IDENTIFICAÇÃO EM LARGA ESCALA DE TRANSCRITOS ANTISENSO E GENES COM EXPRESSÃO ALÉLICA DIFERENCIAL UTILIZANDO BANCOS DE DADOS DE *SAGE* E *MPSS*

DANIEL ONOFRE VIDAL

Tese de doutorado apresentada a Fundação Antônio Prudente para obtenção do Título de Doutor em Ciências

Área de Concentração: Oncologia

Orientadora: Dra. Anamaria Aranha Camargo
Co-orientador: Dr. Sandro José de Souza

São Paulo
2009

FICHA CATALOGRÁFICA

Preparada pela Biblioteca da Fundação Antônio Prudente

Vidal, Daniel Onofre
Identificação em larga escala de transcritos antisenso e genes com expressão alélica diferencial utilizando bancos de dados de SAGE E MPSS / Daniel Onofre Vidal – São Paulo, 2009.
107p.
Tese (doutorado) Fundação Antônio Prudente.
Curso de Pós-Graduação em Ciências-Área de concentração: Oncologia.
Orientadora: Anamaria Aranha Camargo

Descritores: 1. EXPRESSÃO GÊNICA. 2. GENES. 3. TRANSCRITOS ANTISENSO. 4. EXPRESSÃO ALÉLICA DIFERENCIAL. 5. SAGE. 6. MPSS.

DEDICATÓRIA

Dedico esta tese aos meus dois grandes e eternos amores: minha esposa Ana Paula e a minha filha Maria Clara. Hoje tudo o que faço é por vocês e todo o meu amor é dedicado a vocês!!!

Aos meus exemplos de vida, meus pais Dagoberto e Angela e ao meu eterno companheiro, meu irmão Diego e a sua esposa Natália. Obrigado por terem construído esse pilar sólido que é nossa família.

Aos meus sogros, Valentim e Lourdes, ao meu cunhado Valmor e aos meus sobrinhos Enzo e Lucas. Sou muito grato por me acolherem e também por fazer parte dessa família.

Obrigado, vale viver cada dia por vocês!!!

AGRADECIMENTOS

Em primeiro lugar, agradeço a minha orientadora Dra. Anamaria Aranha Camargo por ter me acolhido em um momento difícil e por ter me dado a oportunidade de trabalhar com uma das pessoas mais extraordinárias que provavelmente irão cruzar a minha vida. Muito obrigado, nunca irei me esquecer de tudo que aprendi e vivi nesses anos...

Ao meu co-orientador Dr. Sandro José de Souza, pela imensa contribuição em minha formação e também por todas as discussões que contribuíram para o desenvolvimento desse trabalho.

À Valéria Paixão, uma grande amiga que me acompanha há seis anos e que sempre me ajudou com seus conselhos e também broncas. Serei eternamente grato, pois todo esse tempo você foi uma "mãezona" para mim.

À Lílian Pires, que além de ser uma grande amiga de longa data, foi por um bom tempo parceira neste projeto. Parabéns, sem a sua ajuda, nada disso teria acontecido, essa conquista também é sua.

À Anna Christina M. Salim, que dividiu comigo todos esses anos de bancada, tornando-se uma grande amiga e que foi muito importante em um dos momentos mais difíceis em todo esse tempo. Serei eternamente grato a você.

Ao Pedro Galante e Jorge Estefano por terem desenvolvido toda a análise computacional deste trabalho. Parabéns, o trabalho de vocês é admirável.

À Cibele Masotti, pela enorme contribuição neste trabalho, pelas conversas, discussões científicas e principalmente pela amizade todo este tempo.

À Maria Cristina Ferreira, que foi muito importante para o desenvolvimento deste trabalho.

À Ana Paula Silva, que também teve uma contribuição enorme neste trabalho, principalmente em seu início.

À Dra. Mônica Poli e Dr. Rafael Colella, por possibilitarem a coleta das amostras de sangue dos doadores do departamento de Hemoterapia do Hospital A.C. Camargo.

À Dra. Helena Brentani e Dra. Ana Tereza Vasconcelos, pelas sugestões pertinentes que contribuíram para o desenvolvimento deste projeto.

Aos eternos amigos Raphael e Fabiana, que foram muito importantes na minha chegada ao grupo e também que compartilharam comigo muitas alegrias e também tristeza nesse período e desde a época da graduação.

Ao Érico Costa, que além de ser um grande amigo e pesquisador, também se revelou um exímio pescador. Agradeço por me ensinar a gostar de aquários, uma terapia em muitos momentos.

À Daniela Mattos, pela grande amizade, pelas conversas, pelos conselhos, pelas caronas e pelos agradáveis momentos que vivemos no trabalho.

À Renata Capelucci, pela disposição e preciosa ajuda e pelo seu competente trabalho durante esses anos.

A todo o LMBG, Anna Chris, Ana Paula Silva, Ana Paula Pardo, Alex, Bruna, Cibele, Daniela, Débora, Érico, Fabiana, Felícia, Fabrício, Lílian Pires, Lílian Inoue, Murilo, Natália, Paula, Raphael, Ricardo, Tâmara e Valéria, pela adorável convivência. Tenham certeza que sempre farão parte da minha vida!!!

Aos diretores da pós-graduação da FAP, Dr. Luiz Fernando Lima Reis e Dr. Fernando Augusto Soares. Para mim é um orgulho fazer parte desta instituição e por todo o apoio durante esses anos e por não medirem esforços para alcançar tamanha excelência.

À Ana Maria Kuninari, coordenadora da pós-graduação, uma das pessoas com a qual tive o primeiro contato quando aqui cheguei há seis anos. Muito obrigado por sempre estar disposta a ajudar, pela paciência e por tudo o que fez por mim todos esses anos.

A toda equipe da biblioteca, que aqui represento pela Suely, Francyne e Rosi, por sempre estarem dispostas a ajudar, pela alegria, mas principalmente pela competência com a qual administram a nossa biblioteca. Ainda me lembro das aulas de ginástica na biblioteca.

RESUMO

Vidal DO. **Identificação em larga escala de transcritos antisenso e genes com expressão alélica diferencial utilizando bancos de dados de SAGE e MPSS**. São Paulo; 2009. [Tese de Doutorado-Fundação Antônio Prudente]

Recentes relatos vêm demonstrando a ocorrência de um número crescente de transcritos antisenso naturais (NATs) no genoma humano e a ocorrência de expressão alélica diferencial (ADE) em genes autossômicos não submetidos a *imprinting*. Devido aos diversos mecanismos pelos quais podem afetar a expressão gênica, alterações na transcrição dos NATs podem estar envolvidas no desenvolvimento de patologias, como o câncer. Da mesma forma, genes que apresentam expressão alélica diferencial têm sido associados à variabilidade fenotípica e podem também contribuir para o desenvolvimento de doenças complexas em humanos. Neste trabalho apresentamos duas estratégias inéditas para identificar em larga escala novos transcritos antisenso e genes que apresentam expresssão alélica diferencial. A primeira estratégia baseou-se na utilização de ferramentas computacionais para a identificação de *tags* de MPSS que mapeavam na fita oposta de genes conhecidos representados por uma sequência de mRNA completa, sendo a *tag* de MPSS a única evidência da existência do transcrito antisenso. Assim, de um total de 340.829 *tags* únicas e distintas presentes em 41 bibliotecas de MPSS, 4.308 *tags* indicaram a existência de um NAT. A metodologia de GLGI-MPSS foi aplicada em 96 dessas 4.308 *tags*, permitindo a sua extensão em um fragmento maior de cDNA correspondente a extremidade 3' do transcrito. O alinhamento desses fragmentos contra a sequência do genoma humano utilizando BLAT, confirmou que 46/96 fragmentos de GLGI-MPSS correspondiam a extensões 3' específicas e com orientação antisenso. Interessantemente, observamos que 41,3% (19/46) desses fragmentos de GLGI-MPSS apresentaram em sua extremidade 3' uma cauda poli(A) que alinhava à sequência do genoma humano. Demonstramos que uma fração desses transcritos são artefatos gerados por eventos de pareamento interno em DNA contaminante e que outra fração desses transcritos é real e pode ser atribuída a

eventos de retroposição no genoma humano. A expressão de 25/27 fragmentos de GLGI-MPSS restantes foram avaliados por RT-PCR fita específica, e a existência de 17/25 (68%) foi confirmada através dessa metodologia. A segunda estratégia baseou-se na utilização de ferramentas computacionais que permitiram a integração de dados de sequências expressas (mRNA e SAGE) e polimorfismos (SNPs) à sequência do genoma humano possibilitando a criação de um banco de dados contendo *tags* alelo específicas de SAGE. Dessa maneira, foi possível inferir a expressão de cada um dos alelos de um gene a partir da freqüência das *tags* alelo específicas representadas nas diferentes bibliotecas de SAGE. Assim, de um total de 20.034 genes, 1.295 (6,46%) apresentaram *tags* alelo específicas e, de acordo com o padrão de expressão dessas *tags*, esses genes foram classificados em 3 categorias principais: a) 481 genes (37,2%) foram classificados com expressão alélica diferencial; b) 442 genes (34,1%) foram classificados com expressão monoalélica, dos quais 242 estavam representados em mais de 10 bibliotecas de SAGE; e por fim, c) 372 genes (28,7%) foram classificados com expressão bialélica. Vinte genes foram selecionados para validação experimental por meio do sequenciamento direto do gDNA e cDNA, dos quais 13 apresentaram mais de 5 indivíduos heterozigotos. Desses 13 genes, 10 (77%) apresentaram ADE em pelo menos 20% dos indivíduos heterozigotos analisados. Interessantemente, para o gene *PHC1* observamos a expressão monoalélica em todos os indivíduos heterozigotos. Levando em conta a eficiência de nossa estratégia experimental (77%), nossos dados sugerem que pelo menos 43,0% (481+242 x 0,77/1.295) dos genes humanos apresentam expressão alélica diferencial. Em conjunto, nossos resultados demonstraram que a estratégia computacional utilizando os dados de MPSS foi eficaz para a identificação de novos transcritos antisenso no genoma humano e, ainda, que *tags* alelo específicas de SAGE podem ser eficientemente utilizadas na identificação de genes humanos que apresentam expressão alélica diferencial.

SUMMARY

Vidal DO. [**High-throughput identification of natural antisense transcripts and genes displaying allelic differential expression using SAGE and MPSS databases**]. São Paulo; 2009. [Tese de Doutorado-Fundação Antônio Prudente]

Recent reports have demonstrated the occurrence of an increasing number of natural antisense transcripts (NATs) in the human genome and the occurrence of allelic differential expression (ADE) in non-imprinted autosomal genes. Due to the diverse mechanisms by which NATs can affect gene expression, their abnormal expression may be involved in the development of pathological states, such as cancer. Similarly, genes displaying ADE have been associated with phenotypic variability and may also contribute to the development of complex genetic diseases. In this work, we present two unpublished strategies to high-throughput identification of new NATs and genes displaying ADE. The first strategy was based on the use of computational tools for the identification of MPSS tags that mapped on the opposite strand of known human genes represented by mRNA sequences, and for which the MPSS tag represents the only evidence of the existence of the NAT. Thus, from a total of 340,829 unique and distinct MPSS tags present in 41 MPSS libraries, 4,308 tags indicated the existence of a new NAT. The GLGI-MPSS methodology was applied for 96 out of these 4,308 tags, allowing their extension into a longer cDNA fragment corresponding to the 3' end of the transcript. The alignment of these fragments against the human genome sequence using BLAT, confirmed that 46/96 GLGI-MPSS fragments corresponded to 3' especific extensions with antisense orientation. Interestingly, we observed that 41.3% (19/46) of these GLGI-MPSS presented at their 3' end a poly(A) tail aligned to the human genome sequence. We demonstrated that a fraction of these transcripts are artifacts generated by internal priming in contaminating DNA and that another fraction of these transcripts are real and could be attributed to retroposition events in the human genome. The expression of the remaining 25/27 GLGI-MPSS fragments was evaluated by strand-specific RT-PCR, and the existence of 17/25 was confirmed by this methodology. The second strategy was based on the use of computational

tools that allowed the integration of data from expressed sequences (mRNA and SAGE) and polymorphisms (SNPs) to the human genome sequence for the creation of a database containing allele-specific SAGE tags. In this way, it was possible to infer the expression of each allele of a gene from the frequency of each allele-specific tag represented in different SAGE libraries. So, from a total of 20,034 genes, 1,295 (6.46%) genes presented allele-specific SAGE tags and according to their expression pattern genes were classified into 3 major categories: a) 481 (37.2%) genes were classified with allelic differential expression; b) 442 (34.1%) genes were classified with monoallelic expression, of which 242 were represented in more than 10 SAGE libraries; and c) 372 (28.7%) genes were classified with bialleic expression. Twenty genes were chosen for experimental validation by gDNA and cDNA direct sequencing, of which 13 presented more than 5 individual heterozygotes. From these 13 genes, 10 (77%) demonstrated ADE in at least 20% of the heterozygotes evaluated. Interestingly, for *PHC1* we observed monoallelic expression in all heterozygotes. Taking into account our experimental validation efficiency (77%), our analysis suggests that at least 43% of all human genes (481+242 x 0.77/1,295) display ADE. Taken together, our results demonstrated that the computational strategy using MPSS data was effective in the identification of new NATs in the human genome and that allele-specific SAGE tags can be efficiently used to expedite the identification of human genes displaying ADE.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

µg	micrograma
µl	microlitro
µM	micromolar
ADAR	*adenosine deaminase acting on RNA*
ADE	expressão alélica diferencial
antiCODE	*natural sense-antisense transcripts database*
ASSAGE	*Asymetric Strand-Specific Analysis of Gene Expression*
BLAT	*BLAST-Like Alignment Tool*
cDNA	DNA complementar
CEPH	*Centre d'Etude du Polymorphisme Humain*
dbSNP	*SNP database*
DNA	ácido desoxirribonucléico
dsRNA	dupla fita de RNA
ESTs	etiqueta de sequências expressas
gDNA	DNA genômico
GLGI-MPSS	*Generation of Longer cDNA fragments from MPSS tags for Gene Identification*
GO	*Gene Ontolotogy*
LCL	linhagem celular linfoblastóide
LLA	leucemia linfoblástica aguda
M	molar
miRNAs	*microRNAS*
mM	milimolar
MPSS	*Massively Parallel Signature Sequencing*
mRNA	RNA mensageiro
NATs	transcritos antisenso naturais
NATsDB	*Natural Antisense Transcripts DataBase*
NCBI	*National Center for Biotechnology Information*
ncRNAs	*non coding RNA*

ng	nanogramas
nt	nucleotídeos
ORFs	*open reading frames*
pb	pares de base
PCR	*Polymerase Chain Reaction*
pH	potencial hidrogeniônico
piRNA	*Piwi-interacting RNA*
RNA	ácido ribonucléico
RNAi	*RNA interference*
rpm	rotações por minuto
RT-PCR	*Reverse Transcriptase – Polymerase Chain Reaction*
SAGE	*Serial Analysis of Gene Expression*
siRNAs	*short interfering RNAs*
snoRNAs	*small nucleolar RNAs*
SNP	*Single Nucleotide Polymorphism*
snRNAs	*small nuclear RNAs*
U	unidade

ÍNDICE

CAPÍTULO I

Identificação em larga escala de transcritos antisenso do genoma humano utilizando banco de dados de MPSS

1 CAPÍTULO I

1.1 INTRODUÇÃO

1.1.1 Transcritos Antisenso Naturais (*Natural Antisense Transcripts*) e a complexidade do genoma humano

O número de genes humanos que codificam proteínas tem sido constantemente revisado e parece girar em torno de 20.000 a 30.000 genes, um número muito abaixo do inicialmente estimado (LANDER et al. 2001; VENTER et al. 2001; CLAMP et al. 2007). Estudos comparativos de genomas eucariotos sugerem que o número de genes que codificam proteínas, sozinho, não é suficiente para conferir as diferenças moleculares e celulares encontradas entre os organismos (BEITER et al. 2009). Parece não haver uma correlação direta entre o número de genes codificadores e a complexidade dos organismos, o que fica claro quando observamos exemplos como o camundongo, que apresenta em torno de 30.000 genes codificadores (WATERSTON et al. 2002), a *Drosophila melanogaster* em torno de 15.000 (ADAMS et al. 2000), o *Caenorhabditis elegans* em torno de 19.000 (WATERSTON 1998) e a *Arabidopsis thaliana* em torno de 27.000 (Arabidopsis Genome Initiative 2000).

A ocorrência de eventos pós transcricionais, como poliadenilação alternativa e *splicing* alternativo, pode em parte explicar essa falta de correlação (PENNISI 2005). No entanto, uma forte evidência de que o genoma humano pode ser extensivamente transcrito e que, o número de genes não codificadores (ncRNAs) é

muito maior do que o predito foi apresentada recentemente em trabalhos que utilizaram *tiling arrays* para avaliar a transcrição em cromossomos individuais (KAPRANOV et al. 2002; KAMPA et al. 2004; CHENG et al. 2005). Dessa forma, a explicação para a complexidade dos organismos não tem sido exclusivamente atribuída aos genes codificadores e a seus transcritos, mas também a esse reservatório abundante de ncRNAs, que em estimativas mais recentes parece representar ~90% da capacidade transcricional do genoma humano (GINGERAS 2007).

Em relação a esses ncRNAs, são conhecidos os *microRNAs* (miRNAs) (PASQUINELLI et al. 2005), os *short interfering RNAs* (siRNAs) (KAWASAKI et al. 2005), os *small nuclear RNAs* (snRNAs) (WILL e LURMANN 2001), os *small nucleolar RNAs* (snoRNAs) (BACHELLERIE et al. 2002), os *Piwi-interacting RNA* (piRNA) (LAU et al. 2006) e, finalmente os transcritos antisenso naturais (NATs) (KUMAR e CARMICHAEL 1998).

Os NATs são moléculas de RNAs endógenas que possuem sequências complementares a outros RNAs endógenos. Apesar de estarem inseridos nesta categoria de ncRNAs, os NATs podem apresentar ORFs (*open reading frames*) potenciais e estima-se que 20% sejam codificadores (KIYOSAWA et al. 2003; KATAYAMA et al. 2005). Os NATs podem ser transcritos em *cis*, a partir da fita oposta de DNA de um mesmo locus genômico, ou em *trans*, a partir de diferentes loci genômicos (Figura 1). Em geral, os transcritos em *cis* apresentam uma longa e perfeita região de sobreposição e estão relacionados a um ou poucos transcritos senso (Figura 1A). Por outro lado, os transcritos em *trans* apresentam uma região de sobreposição curta e imperfeita e podem estar associados a vários transcritos senso

(Figura 1B) (KUMAR e CARMICHAEL 1998; VANHÉE-BROSSOLLET e VAQUERO 1998; LAVORGNA et al. 2004; LAPIDOT e PILPEL 2006).

Os NATs também podem ser classificados em diferentes categorias baseado no padrão de sobreposição dos transcritos senso e antisenso. Basicamente, são divididos em três categorias distintas: *head-to-head* ou divergentes (sobreposição das extremidades 5' dos transcritos); *tail-to-tail* ou convergentes (sobreposição das extremidades 3' dos transcritos); e, *embedded* ou contidos (neste caso, um transcrito apresenta sobreposição total dentro do outro transcrito) (Figura 2) (LAPIDOT e PILPEL 2006). Na literatura não há um consenso sobre qual destes tipos ocorre com maior freqüência no genoma de mamíferos. Embora trabalhos anteriores reportem o tipo convergente como o mais frequente (LEHNER et at. 2002; SHENDURE e CHURCH 2002; YELIN et al. 2003; CHEN et al. 2004), trabalhos mais recentes demonstraram que o tipo divergente parece ser o mais comum (KATAYAMA et al. 2005; ZHANG et al. 2006).

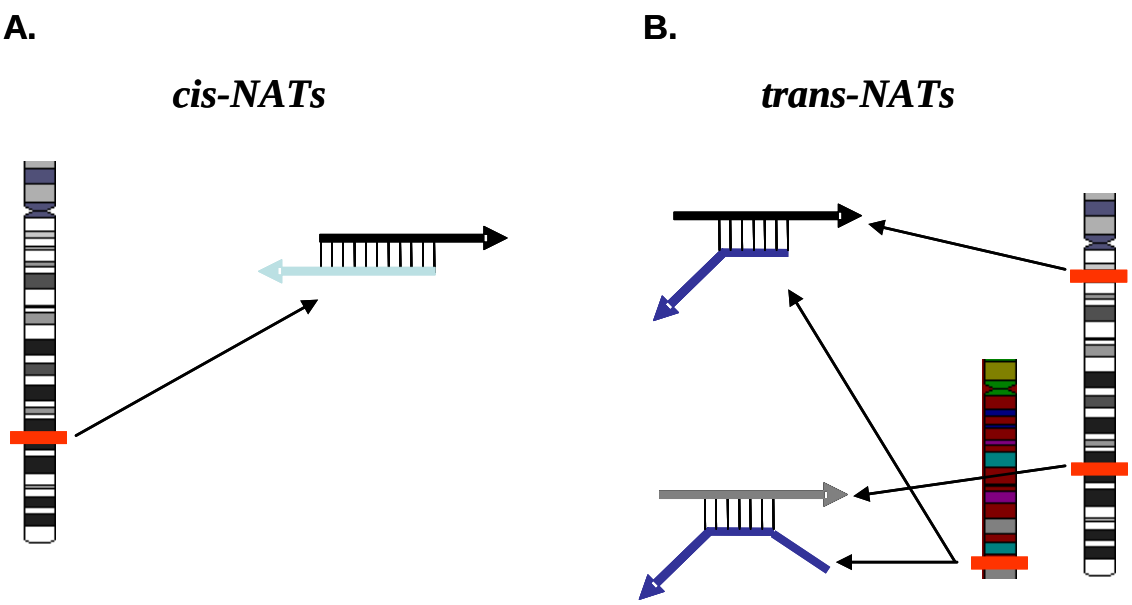


Figura 1 - Tipos de NATs de acordo com sua origem de transcrição. A figura representa a classificação dos NATs de acordo com sua origem de transcrição. Em A, temos um *cis*-NAT, transcrito a partir do mesmo lócus genômico e apresentando sobreposição mais longa e perfeita com apenas um transcrito senso; em B, temos um *trans*-NAT, transcrito a partir de loci genômicos diferentes e apresentando sobreposição curta e imperfeita com mais de um transcrito senso. Setas em preto e cinza: transcrito senso; Setas em azul (claro e escuro): transcrito antisenso.

A análise computacional de dados gerados por projetos de sequenciamento em larga escala tem permitido a identificação de um número surpreendente de NATs no genoma de diferentes organismos (SHENDURE e CHURCH 2002; KIYOSAWA et al. 2003; YELIN et al. 2003; KATAYAMA et al. 2005; WANG et al. 2005). Em humanos, por exemplo, os primeiros trabalhos reportaram a existência de 144 NATs (SHENDURE e CHURCH 2002), seguidos de trabalhos que identificaram 2.667 e 5.880 NATs (YELIN et al. 2003; CHEN et al. 2004) e, mais recentemente 12.320 NATs (ENGSTRÖM et al. 2006).

Esse número abundante de NATs, somado ao fato da maioria ser não codificante, sugere que esses transcritos tenham um papel importante na regulação da expressão gênica, principalmente em células eucariotas (LAPIDOT e PILPEL 2006;

WERNER e SAYER 2009). Essa idéia é reforçada uma vez que esses NATs também apresentam conservação evolutiva entre os reinos maior do que o esperado ao acaso (CHEN et al. 2005; ZHANG et al. 2006).

