-annealing. To verify the absence of priming due to self-annealing or genomic DNA contamination, control reactions were performed without addition of primers or of reverse transcriptase, respectively.

Real-time RT-PCR

One microgram aliquots of DNase-treated total RNA from 11 clinical samples of primary pancreatic tumors (5 already used in the microarray experiments and 6 new samples) and 6 of distant metastases with pancreatic origin (4 already used in the microarray and 2 new samples) were reverse transcribed using SuperScript III™ Super Mix kit (Invitrogen) and random hexamer primers according to manufacturer's recommendation. Relative abundance of selected transcripts primary tumor/metastasis samples was determined by real-time PCR using the ABI PRISM® 7300 Real Time PCR System and the SYBR Green PCR Master Mix kit (Applied Biosystems). Reactions were performed in a final volume of 20 μl containing a 5 μl aliquot of diluted cDNA (1:7) and 1 μM of forward and reverse gene-specific primers. Expression levels of hydroxymethylbilane synthase (HMBS) [82] appeared to be constant and was used as a reference gene to make expression measurements comparable across all different samples tested. For quantitative results, the level of each transcript was normalized by the level of HBMS, and represented as fold change using the $2^{-\Delta\Delta Ct}$ method [83], where ΔΔCt = (Ct candidate gene in sample × - Ct reference gene in sample X)$_{sample}$ - mean ΔC_t of all primary tumor samples tested.

Additional material

Additional File 1: Figure S1: Intronic/intergenic lncRNA are enriched in conserved DNA elements. Genomic coordinates of conserved DNA elements identified in Vertebrate, Placental or Primate sequences (see *Methods* for details) were cross-referenced to those from lncRNAs expressed in pancreatic tissue (blue bars). These presented a higher overlap with evolutionary conserved DNA elements than the observed for a random set of genomic DNA sequences with same length and CG content (gray bars) (Fisher's test p <0.05). Bar heights indicate the number of lncRNAs that overlap conserved DNA elements in each taxonomic group divided by the total number of conserved DNA elements present in each group.

Additional File 2: Table S1: List of transcripts differentially expressed in PDAC relative to chronic pancreatitis and non tumor tissue samples combined.

Additional File 3: Table S2: Validation of protein-coding genes differentially expressed in PDAC by meta-analysis of published data.

Additional File 4: Figure S2: Genes modulated in pancreatic cancer are involved in cellular movement, cell-to-cell signaling interactions and endocrine system. Ingenuity Pathway Analysis was used to identify gene networks over-represented among the set of 104 protein-coding genes differentially expressed in PDAC samples. The most significantly enriched network ($p = 10^{-43}$) is comprised by 23 differentially expressed genes measured in the microarrays. Red indicates higher expression, and green, lower expression in tumor tissues relative to chronic pancreatitis and adjacent non-tumor tissue samples combined.

Additional File 5: Table S3: List of genes differentially expressed between PDAC and metastatic tissue samples.

Additional File 6: Figure S3: Genes modulated in metastatic pancreatic cancer are involved in cellular movement, gene expression and immune cell trafficking. Ingenuity Pathway Analysis was used to identify gene networks over-represented among the set of 221 protein-coding genes differentially expressed in PDAC relative to metastasis tissue samples. The most significantly enriched network ($p= 10^{-41}$) is comprised by 24 differentially expressed genes measured in the microarrays. Red indicates higher expression, and green, lower expression in metastasis relative to PDAC tissue samples.

Additional File 7: Table S4: Genes from the pancreatic cancer metastasis signature related to tumor aggressiveness in other cancer types.

Acknowledgements

The authors wish to thank Yuri J. B. Moreira and Rodrigo L. Borges for their support on bioinformatics analysis. This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil. ACT, VMC and BD received fellowships from FAPESP. SVA and EMR received established investigator fellowships from CNPq.

Author details

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brasil. [2]Departamento de Gastroenterologia (LIM-37), Faculdade de Medicina, Universidade de São Paulo, 01246-903, São Paulo, SP, Brasil.

Authors' contributions

Conceived and designed the experiments: EMR, MCCM, MSK, ACT. Performed the experiments: ACT, MFF, BD,VMC and RSF. Analyzed the data: ACT, SVA, EMR. Wrote the paper: ACT, EMR. All authors read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 1 June 2011 Accepted: 13 November 2011
Published: 13 November 2011

References

1. Hezel AF, Kimmelman AC, Stanger BZ, Bardeesy N, Depinho RA: **Genetics and biology of pancreatic ductal adenocarcinoma.** *Genes Dev* 2006, **20**:1218-1249.
2. Yokoyama Y, Nimura Y, Nagino M: **Advances in the treatment of pancreatic cancer: limitations of surgery and evaluation of new therapeutic strategies.** *Surg Today* 2009, **39**:466-475.
3. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, Kloppel G, Saeger HD, Pilarsky C: **Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.** *Oncogene* 2005, **24**:5079-5088.
4. Lopez-Casas PP, Lopez-Fernandez LA: **Gene-expression profiling in pancreatic cancer.** *Expert Rev Mol Diagn* 2010, **10**:591-601.
5. Harsha HC, Kandasamy K, Ranganathan P, Rani S, Ramabadran S, Gollapudi S, Balakrishnan L, Dwivedi SB, Telikicherla D, Selvan LD, *et al*: **A compendium of potential biomarkers of pancreatic cancer.** *PLoS Med* 2009, **6**:e1000046.
6. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, *et al*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321**:1801-1806.
7. Dinger ME, Amaral PP, Mercer TR, Mattick JS: **Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.** *Brief Funct Genomic Proteomic* 2009, **8**:407-423.
8. Ponting CP, Oliver PL, Reik W: **Evolution and functions of long noncoding RNAs.** *Cell* 2009, **136**:629-641.
9. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
10. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, Triche TJ: **The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA.** *BMC Biol* 2010, **8**:149.
11. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484-1488.
12. Mattick JS: **The genetic signatures of noncoding RNAs.** *PLoS Genet* 2009, **5**:e1000459.
13. Huarte M, Rinn JL: **Large non-coding RNAs: missing links in cancer?** *Hum Mol Genet* 2010, **19**:R152-161.
14. Gibb EA, Brown CJ, Lam WL: **The functional role of long non-coding RNA in human carcinomas.** *Mol Cancer* 2011, **10**:38.
15. Prensner JR, Chinnaiyan AM: **The Emergence of lncRNAs in Cancer Biology.** *Cancer Discovery* 2011, **1**:391-407.
16. Tinzl M, Marberger M, Horvath S, Chypre C: **DD3PCA3 RNA analysis in urine–a new perspective for detecting prostate cancer.** *Eur Urol* 2004, **46**:182-186, discussion 187.
17. Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, Egidio CM, Paquola AC, Machado AA, Festa F, *et al*: **Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer.** *Oncogene* 2004, **23**:6684-6692.
18. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, *et al*: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.** *Nature* 2010, **464**:1071-1076.
19. Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, *et al*: **MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer.** *Oncogene* 2003, **22**:8031-8041.
20. Panzitt K, Tschernatsch MM, Guelly C, Moustafa T, Stradner M, Strohmaier HM, Buck CR, Denk H, Schroeder R, Trauner M, Zatloukal K: **Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA.** *Gastroenterology* 2007, **132**:330-342.
21. Ho AS, Huang X, Cao H, Christman-Skieller C, Bennewith K, Le QT, Koong AC: **Circulating miR-210 as a Novel Hypoxia Marker in Pancreatic Cancer.** *Transl Oncol* 2010, **3**:109-113.
22. Li A, Omura N, Hong SM, Vincent A, Walter K, Griffith M, Borges M, Goggins M: **Pancreatic cancers epigenetically silence SIP1 and hypomethylate and overexpress miR-200a/200b in association with elevated circulating miR-200a and miR-200b levels.** *Cancer Res* 2010, **70**:5226-5237.
23. Hamada S, Shimosegawa T: **Biomarkers of pancreatic cancer.** *Pancreatology* 2011, **11(Suppl 2)**:14-19.
24. Steele CW, Oien KA, McKay CJ, Jamieson NB: **Clinical Potential of MicroRNAs in Pancreatic Ductal Adenocarcinoma.** *Pancreas* 2011, **40**:1165-1171.
25. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S, *et al*: **Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers.** *Science* 2011, **331**:593-596.
26. Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reis EM, 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 Biol* 2007, **8**:R43.
27. Reis EM, Ojopi EP, Alberto FL, Rahal P, Tsukumo F, Mancini UM, Guimaraes GS, Thompson GM, Camacho C, Miracca E, *et al*: **Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer.** *Cancer Res* 2005, **65**:1693-1699.
28. Brito GC, Fachel AA, Vettore AL, Vignal GM, Gimba ER, Campos FS, Barcinski MA, Verjovski-Almeida S, Reis EM: **Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma.** *Mol Carcinog* 2008, **47**:757-767.
29. Louro R, Nakaya HI, Amaral PP, Festa F, Sogayar MC, da Silva AM, Verjovski-Almeida S, Reis EM: **Androgen responsive intronic non-coding RNAs.** *BMC Biol* 2007, **5**:4.
30. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-476.
31. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.** *Nucleic Acids Res* 2007, **35**:W345-349.
32. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36**:D154-158.
33. Lestrade L, Weber MJ: **snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.** *Nucleic Acids Res* 2006, **34**:D158-162.
34. Gruber AR, Neubock R, Hofacker IL, Washietl S: **The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures.** *Nucleic Acids Res* 2007, **35**:W335-338.
35. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, *et al*: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-560.
36. Logsdon CD, Simeone DM, Binkley C, Arumugam T, Greenson JK, Giordano TJ, Misek DE, Kuick R, Hanash S: **Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer.** *Cancer Res* 2003, **63**:2649-2657.
37. Gress TM, Wallrapp C, Frohme M, Muller-Pillasch F, Lacher U, Friess H, Buchler M, Adler G, Hoheisel JD: **Identification of genes with specific expression in pancreatic cancer by cDNA representational difference analysis.** *Genes Chromosomes Cancer* 1997, **19**:97-103.
38. Masamune A, Shimosegawa T: **Signal transduction in pancreatic stellate cells.** *J Gastroenterol* 2009, **44**:249-260.
39. Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I: **Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia.** *Hepatogastroenterology* 2008, **55**:2016-2027.
40. Chelala C, Hahn SA, Whiteman HJ, Barry S, Hariharan D, Radon TP, Lemoine NR, Crnogorac-Jurcevic T: **Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets.** *BMC Genomics* 2007, **8**:439.

41. Crnogorac-Jurcevic T, Missiaglia E, Blaveri E, Gangeswaran R, Jones M, Terris B, Costello E, Neoptolemos JP, Lemoine NR: **Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent.** *J Pathol* 2003, **201**:63-74.
42. Buchholz M, Braun M, Heidenblut A, Kestler HA, Kloppel G, Schmiegel W, Hahn SA, Luttges J, Gress TM: **Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions.** *Oncogene* 2005, **24**:6626-6636.
43. Shekouh AR, Thompson CC, Prime W, Campbell F, Hamlett J, Herrington CS, Lemoine NR, Crnogorac-Jurcevic T, Buechler MW, Friess H, *et al*: **Application of laser capture microdissection combined with two-dimensional electrophoresis for the discovery of differentially regulated proteins in pancreatic ductal adenocarcinoma.** *Proteomics* 2003, **3**:1988-2001.
44. Iacobuzio-Donahue CA, Maitra A, Shen-Ong GL, van Heek T, Ashfaq R, Meyer R, Walter K, Berg K, Hollingsworth MA, Cameron JL, *et al*: **Discovery of novel tumor markers of pancreatic cancer using global gene expression technology.** *Am J Pathol* 2002, **160**:1239-1249.
45. Iacobuzio-Donahue CA, Maitra A, Olsen M, Lowe AW, van Heek NT, Rosty C, Walter K, Sato N, Parker A, Ashfaq R, *et al*: **Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays.** *Am J Pathol* 2003, **162**:1151-1162.
46. Deng H, Shi J, Wilkerson M, Meschter S, Dupree W, Lin F: **Usefulness of S100P in diagnosis of adenocarcinoma of pancreas on fine-needle aspiration biopsy specimens.** *Am J Clin Pathol* 2008, **129**:81-88.
47. Ohuchida K, Mizumoto K, Yu J, Yamaguchi H, Konomi H, Nagai E, Yamaguchi K, Tsuneyoshi M, Tanaka M: **S100A6 is increased in a stepwise manner during pancreatic carcinogenesis: clinical value of expression analysis in 98 pancreatic juice samples.** *Cancer Epidemiol Biomarkers Prev* 2007, **16**:649-654.
48. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**: R70.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
50. Guo S, Hakimi MA, Baillat D, Chen X, Farber MJ, Klein-Szanto AJ, Cooch NS, Godwin AK, Shiekhattar R: **Linking transcriptional elongation and messenger RNA export to metastatic breast cancers.** *Cancer Res* 2005, **65**:3011-3016.
51. **Ingenuity® Systems.** [http://www.ingenuity.com].
52. Tabata T, Tsukamoto N, Fooladi AA, Yamanaka S, Furukawa T, Ishida M, Sato D, Gu Z, Nagase H, Egawa S, *et al*: **RNA interference targeting against S100A4 suppresses cell growth and motility and induces apoptosis in human pancreatic cancer cells.** *Biochem Biophys Res Commun* 2009, **390**:475-480.
53. Aloysius MM, Zaitoun AM, Awad S, Ilyas M, Rowlands BJ, Lobo DN: **Mucins and CD56 as markers of tumour invasion and prognosis in periampullary cancer.** *Br J Surg* 2010, **97**:1269-1278.
54. Vlecken DH, Bagowski CP: **LIMK1 and LIMK2 are important for metastatic behavior and tumor cell-induced angiogenesis of pancreatic cancer cells.** *Zebrafish* 2009, **6**:433-439.
55. Ruckert F, Dawelbait G, Winter C, Hartmann A, Denz A, Ammerpohl O, Schroeder M, Schackert HK, Sipos B, Kloppel G, *et al*: **Examination of apoptosis signaling in pancreatic cancer by computational signal transduction analysis.** *PLoS One* 2010, **5**:e12243.
56. Wilusz JE, Sunwoo H, Spector DL: **Long noncoding RNAs: functional surprises from the RNA world.** *Genes Dev* 2009, **23**:1494-1504.
57. Louro R, Smirnova AS, Verjovski-Almeida S: **Long intronic noncoding RNA transcription: expression noise or expression choice?** *Genomics* 2009, **93**:291-298.
58. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "dark matter" transcripts are associated with known genes.** *PLoS Biol* 2010, **8**:e1000371.
59. Pang KC, Frith MC, Mattick JS: **Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function.** *Trends Genet* 2006, **22**:1-5.
60. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF: **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome.** *Nat Biotechnol* 2005, **23**:1383-1390.
61. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**:223-227.
62. Struhl K: **Transcriptional noise and the fidelity of initiation by RNA polymerase II.** *Nat Struct Mol Biol* 2007, **14**:103-105.
63. Pan S, Chen R, Crispin DA, May D, Stevens T, McIntosh MW, Bronner MP, Ziogas A, Anton-Culver H, Brentnall TA: **Protein Alterations Associated with Pancreatic Cancer and Chronic Pancreatitis Found in Human Plasma using Global Quantitative Proteomics Profiling.** *J Proteome Res* 2011, **10**:2359-2376.
64. Weichert W, Boehm M, Gekeler V, Bahra M, Langrehr J, Neuhaus P, Denkert C, Imre G, Weller C, Hofmann HP, *et al*: **High expression of RelA/p65 is associated with activation of nuclear factor-kappaB-dependent signaling in pancreatic cancer and marks a patient population with poor prognosis.** *Br J Cancer* 2007, **97**:523-530.
65. Dingemans A-M, Vosse B, Van den Boogaart V, Griffioen A, Thijssen V: **Profiling of integrin (ITG) expression in early stage non-small cell lung cancer (NSCLC).** *AACR Meeting Abstracts* 2008, **2008**:2161-.
66. , ClinicalTrials.gov identifier: NCT00077155 [http://clinicaltrials.gov/ct2/show/NCT00077155?intr=Cilengitide&rank = 20].
67. Zhao Y, Shen S, Guo J, Chen H, Greenblatt DY, Kleeff J, Liao Q, Chen G, Friess H, Leung PS: **Mitogen-activated protein kinases and chemoresistance in pancreatic cancer cells.** *J Surg Res* 2006, **136**:325-335.
68. Campagna D, Cope L, Lakkur SS, Henderson C, Laheru D, Iacobuzio-Donahue CA: **Gene expression profiles associated with advanced pancreatic cancer.** *Int J Clin Exp Pathol* 2008, **1**:32-43.
69. Yachida S, Iacobuzio-Donahue CA: **The pathology and genetics of metastatic pancreatic cancer.** *Arch Pathol Lab Med* 2009, **133**:413-422.
70. 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.** *Hum Mol Genet* 2005, **14**:1465-1474.
71. Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG: **A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition.** *Genes Dev* 2008, **22**:756-769.
72. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM: **High-fidelity mRNA amplification for gene profiling.** *Nat Biotechnol* 2000, **18**:457-459.
73. Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, 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.** *Proc Natl Acad Sci USA* 2000, **97**:3491-3496.
74. Camargo AA, Samaia HP, Dias-Neto E, Simao DF, Migotto IA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, *et al*: **The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome.** *Proc Natl Acad Sci USA* 2001, **98**:12103-12108.
75. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**:185-193.
76. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**:118-127.
77. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
78. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841-842.
79. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, *et al*: **CAGE: cap analysis of gene expression.** *Nat Methods* 2006, **3**:211-222.
80. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *J Mol Biol* 1987, **196**:261-282.
81. Deducer:, A GUI for R [www.deducer.org].
82. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3**:RESEARCH0034.
83. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Res* 2001, **29**:e45.

doi:10.1186/1476-4598-10-141
Cite this article as: Tahira *et al.*: Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Molecular Cancer* 2011 **10**:141.

Anais da Academia Brasileira de Ciências (2011) 83(2): 673-694
(Annals of the Brazilian Academy of Sciences)
Printed version ISSN 0001-3765 / Online version ISSN 1678-2690
www.scielo.br/aabc

Non-coding RNAs in schistosomes: an unexplored world

KATIA C. OLIVEIRA[1], MARIANA L.P. CARVALHO[1], VINICIUS MARACAJA-COUTINHO[1], JOÃO P. KITAJIMA[2] and SERGIO VERJOVSKI-ALMEIDA[1]

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo,
Av. Prof. Lineu Prestes, 748, 05508-000 São Paulo, SP, Brasil
[2]Hospital Israelita Albert Einstein, Av. Albert Einstein, 627, 05652-000 São Paulo, SP, Brasil

Manuscript received on February 5, 2011; accepted for publication on April 28, 2011

ABSTRACT

Non-coding RNAs (ncRNAs) were recently given much higher attention due to technical advances in sequencing which expanded the characterization of transcriptomes in different organisms. ncRNAs have different lengths (22 nt to >1,000 nt) and mechanisms of action that essentially comprise a sophisticated gene expression regulation network. Recent publication of schistosome genomes and transcriptomes has increased the description and characterization of a large number of parasite genes. Here we review the number of predicted genes and the coverage of genomic bases in face of the public ESTs dataset available, including a critical appraisal of the evidence and characterization of ncRNAs in schistosomes. We show expression data for ncRNAs in *Schistosoma mansoni*. We analyze three different microarray experiment datasets: (1) adult worms' large-scale expression measurements; (2) differentially expressed *S. mansoni* genes regulated by a human cytokine (TNF-α) in a parasite culture; and (3) a stage-specific expression of ncRNAs. All these data point to ncRNAs involved in different biological processes and physiological responses that suggest functionality of these new players in the parasite's biology. Exploring this world is a challenge for the scientists under a new molecular perspective of host-parasite interactions and parasite development.

Key words: *Schistosoma mansoni*, non-coding RNAs, gene expression profile, genome, transcriptome.

BACKGROUND

Schistosomiasis is a parasitic debilitating disease widespread in the world. The disease occurs in 76 countries especially in America, Africa and Asia. It is estimated that 779 million people in these countries are at risk of this infection and around 207 million people are infected. Three major species of *Schistosoma* are responsible for the disease: *S. mansoni* (that occurs in America and Africa), *S. japonicum* (Asia) and *S. haematobium* (Africa and eastern Mediterranean) (Steinmann et al. 2006).

S. mansoni has a complex life cycle with six distinct developmental stages in two hosts: eggs, miracidia (first larval stage of free living), primary and secondary sporocists (inside the intermediate host – a snail), cercariae (second larval stage of free living, that infects the definitive host), schistosomula (inside the definitive host – a mammalian) and adult worms (after 42 days of infection) (Gryseels et al. 2006). The adult couple starts the oviposition process and migrates to the mesenteric veins. Eggs cross the epithelial layer of veins and intestinal wall and are eliminated in the feces, restarting the biological cycle; many eggs go through blood circulation and cause inflammatory processes especially in the liver and this is the main cause of the pathological process (Gryseels et al. 2006).

Schistosome couples can live for years inside the host, suggesting that they are completely adapted to the host environment. It is already known that schistosomes take advantage of host signals from endocrine and im-

Correspondence to: Sergio Verjovski-Almeida
E-mail: verjo@iq.usp.br

mune system, uptake nutrients for their development and differentiation (Amiri et al. 1992, De Mendonca et al. 2000, Davies et al. 2001, Escobedo et al. 2005, Han et al. 2009), and in addition the parasite has many ortholog genes to human receptors (Agboh et al. 2004, Osman et al. 2006, Khayath et al. 2007, Wu et al. 2007, Oliveira et al. 2009). It is also known that the parasite has a sophisticated alternative splicing mechanism for genes encoding secreted proteins such as micro-exon genes (MEGs) (Demarco et al. 2010), polymorphic mucins genes (SmPoMucs) (Roger et al. 2008) and venom allergen-like (SmVALs) genes (Chalmers et al. 2008). These mechanisms are supposed to increase the repertoire of parasite proteins (Verjovski-Almeida and Demarco 2011), possibly helping to evade the immune response. Understanding the molecular mechanisms that are responsible for such a diverse life cycle and that promote the sophisticated parasite's adaptation is a challenge to the research community.

OVERVIEW ABOUT THE CURRENT KNOWLEDGE ON NON-CODING RNA

In the last years with the advance of sequencing technologies, many genomes of model organisms as well as their transcriptomes have been sequenced and a huge amount of sequence information has become available to the scientific community (The_*C.Elegans*_Sequencing_Consortium 1998, Kaul et al. 2000, Carninci et al. 2003, Begun et al. 2007, Birney et al. 2007, Church et al. 2009). As a consequence of the analysis of these genomes, the central dogma of molecular biology, namely that genetic information flows from DNA to RNA to protein, the final effector in the cell, has been challenged (Mattick 2003). It has been observed that the majority of the mammalian genome is pervasively transcribed and different types of RNAs without protein-coding potential have been identified (Birney et al. 2007, Nakaya et al. 2007). Because of this new finding, nowadays it is much more complicated to define a gene. A current definition of gene is: "The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein et al. 2007).

In general, one protein-coding gene is defined by the presence of an ORF longer than 100 amino acids. It is estimated that up to 90% of the human genome is transcribed, however only 2% of these transcripts are protein-coding genes (Claverie 2005, Johnson et al. 2005, Birney et al. 2007). These data revealed that most of the transcripts are non-coding RNAs. In face of this scenario it has been argued that this may reflect transcription noise in the cell with no biological relevance (Huttenhofer et al. 2005, Werner and Berdal 2005). In the opposite scenario, important biological functions including regulatory networks that directly involve ncRNA molecules have been described in the last years (Mattick 2004, Reis et al. 2005, Dinger et al. 2009, Louro et al. 2009, St. Laurent et al. 2009, Chen et al. 2010a, De Lucia and Dean 2010, Nag and Jack 2010).

Distinct characteristics of ncRNAs lead to the proposal that they have important regulatory functions: (i) conservation of their promoters, splice junctions, exons, predicted structures, genomic positions and expression (less than in protein-coding genes, nevertheless conserved) (Trinklein et al. 2004, Washietl et al. 2005, Louro et al. 2009, Mercer et al. 2009), (ii) their dynamic expression and alternative splicing during differentiation (Rinn et al. 2007, Guttman et al. 2009), (iii) their altered expression in cancer and other diseases (Reis et al. 2004, Brito et al. 2008, Nana-Sinkam et al. 2009, Kumar et al. 2010), (iv) their association with a particular chromatin signature that is indicative of actively transcribed genes (Dindot et al. 2009, Guttman et al. 2009), (v) their regulation by key morphogens and transcription factors (Cawley et al. 2004, Li et al. 2010); (vi) their tissue and cell-specific expression pattern and sub cellular localization (Louro et al. 2007, Nakaya et al. 2007, Birney et al. 2007, Mattick 2009).

In our perspective, one of the most important points of this discussion is the fact that the complexity of organisms along the evolution has been associated with the expansion of genomic elements. Comparison between the increasing number of protein-coding genes and non-protein coding genes clearly reveals that the expansion of ncDNA (especially the intronic regions) is much higher than the expansion of protein coding genes (Mattick 2004). It raises the hypothesis that regulation of complex functions is mediated by sophisticated mechanisms involving ncRNAs. Probably this was not noticed before because of the protein-centered view of

molecular biology. Accumulation of large amounts of sequencing data (Core et al. 2008, Mortazavi et al. 2008, Oliver et al. 2009) and of tiling array experiments (Johnson et al. 2005, Wilhelm et al. 2008) was necessary to show pervasive transcription especially in higher eukaryotes. In consequence of these observations the number of articles related with the description and study of ncRNAs has increased considerably in the last years (Mattick 2009).

Presently many types of ncRNAs have been characterized and described in the literature. **Short ncRNAs** are the most studied class of ncRNA in humans and model organims; they include **microRNAs** which are small RNAs (22 nt long) that regulate gene expression of hundreds to thousands genes by partial complementary base pairing to specific mRNAs (a post-transcription regulation). They direct degradation of target mRNAs through cleavage by Argonaute enzyme (present in RISC complex) (Bartel 2004); and **siRNAs**, which are endogenous small RNAs with 21 nt length produced by Dicer cleavage of perfectly complementary dsRNA duplexes. They form complexes with Argonaute proteins and are involved in gene regulation, transposon control and viral defense; these RNAs can act in *cis* or in *trans* (Brosnan and Voinnet 2009). **Long ncRNAs (lncRNAs)** are defined as RNAs of little protein-coding potential, with a length higher than 100-200 bp (arbitrary limit). They are the least understood transcriptional unit and comprises a heterogeneous group of transcripts (Costa 2010); **Intronic long ncRNAs** can be the product of a splice processing or originate from an independent transcription (Rearick et al. 2010); **Large intergenic ncRNAs** appear to be selected for conservation by evolution and are associated with epigenetic regulation (Guttman et al. 2009).

Another important concept in non-coding RNA is related to **NATs: Natural Antisense Transcripts**. These are generally non-protein coding transcripts, but fully processed, mRNAs that are transcribed from the opposite strand of protein-coding sense transcript (Werner and Swan 2010). Studies reveal a conservation of these transcripts among human, mouse and fish (Dahary et al. 2005, Zhang et al. 2006). These transcripts can act as precursor of siRNA, miRNA, gene silencing, although the roles of NATs are not completely understood.

Different mechanisms are involved in the action of ncRNA; here we summarize in Figure 1 the described mechanisms in model organisms (Brosnan and Voinnet 2009, Mercer et al. 2009, Wilusz et al. 2009, Chen and Carmichael 2010).

THE GENOME AND TRANSCRIPTOME OF SCHISTOSOMES

The genome of schistosomes is organized in eight chromosomes, seven autosomal and one sexual. In 2003 the transcriptomes of *S. mansoni* (Verjovski-Almeida et al. 2003) and *S. japonicum* (Hu et al. 2003) were published in Nature Genetics, giving insights and perspectives for functional genomics (Verjovski-Almeida et al. 2004). Six years later, in 2009, the genome sequences from both parasites were published in Nature (Berriman et al. 2009, Zhou et al. 2009). Additionally, the genome sequencing project of a third *Schistosoma* species, *S. haematobium*, is on the way and will provide a new collection of sequences in a not too distant future (Webster et al. 2010).

Table I summarizes the features of *S. mansoni* and *S. japonicum* published genomes (Berriman et al. 2009, Zhou et al. 2009) and all public EST transcripts. We can clearly see the similarities in the genome structure between these two schistosome species.

EST *vs.* Gene Predictions: How Many *S. mansoni* Genes? How Many Potential Non-coding RNAs?

We performed a comparison between the transcriptome and genome of *S. mansoni* in order to determine the percentage of transcripts that may be related to potentially non-coding RNAs (Fig. 2). We used all 205,892 public *S. mansoni* ESTs and mRNAs available in GenBank at the beginning of December 2010; in the first step of the analysis, we filtered out ESTs that match vectors (133 ESTs). Out of the remaining 205,759 ESTs we found that 154,707 (75.1%) (Fig. 2, upper part) could be mapped to *S. mansoni* annotated genes (i.e. 13,215 Smp protein-coding gene predictions (Berriman et al. 2009) plus 2,842 other described non-coding genes available at the Sanger Institute website (http://www.sanger.ac.uk/resources/downloads/helminths/schistosoma-mansoni.html), such as tRNAs, microRNAs, small nucleolar RNAs and ribosomal RNAs, which will be discussed later in this review).