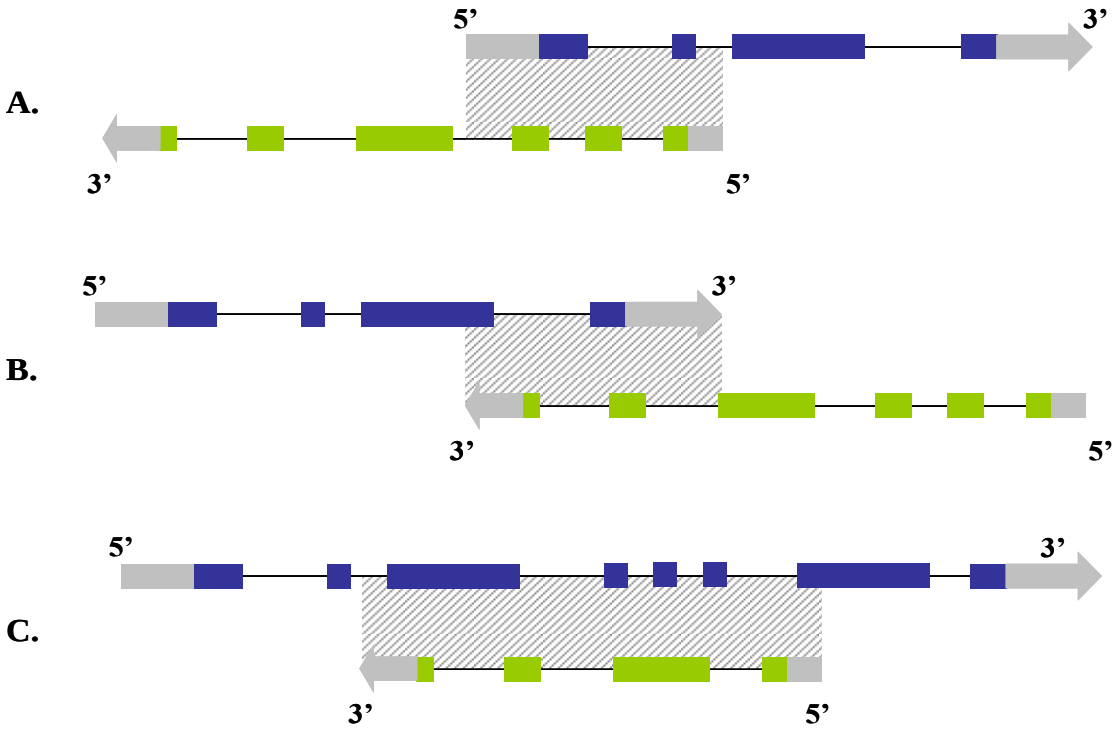


Figura 2 - Classificação dos NATs de acordo com o padrão de sobreposição dos transcritos senso/antisenso. Em A, temos a representação de um caso divergente (*head-to-head*), no qual os transcritos apresentam sobreposição em suas extremidades 5'; em B, temos a representação de um caso convergente (*tail-to-tail*), no qual os transcritos apresentam sobreposição em suas extremidades 3'; em C, temos a representação de um caso contido (*embedded*), no qual um dos transcritos apresenta sobreposição total a região do outro transcrito. As caixas coloridas (azul e verde) representam os exons, as caixas em cinza representam as extremidades 5' e 3' não traduzidas (UTRs) e o retângulo hachurado a região de sobreposição entre os transcritos.

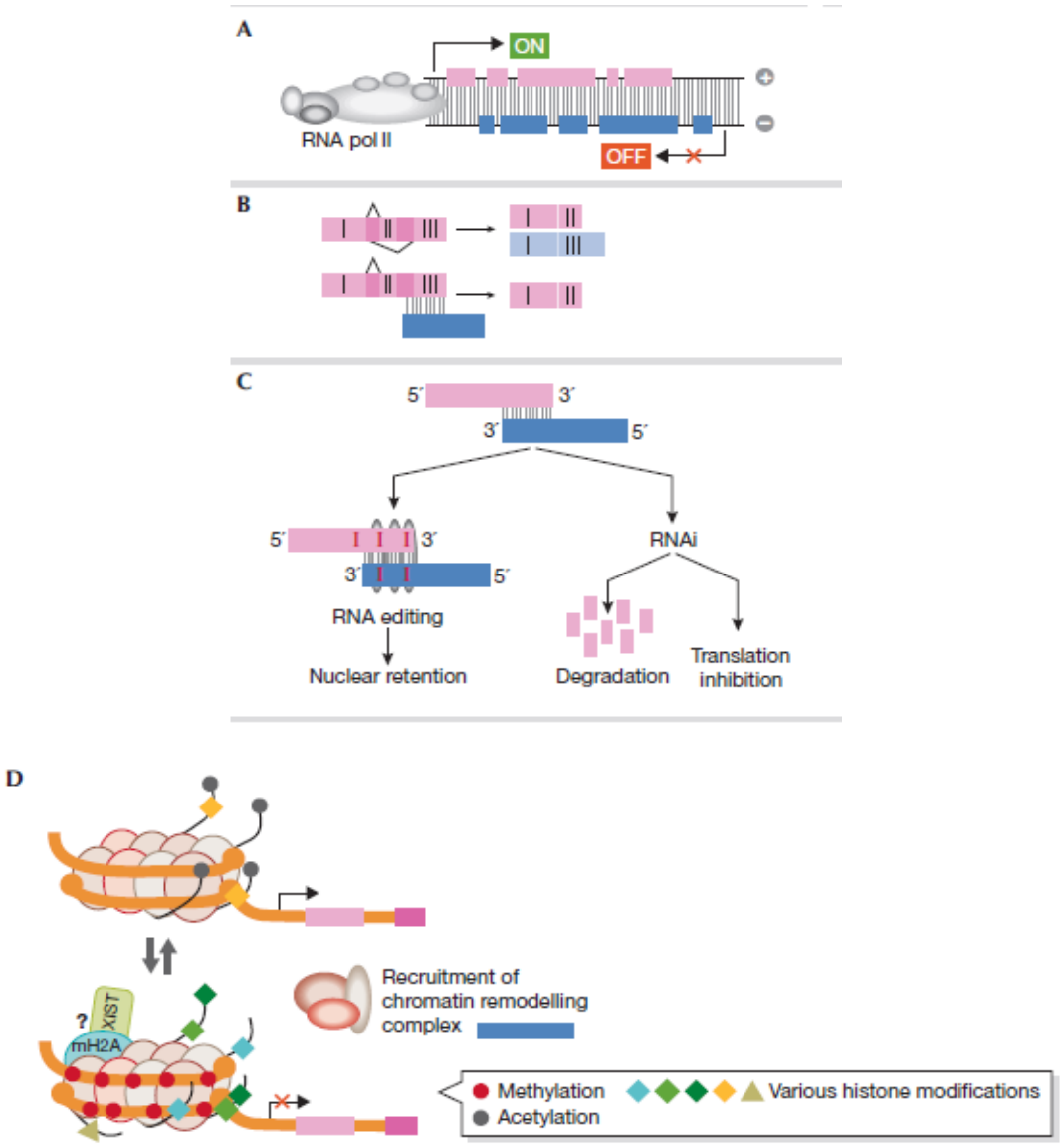
1.1.2 NATS e os mecanismos de regulação da expressão gênica

A regulação da expressão gênica por NATs tem sido associada a quatro mecanismos principais: I) a interferência transcricional, II) o RNA *masking* , III) a mecanismos dependentes de formação de dupla fita de RNA (RNA *editing* ou RNA

interference (RNAi)), e IV) ao remodelamento de cromatina/metilação do DNA (Figura 3) (LAPIDOT e PIPEL 2006).

Na interferência transcricional a regulação ocorre por meio de competição, devido a estrutura tridimensional da maquinaria de transcrição e a necessidade de abertura das fitas de DNA. A transcrição do DNA envolve grandes complexos protéicos (RNA polimerase II e proteínas de enovelamento do DNA) de tal forma que unidades transcricionais que se sobrepõe não possam ser transcritas concomitantemente (Figura 3A). Isso sugere que transcritos envolvidos na formação de NATs, exclusivamente os *cis*, podem sofrer regulação por competição e, portanto, apresentar uma expressão inversamente proporcional, ou mesmo os dois serem silenciados. De fato, isto ocorre e já foi demonstrado para alguns NATs envolvendo, por exemplo, *eIF2α* (SILVERMAN et al. 1992).

RNA *masking* consiste na formação de dupla fita de RNA (senso/antisenso) que pode afetar qualquer passo da expressão gênica envolvendo a interação entre proteína e RNA, incluindo *splicing*, transporte, poliadenilação, tradução ou degradação (Figura 3B). Um exemplo é a regulação exercida pelo antisenso *Rev-ErbAα* na expressão de formas alternativas de *splicing* (*ErbAα1* ou *ErbAα2*) do receptor de hormônio tireóide ErbAα. (HASTINGS et al. 1997).



Fonte: LAPIDOT e PIPEL (2006).

Figura 3 - Mecanismos de regulação da expressão gênica associados a NATs. Em A, temos representado o mecanismo de interferência transcricional; em B, o mecanismo de RNA *masking*; em C, o mecanismo de RNA *editing* e RNA *interference*; em D, o mecanismo de remodelamento da cromatina e metilação do DNA.

RNA *editing* é um mecanismo que faz parte da estratégia de defesa contra a formação de dupla fita de RNA (dsRNA) (Figura 3C). Esse mecanismo consiste na conversão de adenosina para inosina na dsRNA, o que é feito por uma enzima conhecida como ADAR (*adenosine deaminase acting on RNA*) (TONKIN et al.

2002). Esse mecanismo pode atuar antes ou durante o processo de *splicing* e quando ocorre em sequências longas e perfeitas de dsRNA pode levar a retenção do RNA no núcleo ou mesmo a degradação do RNA no citoplasma das células (LEVANON et al. 2004). Entretanto, RNA *editing* não parece ser um evento freqüente que ocorre nas regiões de sobreposição de pares de NATs identificados em humanos e camundongos (NEEMAN et al. 2005).

RNA *interference* (RNAi) é outro mecanismo que faz parte da estratégia de defesa contra a formação de dsRNA. Esse mecanismo consiste na formação de fragmentos de dsRNA que são clivados por um complexo de proteínas, levando a formação de fragmentos menores de dsRNA (21 a 23 nucleotídeos) que servirão de substrato para degradação específica do mRNA complementar (Figura 3C) (LAVORGNA et al. 2004). WATANABE et al. (2008) por meio do sequenciamento de 100.000 RNAs não codificantes pequenos (*small RNA*) de oócitos de camundongo, identificaram 17 *clusters* de siRNAs (*small interfering RNA*) associados a regiões de transcrição de *cis*-NATs. Para o lócus *Pdzd11/Kif4*, os autores demonstraram que em oócitos mutados para Dicer (proteína que processa a dsRNA em siRNA) o nível de siRNAs originados neste lócus decresceu 7 vezes e, ainda, a expressão de ambos os genes aumentou 1,5 vezes. Esses foram os primeiros relatos de que a formação de pares de NATs pode regular a expressão gênica por meio de RNAi em mamíferos.

Por fim, somado aos exemplos anteriores, também tem sido proposto que a transcrição antisenso pode estar envolvida em outros processos que regulam a expressão gênica, como no caso dos genes que apresentam expressão monoalélica. A expressão monoalélica inclui a inativação do cromossomo X, o *imprinting* genômico

e a exclusão alélica que ocorre nos linfócitos B e T. Estudos recentes demonstram que, no caso de expressão monoalélica, a regulação não parece ocorrer devido a formação de dsRNA (senso/antisenso), mas sim pelo fato dos transcritos antisenso promoverem modificações na estrutura da cromatina e no padrão de metilação dos alelos desses genes (Figura 3D) (SLEUTELS et al. 2002; TUFARELLI et al. 2003).

1.1.3 Identificação de NATs a partir de análises computacionais em larga escala

Os NATs foram inicialmente identificados em estudos que avaliaram genes individuais. Contudo, com a disponibilidade da sequência completa do genoma humano, o acúmulo de milhões de sequências expressas em bancos de dados (mRNAs e ESTs) e o desenvolvimento de ferramentas computacionais tornaram possível a predição em larga escala dos NATs. Esses trabalhos têm descrito a ocorrência abundante deste tipo de transcrito no genoma de mamíferos. Os primeiros relatos sugeriram que 20% dos genes humanos apresentam NATs (CHEN et al. 2004), entretanto estudos mais recentes mostraram que essa proporção pode chegar a 50% (ENGSTRÖM et al. 2006).

As primeiras evidências de que a ocorrência de NATs é um fenômeno comum no genoma humano foram apresentadas por LEHNER et al. (2002). Os autores procuraram por complementaridade entre transcritos humanos por meio de comparação direta entre as sequências de mRNA (12.000 sequências) disponíveis em bancos de dados públicos. Essa abordagem permitiu a identificação dos dois tipos (*cis* e *trans*) de NATs e os autores identificaram um total de 167 pares de NATs.

A maioria dos trabalhos subsequentes que visaram a identificação de NATs agregaram os dados de ESTs e a sequência do genoma humano as suas análises (SHENDURE e CHURCH 2002; YELIN et al. 2003; KIYOSAWA et al. 2003; CHEN et al. 2004; KATAYAMA et al. 2005; ZHANG et al. 2006; ENGSTRÖM et al. 2006). O principal desafio na identificação de NATs é definir a correta orientação das sequências expressas disponíveis em bancos públicos, principalmente das ESTs. Para isso, critérios estringentes que levam em consideração algumas características intrínsecas desses transcritos são utilizados nesses trabalhos a fim de garantir a orientação correta das ESTs. Entre esses critérios estão a presença do sítio canônico de *splicing* (GT-AG), a presença de cauda poli(A) e também a presença de sinal de poliadenilação nestas sequências. Utilizando-se esse conjunto de sinais é possível estabelecer a orientação correta para uma grande parte dessas sequências (SHENDURE e CHURCH 2002; CHEN et al. 2004; ENGSTROM et al. 2006). Os resultados desses trabalhos estão sumarizados na Tabela 1.

Tabela 1 - Trabalhos que visaram a identificação de NATs ao longo do tempo.

Referência	Número e Tipo de sequências	Agrupamento (*clusters*)	Pares de NATs	Organismo
Shendure e Church (2002)	1.150.000 (mRNA e ESTs)	-	144	humano
	550.000 (mRNA e ESTs)	-	73	camundongo
Yelin et al. (2003)	82.289 (mRNA) e 2.750.000 (ESTs)	61.048	2.667	humano
Kiyosawa et al. (2003)	60.770 (mRNA)	16.550	2.481	camundongo
Chen et al. (2004)	350.000 (mRNA e ESTs)	22.340	5.880	humano
Katayama et al. (2005)	158.807 (mRNA)	50.110	8.650	camundongo
Zhang et al. (2006)	2.262.000 (mRNA e ESTs)	15.000	3.915	humano
	1.549.000 (mRNA e ESTs)	13.800	3.040	camundongo
Engström et al. (2006)	138.350 (mRNA) e 3.300.000 (ESTs)	42.890	12.320	humano
	107.740 (mRNA) e 2.400.000 (ESTs)	36.600	8.960	camundongo

A tabela descreve os resultados obtidos por trabalhos que tiveram como objetivo a identificação em larga escala de NATs. Em todos esses trabalhos foram utilizadas sequências de mRNA e/ou ESTs e a sequência do genoma, bem como a utilização de ferramentas computacionais para a identificação de NATs.

Podemos observar na Tabela 1, que com o passar dos anos o número de NATs identificados vêm crescendo consideravelmente, fato esse que se deve ao aumento do número de sequências expressas (mRNA ou ESTs) depositadas em bancos de dados públicos. Dessa forma, tornou-se bem consolidada a idéia de que os NATs são um evento freqüente no genoma de mamíferos. É importante ressaltar que a utilização da sequência do genoma nessas análises possibilita apenas a identificação de *cis*-NATs, uma vez que um único (melhor) alinhamento das sequências é considerado nas análises. Em contrapartida, neste caso é possível identificar transcritos antisenso mapeados na região intrônica de transcritos senso. De qualquer forma, o número de NATs identificados até o momento pode estar subestimado e deve ser ainda maior, uma vez que nos trabalhos acima citados não são identificados e considerados a ocorrência dos *trans*-NATs e também que uma porção das sequências expressas (sem poliadenilação e sem sítios de *splicing*) não são consideradas para garantir a estringência das análises (BEITER et al. 2009).

Em alguns dos trabalhos apresentados acima (Tabela 1) foram descritos esforços experimentais no sentido de validar os achados computacionais observados. YELIN et al. (2003) demonstraram, por *microarray*, que 60% (154/264) dos NATs avaliados estavam expressos em um mistura de RNA de diferentes linhagens celulares representando diversos tecidos humanos. SHENDURE e CHURCH (2002) e CHEN et al. (2004), utilizando RT-PCR fita específica, validaram 80% e 96% (33/39 e 24/25, respectivamente) dos NATs avaliados e identificados em suas estratégias computacionais. Para os NATs identificados *in silico*, também por RT-PCR fita específca, ENGSTRÖM et al. (2006) validaram a expressão de 80% (16/20) no RNA de amostras de cérebro de camundongo.

Além das sequências de mRNA e ESTs outras fontes de sequências expressas, como SAGE (*Serial Analysis of Gene Expression*), também têm sido empregadas na identificação de NATs (QUÉRÉ et al. 2004; GE et al. 2006). QUÉRÉ et al. (2006), utilizando dados públicos de SAGE, identificaram as *tags* que apresentaram orientação antisenso a transcritos conhecidos (48.220 UniGene *clusters*) e dessa forma, os autores observaram a existência de 417 pares de NATs, dos quais 198 não haviam sido descritos até aquele momento por outros trabalhos. GE et al. (2006) utilizaram o genoma humano e observaram a ocorrência de 45.321 *tags* de SAGE apresentando orientação antisenso a 9108 sequências referência de transcritos humanos (*RefSeq*). Dessas 9108, para 3198 a *tag* de SAGE era a única evidência que suportava a existência do transcrito antisenso.

Recentemente, a utilização de novas abordagens e o desenvolvimento de novas metodologias tem contribuído para o aumento considerável do número de NATs identificados no genoma de mamíferos. GE et al. (2008) utilizaram uma plataforma de *array* (*Affymetrix Exon array*) em combinação a um protocolo modificado de síntese de cDNA e identificaram 1.516 *clusters* do *UniGene* com orientação antisenso. Desses, apenas 490 já foram descritos por outros trabalhos. Os autores avaliaram por RT-PCR fita específica 24 dos NATs identificados, dos quais 17 (74%) foram validados.

HE et al. (2009) descreveram uma nova metodologia para a identificação de NATs, denominada ASSAGE (*Asymetric Strand-Specific Analysis of Gene Expression*), que consiste no tratamento do RNA com bissulfito, o que promove a troca de todas as citidinas por uracilas. Após o tratamento com bissulfito, quando esse RNA é comparado ao DNA virtualmente convertido com bissulfito, só irá

apresentar similaridade com a fita a qual foi originado. Associado ao sequenciamento em larga escala (*Illumina*) do cDNA essa técnica demonstrou-se muito robusta na geração de *tags* (~4 milhões), que fornecem informação sobre o nível de expressão do transcrito e de qual fita do DNA esse transcrito foi originado. Em células normais do sangue periférico, os autores observaram a existência tanto de *tags* senso como antisenso indicando a formação de NATs para 2.061 (15,9%) dos 12.976 genes (*Ensembl*) avaliados.

Com a gama enorme de dados gerados por trabalhos que visam à identificação de NATs, tem sido proposta a construção de banco de dados que agrupem todas essas informações, e facilitem o estudo desses transcritos. Um deles foi desenvolvido por ZHANG et al. (2007) e disponibiliza os dados descritos por ZHANG et al. (2006) em um banco de dados conhecido como *NATsDB* (*Natural Antisense Transcripts DataBase)*. YIN et al. (2007), utilizando dados de nove trabalhos descritos na literatura, disponibilizaram um banco de dados contendo 30.000 pares senso/antisenso representados em 12 espécies diferentes de eucariotos. Essa plataforma, chamada *antiCODE* (*natural sense-antisense transcripts database)* integra ferramentas de busca eficientes a fim de facilitar o acesso a informações relevantes para cada transcrito, como por exemplo o padrão de sobreposição.

1.1.4 NATs e sua associação com doenças

Devido ao número abundante de NATs descritos na literatura e aos diversos mecanismos pelos quais eles podem regular a expressão gênica, não é surpreendente que alterações no padrão de expressão desses transcritos possam levar a padrões anormais de expressão gênica, contribuindo para o desenvolvimento de patologias,

incluindo o câncer. Apesar deste fato, poucos transcritos antisenso já foram associados a doenças em humanos. A Tabela 2 demonstra alguns exemplos representativos de NATs que possivelmente estejam associados a doenças.

Tabela 2 - Exemplos de doenças humanas possivelmente associadas a NATs.

Referência	NATs	Patologia
Capaccioli et al. (1996)	*BCL2* e *IgH*	linfoma folicular de células B
Rougeulle e Lalande (1998)	*SNURF-SNRP* e *UBE3A*	Síndrome de Prader-Willi e Angelman
Thrash-Binghan e Tartof (1999)	*HIF1α* e *aHIF*	carcinoma renal e de mama
Smilinich et al. (1999)	*KvLQT1*	Síndrome de Beckwith-Wiedemann
Yamamoto et al. (2002)	*Survivin* e *EPR1*	carcinoma coloretal
Tufarelli et al. (2003)	*LUC7L* e *α-globulin*	α-talassemia
Faghihi et al. (2008)	*BACE1* e *BACE-AS*	doença de Alzheimer

A tabela descreve os trabalhos que demonstraram alguns exemplos de NATs que podem estar envolvidos com uma variedade de doenças em humanos.

Como podemos observar na Tabela 2, alguns trabalhos têm sugerido o envolvimento dos transcritos antisenso na tumorigênese. SHENDURE e CHURCH (2002), a partir dos 144 *clusters* formando pares senso/antisenso identificados, demonstraram em tecido neoplásico a presença de uma maior fração de ESTs antisenso relacionadas a um suposto gene supressor de tumor (*DFFA-like effector B*), em relação às ESTs senso correspondentes, sendo que no tecido normal observaram o contrário. Neste mesmo trabalho, um padrão semelhante de expressão de ESTs de transcrito senso/antisenso foi observado para o gene *Burkitt lymphoma receptor 1*.

REIS et al. (2004) demonstraram a expressão de transcritos antisenso intrônicos associados a diferenciação tumoral. Em tumores de próstata com diferentes graus de diferenciação (segundo o escore de G*leason* - GS5 a 10), os autores avaliaram a expressão de transcritos antisenso que mapeavam em regiões intrônicas de genes conhecidos, por meio de *microarray* de cDNA. Após a análise de

~38.000 regiões intrônicas únicas de genes conhecidos, em 27 amostras de tumores de próstata, os autores observaram que uma fração considerada alta (6,6%) desses tipos de transcritos apresentou correlação significativa com o grau de diferenciação do tumor. Observaram ainda que do grupo de 12 genes mais correlacionados com o grau de diferenciação do tumor, 6 eram transcritos antisenso intrônicos mapeados em genes conhecidos. Os autores validaram em seis amostras de tumor de próstata, por RT-PCR fita específica em tempo real, a correlação (p= 0,0024) entre a expressão do transcrito antisenso intrônico para o gene *RASSF1* e o grau de diferenciação do tumor. Este foi o primeiro trabalho a utilizar a técnica de *microarray* para avaliar a expressão de transcritos antisenso em tumores.

Recentemente, MONTI et al. (2009) também sugeriram que a regulação gênica que ocorre envolvendo transcritos que formam pares senso/antisenso está associada ao desenvolvimento de tumores. Utilizando dados de sequências expressas (mRNA e ESTs) anotadas em banco de dados públicos e a sequência do genoma humano, os autores identificaram 101 genes mapeados no cromossomo 6 (6q21 e 6q27), dos quais 24 estavam envolvidos na formação de NATs. Dos 16 genes que seguiram para avaliação experimental por RT-PCR fita específica, a existência do transcrito antisenso foi confirmada em 11 (69%). Para os pares de transcritos envolvendo o gene *RPS6KA2*, os autores observaram que uma menor expressão do transcrito senso foi associada ao aumento de expressão do transcrito antisenso em 6 de 7 linhagens celulares de câncer de mama analisadas. Os autores confirmaram esses dados em 71% (42/59) das amostras de pacientes com câncer de mama por meio de *microarray* de cDNA, sugerindo que esse fato poderia estar associado ao desenvolvimento de tumores de mama. O gene *RPS6KA2* foi recentemente descrito

como um gene supressor de tumor e também apresentando expressão monoalélica (BIGNONE et al. 2007). Os resultados obtidos nestes trabalhos corroboraram a hipótese de que alterações no padrão de expressão de transcritos antisenso estejam relacionadas ao câncer.

Assim, visto o número surpreendente de NATs existentes, o fato de apresentarem-se conservados entre as espécies, evidências de que esses transcritos tenham um papel importante na regulação da expressão gênica e, ainda, recentes relatos de que eles possam estar associados a ocorrência de doenças, torna-se importante o desenvolvimento de novas abordagens para a identificação de novos NATs. Portanto, neste trabalho nós propomos a utilização de uma metodologia de análise da expressão gênica, denominada MPSS (*Massively Parallel Signature Sequencing)*, para a identificação de novos NATs no genoma humano.