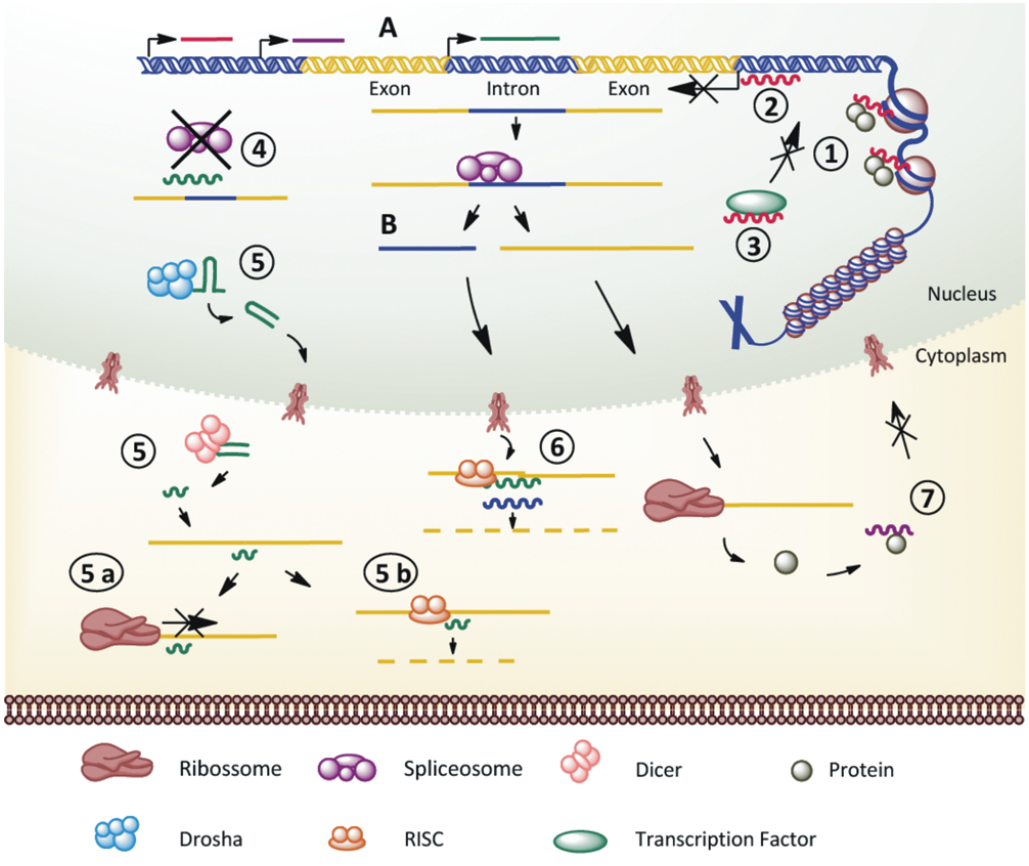


Fig. 1 – Main described mechanisms of ncRNA action in the cell. Pre-transcriptional mechanism: **(1)** ncRNAs acting on protein complexes that are involved in chromatin remodeling lead local regulation of gene expression; **(2)** ncRNA forms a triplex at the promoter region of genes and inhibits gene transcription; **(3)** ncRNA interacts with transcription factors and acts as co-repressor or co-activator of transcription (modulates protein activity). Post-transcriptional mechanism: **(4)** ncRNA may act at the spliceosome interfering in the splicing process; **(5)** ncRNA may generate miRNA (by the processing steps involving Drosha and Dicer) and either **(5a)** Inhibit mRNA translation, or **(5b)** Degrade mRNA target by RISC; **(6)** ncRNA may act as endogenous small-interfering RNA (siRNA) and be cleaved by RISC. Post-translation mechanism: **(7)** ncRNA may interact with target proteins altering protein localization and organizational role in the cell. Biosynthesis of ncRNA: **(A)** ncRNA may be generated by independent transcription, or **(B)** ncRNA may be generated from spliced introns of protein-coding genes.

From the 51,052 ESTs that do not match Smps or other described genes, we filtered out 10,942 ESTs that match repetitive elements, mitochondrial genes, etc. The remaining 40,110 ESTs (19.5%) were assembled using CAP3 with default parameters, generating 5,166 contigs (comprised of 22,553 ESTs, 11.0%) and 17,557 EST singlets (8.5%) (Fig. 2).

The assembled ESTs described above were found to be divided into a major set that matches the genome outside of any predicted Smp gene (15,536 ESTs (7.5%) assembled into 3,311 contigs plus 9,080 singlets (4.4%); a total of 24,616 ESTs, 11.9%) and a smaller set that does not match the genome (7,017 ESTs assembled into 1,855 contigs plus 8,477 singlets; a total of 15,494 ESTs, 7.5%) (Fig. 2). Overall, our analysis shows that 87% of public *S. mansoni* ESTs match the genome and highlights the fact that a considerable fraction (11.9%) shows evidence of transcribed regions in the genome for which no Smp gene prediction was made (Berriman et al. 2009). Additionally, 4,076 Smps and 2,717 other genes were predicted in the genome without *S. mansoni* ESTs evidence.

We analyzed the protein coding potential among the 40,110 ESTs (19.5%) that do not match Smps. In a

TABLE I
Comparison between *S. mansoni* and *S. japonicum* genomes and transcriptomes.

	S. mansoni	*S. japonicum*
Length of genome (Mb)	363	398
Number of genomic scaffolds (>2kb)	5,745	13,235
Number of predicted genes	11,809 genes 13,197 transcripts	13,469
% of repetitive sequences	45	40
CG content (total)	35.3	34.1
CG content coding regions	36.3	36
CG content non-coding regions[#]	35.2	33.8
Average predicted gene size (bp)	11,400*	10,500
Average predicted exon size (bp)	217	n/a
Average intron size (bp)	1,692	n/a
Number of sequenced ESTs in public databases	205,892	105,765

*Calculated by us according to information in the supplementary material not directly mentioned in the text of the genome paper (Berriman et al. 2009). [#]In *S. japonicum* the intergenic CG content was counted and in *S. mansoni* it was not.

first step we looked for a match of the assembled ESTs to a curated protein dataset: UNIPROT (The_Uniprot_Consortium 2010) available at (http://www.uniprot.org/); in a second step, the assembled ESTs that did not match UNIPROT were analyzed for their protein-coding potential using Coding Potential Calculator (CPC) (Kong et al. 2007).

The ESTs that do not match Smps and match the genome (26,616 ESTs) were assembled into 3,311 contigs (15,536 ESTs, 7.5%) (Fig. 2) and we found that 522 of these contigs (2,547 ESTs, 1.2%) have match to 154 UNIPROT known proteins from other organisms and were not predicted in *S. mansoni*; one additional contig (composed of 2 ESTs, 0.00003%) that does not match UNIPROT was predicted by CPC to have a protein-coding potential (Fig. 2). The remaining 2,788 contigs (12,987 ESTs, 6.3%) are potential non-coding RNAs since they do not match UNIPROT proteins and were not predicted by CPC to have protein-coding potential. Out of the 9,080 EST singlets (4.4%) that match the genome outside of Smps, we found that 960 ESTs (0.5%) match 202 UNIPROT known proteins; the remaining 8,120 EST singlets (3.9%) are again potential non-coding RNAs since they do not match UNIPROT and were not predicted by CPC to have protein-coding potential.

Here we conclude that overall, 21,107 ESTs (10.3%) that match the genome have no protein-coding potential and are good candidates for *S. mansoni* non-coding RNAs; these ESTs point to 10,908 genomic regions (2,788 contigs + 8,120 EST singlets) with evidence of ncRNA transcription. These data also point to 356 known UNIPROT proteins (2,547 ESTs assembled into 522 contigs and 960 singlets that match genome) that were expressed in *S. mansoni*, map to the genome sequence and were not predicted by the genome project (see Supplementary Table I).

A total of 15,858 ESTs (7.7%) do not match the genome and are either transcribed from a non-sequenced part of the *S. mansoni* genome or represent contaminants in the transcriptome database, especially the EST singlets. Out of 1855 contigs (7,017 ESTs, 3.4%) we found that 339 contigs (1,857 ESTs, 0.1%) match contaminant sequences (such as *M. musculus*, bacteria, *H. sapiens*, *B. glabata*, *B. taurus* and *R. norvergicus*) (Fig. 2). Among the remaining 1,516 contigs (5,160 ESTs, 2.5%), 448 contigs (2,186 ESTs) match 434 UNIPROT proteins; 1,068 contigs (2,974 ESTs) do not match UNIPROT proteins. From this group, 2 contigs (6 ESTs) have coding potential and 1,066 contigs (2,968 ESTs, 1.4% of all transcripts) do not have protein-coding potential according CPC.

Fig. 2 – Workflow of the genome mapping and annotation of *S. mansoni* ESTs available in GenBank.

From the 8,477 EST singlets that do not match the genome we found that 956 have match to sequences of potential contaminants. From the remaining 7,521 singlets, 1,276 have match 764 UNIPROT proteins and 6,245 ESTs (3.0%) do not match UNIPROT proteins and do not have protein-coding potential (Fig. 2).

We could conclude that 9,213 ESTs (4.4%) between contigs and singlets that do not match the genome are potentially non-coding RNA. These data also point to 1,443 UNIPROT unique proteins that are conserved and are potentially expressed in *S. mansoni* but are not present in the genome (see Supplementary Table I).

In summary, among the 205,892 public *S. mansoni* ESTs a total of 30,320 ESTs (between contigs and singlets) (14.7%) do not have protein-coding potential; of these, 21,107 ESTs (10.2%) match the genome, while only 9,213 ESTs (4.5%) do not match the genome. The fraction of total *S. mansoni* transcription comprised by non-coding RNAs probably reveals a lower level of transcriptional activity of this class of RNAs, compared to the transcriptional activity of protein-coding genes. In fact, it has been reported in humans that the long non-coding RNAs are transcribed at a much lower rate than the protein-coding genes (Kapranov et al. 2007) and the non-coding RNAs are represented by between 10 and 20% of the human EST database collection (Nakaya et al. 2007).

Coverage of ESTs onto the Genome and Gene Predictions: a Surprising Transcription from the Introns?

We calculated the percentage of bases in the genome that is comprised of gene predictions and the number of bases in the genome that are covered by NCBI public ESTs. The genome of *S. mansoni* has 362,876,148 bp distributed in 5,745 scaffolds >2 kbp (Berriman et al. 2009). From all these bases, 165,206,376 bp (45.5%) are loci of predicted genes. From these loci, 15,852,242 bp are exons of gene predictions (4.3% of total bases in the genome and 9.6% of gene prediction loci) and 149,354,134 bp (41% of total bases in the genome and 90.4% of gene prediction loci) are introns.

Based on the public *S. mansoni* ESTs that have so far been accumulated and that mapped to the sequenced part of the genome, we find that a total of 16,516,608 genomic bases were covered by at least one EST, which means that at least 4.6% of the *S. mansoni* genome is transcribed.

From the 16,516,608 transcribed base pairs, a total of 12,717,085 bp (77% of transcribed bases) is located in gene prediction loci (3.5% of genome bases). A total of 7.7% bases in gene prediction loci are covered; the loci include exons and introns of genes.

When looking at the exons in the genome (comprised of 15,852,242 bp), we found that 8,652,015 bp were covered by public ESTs (42% of transcribed bases) which represents coverage of 55% of exon bases of the predicted gene sequences (2.4% of genome bases); out of them 1,557,580 bases are in UTRs (Fig. 3). Looking at the predicted genomic introns (149,354,134 bp) we found that 4,065,070 bp were covered by public ESTs (34% of total transcribed bases) corresponding to coverage of only 2.7% of predicted intron bases (1.1% of genomic bases). We detected that 3,799,523 transcribed bases (1% of the total genomic bases and 23% of transcribed bases) are located in intergenic regions. Figure 3 summarizes these numbers.

Mattick (Mattick 2004) raised the hypothesis that the complexity of an organism probably is derived from the expansion of non-coding regions in the genome, especially because there is no considerable increase in the number of protein-coding genes along the evolution, whereas there is an important expansion of non-coding regions in the genomes of the more complex organisms. This expansion occurred especially in the intronic regions of protein-coding genes. Mattick calculated the ratio between ncDNA/total DNA for a large spectrum of organisms with diverse complexities, and found that complex organisms such as *Mus musculus* and *Homo sapiens* had a value higher than 0.9 (Mattick 2004). Nevertheless, the position of, for example, *Anopheles gambiae*, does not fully concord with the hypothesis. Moreover, within-clade variations can be considerable. For example in the ray-finned fishes, genome size (even allowing for polyploidy) can vary by a factor of 20-fold (Smith and Gregory 2009). Assuming that the number of coding genes is unlikely to vary much within this group, factors other than complexity may govern the amount of ncDNA in these genomes. In fact, factors such as metabolic rates, body size, effective population size are known to affect genome size (Keeling and Slamovits 2005).

Based on the genome and gene predictions (Berriman et al. 2009), here we calculated the ncDNA/Total DNA ratio in *S. mansoni* and found it to be 0.96, a quite high ratio considering the parasite complexity and comparing to the ratio in the human genome. In addition, *S. mansoni* has an unusual intron size distribution (Webster et al. 2010). It is interesting to note that another platyhelminth, the free-living *Schmidtea mediterranea* has a 480 Mb genome (http://genome.wustl.edu/genomes/view/schmidtea_mediterranea/),

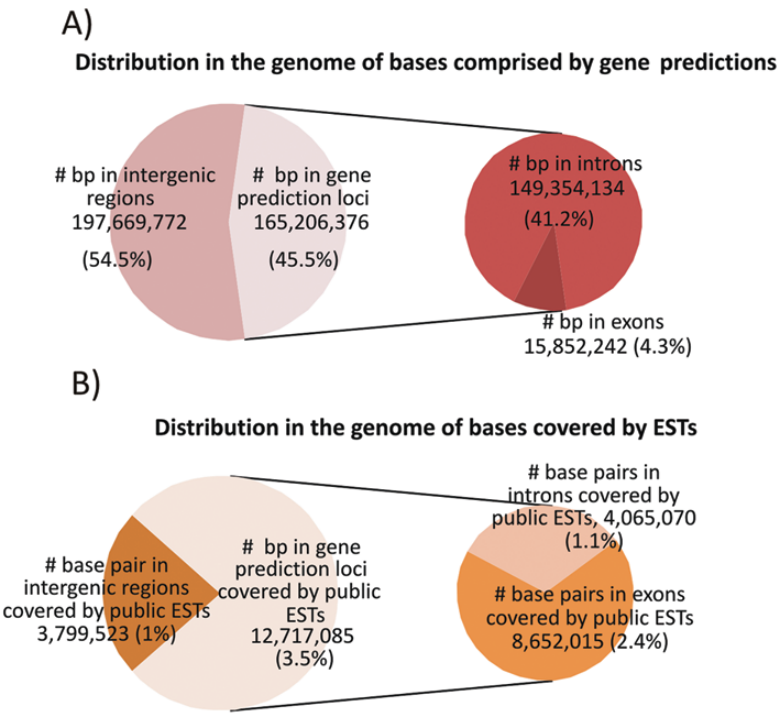


Fig. 3 – Analysis of *S. mansoni* genomic bases. **A)** Distribution of bases comprising gene predictions. **B)** Distribution of bases covered by public ESTs. Percentages on the right-hand part of each figure add up to the total percentage of the corresponding sub-category in the left-hand part.

around 100 Mb bigger than *S. mansoni* and unlikely to have much more coding DNA, which would result in a ncDNA/Total DNA ratio even higher than *S. mansoni*; these observations suggest that the expansion of ncDNA may be one of the mechanisms used by evolution to achieve platyhelminth complexity and to interact with the environment. *S. mansoni* may have undergone genome reduction (Keeling and Slamovits 2005) in comparison to *S. mediterranea* due to parasitism of the former.

So far, there is limited evidence of transcriptional activity along the *S. mansoni* genome; just 4.6% of all genome bases are covered by EST data, as described above. The predicted intronic regions in *S. mansoni* are large; they comprise 41% of total genomic bases and 90% of genomic loci of gene predictions; in *S. mansoni*, transcription detected in intronic regions corresponds to only 1.1% of total genomic bases although it reveals that 34% of the transcribed bases are in introns. In humans 30% of the genome is comprised of introns and a pervasive transcription has been detected (Birney et al. 2007, Kapranov et al. 2007). In humans, an analysis of the 5.3 million public ESTs pointed to the presence of at least one EST mapping to the introns of 74% of all RefSeq human genes (Nakaya et al. 2007). In fact, we re-analyzed the genome mapping of 8 million public human ESTs, and we found approximately 70,000 unique intronic loci covering nearly 42 million genomic bases (1.7% of the human genome) with evidence of transcription.

An intensive effort has been placed in the study of intronic transcription of a segment corresponding to 1% of the human genome by a network of laboratories under the name of ENCODE Project, using different methods such as tiling-arrays, RNA-seq and paired-end sequencing; ENCODE found that 93% of the studied 1% segment was transcribed; both the intronic and intergenic regions showed evidence of transcription (Birney et al. 2007), and the suggestion is that this figure can be extrapolated to the entire human genome. In other higher eukaryotes such as *C. elegans* 70% of the genome is transcribed and an ENCODE analysis is being developed (Gerstein et al. 2010). In *D. melanogaster* 85% of the genome is transcribed (Graveley et al. 2010).