1.1.5 *Massively parallel signature sequencing* (MPSS) e *Generation of Longer cDNA fragments from MPSS tags for Gene Identification* (GLGI-MPSS)

MPSS é uma técnica utilizada para determinar o nível de expressão gênica, baseada nos mesmos princípios de SAGE, ou seja, na produção e quantificação de sequências curtas (*tags)* próximas a extremidade 3' dos transcritos. No entanto, ao contrário do SAGE, o MPSS utiliza a clonagem *in vitro* de fragmentos de cDNA em *microbeads* e o sequenciamento em larga escala dessas partículas sem a necessidade de separação física dos fragmentos a serem sequenciados. A associação dessas duas tecnologias permite a produção de um número surpreendentemente maior de *tags* em relação à técnica de SAGE. Em ambas as técnicas o nível de expressão dos genes é

proporcional à freqüência de *tags* correspondentes (VELCULESCU et al. 1995; BRENNER et al. 2000).

Uma vantagem da técnica de MPSS é o número de *tags* geradas em cada biblioteca. Enquanto uma biblioteca de SAGE, em média, apresenta 120 mil *tags*, para MPSS, em média, as bibliotecas apresentam 1.1 milhões de *tags*, ou seja, um número 10 vezes superior de *tags*. Em potencial, essa enorme quantidade de *tags* geradas por MPSS permite uma cobertura mais abrangente do transcriptoma e, dessa forma, a identificação de transcritos pouco expressos ou mesmo raros ou, ainda, de transcritos não identificados pela técnica de SAGE (BRENNER et al. 2000). Uma observação importante é a de que é possível inferir a orientação das *tags*, sem a necessidade de utilizar critérios estringentes como no caso das sequências de mRNA e ESTs. Em ambas as metodologias, a extremidade 5' da *tag* está associada ao sítio de restrição (GATC ou CATG) utilizados na geração das mesmas.

Na técnica de MPSS são geradas *tags* com boa qualidade em torno de 17 nucleotídeos (nt) (BRENNER et al. 2000). O mapeamento dessas *tags* de MPSS na sequência do genoma humano possibilita a identificação de transcritos antisenso intrônicos, a eliminação de transcritos mapeados em regiões repetitivas e a eliminação de *tags* que apresentam ambigüidade (representam regiões distintas no genoma).

Ao que parece, nenhum trabalho utilizou dados gerados pela construção de bibliotecas de MPSS para a identificação de novos transcritos antisenso em humanos. Em *Arabidopsis thaliana*, os dados de MPSS foram utilizados para a análise de expressão de transcritos antisenso previamente identificados (WANG et al., 2005)

Em contrapartida, o tamanho das *tag*s de MPSS apresenta uma limitação, pois não é possível a caracterização dos transcritos as quais representam caso essas *tags* não estejam associadas a algum gene conhecido. Para resolver tal limitação foi desenvolvida a técnica de GLGI-MPSS (*Generation of Longer cDNA fragments from MPSS tags for Gene Identification*) (SILVA et al. 2004a).

Originalmente, o protocolo do GLGI foi desenvolvido com a finalidade de converter as *tags* de SAGE (13 nt) em um fragmento 3´ de cDNA correspondente (CHEN et al. 2000). Por meio dessa conversão e com base na sequência do fragmento 3´ de cDNA obtido foi possível caracterizar de forma mais precisa e específica os transcritos representados pela *tags* de SAGE. A utilização dessa técnica permite resolver dois problemas freqüentemente associados com a análise da expressão gênica por SAGE. Primeiramente, a produção de um fragmento de cDNA a partir de uma *tag* de SAGE inicialmente não associada a um gene conhecido facilita a caracterização de novos transcritos. De maneira similar, a técnica de GLGI pode ser utilizada com sucesso na caracterização de transcritos diferencialmente expressos nos casos em que uma única *tag* de SAGE está associada a vários transcritos.

Assim, SILVA et al. (2004a) adaptaram o protocolo original de GLGI-SAGE para converter *tags* de MPSS que não foram associadas a transcritos humanos conhecidos (*tags* órfãs). Para o desenvolvimento dessa nova técnica, algumas modificações foram introduzidas ao protocolo original de GLGI. Tais modificações incluíram mudanças na sequência dos adaptadores usados para a amplificação do fragmento 3’ de cDNA, a utilização de um número maior de ciclos na PCR e o rastreamento de um número maior de colônias para cada fragmento de GLGI. A

modificação na sequência dos adaptadores foi necessária devido a utilização de uma enzima de restrição (*DpnII)* diferente para a construção das bibliotecas de MPSS, o aumento no número de ciclos na PCR e o rastreamento de um maior número de colônias foram necessários devido à baixa expressão dos transcritos correspondentes a essas *tags* de MPSS. Utilizando essa metodologia, SILVA et al. (2004a) identificaram novos transcritos antisenso, *tags* alternativas de MPSS geradas por transcritos polimórficos, *tags* derivadas de transcritos apresentando *splicing* alternativo e também *tags* provenientes de artefatos gerados pelo pareamento interno (*internal priming)* do oligo (dT) no momento da construção da biblioteca de MPSS. GLGI-MPSS é uma técnica rápida, específica e eficiente na análise em larga escala de *tags* de MPSS que não estão associadas a transcritos conhecidos (SILVA et al. 2004a). Uma representação esquemática da técnica de GLGI-MPSS pode ser observada na Figura 4.

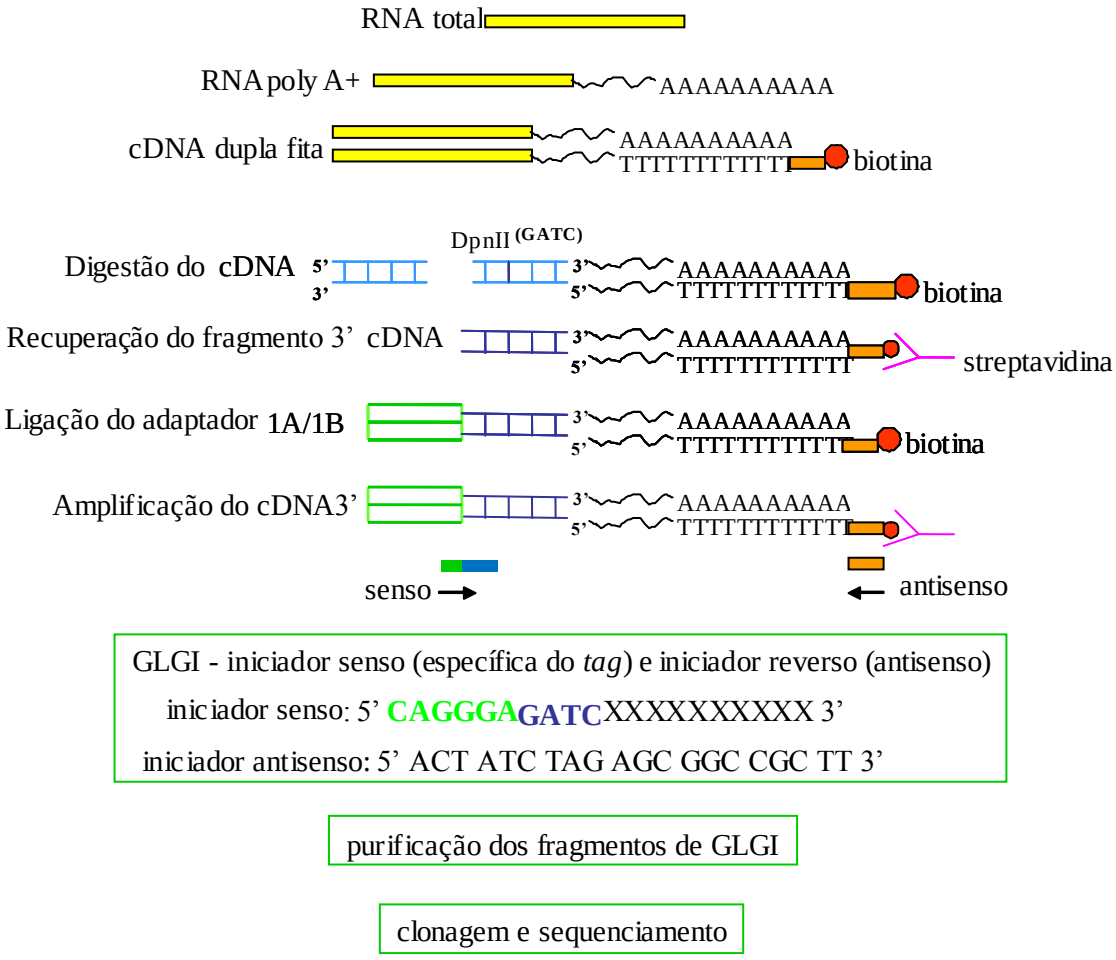


Figura 4 - Representação esquemática da técnica de GLGI-MPSS. Para a construção da biblioteca de GLGI-MPSS, o cDNA foi sintetizado utilizando-se um iniciador oligo (dT) biotinilado, importante para a recuperação do fragmento 3' de cDNA e, em seguida, submetido a digestão com a enzima de restrição *DpnII*, a mesma utilizada para a geração da *tag* de MPSS. Após a digestão, foram ligados adaptadores ao fragmento 3´ de cDNA, que são importantes para a amplificação por PCR em larga escala do fragmento acima mencionado. Desta forma, para a amplificação dos fragmentos de GLGI foram utilizados iniciadores senso que incluem 17 pares de base (pb) da *tag* de MPSS e 6 pb adicionais (CAGGGA) correspondentes à seqüência dos adaptadores, totalizando 23 pb para cada iniciador específico (5´ CAGGAGATCXXXXXXXXXXXXXX 3'). Também foi utilizado um iniciador antisenso (5' ACTATCTAGAGCGGCCGCTT 3') presente na extremidade 3' de todas as moléculas de cDNA que foi incorporado a seqüência do oligo (dT) utilizado na síntese do cDNA dupla fita. Após essa etapa, as reações de GLGI foram purificadas, os fragmentos foram clonados e seqüenciados.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Este projeto tem o objetivo de identificar e validar a existência de novos transcritos antisenso, presentes no genoma humano, a partir de dados de MPSS.

1.2.2 Objetivos específicos

- Identificar por meio de análises computacionais *tags* de MPSS que estão mapeadas em exons ou introns de transcritos humanos e cuja orientação seja compatível com a existência de um transcrito antisenso;
- Estender as *tags* de MPSS correspondentes aos transcritos antisenso por meio da metodologia de GLGI-MPSS, a fim de obter um fragmento de cDNA de maior extensão que permita a melhor caracterização do transcrito antisenso;
- Validar por meio de RT-PCR fita específica a existência dos transcritos antisenso correspondentes as *tags* de MPSS extendidas por meio de GLGI-MPSS.

1.3 MATERIAL E MÉTODOS, RESULTADOS E DISCUSSÃO

Seguindo as normas estabelecidas pela Comissão de Pós-Graduação da Fundação Antônio Prudente, optamos por apresentar esta tese em forma de artigo. Assim, apresentamos uma breve descrição dos tópicos abordados no artigo já publicado e intitulado "*Sense-antisense pairs in mammals: functional and evolutionary considerations*". Os tópicos contendo os materiais e métodos, os resultados e a discussão do trabalho estão descritos no artigo a seguir.

1.3.1 Artigo intitulado: "*Sense-antisense pairs in mammals: functional and evolutionary considerations*".

O objetivo principal deste projeto foi a identificação de novos NATs utilizando dados de expressão gênica gerados pela metodologia de MPSS. NATs para os quais a única evidência de sua existência era a *tag de MPSS*, não sendo representados por nenhum outro tipo de sequências expressas (mRNA ou ESTs).

Assim, em colaboração com o grupo de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer foi desenvolvida uma estratégia computacional que permitiu a identificação de 4.308 genes humanos formando pares com pelo menos um transcrito antisenso evidenciado pela existência de pelo menos uma *tag* de MPSS mapeada na mesma região genômica, entretanto com orientação oposta.

Após essa análise inicial *in silico*, passamos a etapa de validação experimental na qual 96 *tags* de MPSS representando novos transcritos antisenso foram selecionadas e submetidas à técnica de GLGI-MPSS para a melhor caracterização destes transcritos. A técnica de GLGI é mais eficiente quando as

bibliotecas de GLGI – MPSS são construídas a partir da mesma fonte (RNA de tecido ou linhagem celular) as quais foram utilizadas para a construção das bibliotecas de MPSS. Das 41 bibliotecas de MPSS, nós tínhamos disponíveis os RNAs das linhagens celulares de mama Hb4a e Hb4a-C5.2. Assim, 46/96 *tags* avaliadas foram confirmadas como extensões 3' específicas (fragmentos de GLGI) e com orientação antisenso que representavam possivelmente novos transcritos antisenso.

Durante essa etapa, observamos que uma fração (19 dos 46 fragmentos de GLGI validados) apresentaram a cauda poli(A) alinhada diretamente ao genoma humano, o que não era esperado. Assim, desenvolvemos uma estratégia experimental, baseada na utilização de RT-PCR para avaliar transcritos com as mesmas características (poli(A) mapeando no genoma) presentes em bibliotecas de cDNA enriquecidas para este tipo de sequências. Dos 11 transcritos selecionados, para 7 não foi possível confirmar sua existência, dessa forma sugerimos que uma parte destes transcritos sejam artefatos gerados por eventos de *internal priming* no genoma humano. Entretanto, confirmamos a ocorrência de uma parte (4) deste tipo de transcritos e sugerimos que tais transcritos possam ter sido gerados por eventos de retroposição. Esses resultados serão apresentados no artigo já publicado (GALANTE et al. 2007).

Finalmente, 27/46 fragmentos de GLGI-MPSS apresentaram-se como potenciais transcritos antisenso. Esses transcritos foram submetidos à avaliação por meio de RT-PCR fita específica para a confirmação de sua existência, sendo esses resultados apresentados no item 1.3.2.

A estratégia desenvolvida em nosso trabalho para a identificação de novos NATs está representada na Figura 5.

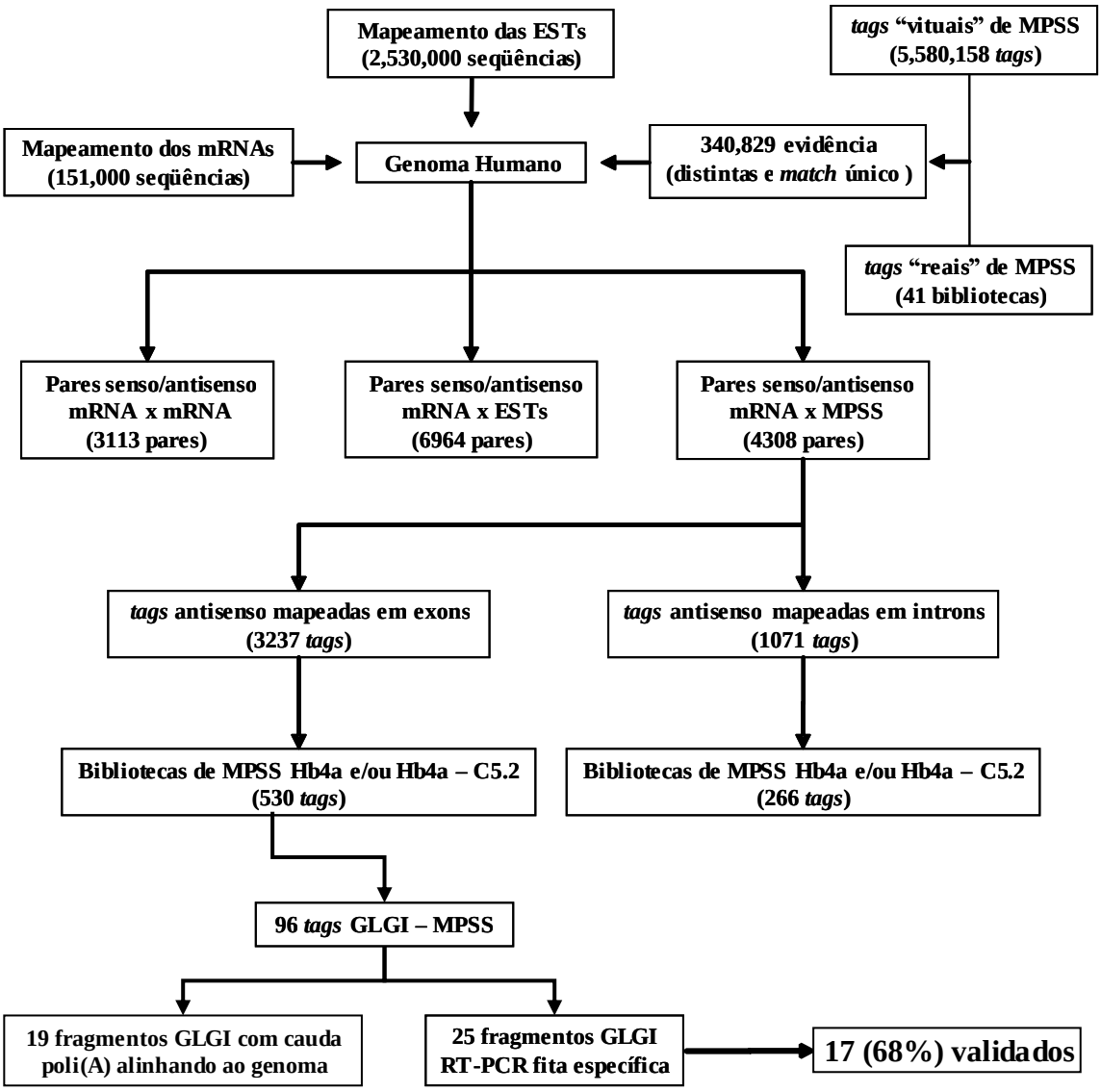


Figura 5 - Representação da estratégia utilizada na identificação de *tags* de MPSS com orientação antisenso a transcritos conhecidos. Nossa estratégia consistiu em mapear no genoma humano as seqüências de mRNA, ESTs e MPSS disponíveis em bancos públicos, o que permitiu a identificação de três tipos diferentes de pares senso/antisenso (mRNA x mRNA, mRNA x ESTs, mRNA x *tags* MPSS). Após essa etapa, os pares formados por *tags* de MPSS foram divididos de acordo com a sua sobreposição ao outro transcrito (mRNA) em exonicos e intronicos. Assim, 96 *tags* de MPSS foram selecionadas e submetidas ao protocolo de GLGI. A técnica de GLGI é mais eficiente quando as bibliotecas de GLGI – MPSS são construídas a partir da mesma fonte (RNA de tecido ou linhagem celular) as quais foram utilizadas para a construção das bibliotecas de MPSS. Das 41 fontes utilizadas para a construção das bibliotecas de MPSS haviam disponíveis os RNAs das linhagens celulares de mama *Hb4a* e *Hb4a-C5.2*. Dessa maneira, o RNA dessas linhagens foi utilizado como molde na construção de nossas bibliotecas de GLGI. Portanto, foi possível analisar com maior eficiência os dados em relação às bibliotecas de MPSS derivadas do RNA dessas linhagens. Dessa maneira, os 25 fragmentos de GLGI

confirmados como extensões 3’ específicas e orientação antisenso foram avaliados por RT-PCR fita específica, dos quais 17 (68%) foram validados.

1.3.2 - Artigo

Research

Highly accessed Open Access

Sense-antisense pairs in mammals: functional and evolutionary considerations

Pedro AF Galante*†, Daniel O Vidal*, Jorge E de Souza*, Anamaria A Camargo* and Sandro J de Souza*

Addresses: *Ludwig Institute for Cancer Research, São Paulo Branch, Hospital Alemão Oswaldo Cruz, Rua João Juliao 245, 1 andar, São Paulo, SP 01323-903, Brazil. †Department Of Biochemistry, University of São Paulo, Av. Prof. Lineu Prestes, 748 - sala 351, São Paulo, SP 05508-900, Brazil.

Correspondence: Sandro J de Souza. Email: sandro@compbio.ludwig.org.br

Published: 19 March 2007

Genome Biology 2007, 8:R40 (doi:10.1186/gb-2007-8-3-r40)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/3/R40

Received: 3 May 2006
Revised: 4 September 2006
Accepted: 19 March 2007

Abstract

Background: A significant number of genes in mammalian genomes are being found to have natural antisense transcripts (NATs). These sense-antisense (S-AS) pairs are believed to be involved in several cellular phenomena.

Results: Here, we generated a catalog of S-AS pairs occurring in the human and mouse genomes by analyzing different sources of expressed sequences available in the public domain plus 122 massively parallel signature sequencing (MPSS) libraries from a variety of human and mouse tissues. Using this dataset of almost 20,000 S-AS pairs in both genomes we investigated, in a computational and experimental way, several putative roles that have been assigned to NATs, including gene expression regulation. Furthermore, these global analyses allowed us to better dissect and propose new roles for NATs. Surprisingly, we found that a significant fraction of NATs are artifacts produced by genomic priming during cDNA library construction.

Conclusion: We propose an evolutionary and functional model in which alternative polyadenylation and retroposition account for the origin of a significant number of functional S-AS pairs in mammalian genomes.

Background

Natural antisense RNAs (or natural antisense transcripts (NATs)) are endogenous transcripts with sequence complementarity to other transcripts. There are two types of NATs in eukaryotic genomes: cis-encoded antisense NATs, which are transcribed from the opposite strand of the same genomic locus as the sense RNA and have a long (or perfect) overlap with the sense transcripts; and trans-encoded antisense NATs, which are transcribed from a different genomic locus of the sense RNA and have a short (or imperfect) overlap with the sense transcripts. Cis-NATs are usually related in a one-to-one fashion to the sense transcript, whereas a single trans-NAT may target several sense transcripts [1-3]. In this manuscript, we describe analyses in which only cis-NATs were considered. From now on, we refer to these loci as sense-antisense (S-AS) pairs.

When evaluated globally, several features related to the distribution of NATs strongly suggest they have a prominent role in antisense regulation in gene expression [4-7]. For instance, expression of S-AS transcripts tends to be positively or negatively correlated and is more evolutionarily conserved than expected by chance [4,5,7]. Although experimental validation of a putative regulatory role has been achieved for a few models [8-10], it is still unknown whether antisense regulation is a rule or an exception in the human genome. NATs have been implicated in RNA and translational interference [11], genomic imprinting [12], transcriptional interference [13], X-inactivation [14], alternative splicing [10,15] and RNA editing [16]. Moreover, an accumulating body of evidence suggests that NATs might have a pivotal role in a range of human diseases [2].

NATs were initially identified in studies looking at individual genes. However, with the accumulation of whole genome and expressed sequences (mRNA and ESTs) in public databases, a significant number of NATs has been identified using computational analysis [17-22]. These studies showed a widespread occurrence of these transcripts in mammalian genomes. The first evidence that antisense transcription is a common feature of mammalian genomes came from analysis of reverse complementarity between all available mRNA sequences [17]. Subsequent studies, using larger collections of mRNA sequences, ESTs and genomic sequences, confirmed and extended these initial observations [18-22]. More recently, other sources of expression data, such as serial analysis of gene expression (SAGE) tags, were used to expand the catalog of NATs present in mammalian genomes [23,24]. At present, it is estimated that at least 15% and 20% of mouse and human transcripts, respectively, might form S-AS pairs [18,22], although a recent analysis [25] reported that 47% of human transcriptional units are involved in S-AS pairing (24.7% and 22.7% corresponding to S-AS pairs with exon and non-exon overlapping, respectively).

The major obstacle in using expressed sequence data for NAT identification is how to determine the correct orientation of the sequences, especially ESTs. Many ESTs were not directionally cloned and even well-known mRNA sequences were registered from both strands of cloned cDNAs or are incorrectly annotated. As done by others [18,22,23], we here established a set of stringent criteria, including the orientation of splicing sites, the presence of poly-A signal and tail as well as sequence annotation, to determine the correct orientation of each transcript relative to the genomic sequence and made a deep survey of NAT distribution in the human and mouse genomes. Using a set of computational and experimental procedures, we extensively explored expressed sequences and massively parallel signature sequencing (MPSS) data mapped onto the human and mouse genomes. Besides generating a catalog of known and new S-AS pairs, our analyses shed some light on functional and evolutionary aspects of S-AS pairs in mammalian genomes.

Results and discussion

Overall distribution of S-AS pairs in human and mouse genomes

To identify transcripts that derive from opposite strands of the same locus, we used a modified version of an in-house knowledgebase previously described for humans [26-28]. This knowledgebase contains more than 6 million expressed sequences mapped onto the human genome sequence and clustered in approximately 111,000 groups. Furthermore, SAGE [29] and MPSS [30] tags were also annotated with all associated information, such as tag frequency, library source and tag-to-gene-assignment (using a strategy developed by us for SAGE Genie [31]). An equivalent knowledgebase was built for the mouse genome (for more details see Materials and methods).