Given the above numbers, it is not surprising that that 34% of the transcribed bases in the public *S. mansoni* EST database are in introns. Nevertheless, the limited evidence of genomic coverage of intronic transcription in *S. mansoni* (only 1.1% of total genomic bases) can be explained by the fact that until now a limited number of sequencing projects were executed (Franco et al. 1995, 2000, Merrick et al. 2003, Verjovski-Almeida et al. 2003); in addition, the sequencing projects in *S. mansoni* were performed with poly-A transcripts, and it is known that many transcripts, especially non-coding RNAs are not poly-adenylated (Kiyosawa et al. 2005). It suggests that a big effort towards deep-sequence and tilling array approaches is warranted to obtain higher transcription coverage of the *S. mansoni* genome.

NON-CODING RNAS AND SCHISTOSOMES

Short Non-coding RNAs

The *S. mansoni* transcriptome project (Verjovski-Almeida et al. 2003) provided data for the identification of several ESTs encoding proteins related to components of the microRNA processing machinery. Various papers have been published in recent years; most of these published works are related to the identification of the RNAi/miRNA pathway as well as to individual miRNA identification, as detailed below.

The current model of RNAi processing involves two cleavage steps, each one centered on a ribonuclease enzyme. The precursor RNA (either a dsRNA or a miRNA primary transcript) is processed into a short inhibitory RNA (siRNA) by RNAse III enzymes called Drosha (in the nucleus) and Dicer (in the cytoplasm), with dsRNA binding domain (dsRBD) protein acting as a cofactor. In the second step, siRNA is loaded into the effector protein complex called RNA-induced silencing complex (RISC). This siRNA is opened in a strand-specific manner during RISC assembly and single-stranded siRNA locates its cognate mRNA target by base pairing. Gene silencing is the result of the nucleolytic degradation of the RNAse H enzyme Argonaute. If the siRNA/mRNA duplex contains mismatches at the scissile site, as is often the case of miRNAs, the mRNA target is not cleaved and gene silencing results from translational inhibition (Pratt and Macrae 2009).

The complete conserved machinery to process microRNA or siRNA has been identified in *S. mansoni* and *S. japonicum*. In 2008 Krautz-Peterson and Skelly (Krautz-Peterson and Skelly 2008a) described the Dicer gene in *S. mansoni*, and observed the highest expression levels in schistosomula (15 days-old) and eggs. Later, in 2009 Gomes et al. (Gomes et al. 2009) described Dicer, Drosha and four different Argonaute proteins in *S. mansoni*; they also observed the highest expression level in eggs, but not in schistosomula.

In *S. japonicum*, two articles were published in 2010; Chen et al. (Chen et al. 2010b) described three Argonaute proteins and observed the highest expression levels in eggs and miracidia; and Luo et al. (Luo et al. 2010) described Dicer and four Argonaute proteins. They also observed that the highest expression levels of Dicer and Argonaute 1 are in eggs and miracidia.

In general these authors are in agreement in their observations and this pattern of expression suggests that the miRNA regulatory pathway might take part in the transformation and development of schistosomes. Another hypothesis is that miRNA could be responsible for the repression of translation in eggs (Schier 2007).

The characterization of miRNA pathway genes points to the mechanism that causes RNA interference in schistosomes. The RNAi approach was already used in several areas of *S. mansoni* biology (Boyle et al. 2003, Skelly et al. 2003, Correnti et al. 2005, Delcroix et al. 2006, Dinguirard and Yoshino 2006, Freitas et al. 2007, Krautz-Peterson et al. 2007, 2010, Ndegwa et al. 2007, Krautz-Peterson and Skelly 2008b, Morales et al. 2008, Pereira et al. 2008, Faghiri and Skelly 2009, Rinaldi et al. 2009, Beckmann et al. 2010) as well as in *S. japonicum* (Cheng et al. 2005, Zhao et al. 2008, Kumagai et al. 2009, Zou et al. 2010). RNAi has proved itself as an important tool to elucidate gene function in schistosomes, in a similar way as in other organisms.

RNAi experiments have mostly used the electroporation method to deliver the dsRNA or RNAi into the parasite, with soaking or biobalistic delivery being also used. However, the mechanism of dsRNA uptake used by the worm remains unclear. Recently Krautz-Perterson (Krautz-Peterson et al. 2010) described a *S. mansoni* homolog protein to SID-1 (Systemic RNA Interference-Defective) of *Caenorhabditis elegans*, a multi-membrane spanning RNA importing protein, that con-

tains 21 exons and potentially encodes a protein with 1018 amino acids. This SID-1 protein has been shown to be required for uptake of dsRNA in *C. elegans* (Winston et al. 2002); probably SmSID-1 may have the same function in *S. mansoni*. Localization of this transport protein on the parasite and its functional characterization are some interesting subjects that remain to be clarified.

The first miRNAs characterized in schistosomes were described in 2008 (Xue et al. 2008). In that work the authors described 227 cloned microRNAs in *S. japonicum*. Among the cloned miRNAs, five have high level of conservation with well characterized microRNAs in more complex organisms, such as human, mouse, *C. elegans*, *Drosophila*: let-7, miR-71, miR-new1, mir-125 and bantam.

Huang et al. (Huang et al. 2009) described 176 miRNAs in *S. japonicum* (including let-7, miR-71, bantam and miR-125), among them 172 novel miRNAs. All these new miRNAs were identified and mapped to the genome by the presence of an inferred RNA hairpin with pairing characteristics of known miRNA structure. The authors also analyzed the differential expression between mixed adult worms and hepatic schistosomula and observed that 35 out of 176 were expressed in adult worms, 60 in schistosomula and 81 in both stages.

In 2010 two more publications focused on the identification and characterization of short ncRNAs in *S. japonicum* using deep-sequencing approach (Hao et al. 2010, Wang et al. 2010); this strategy has proved itself to be a powerful technique to identify small ncRNAs.

Wang et al. (Wang et al. 2010) described 20 species-conserved miRNAs and 16 schistosome-specific miRNAs. These miRNAs were validated using northern blot or stem-loop qRT-PCR approaches. The paper also described the identification of 4,858 putative endogenous siRNAs, 40% of them related to retrotransposons (TE-derived) (Wang et al. 2010) as expected by comparison to reports from other species (Golden et al. 2008).

Hao et al. (Hao et al. 2010) sequenced 5.3 and 4.2 million reads from small RNAs from adult worms and schistosomula respectively; these sequences represent around 1.1 million unique clean sequences. In both stages, the majority of sequences are siRNAs, and they are Transposable-Elements-derived. The authors point to the identification of 38 unique *S. japonicum* transcripts and 16 miRNAs that belong to 13 miRNA families conserved in other metazoan organisms. The amount of siRNAs was at least 4.4 times larger in schistosomula and 1.6 times larger in adult worms than in other stages.

More recently, Simoes et al. (Simoes et al. 2011) performed a bioinformatics homology-based analysis and identified conserved miRNA in *S. mansoni*. The authors also identified 211 novel miRNA in *S. mansoni* by sequencing of small-RNA cDNA libraries from adult worms. Out of these 211 candidates, 11 miRNAs had their expression level validated by northen blot analysis; three out of these miRNAs were already described in *S. japonicum*.

Other Non-coding RNAs

In 1998 Ferbeyre et al. (Ferbeyre et al. 1998) performed an *in silico* search for RNA structural motifs in sequence databases and have found a hammerhead ribozyme domain encoded in the satellite repetitive DNA of *S. mansoni* (Ferbeyre et al. 1998). Transcripts are expressed from these repeats as long multimeric precursor RNAs that cleave *in vitro* and *in vivo* into unit-length fragments. This RNA domain is able to engage in both *cis* and *trans* cleavage typical of the hammerhead ribozyme (Ferbeyre et al. 1998).

Copeland et al. (Copeland et al. 2009) carried an extensive *in silico* search in the genome of *S. mansoni* and *S. japonicum* and performed a homology-based annotation of the "house-keeping" ncRNAs in schistosomes. The authors were able to identify 23 types of ncRNAs with conserved primary and secondary structure; among these we mention rRNA, snRNA, SLRNA, SRP, tRNA and RNase P, and possibly MRP and 7sK RNAs. The previously described hammerhead ribozyme RNA (Ferbeyre et al. 1998) is the most diverse because it originates from repetitive DNA; tRNAs were found to be the next most diverse ncRNAs encoded in the *S. mansoni* genome (tRNAscan-SE predicted a total of 713 tRNAs) (Copeland et al. 2009). The authors focused on the comparison between tRNA populations in other schistosomes and in a free-living platyhelminth organism (*Schmidtea mediterranea*); they also confirmed in *S. mansoni* the first miRNAs described in *S. japonicum* by Xue et al. (Xue et al. 2008).

Until now few articles studied the expression and characterization of ncRNAs in schistosomes, as reviewed above. Specifically, nothing can be found in the literature about long (>200 nt) ncRNAs.

DIFFERENTIAL EXPRESSION OF NON-CODING RNAS IN SCHISTOSOMES: OUR EXPERIENCE AND PERSPECTIVE

Detection of Transcription from both Genomic Strands in the Same Locus in *S. mansoni* Adult Worms

Our group in 2007 studied the transcriptome of *S. mansoni* adult worms using a microarray platform with 44 k oligonucleotide (60-mer) probes (Verjovski-Almeida et al. 2007); we detected 156 genome loci that were represented by probes on both genomic strands for which there was evidence of expression from both probes. From these, 9 loci were selected for validation using strand-specific RT-qPCR and we validated 6 loci. These 156 loci may be sources of ancestral Natural Antisense Transcripts (NATs) (Werner and Swan 2010), that have not been characterized so far.

The above paper (Verjovski-Almeida et al. 2007) was published before the availability of *S. mansoni* gene predictions to the scientific community at Schisto GeneDB website (http://www.genedb.org/Homepage/Smansoni). Here we carefully performed a re-annotation of this array platform and mapped the oligonucleotide probes to *S. mansoni* gene predictions and proteins available in GenBank (nr). This re-annotation is available as Supplementary Table II. With the re-annotation we found that 108 out of 156 loci with expression from both genomic strands do map to gene predictions. From these 108 loci, 18 map on protein-coding exons and therefore only one strand in the pair is an ncRNA. Another 18 of them map to the UTR region and an additional 72 map to intronic regions (2 loci in the same gene prediction). All these 108 are candidates of NAT transcription (Werner and Swan 2010). Out of the remaining 48 loci that do not match *S. mansoni* gene predictions, 27 match conserved proteins in GenBank and again only one strand in the pair is an ncRNA. Finally, 21 still have no match proteins either in GenBank or gene predictions.

This finding is extremely interesting because here we point to the first potential NATs in schistosomes. The list with these 135 loci with evidence of NATs is available in Supplementary Table III. From the 6 loci with expression in both genomic strands, which were validated by strand-specific RT-qPCR (Verjovski-Almeida et al. 2007), we found that 4 are located in gene prediction loci (highlighted in Supplementary Table III). They are: Smp_174720, Smp_136110, Smp_096790 (all of them mapped to intronic regions), and Smp_194860 (mapped to an exonic region). Probes that have the same orientation of the coding message and map to intronic regions may be revealing novel non-predicted exons of that given protein-coding gene. An alternative explanation remains, in that these introns may be genomic loci of independent transcription or the intron spliced from the immature pre-mRNA may be processed as a precursor of ncRNA (i.e. miRNA). Evidence of transcription has been obtained from the opposite strand of protein-coding predicted genes, which certainly points to independent antisense transcriptional events. The future molecular characterization of these candidates should help in understanding the molecular mechanism of gene regulation in schistosomes.

Non-coding RNAs with Expression Changes Induced by Human TNF-α

In December 2009 our laboratory published the ortholog gene of TNF-α receptor in *S. mansoni* and characterized the effect of human TNF-α on the parasite gene expression using a microarray platform of 44k oligonucleotide probes (Oliveira et al. 2009). Here, we re-annotated the 44k oligonucleotide array according to gene predictions that appear in the genome publication (Berriman et al. 2009) in order to highlight the differentially expressed probes that map to the opposite strand of known protein-coding genes (potentially ncRNAs regulated by TNF-α).

Expression changes induced by treatment with human TNF-α had been detected in newly transformed 3 h-old schistosomula in culture (1 h treatment). A set of 755 probes had been identified with a statistically significant (q-value < 0.05) differential expression between TNF-α treated and control early schistosomula (Oliveira et al. 2009); with the re-annotation we conclude that 686 unique genes were affected.

Among these 686 genes, 564 match *S. mansoni* gene predictions, 32 match *S. japonicum* gene predic-

tions, 69 have match to conserved proteins in GenBank, comprising a total of 667 known coding genes, and 21 have no match. Among these 667 genes, 65 have significant changes in the expression level of the anti-sense message of the respective loci and 6 loci have significant expression changes in both sense and anti-sense messages. From the 6 loci with changes in the expression level in both strands, 3 of them have a decreased expression of sense and anti-sense messages in response to human TNF-α while the other 3 have a discrepant expression pattern; 3 of these 6 loci have pairs of probes that map to intronic regions of predicted genes, 2 in UTR regions and one in a coding exon. All the 65 gene loci with detected expression in the anti-sense strand and the 6 loci with expression in both genomic strands are listed in Supplementary Table IV (Part A).

In adult worms treated during 1 h or 24 h with TNF-α we had identified (Oliveira et al. 2009) two distinct expression patterns in treated adult worms: genes with transient expression changes (up-regulated at 1 h treatment and down-regulated at 24 h treatment, or the opposite pattern) and genes with sustained changes (up-regulated at 1 h and 24 h treatment, or down-regulated throughout).

A set of 1,594 probes revealed statistically significant (q-value < 0.05) transient changes in expression (Oliveira et al. 2009). With the microarray re-annotation we conclude that there are 1404 unique genes with transient changes induced by TNF-α; 1048 genes have match to *S. mansoni* gene predictions, 54 match *S. japonicum* gene predictions, 203 match conserved proteins (GenBank), comprising a total of 1305 known protein-coding genes, and 99 have no match. Among these 1305 differentially expressed known protein-coding genes we verified that 177 coding genes have significant changes in the expression level of the anti-sense message from the respective loci, and 45 loci have expression changes in both sense and anti-sense messages. In the group of genes with differential expression in both loci strands, 28 of them have the same pattern of expression change in the sense/anti-sense pair of probes, while 19 genes have an opposite pattern of expression between sense and anti-sense probes.

In consequence of the re-annotation we observed that out of the 45 gene loci that have significant changes in expression in both strands, 29 have pairs of probes that map to intronic regions of predicted genes and 3 pairs of probes map to UTRs. All 45 loci with expression in both genomic strands and 177 gene loci with anti-sense expression are listed in Supplementary Table IV (Part B).

A group of genes had been identified with sustained changes in expression at 1 and 24 h TNF-α treatment (Oliveira et al. 2009). A total of 626 probes had a sustained change in expression; with the present microarray re-annotation we conclude that there are 584 differentially expressed unique genes with sustained changes in the expression pattern induced by TNF-α. From these 584 genes, 471 have match to *S. mansoni* gene predictions, 25 match *S. japonicum* gene prediction, 58 genes have match to conserved proteins in GenBank, comprising a total of 554 annotated genes, and 30 have no match to GenBank. Among the 554 known coding genes, 7 gene loci have evidence of transcription in both strands (4 with expression induced by TNF-α in the messages from both strands and 3 with opposite expression pattern in each strand). Interestingly, in these 3 loci with opposite expression pattern between sense and anti-sense messages, the pair of probes maps to intronic regions of the protein-coding genes. We also observed in the group of 584 known protein-coding genes with sustained changes in the expression level that 3 genes have significant changes in the expression profile just in the anti-sense message of the locus; all these 10 gene loci with expression in anti-sense and in both strands are available in Supplementary Table IV (Part C).