We first designed software that searched the human and mouse genomes extracting gene information from transcripts mapped onto opposite strands of the same locus. Several parameters were used by the software to identify S-AS pairs, such as: sequence orientation given by the respective GenBank entry; presence and orientation of splice site consensus; and presence of a poly-A tail (for more details see Materials and methods). We found 3,113 and 2,599 S-AS pairs in human and mouse genomes, respectively, containing at least one full-insert cDNA (sequences annotated as 'mRNA' in GenBank and referred to here as such) in each orientation (Table 1). Furthermore, we also made use of EST data from both species. A critical issue when using ESTs is the orientation of the sequence, a feature not always available in the respective GenBank entries. We overcame this problem by simply using those ESTs that had a poly-A tail or spanned an intron and, therefore, disclosed their strand of origin by the orientation of a splicing consensus sequence (GT...AG rule). We found 6,964 and 5,492 additional S-AS pairs when EST data were incorporated into the analysis, totaling 10,077 and 8,091 pairs for human and mouse genomes, respectively (Table 1). All of these pairs contained at least one mRNA since we did not analyze EST/EST pairs. It is important to note that we haven't considered in the present analysis non-polyadenylated transcripts and trans-NATs. Thus, the total number of NATs is likely to be even higher in both genomes. Data presented in Table 1 are split in cases where a single S-AS pair is present in a given locus (single bidirectional transcription) and in cases where more than one pair is present per locus (multiple bidirectional transcription). Additional data file 1 lists two representative GenBank entries for all S-AS pairs split by chromosome mapping in the two species. As previously observed [17], S-AS pairs are under-represented in the sex chromosomes of both species (Additional data file 2).

The above numbers confirm that S-AS pairs are much more frequent in mammalian genomes than originally estimated [4,17,18]. Our analyses suggest that at least 21,000 human and 16,000 mouse genes are involved in S-AS pairing. These numbers are more in agreement with those from [32] in their

Table 1

Overall distribution of S-AS pairs in the human and mouse genomes

cDNA type	Single bidirectional transcription		Multiple bidirectional transcription	
	Human	Mouse	Human	Mouse
mRNA-mRNA	2,109	1,879	1,004	720
mRNAs-ESTs	3,299	3,265	3,665	2,227
Total	5,408	5,144	4,669	2,947

Single bidirectional transcription corresponds to those loci in which only one S-AS pair is present. Multiple bidirectional transcription corresponds to those loci in which more than one S-AS pairs is present (at least one gene belongs to more than one S-AS pair).

analysis using tiling microarrays to evaluate gene expression of a fraction of the human genome. For the mouse genome, our numbers are in agreement with those reported by Katayama et al. [8]. A more recent analysis [25] also gives a similar estimate of S-AS pairs in both human and mouse genomes.

Could this high number of S-AS pairs be due to the stringency of our clustering strategy? If the same transcriptional unit is fragmented in close contigs due to 3' untranslated region (UTR) heterogeneity, the total number of clusters would be inflated, leading to an erroneous count of S-AS pairs. To evaluate this possibility, we relaxed our clustering parameters, requiring a minimum of 1 base-pair (bp) same strand overlap for clustering. Furthermore, we collapsed into a single cluster all pairs of clusters located in the same strand and less than 30 bp away from each other. Additional data file 3 shows the total number of clusters and S-AS pairs after this new clustering strategy was employed. As expected, both the total number of clusters and S-AS pairs decreased with the new clustering methodology. The total number of clusters decreased by 2% and 1% for human and mouse, respectively, while the total number of S-AS pairs decreased by 0.3% for both human and mouse. Thus, the small difference observed does not affect the conclusions on the genomic organization of S-AS pairs. For all further analyses, we decided to use the original dataset obtained with a more stringent clustering methodology.

We further explored the genomic organization of S-AS pairs using the subset of 3,113 human and 2,599 mouse pairs that contained mRNAs in both sense and antisense orientations. The genomic organization of S-AS pairs can be further divided into three subtypes based on their overlapping patterns: head-head (5'5'), tail-tail (3'3') or embedded (one gene contained entirely within the other) pairs (Table 2). For a schematic view of the genomic organization of S-AS pairs, see Additional data file 4. Embedded pairs are more frequent in both species, corresponding to 47.8% and 42.5% of all pairs in human and mouse, respectively. If we take into account the intron/exon organization of both genes, we observe that the most frequent overlap involves at least one exon-intron border. In spite of this, a significant amount of NATs maps completely within introns from the sense gene in both human and mouse (category 'Fully intronic' in Table 2). Interestingly, more than three-quarters of all S-AS pairs categorized as 'Fully intronic' fall within the embedded category for human and mouse. How unique is this distribution? Monte Carlo simulations, in which we randomly replaced NATs in relation to sense genes while keeping their 5'5'/embedded/3'3' orientation, show that the distribution of S-AS pairs is quite unique. All three categories of S-AS pairs deviate from a random distribution (chi-square = 11.5, df (degrees of freedom) = 2, $p = 0.003$ for embedded pairs; chi-square = 49, df = 2, $p = 2.3 \times 10^{-11}$ for 5'5' pairs; chi-square = 132, df = 2, $p = 2.1 \times 10^{-29}$ for 3'3' pairs). This peculiar distribution will be further discussed in the light of the expression analyses. Since these intronic NATs have been shown to be over-expressed in prostate tumors [33], our dataset should be further explored regarding differential expression in cancer. Due to their genomic distribution, any putative regulatory role of these intronic NATs would have to be restricted to the nucleus. Interestingly, Kiyosawa et al. [34] observed that a significant amount of NATs in mouse is poly-A negative and nuclear localized.

Another interesting observation is the higher frequency of intronless genes within the set of S-AS pairs (Table 3). About half (47%) of all mRNA/mRNA S-AS pairs in humans contains at least one intronless gene. This number is slightly lower for mouse (44%) (Table 3). Interestingly, intronless genes are significantly enriched within the set of embedded pairs (chi-square = 95.9, $p < 1.2 \times 10^{-22}$ for human and chi-square = 3.98 and $p < 0.045$ for mouse). For humans, 66% of all S-AS pairs containing at least one intronless gene are within the 'embedded' category; Sun et al. [5] found 43.4% of their S-AS pairs as 'embedded'. Furthermore, they found 35% of 3'3' pairs while we found only 25%. These differences are probably due to the fact that Sun et al. [5] included in their analyses pairs containing only ESTs.

All these results clearly show that subsets of S-AS pairs have distinct genomic organization, suggesting that they may play different biological roles in mammalian genomes. Below we will discuss these data in a functional/evolutionary context.

Table 2

Distribution of NATs in relation to the genomic structure of the sense transcript

	Human			Mouse		
	5'5'	Embedded	3'3'	5'5'	Embedded	3'3'
Fully exonic	112 (20%)	32 (3%)	213 (40%)	156 (27%)	14 (2%)	227 (45%)
Exonic/intronic	362 (64%)	372 (37%)	259 (48%)	360 (62%)	338 (42%)	242 (48%)
Fully intronic	92 (16%)	606 (60%)	61 (12%)	61 (11%)	448 (56%)	33 (7%)
Total	566	1,010	533	577	800	502

5'5', head-head orientation; 3'3', tail-tail orientation.

Conservation of S-AS pairs between human and mouse

Using our set of human and mouse S-AS pairs, we measured the degree of conservation between S-AS pairs from human and mouse. Since the numbers reported so far are discrepant, ranging from a few hundred [5,6] to almost a thousand [25], we decided to use different strategies. We first used a strategy based on HomoloGene [35]. The number of S-AS pairs with both genes mapped to HomoloGene is 854 for human and 579 for mouse. Among these, 190 S-AS pairs are conserved between human and mouse. One problem with this type of analysis lies in its dependence on HomoloGene, which, for example, does not take into consideration genes that do not code for proteins. Therefore, we decided to implement a different strategy, in which we identified those pairs that had at least one conserved gene mapped by HomoloGene and tested each known gene's NAT for sequence level conservation. Using this strategy, we found an additional 546 cases, giving a total of 736 (190 + 546) conserved S-AS pairs between human and mouse. Finally, we also applied to our dataset the same strategy used by Engstrom et al. [25], in which they counted the number of human and mouse S-AS pairs that had exon overlap in corresponding positions in a BLASTZ alignment of the two genomes. We applied the same strategy to our dataset and found 1,136 and 1,144 corresponding S-AS pairs in human and mouse, respectively. As observed by Engstrom et al. [25] the numbers from human and mouse slightly differ because a small proportion of mouse pairs corresponded to several human pairs and vice versa. Additional data file 5 lists all S-AS pairs found by the three methodologies discussed above.

There is a predominance of 3'3' pairs in all sets of conserved S-AS pairs. For the first strategy solely based on HomoloGene, 67% of all pairs are 3'3' compared to 19% embedded and 14% 5'5'. For the dataset obtained using the strategy from Engstrom et al. [25], there is also a prevalence of 3'3'pairs (48%) compared to embedded (14%) and 5'5 (38%) pairs. We have also modified the method of Engstrom et al. [25] to take into account all S-AS pairs and not only those presenting exon-exon overlap. These data are shown in Additional Data File 6. We observed that S-AS pairs whose overlap is classified as 'Fully intronic' are less represented in the set of conserved S-AS pairs (18% in this set compared to 29% in the whole dataset of S-AS pairs). The same is true for S-AS pairs containing at least one intronless gene (26% in the set of conserved S-AS pairs compared to 47% in the whole dataset). These last results are in accordance with our previous observation that conserved S-AS pairs are enriched with 3'3' pairs. As seen in Tables 2 and 3, 3'3' pairs are poorly represented in the categories 'Fully intronic' (Table 2) and 'Intron/intronless' (Table 3).

Discovery of new S-AS pairs in human and mouse genomes using MPSS data

Large-scale expression profiling tools have been used to discover and analyze the co-expression of S-AS pairs [5,23,34]. Quéré et al. [23], for instance, recently explored the SAGE

Table 3

Classification of S-AS pairs in reference to their orientation and the presence of introns at the genome level for both genes in a pair

NAT pair	Human			Mouse		
	5'5'	Embedded	3'3'	5'5'	Embedded	3'3'
Both with intron	342 (61%)	351 (35%)	417 (78%)	259 (45%)	394 (49%)	390 (78%)
Intron-intronless	206 (36%)	645 (64%)	103 (19%)	285 (49%)	398 (50%)	96 (19%)
Both intronless	18 (3%)	14 (1%)	13 (3%)	33 (6%)	8 (1%)	16 (3%)
Total	566	1,010	533	577	800	502

5'5', head-head orientation; 3'3', tail-tail orientation.

Table 4

Distribution of MPSS tags in an antisense orientation in human and mouse genomes

	Number of clusters	
	Human	Mouse
One exonic tag	2,212 (51.3%)	124 (57.3%)
One intronic tag	875 (20.3%)	90 (41.7%)
Exonic and intronic tag	707 (16.4%)	2 (1%)
Multiple exonic tags	318 (7.4%)	0
Multiple intronic tags	196 (4.6%)	0
Total	4,308	216

Exonic and intronic refer to the genome organization of the sense gene. For instance, the category 'One exonic tag' corresponds to those genes with only one antisense tag complementary to its exonic region. All identified tags are found at a frequency ≥3 tags per million (see Materials and methods).

repositories to detect NATs. These authors searched for tags mapped on the reverse complement of known transcripts and analyzed their expression pattern on different SAGE libraries. However, no attempt was made to experimentally validate the existence of such NATs. Here, we made use of MPSS data available in public repositories [36,37] to search for new NATs in both human and mouse genomes. Since MPSS tags are longer than conventional SAGE tags, we can use the genome sequence for tag mapping. Furthermore, MPSS offers a much deeper coverage of the transcriptome since at least a million tags are generated from each sample.

We made use of 122 MPSS libraries derived from a variety of human and mouse tissues (81 libraries for mouse, 41 for human; see the list in Additional data file 7). Our strategy was based on the generation of virtual tags from each genome by simply searching the respective genome sequence for DpnII sites. Since these sites are palindromes, we extract, for each one, two virtual tags (13 and 16 nucleotide long tags for human and mouse, respectively), both immediately downstream of the restriction site but in opposite orientations (see Materials and methods for more details). In this way, we could evaluate the expression of transcriptional units present in both strands of DNA. We obtained 5,580,158 and 8,645,994 virtual tags for the human and mouse genomes, respectively. This set of virtual tags was then compared to a list of tags observed in the MPSS libraries. As true for any study using mapped tags, our analysis misses those cases in which a tag maps exactly at an exon/exon border at the cDNA level.

We first evaluated the number of cDNA-based S-AS pairs (shown in Table 1) that were further confirmed by the presence of an MPSS tag. Data for this analysis are presented as Additional data file 8. Roughly, 84% and 51% of all cDNA-based S-AS pairs were confirmed by MPSS data for human and mouse, respectively.

Since we were interested in finding new antisense transcripts, we searched for tags found in the MPSS libraries that were mapped on the opposite strand of both introns and exons of known genes. For this analysis we excluded those genes that were already part of S-AS pairs as described above. For humans, 4,308 genes have at least one MPSS tag derived from the antisense strand (Table 4). For 1,221 human genes there were two or more distinct MPSS tags in the antisense orientation. Another interesting observation is the larger number of MPSS tags antisense to exonic regions of the sense genes. Unexpectedly, we found a much smaller number of antisense tags for mouse (Table 4). Although the number of mouse libraries is larger (81 mouse and 41 human libraries), the number of unique tags is significantly smaller (56,061 for mouse and 340,820 for human). The assignment of these unique tags to known genes shows a smaller representation of known genes in the mouse dataset (51% against 66% for human). It is unlikely, however, that these differences can explain the dramatic difference shown in Table 4. Further analyses are needed to solve this apparent discrepancy.

To experimentally validate the existence of these novel human NAT candidates we used the GLGI (Generation of Longer cDNA fragments from SAGE for Gene Identification)- MPSS technique [38] to convert 96 antisense MPSS tags into their corresponding 3' cDNA fragments. A sense primer corresponding to the antisense MPSS tag was used for GLGI-MPSS amplification as described in Materials and methods. A predominant band was obtained for most of the GLGI-MPSS reactions (Figure 1). Amplified fragments were purified, cloned, sequenced and aligned to the human genome sequence. We were able to generate a specific 3' cDNA fragment for 46 (50.5%) out of 91 novel antisense candidates. Of these 46, the poly-A tail of 19 aligned with stretches of As in the human genome sequence (this finding will be discussed further). The existence of three of these antisense transcripts, out of three that were tested, was further confirmed by orientation-specific RT-PCR (data not shown).

Among the 49.5% (91 - 46 = 45) of candidates that were not considered to be validated, we found 25 that were amplified in the GLGI-MPSS experiment but whose exon-intron organ-

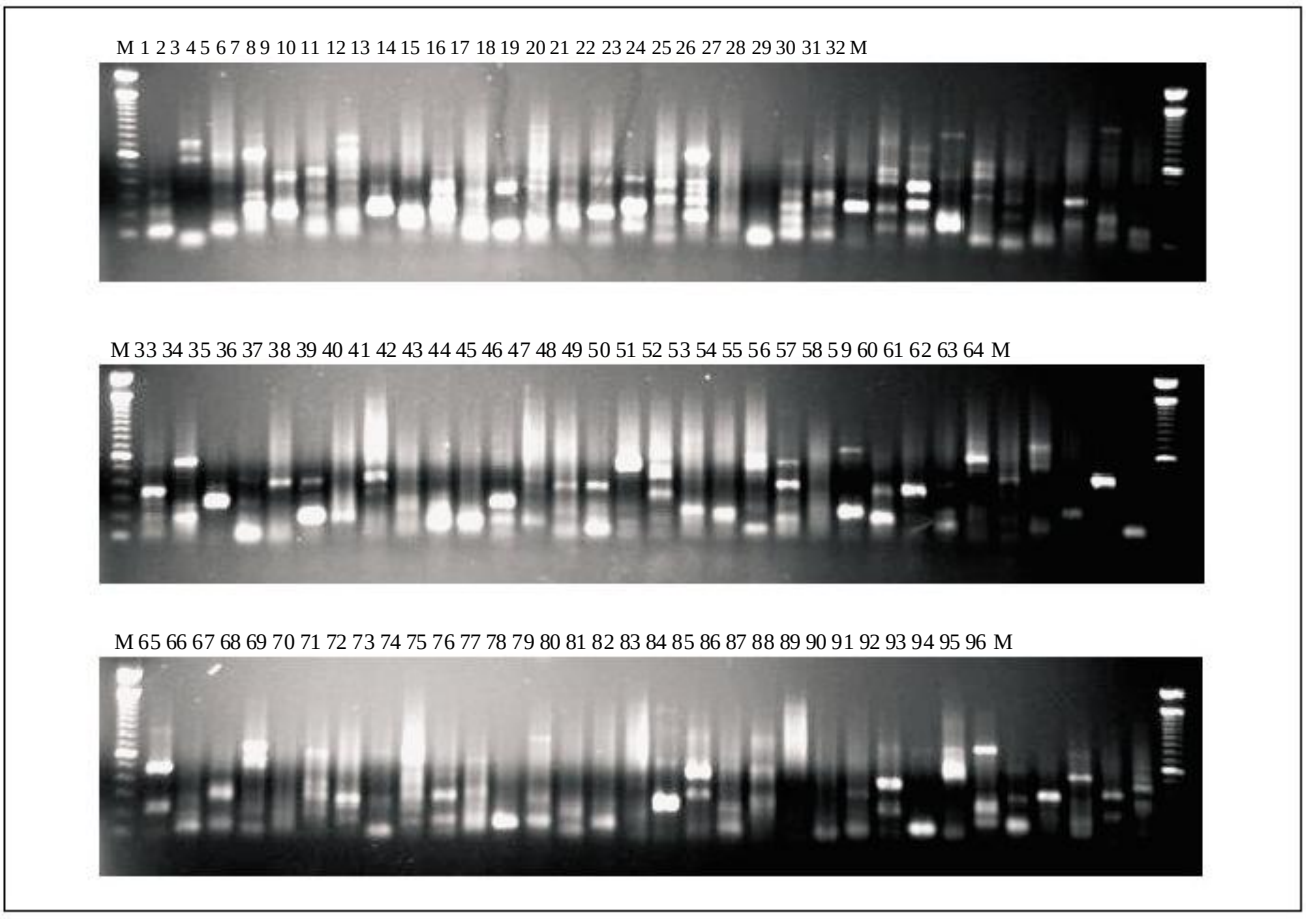


Figure 1
GLGI-MPSS amplification. GLGI amplifications for 96 MPSS antisense tags were analyzed on agarose gels stained with ethidium bromide. Note that some lanes show only a single amplified band whereas others have more than one band and sometimes a smear. A 100 bp ladder (M) was used as molecular weight marker.

ization was identical to the sense gene. Although antisense sequences like these have already been observed [39], we did not consider them as validated antisense transcripts. Orientation-specific RT-PCR confirmed the existence of one transcript, out of two that were tested.

Alternative polyadenylation as a major factor in defining S-AS pairs

Dahary et al. [6] observed that S-AS overlap usually involves transcripts generated by alternative polyadenylation. This observation had already been reported by us and others [40]. We decided to test if these preliminary observations would survive a more quantitative analysis. We found that the S-AS overlap is predominantly due to alternative polyadenylation variants. Roughly, 51% of all S-AS pairs (274 out of 533 3'3' pairs) overlap due to the existence of at least one variant. This number is certainly underestimated since many variants are still not represented in the sequence databases. The above observation raises the exciting possibility that antisense regulation is associated with the regulation of alternative polyadenylation. It is expected that the presence of overlapping genes imposes constraints on their evolution since any mutation will be evaluated by natural selection according to its effect in both genes. Thus, in principle, overlapping genes should impose a negative effect on the fitness of a subject. Alternative polyadenylation has the potential to relax such negative selection since the overlapping is dependent on a post-transcriptional modification.

If alternative polyadenylation is a significant factor in defining S-AS pairs, we would expect a lower rate of alternative polyadenylation in chromosome X, which has the smallest density of S-AS pairs. Indeed, only 20% of all messages from the X chromosome show at least two polyadenylation variants, compared to 27.5%, on average, for the autosomes (chi-square = 34.91, df = 1, $p < 0.0001$).

A fraction of S-AS pairs is generated through internal priming and retroposition events

During the validation of new NATs identified using the MPSS data, we noticed that a significant fraction of GLGI amplicons (19 out of 46 validated fragments) had their 3' ends aligning

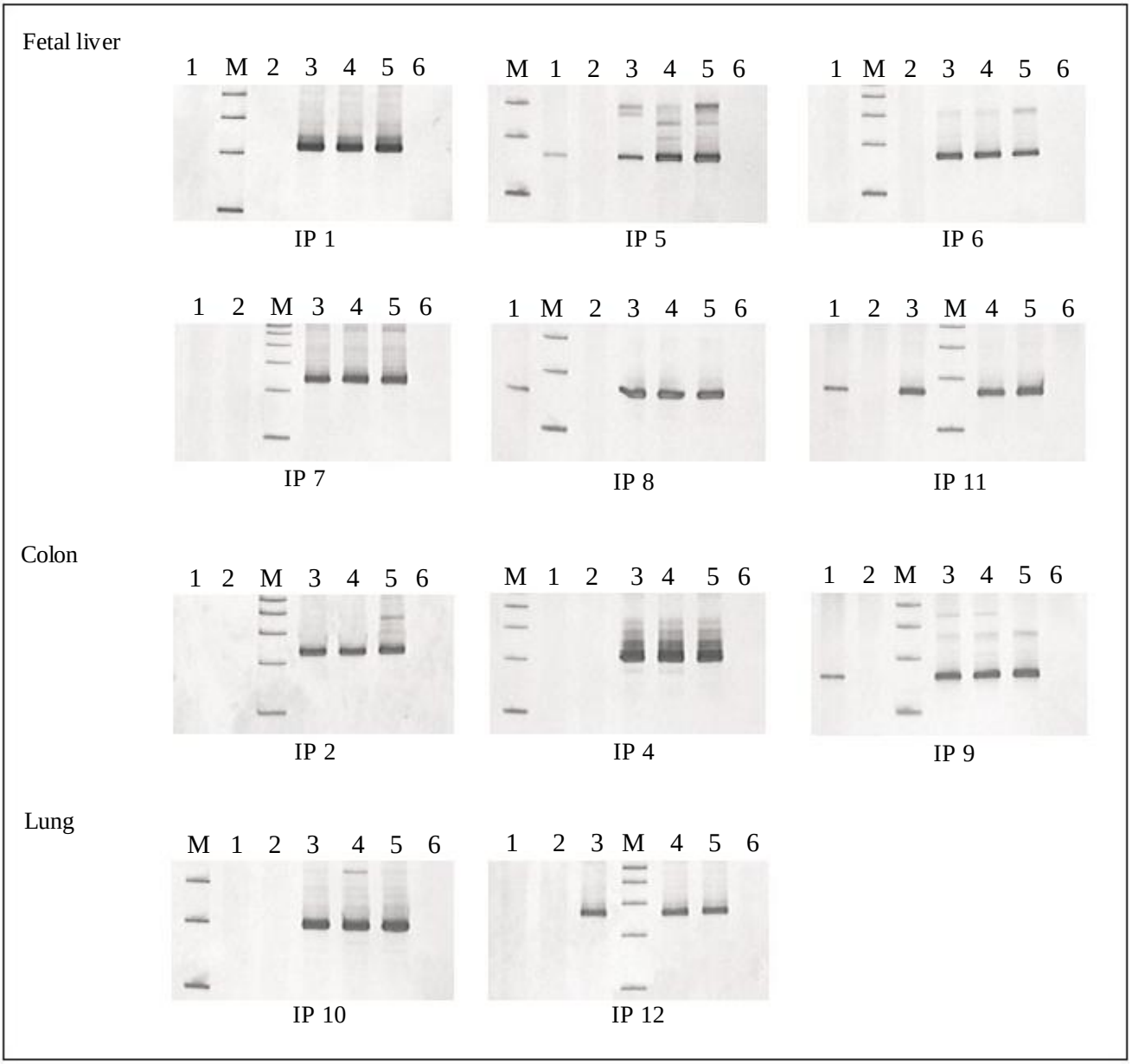


Figure 2
RT-PCR analysis for the internal priming (IP) candidates in fetal liver, colon and lung total RNA. RT-PCR was conducted in DNA-free RNA previously treated with DNAse (lanes 1 and 2) and in untreated RNA, which was, therefore, contaminated with genomic DNA (gDNA; lanes 3 and 4) for each candidate in the corresponding tissue. As a control, RT-PCR was conducted in the presence (lanes 1 and 3) and absence (lanes 2 and 4) of reverse transcriptase. gDNA was used as a positive control of the PCR reaction (lane 5) and no template as a negative control (lane 6). For fetal liver, in 3 IP candidates (5, 8 and 11) the PCR products (152 bp, 153 bp and 160 bp, respectively) were observed in the treated RNA when RT was added (lane 1) or in untreated RNA independent of the RT (lanes 3 and 4). For colon, in 1 IP candidate (9) the PCR product (158 bp) was observed in the treated RNA when RT was added (lane 1) or in untreated RNA independent of the RT (lanes 3 and 4). For the remaining IP candidates (1, 2, 4, 6, 7, 10 and 12), the PCR products (214 bp, 229 bp, 207 bp, 156 bp, 227 bp, 205 bp and 234 bp, respectively) were observed only in untreated RNA independent of the RT (lanes 3 and 4). The PCR products were analyzed on 8% polyacrylamide gels with silver staining. A 100 bp ladder (M) was used as molecular weight marker. In each gel the lower fragment in lane M correspond to 100 bp.

to stretches of As in the human genome. This motivated us to search for similar cases in the set of cDNA-based S-AS pairs identified in this study. We found that 18% and 26% of all S-AS pairs have at least one gene with its 3' end aligning with a stretch of A's in the human and mouse genomes, respectively. This number is certainly inflated by ESTs since it decreases to 11.7% for human and 12.6% for mouse when only mRNA/mRNA S-AS pairs are considered. Two possibilities could

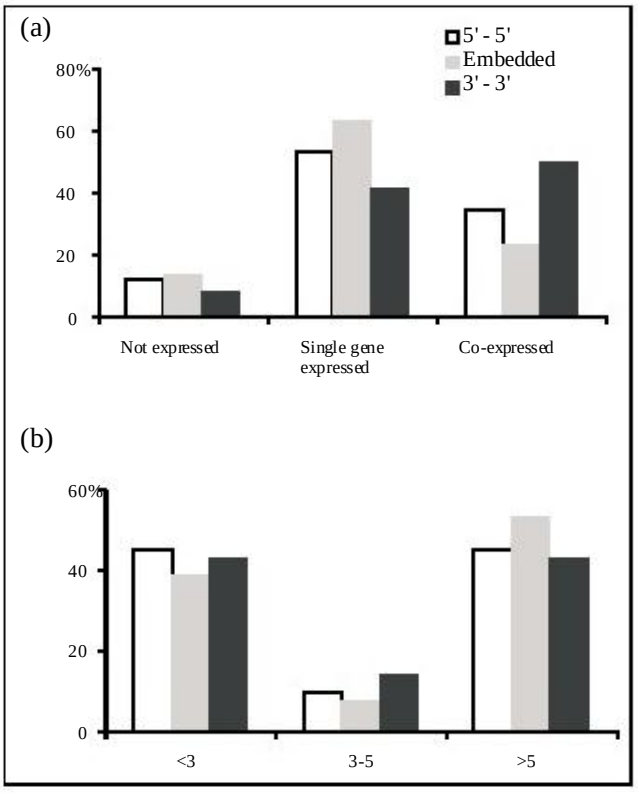


Figure 3
Expression pattern (in a set of 31 tissues covered by MPSS) of genes belonging to all three types of S-AS pairs (3'3', 5'5'and embedded). (a) Categories are as follows: 'no expression', for S-AS pairs whose expression was not detected (see Materials and methods for details); 'single-gene expression', for S-AS pairs in which expression is observed for only one gene in the pair; 'co-expression', for pairs in which expression is seen for both genes in the pair. (b) Rate of differential expression for the set of co-expressed S-AS pairs. Ratio of sense/antisense genes in the pair is shown on the x-axis.

account for this observation. First, a fraction of all antisense transcripts would be artifacts due to genomic priming with contaminant genomic DNA during cDNA library construction. An alternative is the possibility that antisense genes were constructed during evolution by retroposition events. Both possibilities are in agreement with the observation that antisense genes are depleted of introns.

An experimental strategy was developed to evaluate the likelihood of genomic priming as a factor generating artifactual antisense cDNAs. A total of 11 mRNA candidates derived from cDNA libraries from fetal liver, colon and lung with a high proportion of sequences that had their 3' ends aligning to stretches of As in the human genome were selected for experimental validation by RT-PCR. cDNA samples used in these experiments were reverse transcribed from fetal liver, colon and lung total RNA treated or not with DNAse. As can be seen in Figure 2, specific amplifications could not be achieved for 7 (63.6%) out of the 11 selected candidates when cDNA samples used as templates for PCR amplification were prepared from DNA-free RNA. On the other hand, when untreated RNA was used for cDNA synthesis, all candidates could be amplified, suggesting that a significant proportion of these internal priming sequences were indeed generated from contaminant genomic DNA.

Some other features support the artifactual origin of these antisense transcripts. First, cDNAs containing a stretch of As at their 3' genomic end have much less polyadenylation signals than genes in general (17% compared to 85%). Furthermore, these genes have a much narrower and rarer expression pattern when analyzed by SAGE and MPSS than genes in general (data not shown). These observations suggest that a significant fraction of all antisense genes are actually artifacts, due to genomic priming during library construction.

Retroposition generates intronless copies of existing genes through reverse transcription of mature mRNAs followed by integration of the resulting cDNA into the genome (for a review, see Long et al. [41]). Eventually, the cDNA copy can be involved in homologous recombination with the original source gene as has been suggested for yeast [42]. Retroposition was thought to generate non-functional copies of functional genes. However, several groups have shown that retroposition has generated a significant amount of new functional genes in several species [43-45]. Recently, Marques et al. [43] found almost 4,000 retrocopies of functional genes in the human genome. More recently, the same group reported that more than 1,000 of these retrocopies are transcribed, of which at least 120 have evolved as bona fide genes [46].

Retrocopies usually have a poly-A tail at their 3' end because of the insertion of this post-transcriptional modification together with the remaining cDNA. Thus, retroposition can explain the high incidence of antisense transcripts with a poly-A tail at their 3' end. To evaluate the contribution of retrocopies to the formation of S-AS pairs we compared the loci identified by Marques et al. [43] as retrocopies with the list of S-AS pairs identified in this study. Out of 413 retrocopies represented in the cDNA databases, 138 were involved in S-AS pairs (70 mRNA/mRNA and 68 mRNA/EST pairs). For the 70 mRNA/mRNA pairs, 78% were classified as embedded. This is in agreement with our previous observation that embedded pairs are enriched with intronless genes. Thus, retroposition seems to significantly contribute to the origin of embedded S-AS pairs.

Expression patterns within S-AS pairs
A critical issue to effectively evaluate the role of antisense transcripts in regulating distinct cellular phenomena is related to the expression pattern of both sense and antisense transcripts belonging to the same S-AS pair. Several reports have been published based on large-scale gene-expression analyses [5,19,23,47,48]. Similar to Wang et al. [48], we here used MPSS libraries available for human to explore this issue.

Table 5

Frequency of different types of alternative splicing in exon-intron borders with or without an antisense transcript

	Total	Alternative borders	Intron retention	Exon skipping	Alternative 3'/5' site
Borders with antisense					
Terminal donor	2,578	553	130	7	416
Internal donor	7,632	3,100	535	1,616	949
Terminal acceptor	7,749	3,145	493	1,642	1,010
Internal acceptor	2,763	688	208	7	473
Borders without antisense					
Terminal donor	2,200	579	101	32	446
Internal donor	23,414	8,674	1,080	4,997	2,597
Terminal acceptor	23,447	8,787	1,022	5,007	2,758
Internal acceptor	1,732	545	154	16	375

Tag to gene assignment was performed as previously described [31,49]. To ensure the MPSS sequences were unambiguously matched to the assigned transcript, we removed tags mapped to more than one locus. Frequencies for all tags assigned to genes in an S-AS pair were collected from all MPSS libraries.

Figure 3 shows the expression pattern of S-AS pairs for all MPSS libraries for human. We divided the dataset into the following categories as before: 3'3', 5'5' or embedded. Several features are evident. The rate of co-expression in our dataset was 35.1% compared to 44.9% observed by Chen et al. [4]. The differences are probably due to experiment design in both reports (for example, differences in the dataset and in the way the rate was calculated). Second, the rate of co-expression is significantly higher for 3'3' pairs when compared to the frequency of the embedded pairs (50.3%, chi-square = 134, df = 1, $p = 5.4 \times 10^{-31}$). This supports a previous conclusion from Sun et al. [5] that 3'3' S-AS pairs are significantly more co-expressed than other pairs and, therefore, are more prone to be involved in antisense regulation. It is important to mention that 5'5' pairs are also enriched in co-expressed pairs when compared to embedded pairs (chi-square = 23.5, df = 1, $p = 1.2 \times 10^{-6}$). We observed no statistical difference among the three categories regarding differential expression of both genes in a pair.

Influence of antisense transcripts in the splicing of sense transcripts

It is quite clear nowadays that a significant fraction of all human genes undergo regulated alternative splicing, producing more than one mature mRNA from a gene (Galante et al. [27] and references therein). Although several regulatory elements in cis and trans have been identified (for a review see Pagani and Baralle [50]), it is reasonable to say that we are far from a complete understanding of how constitutive and alternative splicing are regulated. One possible regulatory mechanism involves antisense sequences. Since the late 1980s, it is known that antisense RNA can inhibit splicing of a pre-mRNA in vitro [15]. A few years later, Munroe and Lazar [51] observed that NATs could inhibit the splicing of a message derived from the other DNA strand, more specifically the ErbAα gene. More recently, Yan et al. [52] characterized a new human gene, called SAF, which is transcribed from the opposite strand of the FAS gene. Overexpression of SAF altered the splicing pattern of FAS in a regulated way, suggesting that SAF controls the splicing of FAS. With the growing amount of genomic loci presenting both sense and antisense transcripts, a general role for S-AS pairing in splicing regulation has been proposed [47]. However, no systematic large-scale analysis has been reported so far investigating this issue for mammals. We made use of the human dataset described in this report to tackle this problem.

We first tested whether the rate of alternative splicing in the sense gene would be affected by the existence of an antisense transcript. It is expected that the effect of S-AS pairing on splicing would be restricted to those exon-intron borders located in the region involved in pairing. We therefore restricted the analysis to those exon-intron borders spanning the region involved in an S-AS pairing. Our strategy was to compare the number of splicing variants for those borders against all other exon-intron borders (those without an antisense transcript) in the same genes. To make the analysis more informative we split the borders into four categories (terminal donor, internal donor, internal acceptor and terminal acceptor). For both internal donor and acceptor sites, the presence of an antisense transcript slightly increased the rate of alternative splicing (Table 5; 4% and 3% increases, respectively). For the terminal sites, the presence of a NAT had the opposite effect (5% and 6% decrease for donor and acceptor, respectively). Table 5 also shows that these differences are

predominantly due to intron retention. On the other hand, NATs located within the introns and exons (but not spanning the border) have no major effect on the splicing of the respective borders. The observed differences between borders with or without NATs is statistically significant (chi-square = 31.2, df = 1, $p = 2.3 \times 10^{-8}$ for donor sites; and chi-square = 23, df = 1, $p = 1.6 \times 10^{-6}$ for acceptor sites).

Recently, Wiemann et al. [53] reported a new variant of IL4L1 that contains the first two exons of an upstream gene, NUP62. This chimeric transcript was expressed in a tissue and cell-specific manner. The authors speculated that cell type specific alternative splicing was involved in the generation of this chimeric transcript. We speculate that NATs could be involved in the generation of this type of chimeric cDNA. The same antisense message pairing with both sense messages would form a double-stranded RNA that could induce the spliceosome to skip the paired region and join the two sense messages, a process very similar to the one proposed for trans-splicing in mammals [54]. Interestingly, we found five examples in our dataset of S-AS pairs in which the genomic organization of both sense and antisense genes suggest a process like this. Additional data file 9 illustrates one of these cases. It can be seen that two transcripts represented by cDNAs AK095876 and AK000438 join messages from genes SERF2 and HYPK. The antisense transcript is represented by cDNA AK097682. Additional data file 10 lists all other putative cases of chimeric transcripts. The fact that both sense genes share a common antisense transcript raises the possibility that antisense transcripts can mediate trans-splicing of the sense genes, thereby generating the chimeric transcript.

On the evolution of S-AS pairs: functional implications

It is reasonable to assume that a fraction of all S-AS pairs reached this genome organization solely by chance. However, evidence presented here and elsewhere suggest that this fraction is probably small [6,55,56]. For example, Dahary et al. [6] concluded that antisense transcription had a significant effect on vertebrate genome evolution since the genomic organization of S-AS pairs is much more conserved than the organization of genes in general. However, how did this organization come to be? In principle, S-AS genomic organization should carry a negative effect on the overall fitness of a subject. For each gene in an S-AS pair, its evolution is constrained not only by features of its own sequence but also by functional features encoded by the other gene in the pair. The fact that we observed a significant amount of S-AS pairs in mammalian genomes suggests that there are advantages inherent to this organization to counter-balance the negative effects. The proposed role of NATs in gene regulation is certainly advantageous. We propose here two evolutionary scenarios, not mutually exclusive, that would speed up the generation of S-AS pairs. In one scenario, alternative polyadenylation has a fundamental role. Sun et al. [5] observed a preferential targeting of 3' UTRs for NATs. Our observation that 51% of 3'3' S-AS pairs overlap because of polyadenylation variants suggests that selection has favored cases where overlapping occurs only in a time and spatially regulated manner.

In a second scenario, retroposition generates NATs, which lack introns and may even show a polyadenylation tail integrated into the genome. We observe here that retroposition contributed significantly to the origin of S-AS pairs, especially those classified as embedded. What would be the selective advantages of retrocopies as NATs? Chen et al. [56] observed that antisense genes have shorter introns when compared to genes in general. They speculated that this feature was advantageous during evolution since NATs need to be "rapid responsers" to execute their regulatory activities. Although transcription is a slow process in eukaryotes, another bottleneck in the expression of a gene is splicing. Furthermore, Nott et al. [57] observed that the presence of introns in a gene affects gene expression by enhancing mRNA accumulation. Thus, the argument from Chen et al. [56] gets stronger with the data reported here and by Nott et al. [57] since intronless antisense genes would be transcribed even faster; their transcripts would simply skip splicing and the half-life of the respective messages would be shorter. All key features for genes involved in regulatory activities.

An important issue is the conservation of S-AS pairs between human and mouse. Although we found more than a thousand conserved pairs, this number is still small compared to the whole set of S-AS pairs in both species. Several factors, however, suggest that the number reported here is an underestimate. First, as discussed by Engstrom et al. [25], sequence conservation might not be of primary importance for antisense regulation. Furthermore, it is likely that many truly conserved pairs were not detected because transcript sequences have not been discovered yet. This is more critical in the face of our findings that a significant proportion of 3'3' S-AS pairs depend on alternative polyadenylation for an overlap. It is also quite likely that some S-AS pairs are lineage-specific. For instance, our finding that retroposition contributes to the origin of many S-AS pairs could explain the appearance of lineage-specific S-AS pairs, assuming that the retroposition event occurred after the divergence between human and mouse.

These two evolutionary scenarios (alternative polyadenylation and retroposition) might produce S-AS pairs with different functional implications. The expression and evolutionary conservation analyses presented here, together with evidence from others [5,19,23,47,48] suggest that 3'3' overlap achieved by polyadenylation variants was used throughout evolution to regulate gene expression. Those pairs generated through retroposition may be involved in some other types of regulation, such as alternative splicing.

Conclusion

This is the deepest survey so far of S-AS pairs in the human and mouse genomes. We made use of all cDNAs available in the public domain together with 122 MPSS libraries for human and mouse. The major findings of the present report include: as many as 10,077 and 8,091 S-AS pairs were identified for human and mouse respectively; using MPSS data, we found 4,308 and 216 new putative S-AS loci in human and mouse, respectively; a small fraction of all S-AS pairs are artifacts caused by genomic priming during cDNA library construction; a significant amount of S-AS pairs is due to retroposition events of one of the genes in the pair; quantitative analyses suggest that the presence of an antisense gene,complementary to an exon-intron border of the sense gene, increases the rate of retention of the respective intron. Furthermore, we propose an evolutionary model in which alternative polyadenylation and retroposition are important forces in the generation of S-AS pairs.

Taken together, these results offer, up to now, the vastest catalog of S-AS pairs in human and mouse genomes.

Materials and methods

Mapping cDNAs and MPSS tags onto the human and mouse genomes

We used a modified protocol similar to the one described previously to identify transcription clusters in the human and mouse genomes [27,28]. Briefly, genome sequence (NCBI build no. 35 for human and NCBI build no. 33 for mouse), EST collections (5,992,459 sequences for human and 4,246,824 sequences for mouse) and mRNA sequences (186,358 for human and 120,058 for mouse) were downloaded from UCSC [58]. All cDNAs were mapped to the respective genome sequence using BLAT (default parameters) [59]. The best hit for each cDNA in the genome was identified, followed by a pairwise alignment using Sim4 [60]. Only transcripts presenting identity ≥94%, coverage ≥50% and all splice sites in the same orientations were used.

Correct orientation of ESTs was determined by the presence of a poly-A tail (a stretch of 8 As at the 3' end) and/or a splicing donor (GT) and acceptor (AG) sites. All mRNAs were considered in the 'sense' orientation (oriented from 5' end to 3' end). All cDNAs mapped and reliably orientated were assembled into clusters. One cluster contains cDNAs presenting the same orientation and sharing at least one exon-intron boundary or a minimum of 30 nucleotides of overlap (only for those sequences without a common exon/intron organization).

For the mapping of MPSS data, we first extracted 'virtual' tags for both human and mouse genomes by simply finding all DpnII sites and extracting a 13 (human) or 16 (mouse) nucleotide long sequence immediately downstream of the restriction site in both orientations. These 'virtual' tags present only once in the respective genomes were further used and matched against the 'real' tags found in 41 and 81 MPSS libraries for human and mouse, respectively. Only MPSS tags classified as 'reliable' (present in more than one sequencing run) and 'significant' (tags per million >3) were considered as trusted signatures.

Identification of S-AS pairs

S-AS pairs were identified as those cases in which two clusters, in opposite orientations, overlap at the genome level. For the correct orientation of all mapped cDNAs, we took into consideration several parameters, including: sequence annotation as available in the respective GenBank entry; splice junctions; and poly-A tails and poly-T heads. We excluded from our analyses all cDNAs that presented conflicting orientations as defined by the three criteria above. If only two clusters overlap in the opposite orientation, they were classified as a single bidirectional S-AS pair. If a given cluster overlaps with more than one antisense cluster, they were classified as multiple bidirectional S-AS pairs. S-AS pairs were also classified according to their genomic pattern. Parameters evaluated included: pattern of S-AS overlap (exonic, intronic and exonic/intronic); spanning of introns by the components of a pair as defined by their alignment onto the genome; and chromosome localization and relative orientation within the S-AS pairs (tail-tail, head-head and embedded).

Conservation between human and mouse S-AS pairs

We used three strategies to evaluate the degree of conservation between human and mouse S-AS pairs. First, all pairs were searched against the dataset from HomoloGene [35] and those pairs conserved in both species were counted. In our second strategy, we selected those S-AS pairs in which at least one gene was conserved according to HomoloGene. We then used Needle, an alignment algorithm [61], to test sequence conservation between the respective antisense genes. We classified as conserved those global alignments with identity >30%. Finally, we also used the strategy from Engstrom et al. [25]. We used the net alignment between human and mouse genomes (retrieved from the UCSC Genome Browser database) to define the corresponding (synthenic) regions. We considered a human S-AS pair to be conserved in mouse if it had an exon region aligning (>20 bp) to an exon region from a mouse pair.

Investigation of the expression pattern of S-AS transcripts

We evaluated the expression pattern of S-AS pairs at the whole genome level based on their expression profiles obtained from MPSS libraries (available at [36]). The procedure was previously described by us for SAGE and MPSS [27,31,49]. The tag to gene assignment was done by scanning and extracting virtual tags (13 nucleotide-long sequences present downstream to the 3'-most DpnII restriction sites of each mRNA sequence). To accurately represent the 3' end of a transcript, only mRNA sequences containing a poly-A tail were used. All tags mapped to two or more different genes

were excluded and the frequencies of different tags for the same gene (mainly alternative polyadenylation variants) were summed. MPSS tags were normalized to counts-per-million and the expression data were cross-linked to genomic positions by the extraction of virtual tags for both the human and mouse genomes. Only tags showing 100% identity with a genomic locus were used in the analyses.

The classification of the expression pattern of S-AS pairs was done using those tags with ≥3 tags per million across all MPSS libraries. To evaluate the co-expression of all S-AS pairs, both genes in a pair had to be co-expressed in at least 04 libraries. If both genes in a pair were co-expressed in less than four libraries or they were independently expressed in different libraries, the pair was classified as 'single-gene expression'. The remaining S-AS pairs were classified as 'no-expression'.

Identification of antisense MPSS tags

All DpnII sites in the human and mouse genomes were identified and for each site two 'virtual' MPSS tags were extracted from both DNA strands in the correct orientation. All 'virtual' MPSS tags mapped in the opposite strand of known mRNAs in both genomes were identified. Those mRNAs belonging to an S-AS pair previously identified were excluded. Those antisense MPSS tags mapped just once in the respective genome and present in at least one MPSS library were identified and submitted to experimental validation.

Simulations on the genomic organization of S-AS pairs

A random distribution of S-AS pairs was obtained by re-indexing the coordinates of one gene in all the pairs 1,000 times. This was done by randomly selecting a genomic coordinate for the start of mapping of a given gene. All the remaining exon-intron borders were then re-indexed based on this initial coordinate. The relative organization of both genes in all random S-AS pairs was stored and frequencies for each category were calculated. Those frequencies were used as the expectation for chi-square tests of the null hypothesis.

Identification of splicing variants

Using the database mentioned earlier and described elsewhere [26-28] we identified all exon-intron borders complementary to a NAT. We then compared the rate of alternative splicing in these borders against the borders from the same genes without a NAT. We established a set of stringent criteria to identify alternative borders. These criteria are detailed elsewhere [26-28].

Experimental validation of MPSS antisense tags

MPSS tags corresponding to antisense transcripts were converted into their corresponding 3' cDNA fragments using GLGI-MPSS [37]. Antisense tags were selected from a MPSS library derived from the normal breast luminal epithelial cell line HB4a and the same RNA source was used for GLGI amplification. For the GLGI-MPSS amplification, we used a sense primer including 17 bases of the MPSS tag sequence and 6 additional bases (CAGGGA), giving a total of 23 bases for each primer (5'-CAGGGAGATCXXXXXXXXXXXXX-3'). We also used an antisense primer (ACTATCTAGAGCG-GCCGCTT) present in the 3' end of all cDNA molecules that was incorporated from reverse transcription primers in cDNA synthesis. The reaction mixture was prepared in a final volume of 30 μl, including 1× PCR buffer, 2.0 mM $MgCl_2$, 83 μM dNTPs, 2.3 ng/μl antisense primer, 2.3 ng/μl sense primer, 1.5 U of Taq Platinum DNA polymerase (Invitrogen, San Diego, CA, USA) and 0.5-0.8 μl of the same cDNA source used for MPSS library construction. PCR conditions used for amplification were 94°C for 2 minutes, followed by 30 cycles at 94°C for 30 s, 64°C for 30 s, and 72°C for 35 s. Reactions were kept at 72°C for 5 minutes after the last cycle. The amplified products were ethanol precipitated and cloned into the pGEM®-T Easy vector (Promega, Madison, WI, USA). Twelve colonies for each GLGI-MPSS fragment were screened by PCR using pGEM universal primers and positive colonies were sequenced using Big-Dye Terminator (Applied Biosystems, Foster City, CA, USA) and an ABI3100 sequencer (Applied Biosystems).

Experimental validation of genomic primed sequences

Total RNA derived from fetal liver, colon and lung was purchased from Clontech laboratories (Palo Alto, CA, USA). For cDNA synthesis, 2 μg of total RNA were treated (or not) with 100 units of DNAse I (FPLC-pure, Amersham, Piscataway, NJ, USA) and were reverse transcribed using oligo(dT)12-18, random primers and SuperScript II (Invitrogen), following the manufacturers' instructions. After synthesis, the resulting cDNA was subjected to RNase H treatment. The absence of genomic DNA contamination was evaluated for each preparation. DNA-free total RNA was subjected to PCR amplification using primers within intronic sequences flanking exon 12 of the hMLH-1 gene (forward, 5' TGGTGTCTCTAGTTCTGG3'; reverse 5' CATTGTTGTAGTAGCTCTGC 3'). All PCR amplifications were carried out using 2 μl of cDNA as a template to the final volume of 25 μl and 1× buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTP, 0.2 μM of each specific primer and 0.025 U/μl of Taq DNA polymerase (Life Technologies, San Diego, CA, USA). The following cycling protocol was used: initial denaturation of 94°C for 4 minutes; 94°C for 30 s; 55°C for 45 s; 72°C for 1 minute for 35 cycles; along with a final extension at 72°C for 7 minutes. All PCR products were resolved on 8% polyacrylamide gels and sequenced as described above to verify amplification specificity.