Overall, the above data reveal a potential new set of 303 long non-coding RNAs in schistosomes that are anti-sense messages to known protein-coding genes whose expression is regulated by human TNF-α. Expression changes modulated by an exogenous regulatory molecule from the host (TNF-α) suggests some functionality for these ncRNAs; these long ncRNAs may participate in a sophisticated network of gene regulation in consequence of TNF-α signaling that deserves further characterization.

Non-coding Expression Signature Among Life Cycle Stages

A number of papers already describe differences in gene expression among the developmental stages of schistosomes (Dillon et al. 2006, Vermeire et al. 2006, Jolly et al. 2007, Fitzpatrick et al. 2009, Gobert et al. 2009). Here we performed a set of experiments with 5 developmental stages (eggs, miracidia, cercariae, 7-day-old schistosomula and adult worms) using a 4k-element cDNA microarray platform that was designed to have a considerable fraction of probes for non-protein-coding genes (1133 probes); a detailed description of this platform is deposited in GEO under accession number GPL3929 (Demarco et al. 2006). This is the first microarray analysis of gene expression profile among life cycle stages that focuses on non-protein-coding genes in *S. mansoni*.

We analyzed two biological replica samples of each developmental stage; 3 ug amplified RNA (Wang et al. 2000) was labeled with Cy3 or Cy5 and hybridize to the arrays essentially as previously described (Demarco et al. 2006). The combination of samples on an array was: eggs *vs.* miracidia; cercariae *vs.* 7-day-old schistosomula and 7-day-old schistosomula *vs.* adult worms. We used a dye-swap approach to correct for any bias caused by dye incorporation or by intrinsic differential fluorescence yield of the dyes (Demarco et al. 2006). Raw data of this experiment is deposited in GEO under accession number GSE27026.

We used two different analyses approaches. In the first approach we identified differentially expressed genes between two consecutive developmental stages (eggs *vs.* miracidia; cercariae *vs.* 7-day-old schistosomula and 7-day-old schistosomula *vs.* adult worms) using SAM (Significance Analysis of Microarray) software (Tusher et al. 2001). Overall, we were able to find 1,423 differentially expressed genes between two developmental stages among all previously indicated comparisons (Table II). A detailed description follows below.

In the second approach we identified genes with increased expression levels in at least one developmental stage (for example more highly expressed in cercariae than in all other stages) using ANOVA statistical test (Churchill 2004) corrected for multiple sampling using Bonferroni correction (Shaffer 1995). Overall, with this approach we identified 577 differentially expressed genes with increased expression in at least one specific stage (Table III). A description of affected genes for each stage is given below.

TABLE II
Number of differentially expressed genes in the comparison of two sequential stages (FDR<0.001).

	Eggs	Miracidia
Number of genes	117	636
Protein-coding genes	104	552
Non protein-coding	13	84
Intronic	0	0
Intronic/Exonic	1	5
Intergenic	5	63
Not mapped	7	16
	Cercariae	Schistosomula (7-day-old)
Number of genes	271	130
Protein-coding genes	219	117
Non protein-coding	52	13
Intronic	1	0
Intronic/Exonic	2	3
Intergenic	32	6
Not mapped	17	4
	Schistosomula (7-day-old)	Adults
Number of genes	86	183
Protein-coding genes	75	155
Non protein-coding	11	28
Intronic	0	0
Intronic/Exonic	1	1
Intergenic	2	13
Not mapped	8	14

TABLE III
Number of genes with enriched expression in one developmental stage (Bonferroni adjusted p-value<0.005).

Genes with enriched expression in one stage	
Number of genes	577
Protein-coding genes	473
Non protein-coding genes	104
Intronic	1
Intronic/Exonic	6
Intergenic	61
Not mapped	36

In general, all observed patterns of expression of protein-coding genes among life cycle stages that will be described here, are not inconsistent with the previously published microarray results (Dillon et al. 2006, Jolly et al. 2007, Fitzpatrick et al. 2009, Gobert et al. 2009). Here we would like to especially highlight the non-coding genes that were never the focus of study before.

Differentially expressed non-coding genes identified here were mapped to the genome (using SchistoDB, available at http://schistodb.net/schistodb20/) and annotated as mapping to Intronic, Intronic/Exonic, Intergenic regions or as "Not mapped" (in case of multiple hits or no hit to the genome) (Tables II and III). These non-coding genes were re-confirmed as having no protein-coding potential by using the CPC tool (Kong et al. 2007). In addition, some previously annotated non-coding genes were now mapped to exons of predicted genes; they received an Smp re-annotation and were no longer counted as non-coding.

We are not able to assign a genomic strand for the observed expression of non-coding genes, since the probes on the 4k-microarray platform were generated by PCR amplification of selected double-stranded cDNA clones from the *S. mansoni* EST sequencing project (Verjovski-Almeida et al. 2003). These probes detect expression on either strand of a given locus.

Using the first analysis approach described above, we were able to find a set of 753 differentially expressed genes between eggs and miracidia; out of them, 117 genes with higher expression in egg and 636 genes with the opposite pattern. The complete list of differentially expressed genes between eggs and miracidia are available in Supplementary Table V (Part A). Among the 753 genes, there were 656 protein-coding genes (104 genes in eggs and 552 in miracidia); we highlight that there was significant enrichment (according to Gene Ontology analysis) of genes involved in amino acid and RNA metabolism in miracidia. The GO results are summarized in Supplementary Table V (Part B). We observed a set of 97 non-coding genes; 13 with higher expression in eggs and 84 with higher expression in miracidia (Table II). The non-coding expression profile is represented in Figure 4A. Description of genomic mapping coordinates is available in the supplementary material.

In the comparison between cercariae and 7-day-old schistosomula we found 401 differentially expressed genes; 271 in cercariae and 130 in schistosomula (Table II). From the 271 with higher expression in cercariae, 219 are protein-coding genes, and 52 are non-coding genes. Here, it is interesting to note the expression of a message that maps to the intronic region of the Smp_15-4340 gene, annotated as "nuclear factor Y transcription factor subunit B homolog, putative". This gene has three isoforms, and the intronic transcript could eventually act in *cis* modulating the splicing pattern of this transcript.

From the 130 genes with higher expression in 7-day-old schistosomula, 117 genes are protein-coding genes and 13 are non-protein coding (Table II). Figure 4B illustrates this profile. The list of all differentially expressed genes between cercariae and schistosomula is in Supplementary Table V (Part C). In this comparison we have found enriched GO categories both in cercariae and schistosomula; they are listed in Supplementary Table V (Part D and E), respectively.

In the third comparison, 7-day-old schistosomula *vs.* adults, we found 269 differentially expressed genes, 86 in schistosomula and 183 in adult worms. From the 86 genes with higher expression in schistosomula, 75 are protein-coding genes and 11 non-coding genes. In the opposite scenario, comprising the group of genes with higher expression in adult we have 155 protein-coding genes and 28 non-coding genes. This non-coding expression profile is represented in Figure 4C. No enriched GO categories were found among the protein-coding genes. The list of all differentially expressed genes is in Supplementary Table V (Part F).

Here we point to a set of non-coding genes that may be involved in molecular mechanisms of transformation that occur in each step of *S. mansoni* development.

In the second analysis approach we looked for differentially expressed genes with enriched expression in at least one stage, with the strategy explained before. We found 577 differentially expressed genes (Table III). Among these genes we found 473 protein-coding genes, and 104 non-protein coding genes. All 577 differentially expressed genes are listed in Supplementary Table V (Part G); the non-protein-coding genes signature is represented in Figure 5. In this profile of non-protein-coding genes we highlight one transcript that maps to an

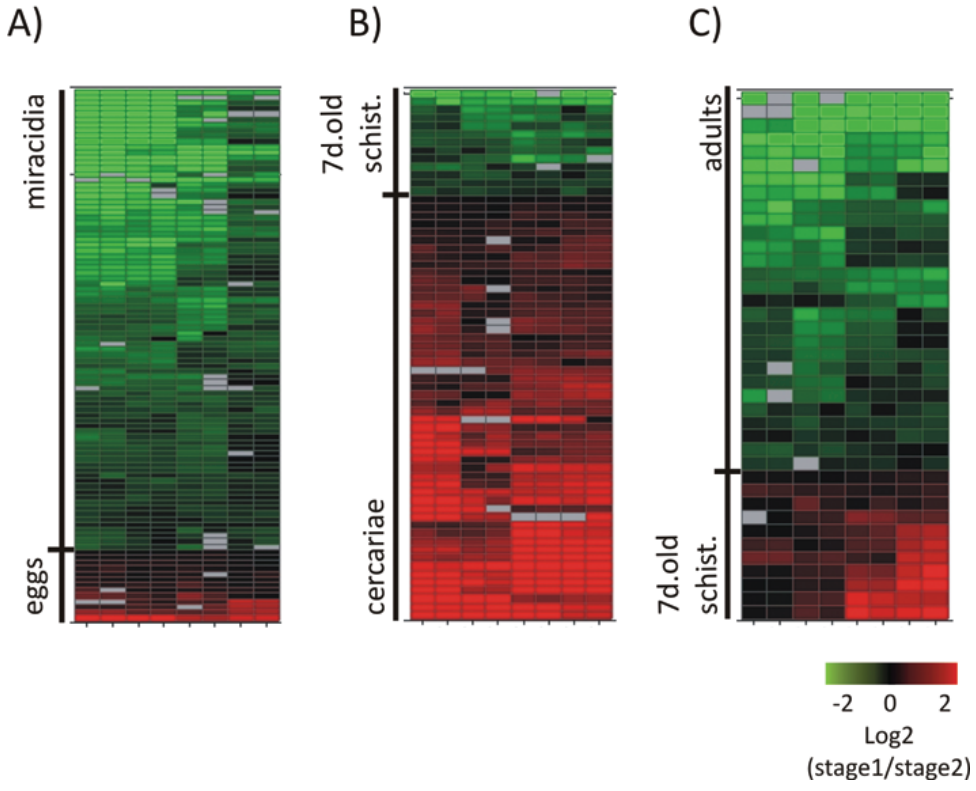


Fig. 4 – Heat map of differential expression of non-protein coding genes between two developmental stages. **A)** eggs *vs.* miracidia; **B)** cercariae *vs.* 7-day-old schistosomula; **C)** 7-day-old schistosomula *vs.* adult worms. Each line represents a gene and each column represents a replica (4 technical replicas for each one of two biological replicas). Color is proportional to expression levels of the gene in a given stage compared to the next, according to the range indicated in the figure insert, and it is calculated as the log2 of the expression ratio in stage 1 (first in the pairwise comparison)/stage 2 (second in the pairwise comparison).

intron of Smp_014290; the Smp_014290 protein-coding gene has 3 alternatively spliced isoforms and it is possible that the non-coding RNA transcribed from its intron eventually acts in alternative splicing modulation; to confirm this hypothesis further studies are necessary.

The signature of stage-enriched expression comprises a new set of transcripts that should receive special attention in the future. This set reveals new molecular targets of specific mechanisms involved in the biology of each stage that should be explored to understand the parasite complexity.

FINAL CONSIDERATIONS / PERSPECTIVES

This is the first review on non-coding RNAs in schistosomes in the literature. Very limited knowledge about schistosomes non-coding RNAs is available. A large further effort using deep-sequencing and tilling array approaches is necessary to finish the genome, to increase the amount of transcriptome data and to identify new non-coding RNAs, especially in *S. mansoni* that until now has been less studied than *S. japonicum* with the non-coding perspective.

Extending genome and transcriptome deep sequencing to invertebrate species other than the *C. elegans* and *D. melanogaster* model organisms may help to identify the classes of ncRNAs in the ancient species. In addition, functional studies are necessary to clarify the mechanisms involved in regulation of these ncRNA and their role in protein-coding gene expression regulation. Because of the peculiarities of schistosomes and the current limited ability to obtain transgenic parasites, characterization of the molecular mechanisms of ncRNA function in schistosomes will be a big challenge.

Here we showed experimental evidence of regulation of non-coding RNAs expression in different bio-

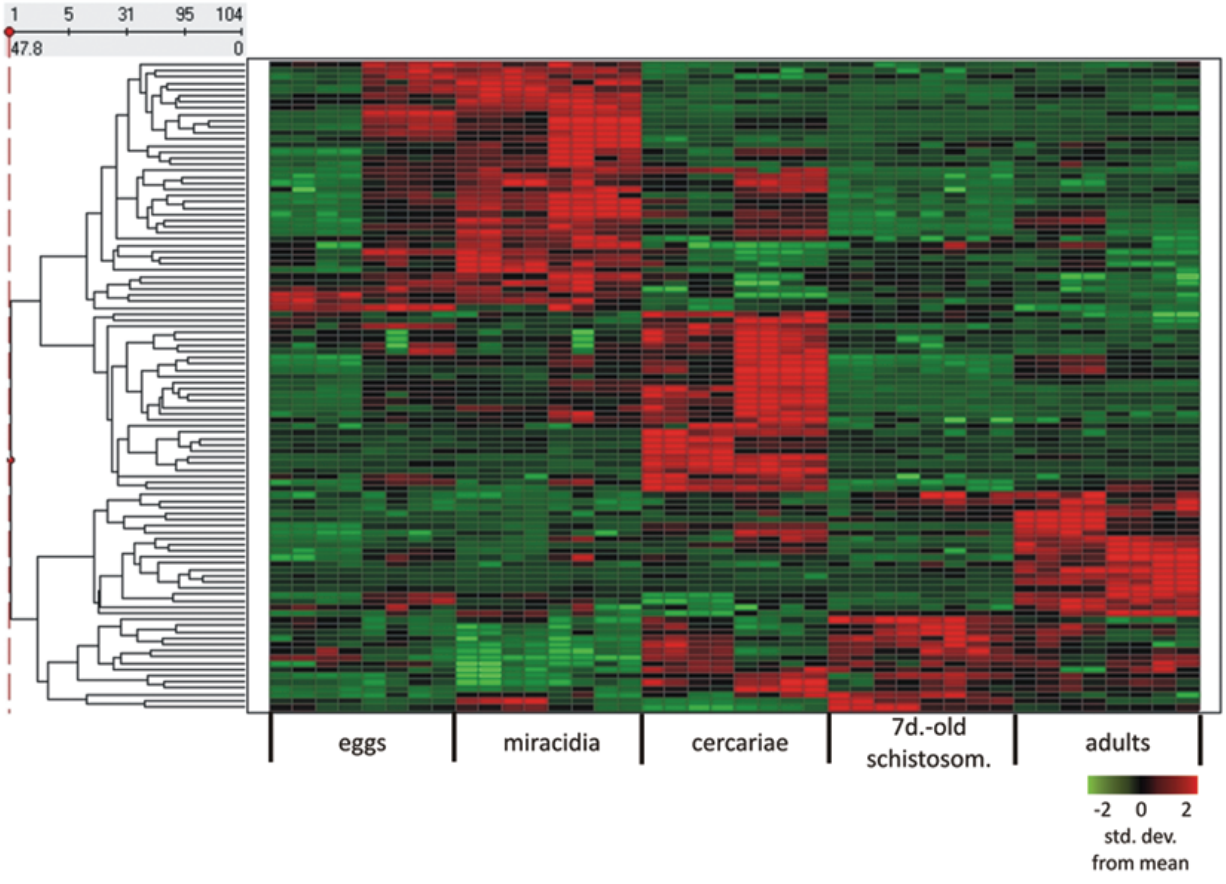


Fig. 5 – Heat map of differential expression of non-protein coding genes among 5 developmental stages. Each line represents a gene and each column represents a replica (4 technical replicas for each one of two biological replicas). Color is proportional to expression levels of the gene in a given stage, according to the range shown in the figure insert, and indicates the number of standard deviations below (green) or above (red) the average expression of that gene across all stages.

logical situations: adult parasites, parasite response to TNF-α host molecule and parasite life cycle stages. These data collections are important evidence of functionality of these non-coding RNAs, and a detailed further characterization of the mechanisms of action is needed. Understanding of the non-coding genes as new players in the biology of schistosomes will shed light on the complexity of processes involved in host-parasite interaction and parasite development.