Strand-specific RT-PCR

In the strand-specific RT-PCR, orientation of the transcript is accessed by restricting which gene-specific primer is present during first-strand cDNA synthesis. For each candidate, 1 μg of total RNA was treated with Promega RQ1 RNAse-free DNAse and tested for remaining DNA contamination as described above. First-strand cDNA synthesis was carried out at 50°C for 2 h using 200 U of SuperScript II (Invitrogen) and 0.9 μM of a primer complementary to the antisense tran-

script. PCR amplifications were performed using 1 μl of the first-strand cDNA as a template in a final volume of 25 μl and 1× buffer, 1.5 mM MgCl2, 0.1 mM dNTP, 0.4 μM of gene spe- cific primers and 1 U of Platinum Taq DNA polymerase (Inv- itrogen). The following cycling conditions were used for amplification: initial denaturation of 95°C for 2 minutes; 94°C for 40 s; reaction-specific annealing temperature for 40 s and 72°C for 1 minute for 35 cycles; followed by a final extension step at 72°C for 7 minutes. All PCR products were resolved on 8% polyacrylamide gels. Controls for the absence of self-priming during cDNA synthesis were done with reverse transcriptase in the absence of primers, and controls for the absence of DNA were done by incubation with primers but with no reverse transcriptase.

Availability

To make our dataset fully accessible to the community we have set up a worldwide web portal [62] containing all raw data generated in this study and a series of tools to explore the data.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a list of representative GenBank entries for all S-AS pairs in both human and mouse. Additional data file 2 is a table showing the total number of S-AS pairs by chromosome for both human and mouse. Additional data file 3 shows the number of clusters and S-AS pairs when a less stringent clustering methodology is applied. Additional data file 4 shows a schematic view of all possible genomic organizations of S-AS pairs. Additional data file 5 lists all S-AS pairs conserved between human and mouse using the three strategies described in the text. Additional data file 6 shows the fraction of S-AS pairs conserved between human and mouse that are classified as 'Fully intronic' and the fraction of conserved S-AS pairs that contain at least one intronless gene. Additional data file 7 is a list of all MPSS libraries used in this study. Additional data file 8 presents the number of cDNA-based pairs that were further confirmed by the MPSS data. Additional data file 9 is a figure illustrating chimeric transcripts joining two adjacent genes (SERF2 and HYPK) with a NAT located between them. Additional file 10 lists all cases of chimeric transcripts identified in our dataset.

Acknowledgements

We would like to thank Artur Ramos de Souza for the design and maintenance of the web portal. We also thank Henrik Kaessmann for making available the data on human retrocopies. We are also indebt to Andrew Simpson for a critical review of the manuscript and to three anonymous reviewers for critical and constructive comments/suggestions.

References

1. Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G: In search of antisense. Trends Biochem Sci 2004, 29:88-94.
2. Kumar M, Carmichael GG: Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. Microbiol Mol Biol Rev 1998, 62:1415-1434.
3. Vanhee-Brossollet C, Vaquero C: Do natural antisense transcripts make sense in eukaryotes? Gene 1998, 211:1-9.
4. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 2005, 21:326-329.
5. Sun M, Hurst LD, Carmichael GG, Chen J: Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. Nucleic Acids Res 2005, 33:5533-5543.
6. Dahary D, Elroy-Stein O, Sorek R: Naturally occurring antisense: transcriptional leakage or real overlap? Genome Res 2005, 15:364-368.
7. Zhang Y, Liu XS, Liu QR, Wei L: Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. Nucleic Acids Res 2006, 34:3465-3475.
8. Katayama S, Tomaru Y, Kasukawa T, Kaki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al.: Antisense transcript in the mammalian transcriptome.Science 2005, 309:1564-1566.
9. Li AW, Murphy PR: Expression of alternatively spliced FGF-2 antisense RNA transcript in the central nervous system: regulation of FGF-2 mRNA translation. Mol Cell Endocrinol 2000, 162:69-78.
10. Hastings ML, Ingle HA, Lasar MA, Munroe SH: Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally-occurring antisense RNA. J Biol Chem 2000, 275:11507-11513.
11. Brantl S: Antisense-RNA regulation and RNA interference. Biochim Biophys Acta 2002, 1575:15-25.
12. Rougeulle C, Heard E: Antisense RNA in imprinting: spreading silence through Air. Trends Genet 2002, 18:434-437.
13. Prescott EM, Proudfoot NJ: Transcriptional collision between convergent genes in budding yeast. Proc Natl Acad Sci USA 2002, 99:8796-8801.
14. Ogawa Y, Lee JT: Antisense regulation in X inactivation and autosomal imprinting. Cytogenet Genome Res 2002, 99:59-65.
15. Munroe SH: Antisense RNA inhibits splicing of pre-mRNA in vitro. EMBO J 1988, 7:2523-2532.
16. Peters NT, Rohrbach JA, Zalewski BA, Byrkett CM, Vaughn JC: RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. RNA 2003, 9:698-710.
17. Lehner B, Willians G, Campbell RC, Sanderson CM: Antisense transcripts in the human genome. Trends Genet 2002, 18:63-65.
18. Kiyosawa H, Yamanaka I, Osato N, Kondo S, RIKEN GER Group, GSL Members: Antisense transcripts with FANTON2 clone set and their implications for gene regulation. Genome Res 2003, 13:1324-1334.
19. Yelin R, Dahary D, Rorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, et al.: Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 2003, 21:379-386.
20. Fahey ME, Moore TF, Higgins DG: Overlapping antisense transcription in the human genome. Comp Funct Genomics 2002, 3:244-253.
21. Shendure J, Church GM: Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol 2002, 3:R44.
22. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD: Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 2004, 32:4812-4820.
23. Quere R, Manchon L, Lejeune M, Clement O, Pierrat F, Bonafoux B, Commes T, Piquemal D, Marti J: Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. Nucleic Acids Res 2004, 32:e163.
24. Wahl MB, Heinzmann U, Imai K: LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. Bioinformatics 2004, 21:1389-1392.
25. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, et al.: Complex loci in human and mouse genomes. PLoS Genetics 2006, 2:e47-.
26. Sakabe NJ, de Souza JE, Galante PAF, de Oliveira PS, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, et al.: ORESTES are enriched in rare exon usage variants affecting the encoded proteins. C R Biol 2003, 326:979-985.

27. Galante PAF, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ: Detection and evaluation of intron retention in the human transcriptome. RNA 2004, 10:757-765.
28. Kirschbaum-Slager N, Parmiggiani RB, Camargo AA, de Souza SJ: Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data. Physiol Genomics 2005, 21:423-432.
29. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. Science 1995, 270:484-487.
30. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al.: Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000, 18:630-634.
31. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, et al.: An anatomy of normal and malignant gene expression. Proc Natl Acad Sci USA 2002, 99:11287-11292.
32. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al.: Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 2005, 308:1149-1154.
33. Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, Egidio CM, Paquola AC, Machado AA, Festa F, et al.: Anti-sense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene 2004, 23:6684-6692.
34. Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K: Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 2005, 15:463-474.
35. HomoloGene [http://www.ncbi.nlm.nih.gov/HomoloGene/]
36. LICR MPSS Repository [http://mpss.licr.org/]
37. NCBI:MouseTranscriptomeProject [http:www.ncbi.nlm.nih.gov/genome/guide/mouse/MouseTranscriptome.html]
38. Silva AP, Chen J, Carraro DM, Wang SM, Camargo AA: Generation of longer 3' cDNA fragments from massive parallel signature sequencing tags. Nucleic Acids Res 2004, 32:e94.
39. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 2005, 15:987-997.
40. Iseli C, Stevenson BJ, de Souza SJ, Samaia HB, Camargo AA, Buetow KH, Strausberg RL, Simpson AJ, Bucher P, Jongeneel CV: Long-range heterogeneity at the 3' ends of human mRNAs. Genome Res 2002, 12:1068-1074.
41. Long M, Betran E, Thornton K, Wang W: The origin of new genes: glimpses from the young and old. Nat Rev Genet 2003, 4:865-875.
42. Fink GR: Pseudogenes in yeast? Cell 1987, 49:5-6.
43. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 2005, 3:e357.
44. Burki F, Kaessmann H: Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet 2004, 36:1061-1063.
45. Emerson JJ, Kaessmann H, Betran E, Long M: Extensive gene traffic on the mammalian X chromosome. Science 2004, 303:537-540.
46. Vinckenbosch N, Dupanloup I, Kaessmann H: Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 2006, 103:3220-3225.
47. Jen C-H, Michalopoulos I, Westhead DR, Meyer P: Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. Genome Biol 2005, 6:R51.
48. Wang X-J, Gaasterland T, Chua N-H: Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. Genome Biol 2005, 6:R30.
49. Silva AP, de Souza JE, Galante PA, Riggins GJ, de Souza SJ, Camargo AA: The impact of SNPs on the interpretation of SAGE and MPSS experiments. Nucleic Acids Res 2004, 32:6104-6110.
50. Pagani F, Baralle FE: Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 2004, 5:389-396.
51. Munroe SH, Lazar MA: Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. J Biol Chem 1991, 266:22083-22086.
52. Yan M-D, Hong C-C, Lai G-M, Cheng A-L, Lin Y-W, Chuang SE: Identification and characterization of a novel gene SAF transcribed from the opposite strand of FAS. Hum Mol Gen 2005, 14:1465-1474.
53. Wiemann S, Kolb-Kokocinski A, Poustka A: Alternative pre-mRNA processing regulates cell-type specific expression of the IL4I1 and NUP62 genes. BMC Biol 2005, 3:16.
54. Takahara T, Kanazu S, Yanagisawa S, Akanuma H: Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic trans-splicing. J Biol Chem 2000, 275:38067-38072.
55. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 2005, 21:326-329.
56. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: Human antisense genes have short introns: evidence for selection for rapid transcription. Trends Genet 2005, 21:203-207.
57. Nott A, Meislin SH, Moore MJ: A quantitative analysis of intron effects on mammalian gene expression. RNA 2003, 9:607-617.
58. UCSC Genome Browser: Download Page [http://hgdownload.cse.ucsc.edu/]
59. Kent WJ: BLAT - the BLAST-like alignment tool. Genome Res 2002, 12:656-664.
60. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 1998, 8:967-974.
61. Rice P, Longden J, Bleasby A: EMBOSS: The European Molecular Biology open software suite. Trends Genet 2000, 16:276-277.
62. LICR Sense/Antisense Portal [http://www.compbio.ludwig.org.br/sense-antisense]

Supplemental Material

Sense-antisense pairs in mammals: functional and evolutionary consideratios.

1.3.3 Validação dos transcritos antisenso por RT-PCR fita específica

Para confirmar a existência dos 27 potenciais transcritos antisenso, todos (*tag* 03, 09, 13, 17, 19, 20, 24, 25, 28, 34, 40, 41, 43, 44, 49, 52, 53, 58, 64, 65, 70, 72, 77, 83, 87, 94 e 95) foram submetidos a RT-PCR fita específica (ver material e métodos – Anexo 1). A organização genômica dos fragmentos antisenso de GLGI-MPSS e o transcrito senso ao qual estão relacionados são demonstrados na Figura 6.

Para a análise por RT-PCR fita específica, quatro reações de síntese de cDNA fita específica foram realizadas para cada transcrito. Na síntese de cDNA foram utilizados como molde o RNA total das linhagens celulares *HB4a* ou *HB4a-C5.2*, para cada candidato de acordo com sua maior expressão segundo os dados de MPSS (Tabela 3).

Tabela 3 - *Tags* antisenso de MPSS.

Número *Tag*	Sequência da *tag* de MPSS	*Hb4a* (*tag*/milhão)	*Hb4a-C5.2* (*tag*/milhão)
3	GATCCACTCAACAAAAT	7.64	9.30
9	GATCAGAAAAAATCAGC	3.05	11.62
13	GATCTTCATGATGGAGG	8.4	3.87
17	GATCTGCTCATGAAATC	31.32	16.28
19	GATCAGGCAGGATAGGG	12.22	21.7
20	GATCAAAACGTGTCACA	6.87	3.87
24	GATCATCTGTAGAGGGA	16.81	10.07
25	GATCTGGATGAGCATAT	6.11	10.07
28	GATCAGATTTACCATTT	3.05	4.65
34	GATCACAACAACTGTCT	5.34	3.87
40	GATCAAATCAGTGTCGG	3.05	3.87
41	GATCAAGAGCAGAGGAG	15.28	19.38
43	GATCGTGCAGAAGGAGG	14.51	11.62
44	GATCTGAAATACAATTC	31.32	38.76
49	GATCTGTACAGTTAGTG	4.58	3.87
52	GATCTCTGCTACAGTAA	4.58	3.87
53	GATCCACATCACCGCCT	4.58	3.87
58	GATCCACGGCAAAACTA	5.34	4.65
64	GATCGGCTGACTATATT	9.93	9.3
65	GATCTGTGGGTTAGCTT	17.57	18.6
70	GATCATCTGGGGTGCGG	3.82	3.87
72	GATCTTCTCAGGGCGAT	3.05	3.1
77	GATCTGGTTGTCAGTTG	-	10.07
83	GATCTGTAAGATGTGAG	-	6.2
87	GATCGATGGTGTTACTC	-	6.2
94	GATCGTCAGGCGGCTTG	-	4.65
95	GATCGAGGGCCACAAAA	-	4.65

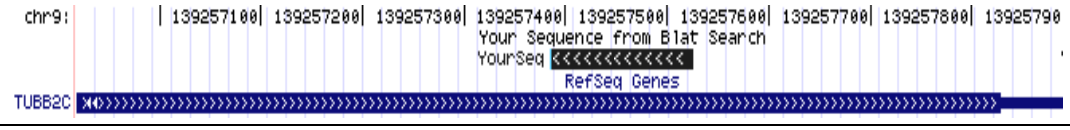
Sequências das *tags* antisenso de MPSS e freqüência que cada *tag* foi encontrada nas bibliotecas *Hb4a* e *Hb4a-C5.2*. **azul**: sítio reconhecido pela enzima de restrição *Dpn*II; **preto**: seqüência da *tag* antisenso de MPSS.

***Tag* 3**

A – fragmento de GLGI

GATCCACTCAACAAAATAGCTGCTGTTTTTGTTTTGGACATTAAGCATTTGCTCATCCACCTCCTTCATGGACATGCGGC
CCCTGAACACGGCGGCAACCGTCAGGTAGCGGCCATGGCGGGGGTCGCAGGCAGCCATCATAAAAAAAA

B – Organização genômica – BLAT

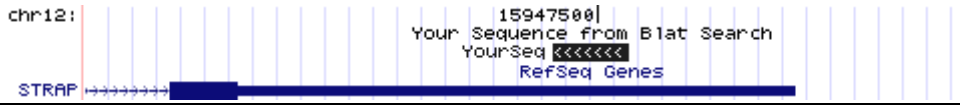


***Tag* 9**

A – fragmento de GLGI

GATCAGAAAAAATCAGCCAATATAGTTTTCTTCTCCATATTAACATAAACAGCTACAAGAACCCTAAAAAAAAA

B – Organização genômica – BLAT



***Tag* 13**

A – fragmento de GLGI

GATCTTCATGATGGAGGCGGCCACTGGGTCGGTGGAGAAGTGCTGGGAGGAGGGAAGCAGGGAAAGCCACTCACTGAGAG
CCCGGGCCGCACGGCCCTCCGCTGCCCAGGTAACAGGGGCAGGAAAAGGAAGGCAGAACAGCGGCTTGGAGCGGCCCCAG
CTGTGCGCTCATCACAAAAAAAAAAAAAAA

B - Organização genômica – BLAT

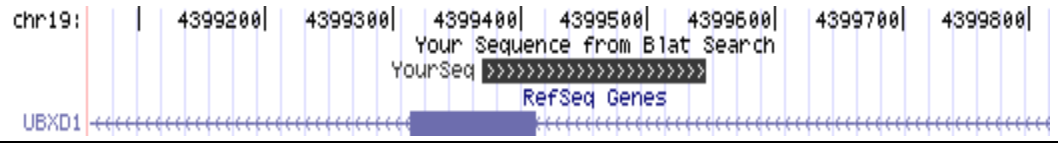


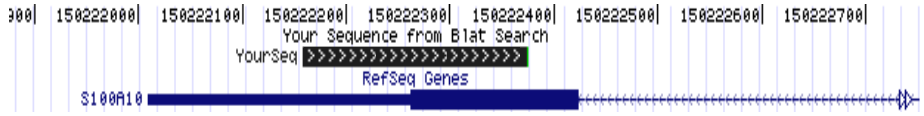
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 17**

A – fragmento de GLGI

GATCTGCTCATGAAATCCTTCTATGGGGGAAGCTGTGGGGCAGATTCCTTAAGCGACCCTTTGGGACAACTCTTATCAGG
GAGGAGCGAACTGCTCATTTCTGCCTACTTCTTTCCCTTCTGCTTCATGTGTACTACAAAATAGTCATTGCATGCAATGG
TGAGGCCCGCAATTAGGGAAAAGAAGCTCTGGAAGCCCACTTTGCCATCTCTACACcAAAAAAAAAAAAA

B – Organização genômica – BLAT



***Tag* 19**

A – fragmento de GLGI

GATCAGGCAGGATAGGGATGGGGCAGGACATGAGGCTGGGGGATGCAGAGGTTAGGTGGGAGAGGCTACCGGAGGAAGAA
TGAGGCTGGTAGGGGAGGGAGAAAGAGAGCAAAGAGAGAGAGGAGCAATTGGGGGCCAGCTGGAGAGCTCAGATGGAGCA
GGTCAGGAGGTGGAACAATGGCAGAGTGAGGGTGGAGGGCGCAGTGTCTGGAGAGGCGGAAATGAGAAGGCTGGAAAAAA
AAA

B - Organização genômica – BLAT

***Tag* 20**

A – fragmento de GLGI

GATCAAAACGTGTCACAGCTGGGAACCAAGAGGAAGTAAAAAAAATACCCACAAATATTCCCCGACTCCCCATTAGTCCT
CCCCCAACAAAAAAAAAAAA

B – Organização genômica – BLAT

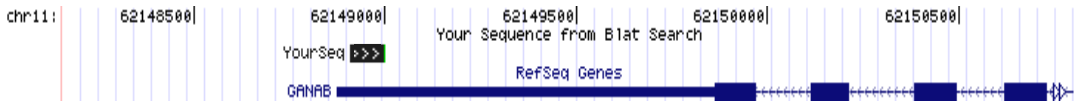


Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 24**

A – fragmento de GLGI

GATCATCTGTAGAGGGAACATCACCAGCGCGCCCAGCCGTTTCCACACCATCCTCCGCGCCGCGGAGCCGGGCAGGGGGC
GAAACTCCCCaGGAAAAAAAAAAAA

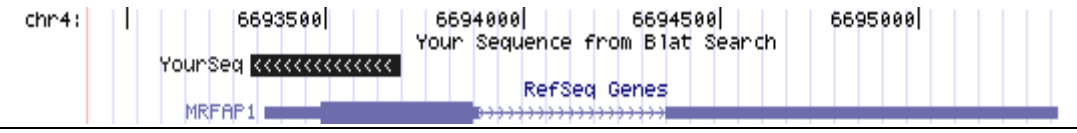
B – Organização genômica – BLAT

***Tag* 25**

A – fragmento de GLGI

GATCTGGATGAGCATATTGTCCATCTCCCACAGCTTGCTCCGGTTCCGCAGGTACGCCCGCCCGTGCTCGCGCGTCAGCG
ACGCGATGTCCTCGCGCATCTCGTTGATGACCGGGAGCAGAAACTGCTCGAAATCCTCCTCGGGCTCCAGCACCTCCACT
TCCTCCGGTTCCGCCAGCTCGACGATGTCCAGGGGCCGCATCTCTTCCCACTGCCTCGGAACCGCAATAGCGATGTCTGT
TGGAGAGAGAAAACCGACACTCGCTATGCTTAGCAATAGAGAGCCCGAATATTCCTGAAAACTTTTACCCTTTTTCAACT
TTTCTTCTCAGAGAAGCCACGGACAAGCGACTTACGCGCAGAATAACGGCGAACAAAAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 28**

A – fragmento de GLGI

GATCAGATTTACCATTTTTAGGGGGAAGAGGGAAAAAAAGGTGTATTTATCATCAGCTAGATGTGCTCACTGTATGCTCC
GTTATTTATATGCAAGGCCCGGGTGACTGGAAGTGCAGTTGTCAGGCATTTTAATAAACTGGACAGCCATTTGTTTCTGC
ACGACAAGGCATCTTTACACAGGAGCAATCGGGAGATAACAGGAAACAGCCAAGCACTCTGCACTGCAACACGCCACCTT
AACAGCTAACCAGCATTACTCAACTGCTACACAACTGCGCCTAGTGCACAAAAAAAAAAAAAA

B - Organização genômica – BLAT

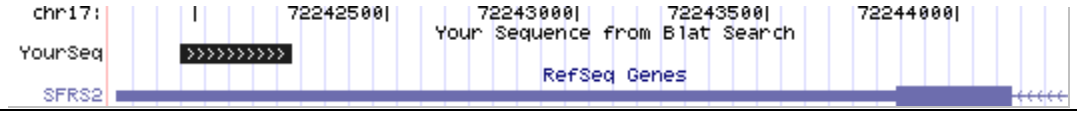


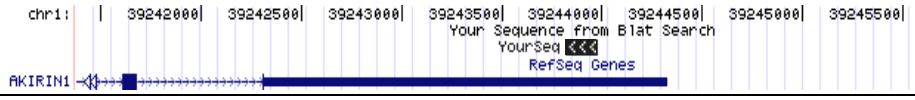
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 34**

A – fragmento de GLGI

GATCACAACAACTGTCTCTGAATATTACCTTACTTTGGAAGGTATTGGCATACATTTGAATTAAACTCAGAATTTTTCAA
AATAAGCCCTAATTAAAAAAACTATTCTATATTCAAAGTAAAACAATATTAATAGAACTTTCCTTTCCAGAAACAATGAA
GGAAACAAAAAAAAAAAA

B – Organização genômica – BLAT

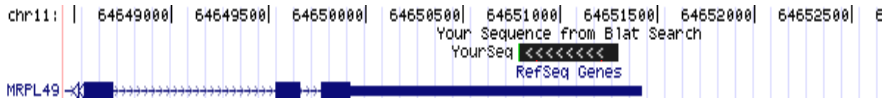


***Tag* 40**

A – fragmento de GLGI

GATCAAATCAGTGTCGGGCTCCTTGCTGTGGCTGTACACTTGAAAATGGAATGGAATCTCCAGGAGTGAGGGTTTCCATG
GAAGCCAGCACTATCTTTGGGGACACCACTGACCATTAACAGATTAGGAAGATGTGCCCATTTTACACACAGAGAAGCAG
CCCAGAGTTCAGTGCCCTGCTTCAGGGCAGAGACCAAGTTGGGGAGTCAGGAACTGAAAGCTGTCTTAGCCAGTTTGGGG
GCAGGGTGGGGTCTCAGGCCTTTGGCCCCTGTGACTCCAAGTATAAGGCATATATGAAGGCTAGTGAGCTGGGGTTGGGA
TGGGATGGCAGGGTGAAAGGTCAGCATATTCCTGCCTGCAAACCCCTCCCCAGCCTTGATTGCAGGTCAGAAGGCTGCAG
GGGGATGTCCCTATGCTGGCTGcGAGTTCCCAGCCACTCCCTTGTGGTCAGGCCAGCAAGTCTGGTCTGCAATGGAACAG
TGACCAAGAACGTGCCTCCTACCTCTCTATCAGAGCCCAAAAAAAAAAAA

B – Organização genômica – BLAT



***Tag* 41**

A – fragmento de GLGI

GATCAAGAGCAGAGGAGGAAGGGGAGAGGCAGGAAGGCAGTTGAAAAAAAAAA

B – Organização genômica – BLAT

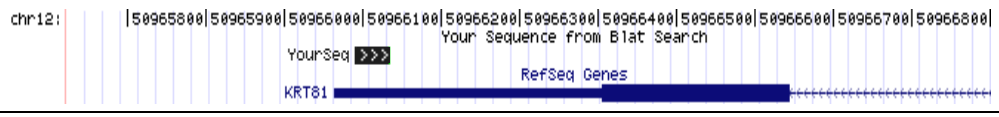


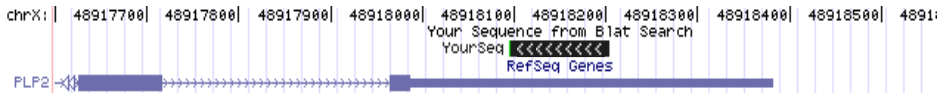
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 43**

A – fragmento de GLGI

GATCGTGCAGAAGGAGGACCTAGGCACACACATATGGTGGCCACACCCAGGAGGGTTGATTGGCAGGCTGGAAGACAAAA
GTCTCCCAATAAAGGCACTTTTACCTCAAAAAAAAAAAAA

B – Organização genômica – BLAT

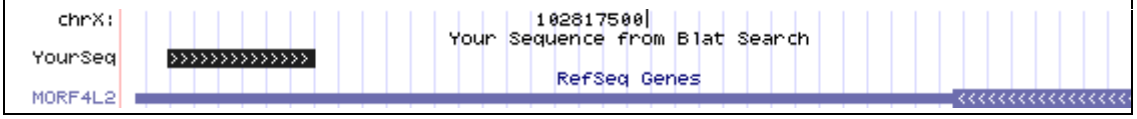


***Tag* 44**

A – fragmento de GLGI

GATCTGAAATACAATTCCTAAAATATCAACTTTTCCAGAAAACCGTGGCTACACAATAATGCATTGCCTCTATCATGTTA
GAACGTGCATTAGACTCAAATACAAAAACCATGAAACAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 49**

A – fragmento de GLGI

GATCTGTACAGTTAGTGGGAGCACACTATATAAATCCTTTGCATGACACCATTCAACAATATTTTTTAAAATCTACAAAA
TTTTAAAGTTTCACTTCCCTAGCAAAATATCTTCAGTCAAGAAATTAGTCTTTGAAAAAAAAAAAAAA

B - Organização genômica – BLAT

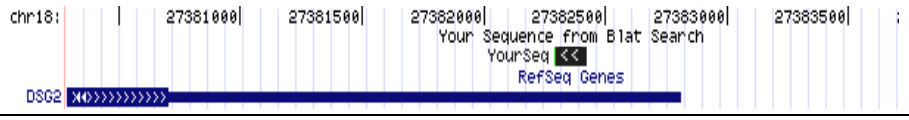


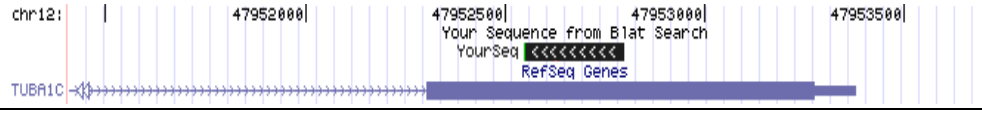
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 52**

A – fragmento de GLGI

GATCTCTGCTACAGTAAGCTGTTCGTGGTAGGCTTTCTCAGCAGAGATGACAGGGGCATATGTGGCCAGAGGGAAGTGGA
TGCGGGGGTAGGGCACCAGGTTGGTCTGGAATTCTGTCAGGTCAACATTCAGGGCTCCATCAAATCTCAGGGAAGCAGTG
ATGGAGGACACAATATGGCTAATAAGGCGGTTAAGGTTAGTGTAGGTTGGGCGCTCGATATCGAGGTTTCTACGACAGAT
GTCATAGATGGCCCAAAAAAAAAAAAAAAAA

B - Organização genômica – BLAT

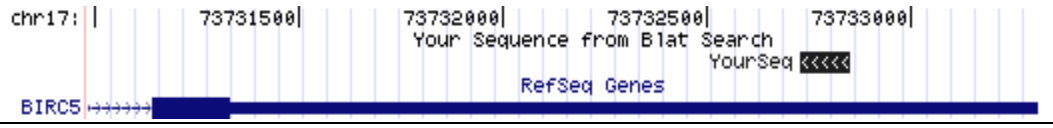


***Tag* 53**

A – fragmento de GLGI

GATCCACATCACCGCCTGGCATGCAAAGGAGTTGAAGACAAAACCATTTTTTTCCAGCTTCTCTACAAAGCCAATTACTA
AGCAACAGTTATTAAAGTGAAGTATCCATTTACAGACTGACAAAAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 58**

A – fragmento de GLGI

GATCCACGGCAAAACTAACTGGTACCATACAAAGATATGTCTATTATTGTCATGTTTAGATTTTCTATTCCACATTCATT
GACAAAAAAAA

B - Organização genômica – BLAT

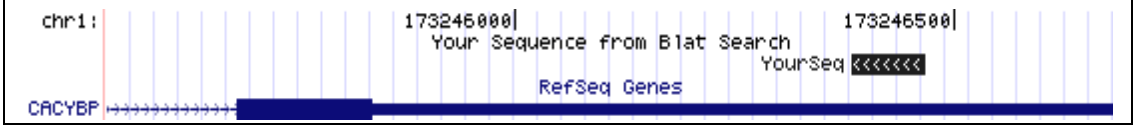


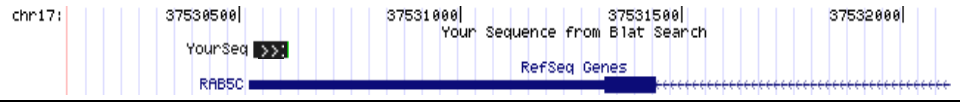
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 64**

A – fragmento de GLGI

GATCGGCTGACTATATTGACAAGATACTGATTGGTTACATGTTGAAGAAAACATACAATACAAAATACAGAAAAAGAAAA
AAAAAAA

B - Organização genômica – BLAT

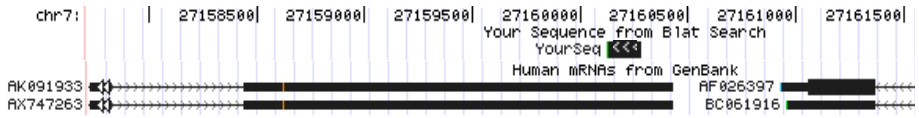


***Tag* 65**

A – fragmento de GLGI

GATCTGTGGGTTAGCTTCTGCTTAGCAGGACTGTGGAGATGCTTCCAGCTTCGCTGTCCTTTCCTCTGGCTCCTGTATCT
TACTGTTCAGCTGTGTTAAATATGTACGCCCTGATGTTTCCTATAATAGCAGATACTGTATATTTGAACAAGATTTTTTT
TTATCAAAAAAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 70**

A – fragmento de GLGI

GATCATCTGGGGTGCGGAGTACAAAGCTTTGCAAGGGTGTGGTTTTGGAATGACGCTAAACTGAAGGTGGAGAGAACAGA
TAAAAAGGTTGGAAGTTGCACACTGTACACTGTTAAGAAGTTGAGCTTTTATCTTAGAGGCAGCAGAAGGTTTGGAGCCA
AGGAATGAAATGATGAGGCGTCCTTCAGGTAATGAACTTCAGCTGCAGTGTGAAAGGGGCAGGAAGACTGGCAtCTGTCA
AAACTGGAACAGTCCAGTGAGTATGTGCAGGCCCGGGCTTGGGCAGTGACGAGGGCAGGGAGCACACATCAATTTCTGCc
GGAAA

B - Organização genômica – BLAT

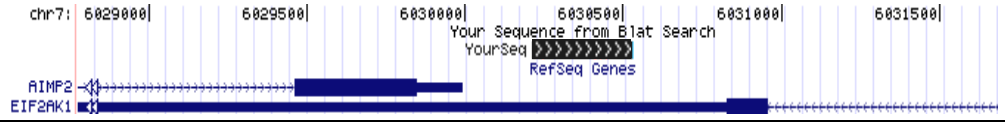


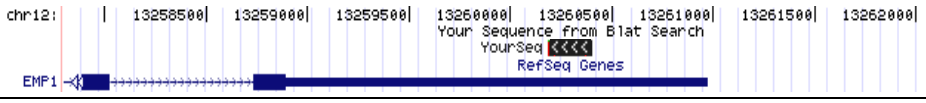
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 72**

A – fragmento de GLGI

GATCTTCTCAGGGCGATAAAGAAAACGTCTTCCCAAAAAAaCCCTTTAGACAGAATCTGAGATGGCTTAAAATAGCTGCT
CCAGGGCAAAGCAATGCCTGCTTAACCAGGAATAACATACTACTTCTAAGTCCTGTATAGAGGTGCTCCTGCGGCAGCAG
ACAGCTCGTCTACCATCTGACTAGGTACCcCAAGAGTTTTTGCTTTTTTATTTTAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 77**

A – fragmento de GLGI

GATCTGGTTGTCAGTTGGGATGGACTTGCCCAAGTGATGGCCCACAGAAAGGCCAAATTTCTTGTTTTTCTCCTCATCCT
GTACCTCTTTTTTCATTAAGAATCCTGCCTGGAAGTTTAGGTCAAAGAGGCTGCTTGGAGCAAAATACAGTGGTGTCTCA
TCCCAAATATTCTCCAGGCGTTTCTTCCATCCTTCATTTGAAAAA

B - Organização genômica – BLAT

***Tag* 83**

A – fragmento de GLGI

GATCTGTAAGATGTGAGAGGTGTTGAATAATCTTTAATATTACACATAAACCACACTAAAATGCCTTTCAATAAGTAAAA
GAAACCATTTTAAATACAGGGAATTATAATTAGATTGGCATAGTTAAGGCCAAAACTATAGACATTGCTACCTTATTTAT
CTTCAACCCTTGCCTTTAAGAGGCAAATGAACATGAAAAAAAAAAAAA

B - Organização genômica – BLAT

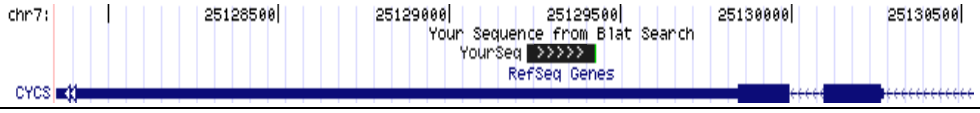


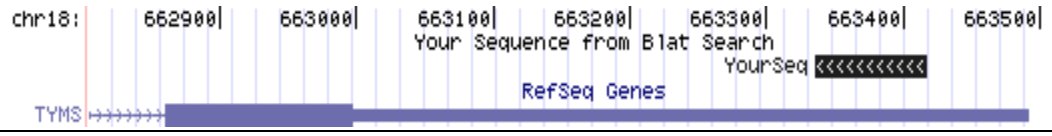
Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

***Tag* 87**

A – fragmento de GLGI

GATCGATGGTGTTACTCAGCTCCCTCAGATTTGAGGGAATAGCTTGTGAAATTCTTAAAATATTCTAAAAATATTCCAAA
AAAAAAAAAAAAA

B - Organização genômica – BLAT

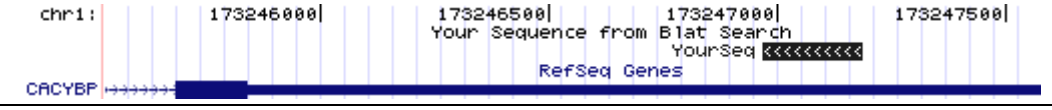


***Tag* 94**

A – fragmento de GLGI

GATCGTCAGGCGGCTTGGAATATATCTATATGCCTATTTTCAGTTATCTTAGTATCATTAAAAACCCATCTTTACCCCAC
TTTTACCATATTCTGTAACTTTCATATGCATTGTGTCTGCTATTCTGTTCATGTGTCAACCCCAATTACAGAGGCATAAT
GTGTTTCAGTATTTACTTATAGCTATCCTATACATAAACTGTTTCTGTATTTTAAATCCTAAAAAAAAAAAAAA

B - Organização genômica – BLAT



***Tag* 95**

A – fragmento de GLGI

GATCGAGGGCCACAAAAAAAGATGGAAGTGGTAAAATTGAAGAAAACATTTTAACCGGAATAAAGGCTTCCATTCAAAAC
CCAAGAAGGGTCCTTTGAGGGATAAAAGGTACAAAAAAAAAAAAAAA

B - Organização genômica – BLAT

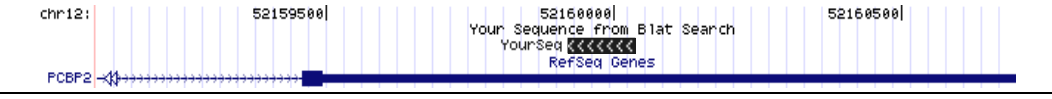


Figura 6 - Organização genômica dos fragmentos antisenso de GLGI em relação ao transcrito senso correspondente. Em (A) está representada a sequência do fragmento gerado por GLGI. Em vermelho está representado o sítio da enzima *DpnII* e em azul a sequência da *tag* de MPSS. Na sequência da *tag* 13, as bases destacadas em verde indicam a borda exon/intron. Em (B) está representado o alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT. Alinhamento obtido por meio da análise do fragmento de GLGI com a utilização da ferramenta BLAT.

Para 2 (*tag* 40 e 41) dos 27 candidatos, não foi possível a padronização da RT-PCR, uma vez que as reações apresentaram uma alta inespecificidade, mesmo em diferentes condições de amplificação. Como podemos observar na Figura 6, o fragmento de GLGI representado pela *tag* 41 tem menos de 50 pb, o que dificultou o desenho de iniciadores para a análise por RT-PCR fita específica.

Portanto, dos 25 candidatos avaliados por RT-PCR fita específica, observamos a presença do fragmento referente ao transcrito antisenso em 17 (*tag* 03, 13, 17, 19, 24, 25, 49, 53, 58, 65, 70, 72, 77, 83, 87, 94 e 95), confirmando desta forma a existência desses transcritos antisenso (Figura 7 – reação 2). Para 3 (*tag* 28, 43 e 52) dos 25 candidatos, além do fragmento correspondente ao antisenso, também observamos a presença do fragmento correspondente a ocorrência de *self-priming*. Dessa forma, não é possível garantir a existência destes transcritos antisenso, uma vez que eles podem ter sido gerados pela formação de estruturas secundárias, durante a síntese de cDNA, as quais podem servir como iniciadores para a ação da transcriptase reversa (Figura 7 - reação 3).

Interessantemente, observamos que dois fragmentos de GLGI representados pela *tag* 58 e *tag* 94 estavam mapeados na porção 3' do gene *CACYBP* (*Calcyclin binding protein*), entretanto em regiões distintas, sendo ambos os fragmentos validados por RT-PCR fita específica (Figura 7 e Figura 8). Isso pode indicar que esses transcritos tenham sido gerados por eventos de poliadenilação alternativa. De fato, demonstramos que 51% de todos os pares senso/antisenso (mRNA x mRNA e mRNA x ESTs) foram gerados devido a existência de pelo menos uma variante de poliadenilação (GALANTE et al. 2007).

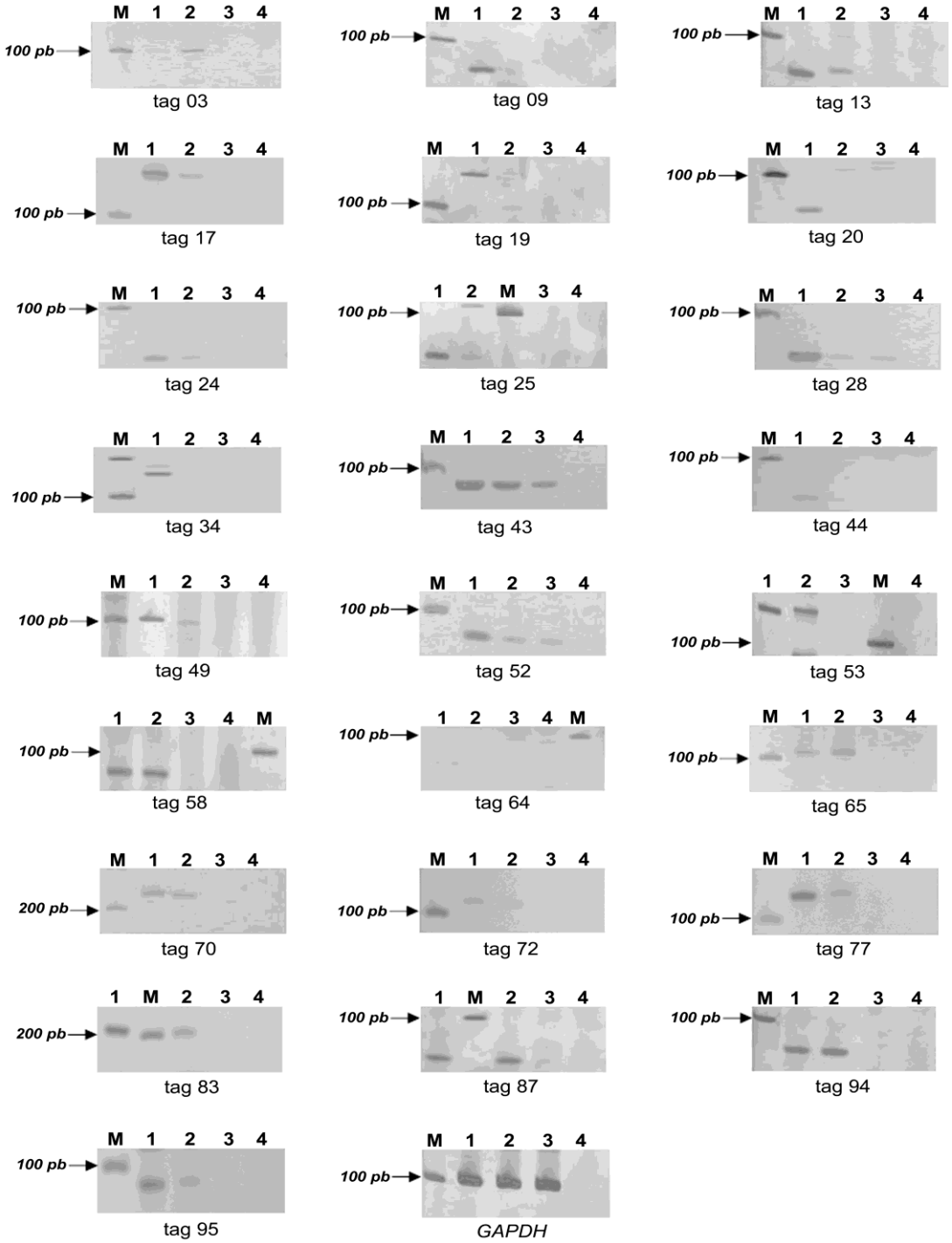


Figura 7 - RT-PCR fita específica. Validação experimental dos transcritos antisenso identificados e do gene controle *GAPDH*. Para os candidatos: (1) presença de transcriptase reversa e iniciador complementar ao transcrito senso; (2) presença de transcriptase reversa e iniciador complementar ao transcrito antisenso; (3) presença de transcriptase reversa e ausência de iniciadores (4) ausência de transcriptase reversa e presença de iniciador complementar ao transcrito antisenso. Para o *GAPDH*: (1,2 e 3) presença de transcriptase reversa e iniciador complementar ao transcrito senso, e em (4) ausência de transcriptase reversa e presença de iniciador complementar ao transcrito senso. Os fragmentos de PCR foram visualizados em gel de acrilamida 8% corados com nitrato de prata. Um marcador molecular de 100pb (M) foi utilizado.

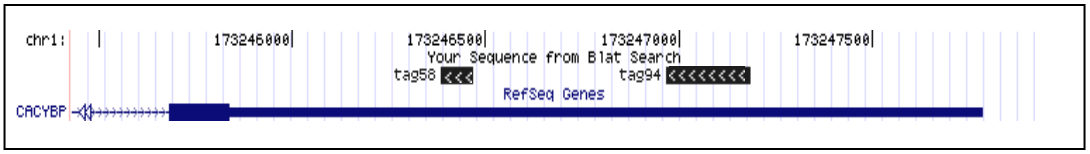


Figura 8 - Mapeamento das *tags* 58 e 94 no genoma humano. Representação esquemática do mapeamento das *tags* 58 e 94 no genoma humano, demonstrando a sobreposição que ocorre na porção 3' do gene *CACYBP*.

Podemos observar que a maioria dos transcritos antisenso validados são representados no gel por fragmentos de baixa intensidade, indicando uma baixa expressão desses transcritos nas linhagens celulares analisadas. Relatos têm demonstrado que a maioria dos NATs apresenta baixos níveis de expressão quando comparados a expressão do transcrito senso correspondente (GE et al. 2006; WERNER et al. 2007). Os dados de MPSS (Tabela 3) corroboram a baixa expressão dos transcritos antisenso validados por RT-PCR fita específica.

Portanto, 68% (17/25) dos fragmentos de GLGI com orientação antisenso são reais e transcritos a partir da fita oposta de um gene senso correspondente. Nossa validação experimental, utilizando a RT-PCR fita específica corrobora os dados de outros estudos. GE et al. (2008) também confirmaram a existência de 71% (17/24) dos transcritos antisenso avaliados por RT-PCR fita específica. Da mesma forma, MONTI et al. (2009) validaram 69% (11/16) dos transcritos antisenso identificados utilizando a mesma técnica.

1.4 CONCLUSÕES

Em conjunto, os dados apresentados neste projeto nos permitem concluir que:

- Dados de MPSS podem ser utilizados para identificação de novos transcritos antisenso naturais;

- Análises computacionais a partir de dados de MPSS permitiram a identificação de 4.308 novos transcritos antisenso naturais não suportados por sequências de mRNA e ESTs;

- A validação experimental desses dados nos permite estimar que pelo menos 52% (2.240 / 4.308) dos novos transcritos antisenso naturais são reais.

CAPÍTULO II

Identificação de genes com expressão alélica diferencial utilizando *tags* alelo específicas de SAGE.

2 CAPÍTULO II

2.1 INTRODUÇÃO

2.1.1 Expressão Alélica Diferencial

Em organismos eucariotos diplóides, por muito tempo assumiu-se que ambos os alelos (materno e paterno) de cada um dos genes eram expressos simultaneamente e em níveis semelhantes, segundo o proposto pela herança Mendeliana clássica. Entretanto, com a descoberta do processo de inativação do cromossomo X (LYON 1961) e do *imprinting* genômico (REIK e WALTER 2001) em mamíferos, ficou claro que alguns genes não seguem este padrão de expressão. Os genes que sofrem *imprinting* apresentam um padrão de expressão no qual somente um alelo é expresso enquanto o outro está silenciado, sendo isso dependente da origem parental (REIK e WALTER 2001). Já a inativação do cromossomo X é um processo que ocorre em células de fêmeas de mamíferos, no qual uma das cópias do cromossomo X é inativada, aleatoriamente. Esta inativação acontece no período embrionário, no momento da implantação do blastocisto no útero e, por isso, acredita-se que metade das células apresente o cromossomo X materno inativado e a outra metade o cromossomo X paterno. Em mamíferos esse fenômeno foi sugerido como um mecanismo de compensação de dose para igualar a expressão dos genes localizados no cromossomo X entre macho e fêmea de uma mesma espécie (LYON 1961).

Diferenças de expressão gênica entre os alelos de um gene, até então, eram atribuídas apenas ao *imprinting* genômico e à inativação do cromossomo X.

Entretanto, recentes relatos têm demonstrado que a expressão alélica diferencial (ADE), independente dos mecanismos de *imprinting* e inativação do cromossomo X, ocorre com frequência entre os indivíduos (KNIGHT 2004). A expressão alélica diferencial que ocorre nesses genes foi descrita como sendo aleatória, ou seja, algumas células expressam apenas um dos alelos (expressão monoalélica), enquanto outras apresentam uma expressão variável entre os dois alelos, independente da origem parental (GIMELBRANT et al. 2007).

Esses mesmos trabalhos têm relatado que de 20 a 50% dos genes humanos apresentam expressão alélica diferencial e sugerem que essa diferença de expressão entre os alelos está envolvida na variabilidade fenotípica observada entre indivíduos e populações, e que pode contribuir para o desenvolvimento tanto de doenças Mendelianas como doenças geneticamente complexas, como o câncer (YAN et al. 2002; LO et al. 2003; GE et al. 2005; GIMELBRANT et al. 2007). A diferença de expressão gênica entre os alelos parece ser hereditária (YAN et al. 2002; GIMELBRANT et al. 2007), envolvendo tanto eventos genéticos (presença de polimorfismos em regiões regulatórias dos genes) como epigenéticos (metilação diferencial entre os alelos) (KNIGHT 2004).

Devido às implicações para a saúde humana, sua associação com desenvolvimento de doenças e por ser um evento mais comum do que estimado inicialmente, trabalhos em larga escala têm sido desenvolvidos com o intuito de identificar genes que apresentam expressão alélica diferencial (LO et al. 2003; GE et al. 2005; PANT et al. 2006; GIMELBRANT et al. 2007; SERRE et al. 2008; PALACIOS et al. 2009).