ACKNOWLEDGMENTS

Funded in part by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and from the SEtTReND grant agreement number 241865 of the FP-7 European Community. KCO and VMC received fellowships from FAPESP; MLPC received a fellowship from CAPES-Brasil, and SVA received an established investigator fellowship award from CNPq.

SUPPLEMENTARY MATERIALS

All supplementary materials are available for download from the following site:
http://www2.iq.usp.br/docente/verjo/downloads/oliveira_kc/2011/

RESUMO

RNAs não codificadores (ncRNAs) têm sido recentemente objeto de atenção muito maior devido aos avanços técnicos no sequenciamento que expandiram a caracterização dos transcritomas em diferentes organismos. ncRNAs possuem diferentes comprimentos (22 nt a >1.000 nt) e mecanismos de

ação que essencialmente compreendem uma sofisticada rede de regulação de expressão gênica. A publicação recente dos genomas e transcritomas dos esquistossomos aumentou a descrição e caracterização de um grande número de genes do parasita. Aqui nós revisamos o número de genes preditos e a cobertura das bases do genoma em face dos ESTs públicos disponíveis, incluindo uma avaliação crítica da evidência e caracterização de ncRNAs em esquistossomos. Nós mostramos dados de expressão de ncRNAs em *Schistosoma mansoni*. Nós analisamos três conjuntos diferentes de dados de experimentos com microarranjos: (1) medidas de expressão em larga escala de vermes adultos; (2) genes diferencialmente expressos de *S. mansoni* regulados por uma citocina humana (TNF-α) no parasita em cultura; e (3) expressão estágio-especifica de ncRNAs. Todos estes dados apontam para ncRNAs envolvidos em diferentes processos biológicos e respostas fisiológicas que sugerem funcionalidade destes novos personagens na biologia do parasita. Explorar este mundo é um desafio para os cientistas sob uma nova perspectiva molecular da interação parasita-hospedeiro e do desenvolvimento do parasita.

Palavras-chave: *Schistosoma mansoni*, RNAs não-codificadores, perfil de expressão gênica, genoma, transcritoma.

REFERENCES

AGBOH KC, WEBB TE, EVANS RJ AND ENNION SJ. 2004. Functional characterization of a P2X receptor from *Schistosoma mansoni*. J Biol Chem 279: 41650–41657.

AMIRI P, LOCKSLEY RM, PARSLOW TG, SADICK M, RECTOR E, RITTER D AND MCKERROW JH. 1992. Tumour necrosis factor alpha restores granulomas and induces parasite egg-laying in schistosome-infected SCID mice. Nature 356: 604–607.

BARTEL DP. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281–297.

BECKMANN S, BURO C, DISSOUS C, HIRZMANN J AND GREVELDING CG. 2010. The Syk kinase SmTK4 of *Schistosoma mansoni* is involved in the regulation of spermatogenesis and oogenesis. PLoS Pathog 6: e1000769.

BEGUN DJ ET AL. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5: e310.

BERRIMAN M ET AL. 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460: 352–358.

BIRNEY E ET AL. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

BOYLE JP, WU XJ, SHOEMAKER CB AND YOSHINO TP. 2003. Using RNA interference to manipulate endogenous gene expression in *Schistosoma mansoni* sporocysts. Mol Biochem Parasitol 128: 205–215.

BRITO GC, FACHEL AA, VETTORE AL, VIGNAL GM, GIMBA ER, CAMPOS FS, BARCINSKI MA, VERJOVSKI-ALMEIDA S AND REIS EM. 2008. Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. Mol Carcinog 47: 757–767.

BROSNAN CA AND VOINNET O. 2009. The long and the short of noncoding RNAs. Curr Opin Cell Biol 21: 416–425.

CARNINCI P ET AL. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13: 1273–1289.

CAWLEY S ET AL. 2004. Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. Cell 116: 499–509.

CHALMERS IW, MCARDLE AJ, COULSON RM, WAGNER MA, SCHMID R, HIRAI H AND HOFFMANN KF. 2008. Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the *Schistosoma mansoni* venom allergen-like (SmVAL) gene family. BMC genomics 9: 89.

CHEN D, FARWELL MA AND ZHANG B. 2010a. MicroRNA as a new player in the cell cycle. J Cell Physiol 225: 296–301.

CHEN J, YANG Y, GUO S, PENG J, LIU Z, LI J, LIN J AND CHENG G. 2010b. Molecular cloning and expression profiles of Argonaute proteins in *Schistosoma japonicum*. Parasitol Res 107: 889–899.

CHEN L-L AND CARMICHAEL GG. 2010. Long noncoding RNAs in mammalian cells: what, where, and why? WIREs RNA 1: 19.

CHENG GF, LIN JJ, SHI Y, JIN YX, FU ZQ, JIN YM, ZHOU YC AND CAI YM. 2005. Dose-dependent inhibition of gynecophoral canal protein gene expression *in vitro* in the schistosome (*Schistosoma japonicum*) by RNA interference. Acta Biochim Biophys Sin (Shanghai) 37: 386–390.

CHURCH DM ET AL. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7: e1000112.

CHURCHILL GA. 2004. Using ANOVA to analyze microarray data. Biotechniques 37: 173–175, 177.

CLAVERIE JM. 2005. Fewer genes, more noncoding RNA. Science 309: 1529–1530.

COPELAND CS, MARZ M, ROSE D, HERTEL J, BRINDLEY PJ, SANTANA CB, KEHR S, ATTOLINI CS AND STADLER PF. 2009. Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. BMC Genomics 10: 464.

CORE LJ, WATERFALL JJ AND LIS JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322: 1845–1848.

CORRENTI JM, BRINDLEY PJ AND PEARCE EJ. 2005. Long-term suppression of cathepsin B levels by RNA interference retards schistosome growth. Mol Biochem Parasitol 143: 209–215.

COSTA FF. 2010. Non-coding RNAs: Meet thy masters. Bioessays 32: 599–608.

DAHARY D, ELROY-STEIN O AND SOREK R. 2005. Naturally occurring antisense: transcriptional leakage or real overlap? Genome Res 15: 364–368.

DAVIES SJ, GROGAN JL, BLANK RB, LIM KC, LOCKSLEY RM AND MCKERROW JH. 2001. Modulation of blood fluke development in the liver by hepatic CD4+ lymphocytes. Science 294: 1358–1361.

DE LUCIA F AND DEAN C. 2010. Long non-coding RNAs and chromatin regulation. Curr Opin Plant Biol 14: 168–173.

DE MENDONCA RL, ESCRIVA H, BOUTON D, LAUDET V AND PIERCE RJ. 2000. Hormones and nuclear receptors in schistosome development. Parasitol Today 16: 233–240.

DELCROIX M, SAJID M, CAFFREY CR, LIM KC, DVORAK J, HSIEH I, BAHGAT M, DISSOUS C AND MCKERROW JH. 2006. A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. J Biol Chem 281: 39316–39329.

DEMARCO R, MATHIESON W, MANUEL SJ, DILLON GP, CURWEN RS, ASHTON PD, IVENS AC, BERRIMAN M, VERJOVSKI-ALMEIDA S AND WILSON RA. 2010. Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts. Genome Res 20: 1112–1121.

DEMARCO R, OLIVEIRA KC, VENANCIO TM AND VERJOVSKI-ALMEIDA S. 2006. Gender biased differential alternative splicing patterns of the transcriptional cofactor CA150 gene in *Schistosoma mansoni*. Mol Biochem Parasitol 150: 123–131.

DILLON GP, FELTWELL T, SKELTON JP, ASHTON PD, COULSON PS, QUAIL MA, NIKOLAIDOU-KATSARIDOU N, WILSON RA AND IVENS AC. 2006. Microarray analysis identifies genes preferentially expressed in the lung schistosomulum of *Schistosoma mansoni*. Int J Parasitol 36: 1–8.

DINDOT SV, PERSON R, STRIVENS M, GARCIA R AND BEAUDET AL. 2009. Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. Genome Res 19: 1374–1383.

DINGER ME, AMARAL PP, MERCER TR AND MATTICK JS. 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic 8: 407–423.

DINGUIRARD N AND YOSHINO TP. 2006. Potential role of a CD36-like class B scavenger receptor in the binding of modified low-density lipoprotein (acLDL) to the tegumental surface of *Schistosoma mansoni* sporocysts. Mol Biochem Parasitol 146: 219–230.

ESCOBEDO G, ROBERTS CW, CARRERO JC AND MORALES-MONTOR J. 2005. Parasite regulation by host hormones: an old mechanism of host exploitation? Trends Parasitol 21: 588–593.

FAGHIRI Z AND SKELLY PJ. 2009. The role of tegumental aquaporin from the human parasitic worm, *Schistosoma mansoni*, in osmoregulation and drug uptake. FASEB J 23: 2780–2789.

FERBEYRE G, SMITH JM AND CEDERGREN R. 1998. Schistosome satellite DNA encodes active hammerhead ribozymes. Mol Cell Biol 18: 3880–3888.

FITZPATRICK JM, PEAK E, PERALLY S, CHALMERS IW, BARRETT J, YOSHINO TP, IVENS AC AND HOFFMANN KF. 2009. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. PLoS Negl Trop Dis 3: e543.

FRANCO GR, ADAMS MD, SOARES MB, SIMPSON AJ, VENTER JC AND PENA SD. 1995. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. Gene 152: 141–147.

FRANCO GR, VALADAO AF, AZEVEDO V AND RABELO EM. 2000. The *Schistosoma* gene discovery program: state of the art. Int J Parasitol 30: 453–463.

FREITAS TC, JUNG E AND PEARCE EJ. 2007. TGF-beta signaling controls embryo development in the parasitic flatworm *Schistosoma mansoni*. PLoS Pathog 3: e52.

GERSTEIN MB, BRUCE C, ROZOWSKY JS, ZHENG D, DU J, KORBEL JO, EMANUELSSON O, ZHANG ZD, WEISSMAN S AND SNYDER M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res 17: 669–681.

GERSTEIN MB ET AL. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330: 1775–1787.

GOBERT GN, MOERTEL L, BRINDLEY PJ AND MCMANUS DP. 2009. Developmental gene expression profiles of the human pathogen *Schistosoma japonicum*. BMC Genomics 10: 128.

GOLDEN DE, GERBASI VR AND SONTHEIMER EJ. 2008. An inside job for siRNAs. Mol Cell 31: 309–312.

GOMES MS, CABRAL FJ, JANNOTTI-PASSOS LK, CARVALHO O, RODRIGUES V, BABA EH AND SA RG. 2009. Preliminary analysis of miRNA pathway in *Schistosoma mansoni*. Parasitol Int 58: 61–68.

GRAVELEY BR ET AL. 2010. The developmental transcriptome of *Drosophila melanogaster*. Nature 471: 473–479.

GRYSEELS B, POLMAN K, CLERINX J AND KESTENS L. 2006. Human schistosomiasis. Lancet 368: 1106–1118.

GUTTMAN M ET AL. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.

HAN ZG, BRINDLEY PJ, WANG SY AND CHEN Z. 2009. *Schistosoma* genomics: new perspectives on schistosome biology and host-parasite interaction. Annu Rev Genomics Hum Genet 10: 211–240.

HAO L, CAI P, JIANG N, WANG H AND CHEN Q. 2010. Identification and characterization of microRNAs and endogenous siRNAs in *Schistosoma japonicum*. BMC Genomics 11: 55.

HUANG J, HAO P, CHEN H, HU W, YAN Q, LIU F AND HAN ZG. 2009. Genome-wide identification of *Schistosoma japonicum* microRNAs using a deep-sequencing approach. PLoS One 4: e8206.

HU W ET AL. 2003. Evolutionary and biomedical implications of a *Schistosoma japonicum* complementary DNA resource. Nat Genet 35: 139–147.

HUTTENHOFER A, SCHATTNER P AND POLACEK N. 2005. Non-coding RNAs: hope or hype? Trends Genet 21: 289–297.

JOHNSON JM, EDWARDS S, SHOEMAKER D AND SCHADT EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21: 93–102.

JOLLY ER, CHIN CS, MILLER S, BAHGAT MM, LIM KC, DERISI J AND MCKERROW JH. 2007. Gene expression patterns during adaptation of a helminth parasite to different environmental niches. Genome Biol 8: R65.

KAPRANOV P ET AL. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.

KAUL S ET AL. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

KEELING PJ AND SLAMOVITS CH. 2005. Causes and effects of nuclear genome reduction. Curr Opin Genet Dev 15: 601–608.

KHAYATH N, VICOGNE J, AHIER A, BENYOUNES A, KONRAD C, TROLET J, VISCOGLIOSI E, BREHM K AND DISSOUS C. 2007. Diversification of the insulin receptor family in the helminth parasite *Schistosoma mansoni*. Febs J 274: 659–676.

KIYOSAWA H, MISE N, IWASE S, HAYASHIZAKI Y AND ABE K. 2005. Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 15: 463–474.

KONG L, ZHANG Y, YE ZQ, LIU XQ, ZHAO SQ, WEI L AND GAO G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35: W345–349.

KRAUTZ-PETERSON G, BHARDWAJ R, FAGHIRI Z, TARARAM CA AND SKELLY PJ. 2010. RNA interference in schistosomes: machinery and methodology. Parasitology 137: 485–495.

KRAUTZ-PETERSON G, RADWANSKA M, NDEGWA D, SHOEMAKER CB AND SKELLY PJ. 2007. Optimizing gene suppression in schistosomes using RNA interference. Mol Biochem Parasitol 153: 194–202.

KRAUTZ-PETERSON G AND SKELLY PJ. 2008a. *Schistosoma mansoni*: the dicer gene and its expression. Exp Parasitol 118: 122–128.

KRAUTZ-PETERSON G AND SKELLY PJ. 2008b. Schistosome asparaginyl endopeptidase (legumain) is not essential for cathepsin B1 activation *in vivo*. Mol Biochem Parasitol 159: 54–58.

KUMAGAI T, OSADA Y, OHTA N AND KANAZAWA T. 2009. Peroxiredoxin-1 from *Schistosoma japonicum* functions as a scavenger against hydrogen peroxide but not nitric oxide. Mol Biochem Parasitol 164: 26–31.

KUMAR M, MABALIRAJAN U, AGRAWAL A AND GHOSH B. 2010. Proinflammatory role of let-7 miRNAs in experimental asthma? J Biol Chem 285: le19: author reply le20.

LI N, MUTHUSAMY S, LIANG R, SAROJINI H AND WANG E. 2010. Increased expression of miR-34a and miR-93

in rat liver during aging; and their impact on the expression of Mgst1 and Sirt1. Mech Ageing Dev 132: 75–85.

LOURO R, NAKAYA HI, AMARAL PP, FESTA F, SOGAYAR MC, DA SILVA AM, VERJOVSKI-ALMEIDA S AND REIS EM. 2007. Androgen responsive intronic non-coding RNAs. BMC biology 5: 4.

LOURO R, SMIRNOVA AS AND VERJOVSKI-ALMEIDA S. 2009. Long intronic noncoding RNA transcription: expression noise or expression choice? Genomics 93: 291–298.

LUO R, XUE X, WANG Z, SUN J, ZOU Y AND PAN W. 2010. Analysis and characterization of the genes encoding the Dicer and Argonaute proteins of *Schistosoma japonicum*. Parasit Vectors 3: 90.

MATTICK JS. 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays 25: 930–939.

MATTICK JS. 2004. RNA regulation: a new genetics? Nat Rev Genet 5: 316–323.

MATTICK JS. 2009. The genetic signatures of noncoding RNAs. PLoS Genet 5: e1000459.

MERCER TR, DINGER ME AND MATTICK JS. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.

MERRICK JM, OSMAN A, TSAI J, QUACKENBUSH J, LOVERDE PT AND LEE NH. 2003. The *Schistosoma mansoni* gene index: gene discovery and biology by reconstruction and analysis of expressed gene sequences. J Parasitol 89: 261–269.

MORALES ME, RINALDI G, GOBERT GN, KINES KJ, TORT JF AND BRINDLEY PJ. 2008. RNA interference of *Schistosoma mansoni* cathepsin D, the apical enzyme of the hemoglobin proteolysis cascade. Mol Biochem Parasitol 157: 160–168.

MORTAZAVI A, WILLIAMS BA, MCCUE K, SCHAEFFER L AND WOLD B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.

NAG A AND JACK T. 2010. Sculpting the flower; the role of microRNAs in flower development. Curr Top Dev Biol 91: 349–378.

NAKAYA HI, AMARAL PP, LOURO R, LOPES A, FACHEL AA, MOREIRA YB, EL-JUNDI TA, DA SILVA AM, REIS EM AND VERJOVSKI-ALMEIDA S. 2007. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 8: R43.

NANA-SINKAM SP, KARSIES T, RISCILI B, EZZIE M AND PIPER M. 2009. Lung microRNA: from development to disease. Expert Rev Respir Med 3: 373–385.

NDEGWA D, KRAUTZ-PETERSON G AND SKELLY PJ. 2007. Protocols for gene silencing in schistosomes. Exp Parasitol 117: 284–291.

OLIVEIRA KC, CARVALHO ML, VENANCIO TM, MIYASATO PA, KAWANO T, DEMARCO R AND VERJOVSKI-ALMEIDA S. 2009. Identification of the *Schistosoma mansoni* TNF-alpha receptor gene and the effect of human TNF-alpha on the parasite gene expression profile. PLoS Negl Trop Dis 3: e556.

OLIVER HF, ORSI RH, PONNALA L, KEICH U, WANG W, SUN Q, CARTINHOUR SW, FILIATRAULT MJ, WIEDMANN M AND BOOR KJ. 2009. Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. BMC Genomics 10: 641.

OSMAN A, NILES EG, VERJOVSKI-ALMEIDA S AND LOVERDE PT. 2006. *Schistosoma mansoni* TGF-beta receptor II: role in host ligand-induced regulation of a schistosome target gene. PLoS Pathog 2: e54.

PEREIRA TC, PASCOAL VD, MARCHESINI RB, MAIA IG, MAGALHAES LA, ZANOTTI-MAGALHAES EM AND LOPES-CENDES I. 2008. *Schistosoma mansoni*: evaluation of an RNAi-based treatment targeting HGPRTase gene. Exp Parasitol 118: 619–623.

PRATT AJ AND MACRAE IJ. 2009. The RNA-induced silencing complex: a versatile gene-silencing machine. J Biol Chem 284: 17897–17901.

REARICK D, PRAKASH A, MCSWEENY A, SHEPARD SS, FEDOROVA L AND FEDOROV A. 2010. Critical association of ncRNA with introns. Nucleic Acids Res 39: 2357–2366.

REIS EM, LOURO R, NAKAYA HI AND VERJOVSKI-ALMEIDA S. 2005. As antisense RNA gets intronic. Omics 9: 2–12.

REIS EM ET AL. 2004. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene 23: 6684–6692.

RINALDI G, MORALES ME, ALREFAEI YN, CANCELA M, CASTILLO E, DALTON JP, TORT JF AND BRINDLEY PJ. 2009. RNA interference targeting leucine aminopeptidase blocks hatching of *Schistosoma mansoni* eggs. Mol Biochem Parasitol 167: 118–126.

RINN JL ET AL. 2007. Functional demarcation of active

and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.

ROGER E, GRUNAU C, PIERCE RJ, HIRAI H, GOURBAL B, GALINIER R, EMANS R, CESARI IM, COSSEAU C AND MITTA G. 2008. Controlled chaos of polymorphic mucins in a metazoan parasite (*Schistosoma mansoni*) interacting with its invertebrate host (*Biomphalaria glabrata*). PLoS Negl Trop Dis 2: e330.

SCHIER AF. 2007. The maternal-zygotic transition: death and birth of RNAs. Science 316: 406–407.

SHAFFER JP. 1995. Multiple Hypothesis Testing. Annu Rev Psychol 46: 23.

SIMOES MC, LEE J, DJIKENG A, CERQUEIRA GC, ZERLOTINI A, DA SILVA-PEREIRA RA, DALBY AR, LOVERDE P, EL-SAYED NM AND OLIVEIRA G. 2011. Identification of *Schistosoma mansoni* microRNAs. BMC Genomics 12: 47.

SKELLY PJ, DA'DARA A AND HARN DA. 2003. Suppression of cathepsin B expression in *Schistosoma mansoni* by RNA interference. Int J Parasitol 33: 363–369.

SMITH EM AND GREGORY TR. 2009. Patterns of genome size diversity in the ray-finned fishes. Hydrobiologia 625: 1–25.

ST LAURENT G 3RD, FAGHIHI MA AND WAHLESTEDT C. 2009. Non-coding RNA transcripts: sensors of neuronal stress, modulators of synaptic plasticity, and agents of change in the onset of Alzheimer's disease. Neurosci Lett 466: 81–88.

STEINMANN P, KEISER J, BOS R, TANNER M AND UTZINGER J. 2006. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis 6: 411–425.

THE_*C.elegans*_SEQUENCING_CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.

THE_UNIPROT_CONSORTIUM. 2010. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.

TRINKLEIN ND, ALDRED SF, HARTMAN SJ, SCHROEDER DI, OTILLAR RP AND MYERS RM. 2004. An abundance of bidirectional promoters in the human genome. Genome Res 14: 62–66.

TUSHER VG, TIBSHIRANI R AND CHU G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116–5121.

VERJOVSKI-ALMEIDA S AND DEMARCO R. 2011. Gene structure and splicing in schistosomes. J Proteomics doi: 10.1016/j.jprot.2011.03.022.

VERJOVSKI-ALMEIDA S, LEITE LC, DIAS-NETO E, MENCK CF AND WILSON RA. 2004. Schistosome transcriptome: insights and perspectives for functional genomics. Trends Parasitol 20: 304–308.

VERJOVSKI-ALMEIDA S, VENANCIO TM, OLIVEIRA KC, ALMEIDA GT AND DEMARCO R. 2007. Use of a 44k oligoarray to explore the transcriptome of *Schistosoma mansoni* adult worms. Exp Parasitol 117: 236–245.

VERJOVSKI-ALMEIDA S ET AL. 2003. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nature genetics 35: 148–157.

VERMEIRE JJ, TAFT AS, HOFFMANN KF, FITZPATRICK JM AND YOSHINO TP. 2006. *Schistosoma mansoni*: DNA microarray gene expression profiling during the miracidium-to-mother sporocyst transformation. Mol Biochem Parasitol 147: 39–47.

WANG E, MILLER LD, OHNMACHT GA, LIU ET AND MARINCOLA FM. 2000. High-fidelity mRNA amplification for gene profiling. Nat Biotechnol 18: 457–459.

WANG Z, XUE X, SUN J, LUO R, XU X, JIANG Y, ZHANG Q AND PAN W. 2010. An "in-depth" description of the small non-coding RNA population of *Schistosoma japonicum* schistosomulum. PLoS Negl Trop Dis 4: e596.

WASHIETL S, HOFACKER IL, LUKASSER M, HUTTENHOFER A AND STADLER PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23: 1383–1390.

WEBSTER JP, OLIVIERA G, ROLLINSON D AND GOWER CM. 2010. Schistosome genomes: a wealth of information. Trends Parasitol 26: 103–106.

WERNER A AND BERDAL A. 2005. Natural antisense transcripts: sound or silence? Physiol Genomics 23: 125–131.

WERNER A AND SWAN D. 2010. What are natural antisense transcripts good for? Biochem Soc Trans 38: 1144–1149.

WILHELM BT, MARGUERAT S, WATT S, SCHUBERT F, WOOD V, GOODHEAD I, PENKETT CJ, ROGERS J AND BAHLER J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239–1243.

WILUSZ JE, SUNWOO H AND SPECTOR DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23: 1494–1504.

WINSTON WM, MOLODOWITCH C AND HUNTER CP. 2002. Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. Science 295: 2456–2459.

WU W, NILES EG AND LOVERDE PT. 2007. Thyroid

hormone receptor orthologues from invertebrate species with emphasis on *Schistosoma mansoni*. BMC Evol Biol 7: 150.

XUE X, SUN J, ZHANG Q, WANG Z, HUANG Y AND PAN W. 2008. Identification and characterization of novel microRNAs from *Schistosoma japonicum*. PLoS One 3: e4034.

ZHANG Y, LIU XS, LIU QR AND WEI L. 2006. Genome-wide *in silico* identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. Nucleic Acids Res 34: 3465–3475.

ZHAO ZR, LEI L, LIU M, ZHU SC, REN CP, WANG XN AND SHEN JJ. 2008. *Schistosoma japonicum*: inhibition of *Mago nashi* gene expression by shRNA-mediated RNA interference. Exp Parasitol 119: 379–384.

ZHOU Y ET AL. 2009. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. Nature 460: 345–351.

ZOU X, JIN YM, LIU PP, WU QJ, LIU JM AND LIN JJ. 2010. RNAi silencing of calcium-regulated heat-stable protein of 24 kDa in *Schistosoma japonicum* affects parasite growth. Parasitol Res 108: 567–572.

Janus - Sistema Administrativo da Pós-Graduação

Universidade de São Paulo

Documento sem validade oficial

FICHA DO ALUNO

95131 - 6210217/1 - Vinicius Ramos Henriques Maracajá Coutinho

Email:	maracaja@iq.usp.br
Data de Nascimento:	23/07/1984
Cédula de Identidade:	RG - 2596069 - PB
Local de Nascimento:	Estado da Paraíba
Nacionalidade:	Brasileira
Graduação:	Ciências Biológicas - Universidade Federal da Paraíba - Paraíba - Brasil - 2006

Curso:	Doutorado Direto
Programa:	Bioinformática (1)
Data de Matrícula:	30/07/2007
Início da Contagem de Prazo:	30/07/2007
Data Limite:	30/07/2013
Orientador:	Prof(a). Dr(a). Sergio Verjovski de Almeida - 30/07/2007 até o presente. E.Mail: verjo@iq.usp.br
Coorientador:	Prof(a). Dr(a). João Carlos Setubal - 22/04/2009 até o presente. E.Mail: setubal@iq.usp.br
Data de Aprovação no Exame de Qualificação:	Aprovado em 16/12/2010
Data do Depósito do Trabalho:	
Título do Trabalho:	
Data Máxima para Aprovação da Banca:	
Data de Aprovação da Banca:	
Data Máxima para Defesa:	
Data da Defesa:	
Resultado da Defesa:	
Histórico de Ocorrências:	Ingressou no Doutorado Direto em 30/07/2007 Matrícula de Acompanhamento em 18/02/2013

Última ocorrência: Matrícula de Acompanhamento em 18/02/2013

Impresso em: 01/07/13 10:22:47

Janus - Sistema Administrativo da Pós-Graduação

Universidade de São Paulo

Documento sem validade oficial

FICHA DO ALUNO

95131 - 6210217/1 - Vinicius Ramos Henriques Maracajá Coutinho

Sigla	Nome da Disciplina	Início	Término	Carga Horária	Cred.	Freq.	Conc.	Exc.	Situação
IBI5031-2/3	Reconhecimento de Padrões I	06/08/2007	28/10/2007	120	8	85	A	N	Concluída
MAE5755-5/2	Métodos Estatísticos Aplicados às Ciências Biológicas (Instituto de Matemática e Estatística - Universidade de São Paulo)	06/08/2007	23/11/2007	120	8	80	C	N	Concluída
BIO5784-2/3	Organização do Genoma Humano (Instituto de Biociências - Universidade de São Paulo)	15/08/2007	28/11/2007	120	8	86,6	A	N	Concluída
IBI5011-2/1	Introdução à Computação para Bioinformática	03/03/2008	13/06/2008	120	8	100	A	N	Concluída
MAC5719-1/1	Tópicos em Bioinformática (Instituto de Matemática e Estatística - Universidade de São Paulo)	03/03/2008	13/06/2008	120	8	100	A	N	Concluída
MAC5760-6/1	Introdução aos Sistemas de Bancos de Dados (Instituto de Matemática e Estatística - Universidade de São Paulo)	03/03/2008	13/06/2008	120	0	0	-	N	Matrícula cancelada
BMP5762-2/1	Bioinformática Aplicada ao Estudo de Doenças Parasitárias (Instituto de Ciências Biomédicas - Universidade de São Paulo)	06/08/2008	14/10/2008	60	4	87	A	N	Concluída
IBI5013-2/1	Banco de Dados para Bioinformática	02/03/2009	24/05/2009	120	8	100	B	N	Concluída
BIO5713-3/1	O Mundo dos RNAs (Instituto de Biociências - Universidade de São Paulo)	06/03/2009	03/07/2009	120	8	75	A	N	Concluída
PSA5866-3/5	Preparação Pedagógica (Instituto de Psicologia - Universidade de São Paulo)	10/03/2009	20/04/2009	30	0	0	-	N	Pré-matrícula indeferida
QBQ5825-8/1	Prática de Ensino de Química e Bioquímica (Instituto de Química - Universidade de São Paulo)	15/03/2010	30/06/2010	45	0	0	-	N	Matrícula cancelada
EDM5102-1/2	Preparação Pedagógica PAE (Faculdade de Educação - Universidade de São Paulo)	13/05/2010	14/07/2010	90	6	100	A	N	Concluída

	Créditos mínimos exigidos		Créditos obtidos
	Para exame de qualificação	Para depósito de tese	
Disciplinas:	62	62	66
Atividades Programadas:			
Seminários:			
Estágios:			
Total:	62	62	66

Créditos Atribuídos à Tese: 130

Observações:

1) Unidades de Ensino responsáveis pelo programa: Escola Superior de Agricultura "Luiz de Queiroz" - Instituto de Biociências - Instituto de Ciências Biomédicas - Instituto de Matemática e Estatística - Instituto de Química - Faculdade de

Filosofia, Ciências e Letras de Ribeirão Preto - Instituto de Física de São Carlos.

Conceito a partir de 02/01/1997:

A - Excelente, com direito a crédito; B - Bom, com direito a crédito; C - Regular, com direito a crédito; R - Reprovado; T - Transferência.

Um(1) crédito equivale a 15 horas de atividade programada.

Última ocorrência: Matrícula de Acompanhamento em 18/02/2013

Impresso em: 01/07/13 10:22:47

Vinicius Ramos Henriques Maracajá Coutinho

São Paulo, SP, Brazil

Address: Rua Aimbere, 2049, Apto 3 – Perdizes – São Paulo, SP – Brazil.
Mobile: +55 11 8316 2500 / +55 83 8873 9721
Website: https://sites.google.com/site/maracajacoutinho/
E-mail: viniciusmaracaja@yahoo.com / maracaja.coutinho@gmail.com
Twitter: **@vin_maracaja**

Degrees

Mar.'07 – Present **PhD in Bioinformatics**
Instituto de Química, Universidade de São Paulo – USP
http://www.iq.usp.br/
Thesis: *In silico characterization of long intronic and antisense ncRNAs in the human genome.*
Advisors: Sergio Verjovski-Almeida & João C. Setubal

Out.'02 – Dez.'06 **B.Sc. in Biological Sciences**
Depto. Biologia Molecular, Universidade Federal da Paraíba – UFPB
http://www.ufpb.br/
Monograph: *Degenerate primers design for NHX1 Antiporter gene in Sorghum bicolor (L.) Moench*
Advisor: Rómulo M. Llamoca-Zárate

Publications

[9] *Paschoal, AR; ***Maracaja-Coutinho, V**; Setubal, JC; Simões, ZLP; Verjovski-Almeida, S; Durham, AM. *Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases*. **RNA Biology**. 2012.
*** Both authors contributed equally to this work.**

[8] Tahira, AC; Kubrusly, MS; Faria, MF; Dazzani, B; Fonseca, RS; **Maracaja-Coutinho, V**; Verjovski-Almeida, S; Machado, MCC; Reis, EM. *Long non-coding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer.* **Molecular Cancer**. 2011

[7] Oliveira, KPO; Carvalho, MLP; **Maracaja-Coutinho, V**; Kitajima, JP; Verjovski-Almeida, S. *Non-coding RNAs in Schistosomes: an unexplored world*. **An. Acad. Bras. Cienc.** 2011.

[6] Queiroz, ATL; **Maracaja-Coutinho, V**; Jardim, ACG; Rahal, P; Matioli, SR; Mello, IMVGC. *Relation of pretreatment sequence diversity in NS5A region of Hepatitis C virus gen 1 with immune response between peg-INF/rib therapy outcomes*. **Journal of Viral Hepatitis**, 2010.

[5] Dias, AM; Queiroz, ATL; **Maracaja-Coutinho, V**. *Schizophrenia, brain disease and meta-analyses: integrating the pieces and testing Fusar-Poli's hypothesis*. **Medical Hypotheses**, 2010.

[4] Dias, AM; **Maracaja-Coutinho, V**; Queiroz, ATL. *The role of Neuregulin 1 in schizophrenia: a bioinformatics approach*. **Nature Precedings**, 2009.

[3] Padilha, IQM; Durbano, JP; Martins, AB; Almeida, RS; **Maracaja-Coutinho, V**; Araujo, DAM. *Bioinformatics as an instrument for digital inclusion and biotechnology diffusion*. **Revista Extensão Cidadã**, 2008. (In portuguese)

[2] Araujo, DAM; **Maracaja-Coutinho, V**; Rego, TG; Padilha, IQM. *Genomics and Bioinformatics: importance and perspectives for Brazilian Northeast region*. **Ciência & Cotidiano**, 2005. (In portuguese)

[1] Costa, DA; Ribeiro, ILAC; Amaral CMM. Llamoca-Zarate, RM; Maracaja-Coutinho, V. *In vitro regeneration of shoots through the shoot apex of sorghum (Sorghum bicolor (L.) Moench).* ***XII Latin-American Plant Physiology Meeting & X Brazilian Plant Physiology Meeting***, 2005. (In Portuguese) [Expanded Abstract].

Grants

- Start-Up Chile Program 4th Round (www.startupchile.org) – US$ 40,000.00 (2012)
- SoftLayer Catalyst Program – US$ 24,000.00 in cloud services (2012)

Reviewer of Journals & Meetings

- 2nd ISCB European Student Council Symposium – Basel, Switzerland (2012)
- 8th ISCB Student Council Symposium – Long Beach, USA (2012)
- 7th ISCB Student Council Symposium – Vienna, Austria (2011)
- Journal of Medicinal Plants Research – JMPR (2010)

Awards & Fellowships

Awards

- 2012 – **Honorable Mention: Databases & Data Integration, Text Mining and Information Extraction**. X-Meeting 2012, Associação Brasileira de Bioinformática e Biologia Computacional – AB³C.
- 2007 – **Elo Cidadão Award**. Universidade Federal da Paraíba.
- 2007 – **Elo Cidadão Award**. Universidade Federal da Paraíba.
- 2005 – **2nd Best Place Poster Award**. X-Meeting 2005, Associação Brasileira de Bioinformática e Biologia Computacional – AB³C.
- 2005 – **Elo Cidadão Awards**. Universidade Federal da Paraíba.
- 2005 – **Expociências 2005 (1º Lugar, Etapa João Pessoa)** – Science Fair of Paraíba State.

Fellowships

- **Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP**

PhD Fellowship.

- **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES**

PhD Fellowship.

- **Brazilian Academy of Sciences – ABC**

Short-time Fellowship - Programa Aristídes Pacheco Leão de Apoio a Vocações Científicas.

- **Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq**

Undergraduate Research Fellowship.

Oral Presentations & Courses as Tutor

Events Oral Presentation

- **Hackers & Founders Córdoba**

ProCórdoba – Agencia para la promoción de las exportaciones.
Córdoba, Argentina, November 2012.

- ***NHX1* antiporter gene: salt resistance**

I Simpósio de Bioinformática da UFPB.
João Pessoa, December 2005.

- **Salt resistance: *primers* for Na+/H+ Antiporter gene in *Sorghum sudanense***

Econtro de Iniciação Científica da UFPB.
João Pessoa, December 2005.

- **Molecular characterization of intronic transcripts regulated by androgen in prostate cancer cell line**

XXXIV Reunião Anual da SBBq (Sociedade Brasileira de Bioquímica e Biologia Molecular)
Águas de Lindóia, July 2005.

Courses as Tutor

- **2010 Bioinformatics Summer Course**

Universidade de São Paulo / Associação Brasileira de Bioinformática e Biologia Computacional
São Paulo, 25 January - 02 February 2010. 75 *hours.*

- **2009 Bioinformatics Summer Course**

Universidade de São Paulo / Associação Brasileira de Bioinformática e Biologia Computacional
São Paulo, 02-06 February 2010. *75 hours.*

- **III Bioinformatics & Molecular Biology Course: Genomics, Transcriptomics and Proteomics**

Universidade Federal da Paraíba
João Pessoa, 11-15 September de 2006. *35 hours.*

- **Introduction to Bioinformatics**

Universidade Federal da Paraíba / Semana da Biologia da UFPB
João Pessoa, 11-22 September 2006. *10 hours.*

- **II Bioinformatics & Molecular Biology Course**

Universidade Federal da Paraíba
João Pessoa, 27 March - 11 April 2006. *30 hours.*

Events Organization

- **Startup Weekend Santiago**

Kauffman Foundation
Santiago, Chile, 22-24 March 2005.

- **I Bioinformatics & Molecular Biology Course**

Universidade Federal da Paraíba
João Pessoa, 25 February - 04 March 2005. *30 hours.*

Supplementary Education

30 Mar.'09 – 02 Apr.'09	**RNA Structure and Function** Internat. Centre for Genetic Engineering and Biotechnology – ICGBE Trieste, Italy. *20 hours*
25 Aug.'08 – 28 Aug.'08	**RNA Bioinformatics** Brazilian Symposium on Bioinformatics Santo André, SP. *08 hours*
25 Aug.'08 – 28 Aug.'08	**Comparative Genomics** Brazilian Symposium on Bioinformatics Santo André, SP. *08 hours*
10 Jul.'06 – 14 Jul.'06	**Nucleic Acids manipulation: PCR, RT-PCR and qPCR** Universidade Estadual Paulista – UNESP Botucatu, SP. *40 hours*
06 Jul.'06 – 07 Jul.'06	**Introduction to the methods for structural protein characterization**

	Laboratório Nacional de Luz Síncroton – LNLS Campinas, SP. *16 hours*
16 Jan.'06 – 28 Jan.'06	**Summer Course: Genomics, Proteomics and Cellular Universe** Universidade de São Pauo – USP Ribeirão Preto, SP. *80 hours*
19 Feb.'04 – 23 Feb.'04	**Extremophiles Biology** Universidade Federal da Paraíba – UFPB João Pessoa, PB. *10 hours*
03 Aug.'03 – 06 Aug.'03	**Molecular Biology Techniques** Encontro Nacional de Biólogos Natal, RN. *4h30 hours*
03 Aug.'03 – 06 Aug.'03	**Genetically Modified Organisms** Encontro Nacional de Biólogos Natal, RN. *4h30 hours*
31 Mar.'03 – 04 Apr.'03	**Introductoin to Perl** Universidade Federal da Paraíba – UFPB João Pessoa, PB. *12 hours*
17 Feb.'03 – 22 Feb.'03	**Introduction to Bioinformática** Universidade Federal da Paraíba – UFPB João Pessoa, PB. *10 hours*

Abstracts in Meetings

[19] *Paschoal, AR; ***Maracaja-Coutinho, V**; Setubal, JC; Simões, ZLP; Verjovski-almeida, S; Durham, AM. *Non-coding transcription and annotation: a guide and web resource for non-coding RNAs databases*. **X-Meeting 2012**, Campinas, Brazil, 2012.

[18] **Maracaja-Coutinho, V**; Beckedorff, FCF; Amaral, MS; Moreira YB; Setubal, JC; Reis, EM; Verjovski-almeida, S. *Expression analysis of the non-coding transcriptome in a prostate cancer cell line revealed novel androgen-regulated intronic ncRNAs*. **ISMB 2011 - 19th International Conference on Intelligent Systems for Molecular Biology & ECCB 2011 - 10th European Conference on Computational Biology**, Vienna, Austria, 2011.

[17] **Maracaja-Coutinho, V**; Setubal, JC; Verjovski-Almeida, S. *Genome-wide in-silico identification of human intronic non-coding RNAs and expression analysis of these transcripts in liver*. **RECOMB 2010 - 14th International Conference on Research in Computational Biology**, Lisboa, Portugal, 2010.

[16] Fachel, AA; Tahira, AC; **Maracaja-Coutinho, V**; Gimba, ERP; Campos, FS; Louro, R; Reis, EM; Verjovski-Almeida, S. *Long intronic noncoding RNA signatures of malignancy and survival outcome in clear cell renal cell carcinoma*. **System Biology: global regulation of gene expression**, Cold Spring Harbor Laboratory, NY, United States, 2010.

[15] Camargo, L; **Maracaja-Coutinho, V**; Verjovski-Almeida, S; Reis, EM. *Effect of DNA methylation on the transcription of intronic noncoding RNAs in cancer cell lines*. **System Biology: global regulation of gene expression**, Cold Spring Harbor Laboratory, NY, United States, 2010.

[14] Fachel, AA; Tahira, A; **Maracaja-Coutinho, V**; Gimba, ERP; Vignal, GM; Campos, FS; Louro, R; Reis, EM; Verjovski-almeida, S. *RNAs não-codificadores intrônicos longos correlacionados com carcinogênese e sobrevida em câncer de rim*. **56º Congresso Brasileiro de Genética**, Guarujá, SP, Brazil, 2010.

[13] **Maracaja-Coutinho, V**; Setubal, JC; Verjovski-almeida, S. *Genome-wide in-silico identification of human intronic non-coding RNAs and expression analysis of these transcripts in liver*. **X-Meeting 2009 - 5th International Conference of the AB³C**, Angra dos Reis, RJ, Brazil, 2009

[12] Grisi, TCSL; Padilha, IQM, Rangel, LTLD; **Maracaja-Coutinho, V**; Lima, LFA; Araujo, DAM. *Microbe diversity in Brazilian Cariri region (Paraíba) soil through 16S rDNA gene metagenomic library*. **55th Brazilian Genetics Meeting**, Águas de Lindóia, SP, Brazil, 2009

[11] **Maracaja-Coutinho, V**; Llamoca-Zarate, RM. *Degenerate primers construction for gene*

identification based on protein sequences. **I Simpósio Brasileiro de Genética Molecular de Plantas**, Natal, RN, Brazil, 2007.

[10] Padilha, IQM; Martins, AB; Durbano, JP; Almeida RS; Melo, LHM; **Maracaja-Coutinho, V**; Araujo, DAM. *Construção de recursos didáticos como instrumento de apoio para divulgação da Bioinformática*. **IX Encontro de Extensão e X Encontro de Iniciação à Docência da UFPB**, João Pessoa, PB, Brasil, 2007.

[9] Almeida, RS; **Maracaja-Coutinho, V**; Padilha, IQM; Araujo, DAM; Llamoca-Zarate, RM. *Patterns of similarity between Inulin synthesis key enzymes.* **XXI FeSBE Meeting**, Águas de Lindóia, SP, Brazil, 2006.

[8] *Silva, JC; ***Maracaja-Coutinho, V**; Nakaya, HI; Louro, R; Amaral, PP; Reis, RM; Verjovski-Almeida, S. *Molecular characterization of intronic transcripts regulated by androgen in prostate cancer cell line*. **XXXIV Reunião Anual da SBBq**, Águas de Lindóia, SP, Brazil, 2005.
* Both contributed equally to this work.

[7] **Maracaja-Coutinho, V**; Costa, DA; Ribeiro, ILAC; Amaral, CMM; Araujo, DAM; Llamoca-Zarate, RM. *Degenerate primer design targeting the NHX1 Antiporter gene from twenty plant species*. **X-Meeting 2005 - 1st International Conference of the AB³C**, Caxambu, MG, Brazil, 2005.

[6] Martins, AB; Faucher N; Padilha, IQM; **Maracaja-Coutinho, V**; Llamoca-Zarate, RM; Araujo, DAM. *Degenerate primer design targeting Glyoxylate pathway enzymes: Isocitrate lyase and Malate synthase in Leishmania chagasi*. **X-Meeting 2005 - 1st International Conference of the AB3C**, Caxambu, MG, Brazil, 2005.

[5] Costa, DA; Ribeiro, ILAC; Amaral CMM. Llamoca-Zarate, RM; **Maracaja-Coutinho, V**. *In vitro regeneration of shoots through the shoot apex of sorghum (Sorghum bicolor (L.) Moench)*. **XII Latin-American Plant Physiology Meeting & X Brazilian Plant Physiology Meeting**, Recife, PE, Brazil, 2005.

[4] **Maracaja-Coutinho, V**; Araujo, DAM; Llamoca-Zarate, RM. *Salt resistance: primers for NHX1 Antiporter gene in Sorghum sudanense*. **VI Encontro Unificado de Ensino, Pesquisa e Extensão da UFPB**, João Pessoa, PB, Brazil, 2005.

[3] Almeida, RS; **Maracaja-Coutinho, V**; Padilha, IQM; Araujo, DAM; Llamoca-Zarate, RM. *In silico analysis of Inulin synthesis key enzymes*. **VI Encontro Unificado de Ensino, Pesquisa e Extensão da UFPB**, João Pessoa, PB, Brazil, 2005.

[2] Durbano, JP; Padilha, IQM; **Maracaja-Coutinho, V**; Almeida, RS; Martins, AB; Van Der Linden, MG; Rego, TG; Araujo, DAM. *Bioinformatics in Academic Extension*. **VI Encontro Unificado de Ensino, Pesquisa e Extensão da UFPB**, João Pessoa, PB, Brazil, 2005.

[1] **Maracaja-Coutinho, V**; *Llamoca-Zarate, RM. Patterns of similarity of putative NHX1 Antiporter proteins from twenty plant species*. **VII Reunião Regional da SBBq & 2nd International Symposium in Biochemistry of Macromolecules**, Recife, PE, Brazil, 2004.

Other Production

- Design and implementation of the IntromeDB: a database for the expression of long intronic ncRNAs (lncRNAs) in Eukaryotes (www.intromedb.org/)
- Design and implementation of the NRDR: Non-coding RNA Databases Resource (www.ncrnadatabases.org/)
- Design of a 244k customized human intronic-exonic-intergenic Agilent oligoarray platform (AMADID: 05248).
- Design of a 244k customized human intronic-exonic-intergenic Agilent oligoarray platform (AMADID: 028681).

July, 2013