2.1.2 Identificação em larga escala de genes com expressão alélica diferencial

Um dos maiores desafios para a identificação de genes com expressão alélica diferencial é a dificuldade de avaliar a expressão atribuída a cada um dos alelos de determinado gene. A utilização de SNPs (*Single Nucleotide Polymorphisms*) localizados em regiões gênicas tem sido uma estratégia muito utilizada na identificação em larga escala de genes com expressão alélica diferencial (YAN et al. 2002; GIMELBRANT et al. 2007; PALACIOS et al. 2009). Os SNPs ocorrem de forma abundante no genoma humano (SACHIDANANDAM et al. 2001) e, atualmente, mais de 17 milhões de SNPs humanos estão descritos em bancos de dados públicos (dbSNP - http://www.ncbi.nlm.nih.gov/projects/SNP/). Os SNPs presentes em regiões transcritas representam uma maneira simples de diferenciar alelos de um mesmo gene e permitem de forma eficiente avaliar o padrão de expressão dos dois alelos a partir do RNA mensageiro de indivíduos heterozigotos. A utilização desses SNPs possibilita a comparação da expressão relativa entre os alelos de um mesmo gene para cada indivíduo na mesma amostra biológica, evitando assim distorções atribuídas ao *background* genético dos diferentes indivíduos e também a outros fatores não genéticos como, por exemplo, a ação de fatores ambientais (YAN e ZHOU 2004).

Os primeiros esforços visando à identificação em larga escala de genes com expressão alélica diferencial utilizando SNPs foram desenvolvidos por meio de um método baseado na discriminação alélica de fragmentos de PCR pela incorporação de nucleotídeos marcados no momento da extensão do fragmento (YAN et al. 2002; BRAY et al. 2003). Dessa forma, YAN et al. (2002) observaram que 46% (6/13) dos genes avaliados em amostras de linhagens celulares linfoblastóides (LCLs)

apresentaram diferenças de 1,3 a 4,3 vezes de expressão entre os alelos, sugerindo que a variação na expressão entre os alelos era muito comum. Apesar disso, poucos indivíduos heterozigotos (3 a 30%) apresentaram expressão diferencial entre os alelos dos genes estudados, o que não é condizente com o mecanismo de *imprinting*. Analisando indivíduos de famílias CEPH (*Centre d'Etude du Polymorphisme Humain*), os autores demonstraram que a expressão alélica diferencial desses genes é compatível com padrões de herança Mendeliana. Resultados semelhantes foram descritos por BRAY et al. (2003) em amostras de cérebro normal. Neste estudo foi observado que 46% (7/15) dos genes avaliados apresentaram expressão diferencial (>20%) entre os alelos. Assim como no estudo de YAN et al. (2002), também observou-se que apenas uma minoria dos indivíduos heterozigotos apresentava tal padrão de expressão.

Apesar de ser uma técnica com excelente propriedade quantitativa (BRAY et al. 2003) a metodologia utilizada nos trabalhos descritos acima não permite a avaliação de um número muito grande de genes. Assim, outro desafio foi o desenvolvimento de novas estratégias visando elevar em nível genômico a análise de expressão alélica diferencial (KNIGHT 2004). Dentre elas, a metodologia de *microarrays* tem sido bastante utilizada e demonstrou ser uma plataforma muito eficaz na identificação em larga escala de genes que apresentam expressão alélica diferencial (Tabela 4). Essa técnica permite distinguir a expressão dos alelos de um gene com base na hibridização específica a sondas distintas que contêm os nucleotídeos variáveis que formam o SNP na posição genômica correspondente (LO et al. 2003; PANT et al. 2006; POLLARD et al. 2007; BJORNSSON et al. 2008; SERRE et al. 2008).

Tabela 4 - A metodologia de *microarray* na identificação em larga escala de genes com expressão alélica diferencial.

Referência	SNPs avaliados	Genes Avaliados		
		Total	com indivíduos heterozigotos	com expressão alélica diferencial (%)
Lo et al. (2003)	1.494	1063	602	326 (54%)
Pant et al. (2006)	8.406	4102	1.389	731 (53%)
Pollard et al. (2007)	7.109	-	2.625	460 (17,5%)
Bjornsson et al. (2008)	12.000	5770	2.885	288 (10%)
Serre et al. (2008)	2.968	1380	643	140 (22%)
Palacios et al. (2009)	11.500	-	1.632	1.195 (72%)

Como pode ser observado na Tabela 4, os trabalhos descritos apresentam uma grande variação na proporção de genes avaliados e identificados com expressão alélica diferencial que pode ser explicada por dois motivos principais. O primeiro seria que esses trabalhos utilizaram diferentes plataformas de *microarray*, que apresentam diferenças na capacidade de discriminar SNPs e genes para avaliar a expressão alélica diferencial. O segundo motivo seria o fato desses trabalhos utilizarem amostras de tecidos diferentes na realização de suas análises, uma vez que para uma parte dos genes que apresentam expressão alélica diferencial, a variação de expressão entre os alelos já foi descrita como tecido específica (YAN e ZHOU 2004; KHATIB 2007).

Dentre os genes que apresentaram expressão alélica diferencial nesses trabalhos estão descritos genes conhecidamente submetidos à *imprinting* genômico como, por exemplo, *SNRPN*, *IPW*, *PEG3*, *PEG10* e *KCNQ1* (LO et al. 2003; POLLARD et al. 2007; BJORNSSON et al. 2008; SERRE et al. 2008; PALACIOS et al. 2009). Da mesma forma, esses trabalhos também identificaram genes com expressão monoalélica e demonstraram que alguns deles são regulados por

imprinting como, por exemplo, *FLJ3371*, *PRIM2A* e *ZNF463* (PANT et al. 2006; BJORNSSON et al. 2008). Geralmente, genes controlados pelo mecanismo de *imprinting* encontram-se agrupados em regiões (*clusters*) no genoma humano (REIK e WALTER 2001), entretanto LO et al. (2003) observaram que a maioria dos genes que apresentaram expressão alélica diferencial não estava localizada em regiões conhecidas de *imprinting*, corroborando que a expressão alélica diferencial é um evento muito comum também em genes que não são regulados por esse mecanismo.

Em uma análise semelhante aos trabalhos descritos, GIMELBRANT et al. (2007) identificaram genes que apresentaram expressão monoalélica aleatória, utilizando uma plataforma de *microarray* (*Affymetrix 500 K SNP array*) contendo sondas para 250.000 SNPs (11.401 genes) do genoma humano e amostras de LCLs. Dos 3.939 genes que puderam ser avaliados, 371 (9,5%) apresentaram expressão monoalélica aleatória, ou seja, um clone de células expressava o alelo materno e outro clone o alelo paterno. Interessantemente, para a grande maioria (80%) desses genes foi observado a expressão alélica diferencial em alguns clones celulares. Em uma extrapolação conservadora dos dados, os autores sugeriram que aproximadamente 1.000 genes humanos apresentam expressão monoalélica aleatória, ou seja, independente da origem parental.

Recentemente, PALACIOS et al. (2009) analisaram a expressão alélica diferencial de ncRNAs em amostras de células mononucleares de sangue periférico. Apesar do fato de estarem utilizando RNA mensageiro em suas análises, o *microarray* (*Mapping 10 k array Affymetrix)* utilizado neste estudo continha sondas (92,8%) que discriminavam 2.311 SNPs intrônicos e 2.455 SNPs intergênicos, dos quais 65% (1.511/2.311) e 48% (1.190/2.455) respectivamente, apresentaram

expressão alélica diferencial. Mediante esses resultados, os autores sugeriram que expressão alélica diferencial também é freqüente em ncRNAs. Esses achados revelam uma complexidade ainda maior nos mecanismos de regulação da transcrição gênica no genoma humano (PALACIOS et al. 2009). Resultados semelhantes para ncRNAs, incluindo os NATs, foram recentemente demonstrados em leveduras (GAGNEUR et al. 2009).

Uma limitação na utilização da metodologia de *microarray* é a quantidade de genes que podem ser avaliados em um único experimento. Assim, a expressão alélica diferencial também tem sido avaliada utilizando-se ferramentas computacionais que permitem a comparação entre as frequências dos alelos associados a um SNP localizado em sequências expressas, principalmente ESTs, como descrito na Tabela 5 (YANG et al. 2003; LIN et al. 2005; GE et al. 2005).

Tabela 5 - O uso de abordagens computacionais para a identificação de expressão alélica diferencial.

Referência	SNPs (total)	ESTs	Agrupamento (*clusters*)	SNPs avaliados (genes)	SNPs ou Genes com expressão alélica diferencial *in silico* (%)
Yang et al. (2003)	-	3.569.546	101.602	19.312 (-)	194 SNPs (1%)
Li et al. (2005)	227.106	1.208.103	19.954	1007 (633)	524 SNPs (47%)
Ge et al. (2005)	11.822	3.856.000	2.500	11.822 (2.500)	976 genes (39%)

A introdução de milhões de sequências expressas em bancos de dados públicos (ESTs) e o desenvolvimento de ferramentas computacionais que permitem a associação dessas sequências aos dados de SNPs também possibilitaram a identificação de um número expressivo de genes que apresentaram expressão alélica diferencial (GE et al. 2005).

As primeiras análises computacionais foram desenvolvidas com intuito de identificar novos genes regulados por *imprinting*. Dessa forma, YANG et al. (2003) procuraram por genes para os quais ESTs representando os diferentes alelos nunca estavam presentes em um mesma biblioteca de cDNA. Em uma análise diferente, LI et al. (2005) identificaram bibliotecas de ESTs que apresentavam a expressão de ambos alelos e então avaliaram a expressão alélica diferencial baseado no número de ESTs que representava cada alelo. Neste trabalho foram consideradas apenas bibliotecas de tecidos normais e ainda foram excluídas as bibliotecas provenientes de *pool* de amostras. Ambos os trabalhos utilizaram os dados contidos em banco de dados públicos de SNPs (dbSNP) e ESTs (http://www.ncbi.nlm.nih.gov/unigene). A diferença observada entre o número de ESTs e SNPs utilizados nestes trabalhos ocorre devido aos diferentes parâmetros utilizados para garantir a qualidade das ESTs e a inclusão dos SNPs.

Uma análise semelhante às descritas acima foi desenvolvida por GE et al. (2005). A fim de confirmar os dados observados *in silico*, os autores desenvolveram uma estratégia experimental utilizando o sequenciamento direto do DNA e cDNA de LCLs. Neste trabalho foi desenvolvido o programa *PeakPicker* para facilitar as análises dos dados de expressão alélica diferencial provenientes do sequenciamento. O programa permite a normalização das bases polimórficas em relação às bases não variáveis que flanqueiam o SNP, garantindo a qualidade da análise. Nesta estratégia, o DNA genômico e o cDNA foram sequenciados em paralelo e as sequências foram analisadas no programa. A razão entre os picos referentes à posição do SNP no DNA de indivíduos heterozigotos foi utilizada para estabelecer a proporção esperada de 1:1 para os alelos no DNA. Assim, quando a proporção dos alelos no cDNA enquadrava-

se fora do intervalo de confiança, determinado a partir dos dados do DNA, a amostra foi considerada com expressão alélica diferencial. GE et al. (2005) submeteram a essa análise 39 dos 976 genes identificados com expressão alélica diferencial, para os quais 14 (36%) foram confirmados com expressão alélica diferencial. Diversos trabalhos na literatura têm utilizado a estratégia experimental desenvolvida por GE et al. (2005) para validar dados *in silico* ou obtidos por outras metodologias como, por exemplo, *microarrays* (POLLARD et al. 2007; SERRE et al. 2008; MILANI et al. 2009; GAGNEUR et al. 2009).

Pouco se sabe sobre os mecanismos de regulação de genes que apresentam expressão alélica diferencial e se os mesmos estão associados a alguma função biológica específica. Assim, com intuito de avaliar se tais estão associados a algum processo biológico em particular, PALACIOS et al. (2009) compararam a distribuição dos genes identificados com ou sem expressão alélica diferencial de acordo com a classificação funcional do banco de dados *Gene Ontology* (GO - http://www.geneontology.org/). Os autores observaram que não ocorrem diferenças significativas entre esses grupos de genes em diferentes processos biológicos, sugerindo que genes que apresentam expressão alélica diferencial podem participar de vários desses processos. Em uma análise semelhante, GIMELBRANT et al. (2007) observaram uma representação significativamente maior de genes que apresentaram expressão monoalélica aleatória na categoria (GO) receptores transmembrana, sugerindo um possível envolvimento desses genes em processos envolvendo mecanismos de comunicação celular.

Como podemos observar, uma série de diferentes estratégias foram utilizadas na busca por genes que apresentam expressão alélica diferencial. Devido a variações

atribuídas à sensibilidade e especificidade de cada método, o limitado número de amostras ou SNPs avaliados, esses trabalhos indicam frequências muito divergentes de genes com expressão alélica diferencial, que variam de 18 a 70% dos genes (POLLARD et al. 2007; PALACIOS et al. 2009). Na verdade, o número de genes com expressão alélica diferencial deve ser ainda maior, uma vez que apenas uma pequena parte dos genes conhecidos (10 a 20%) foi avaliada e poucos dos trabalhos levaram em consideração a ocorrência de expressão alélica diferencial em genes não codificantes (ncRNAs) (POLLARD et al. 2007). Corroborando essa idéia, KHATIB (2007) demonstrou que de 50 genes murinos descritos como regulados por *imprinting*, 26 (52%) apresentaram expressão alélica diferencial em diferentes tecidos.

Portanto, existem evidências de que a expressão alélica diferencial é um evento comum e que ocorre frequentemente no genoma humano. A maioria dos trabalhos acima citados sugere que esse mecanismo seja importante para estabelecer a diversidade dos indivíduos e das células, uma vez que entre 3 e 30% dos indivíduos (heterozigotos) avaliados nesses trabalhos apresentaram expressão alélica diferencial (POLLARD et al. 2007; GIMELBRANT et al. 2007; KHATIB 2007).

2.1.3 Genes com expressão alélica diferencial e sua associação com doenças

Recentemente, alguns trabalhos vêm demonstrando evidências de associação de genes que apresentam expressão alélica diferencial com o desenvolvimento de doenças, como descrito na Tabela 6 (MAHR et al. 2006; LI et al. 2006; WILKINS et al. 2007).

Tabela 6 - Exemplos de doenças humanas possivelmente associadas a genes que apresentam expressão alélica diferencial.

Referência	Genes	Patologia
Mahr et al. (2006)	*RHOB* e *TXNDC3*	Osteoartrite
Li et al. (2006)	*DAPK1*	doença de Alzheimer
Wilkins et al. (2007)	*BMP5*	Osteoartrite

Em relação ao câncer, ainda existem poucos trabalhos que avaliam a ocorrência de expressão alélica diferencial em amostras tumorais (MILANI et al. 2007, 2009).

MILANI et al. (2007) avaliaram a expressão alélica diferencial em um painel de 13 diferentes linhagens celulares tumorais resistentes ao tratamento com drogas antitumorais. Baseado em resultados provenientes de outros estudos que avaliaram a expressão de 7.400 transcritos humanos por cDNA *microarray* nessas mesmas linhagens, os autores analisaram 160 genes (237 SNPs em regiões codificantes) que estavam expressos em pelo menos uma dessas linhagens celulares, dos quais 79 continham SNPs em heterozigose em pelo menos uma das linhagens. Desses 79 genes, 60 (105 SNPs) foram avaliados no cDNA das linhagens. Por meio de *microarray*, os autores observaram a existência de expressão alélica diferencial em 68% (41/60) desses genes.

Em outro estudo, MILANI et al. (2009) analisaram a expressão alélica diferencial em 197 amostras (medula óssea e sangue periférico) de pacientes acometidos por leucemia linfoblástica aguda (LLA). Partindo da análise de 8.000 genes (13.917 SNPs) por *microarray*, os autores encontraram indivíduos heterozigotos para 2.529 genes (3.531 SNPs), dos quais 16% (400/2.529) demonstraram expressão alélica diferencial.

Os trabalhos apresentados acima confirmaram que a ocorrência de genes com expressão alélica diferencial também é frequente em tumores (GIMELBRANT et al. 2007; MILANI et al. 2009), entretanto, ainda há poucos esforços no sentido de associar alterações na expressão alélica diferencial de genes individuais com o desenvolvimento de tumores (JORDHEIM et al. 2008).

Assim, visto que a expressão alélica diferencial é um evento frequentemente observado nos genes humanos e ainda, que evidências sugerem que esse mecanismo é importante para estabelecer a diversidade entre os indivíduos e possa estar associado ao desenvolvimento de doenças, torna-se importante a identificação de genes que apresentam expressão alélica diferencial. Portanto, neste trabalho nós propomos a utilização de uma metodologia de análise da expressão gênica, denominada SAGE (*Serial Analysis of Gene Expression)*, em associação com dados de SNPs disponíveis em bancos de dados públicos (dbSNP) para a identificação de genes que apresentam expressão alélica diferencial.

2.1.4 Banco de dados de *tags* alelo específicas de SAGE

SAGE é uma técnica quantitativa que permite avaliar de maneira global o perfil de expressão de um determinado tecido. A técnica baseia-se na produção de uma sequência curta (*tag*) de 10 a 17 nt adjacente ao último sítio de restrição da enzima *Nla III* em relação a extremidade 3´ de cada transcrito. As *tags* geradas para cada transcrito são clonadas e sequenciadas em larga escala (VELCULESCU et al. 1995; SAHA et al. 2002). A interpretação dos dados de SAGE depende da análise das sequências e da extração e contagem das *tags* dentro de uma biblioteca. Em um segundo momento é preciso associar a sequência de uma determinada *tag* obtida

experimentalmente com o transcrito correspondente. Ao final desse processo é possível inferir o nível de expressão de um determinado transcrito a partir do número de vezes que a sua *tag* foi encontrada na biblioteca de SAGE (VELCULESCU et al. 1995).

Recentemente, nosso grupo avaliou o impacto da presença de SNPs na geração de *tags* alelo específicas de SAGE criando um banco de dados de *tags* alelo específicas (SILVA et al. 2004b). Para a criação desse banco de dados, as informações sobre as sequências de mRNA armazenadas no *Unigene* (http://www.ncbi.nlm.nih.gov/unigene) foram cruzadas com as informações sobre SNPs disponíveis no *NCBI SNP database* (dbSNP). Na construção deste banco de dados inicial foram utilizadas 586.144 *tags* únicas geradas a partir de 260 bibliotecas de SAGE derivadas de 25 tecidos humanos e aproximadamente um total de 5.800.000 SNPs. Neste trabalho também foram consideradas as *tags* geradas pela metodologia de MPSS em um total de 84.555 *tags* distintas encontradas em seis bibliotecas de dois tecidos (cólon e mama). Foram identificadas *tags* alelo específicas para 1.746 genes humanos. As *tags* alelo específicas de SAGE podem ser formadas de três maneiras distintas (Figura 9). A primeira, quando o SNP cria um novo sítio de restrição da enzima *Nla III* mais 3' da *tag* original de SAGE, dando origem a uma nova *tag* (Figura 9A). A segunda, quando o SNP destrói o sítio de restrição da enzima *Nla III* na *tag* original de SAGE, fazendo com que um outro sítio mais 5' ao original dê origem à uma nova *tag* (Figura 9B). Em ambas as situações são geradas *tags* alelo específicas com sequências completamente distintas. Na terceira, quando o SNP não afeta o sítio de restrição da enzima *Nla III*, mas altera a

sequência adjacente ao mesmo, gerando *tags* alelo específicas que diferem em apenas um nucleotídeo (Figura 9C) (SILVA et al. 2004b).

A presença de *tags* alelo específicas para esses genes possibilita então a realização de um estudo em larga escala para a identificação de genes que apresentam expressão alélica diferencial, uma vez que a expressão de cada um dos alelos pode ser inferida a partir da frequência de cada uma das *tags* alelo específicas nas diferentes bibliotecas avaliadas.

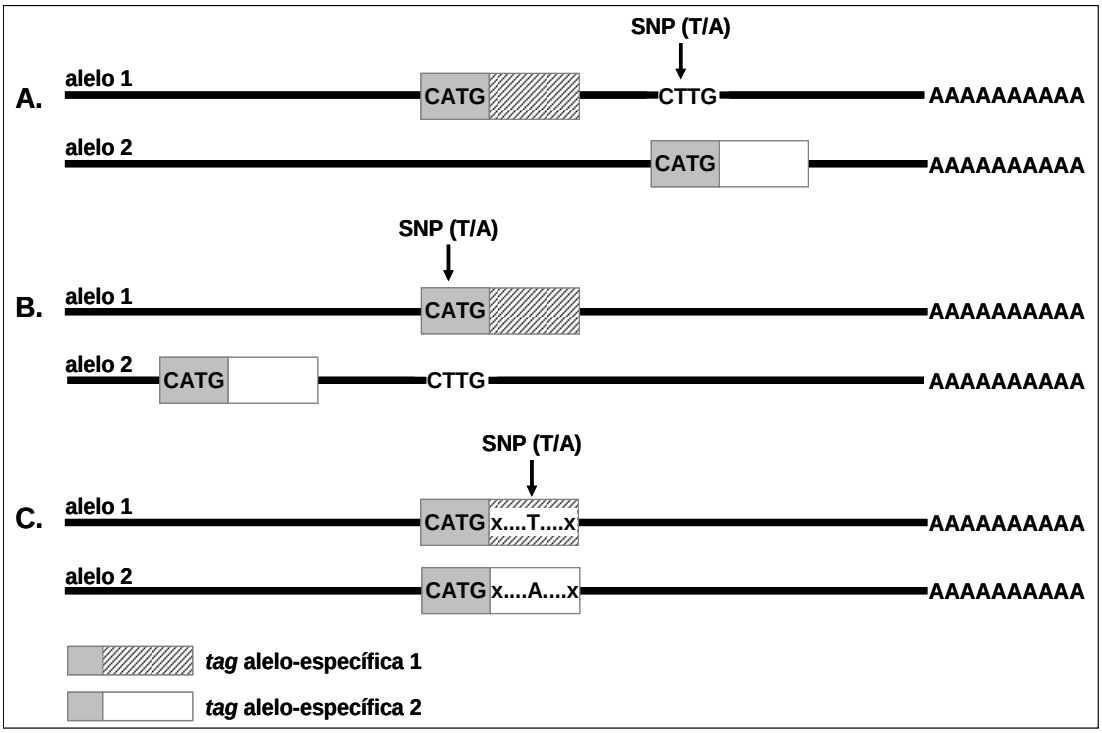


Figura 9 - Formação de *tags* alelo específicas de SAGE decorrentes da presença de SNPs. Em A, o SNP gera um novo sítio de restrição mais 3'; em B, o SNP destrói o sítio de restrição mais próximo da região 3'; em C, o SNP não afeta sítio de restrição mas altera a sequência da *tag* adjacente.

2.2 OBJETIVOS

2.2.1 Objetivo Geral

Nosso principal objetivo foi identificar e validar novos genes humanos que apresentam expressão alélica diferencial, utilizando um banco de dados de *tags* alelo específicas de SAGE.

2.2.2 Objetivos Específicos

- Identificar, dentre os 1.295 genes humanos que possuem *tags* alelo específicas de SAGE, aqueles cujas *tags* alelo específicas apareçam concomitantemente em pelo menos uma biblioteca e que a diferença entre a frequência dessas *tags* seja superior a 3, sugerindo a ocorrência de expressão alélica diferencial;
- Identificar, dentre os mesmos 1.295 genes humanos, aqueles cujas *tags* alelo específicas nunca apareçam concomitantemente em uma biblioteca, sugerindo um padrão de expressão monoalélico;
- Validar um grupo de 20 genes que apresentaram evidência de expressão alélica diferencial na análise *in silico*, por meio do sequenciamento direto do DNA e cDNA.

2.3 MATERIAL e MÉTODOS, RESULTADOS E DISCUSSÃO

Seguindo as normas estabelecidas pela Comissão de Pós-Graduação da Fundação Antônio Prudente, optamos por apresentar este trabalho na forma de artigo. Assim, apresentamos uma breve descrição dos tópicos abordados no manuscrito submetido para publicação e intitulado "*Analysis of allelic differential expression in the human genome using allele-specific SAGE tags*". Os tópicos contendo os materiais e métodos, os resultados e a discussão do trabalho estão descritos no manuscrito a seguir.

2.3.1 Manuscrito intitulado: "*Analysis of allelic differential expression in the human genome using allele-specific SAGE tags*".

O objetivo principal deste projeto foi a identificação de genes que apresentam expressão alélica diferencial utilizando dados de expressão gênica gerados pela metodologia de SAGE.

Assim, em colaboração com o grupo de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer foi desenvolvida uma estratégia computacional que permitiu a integração das informações sobre as sequências de mRNA armazenadas no *Unigene* com as informações sobre SNPs disponíveis no dbSNP possibilitando a criação de um banco de dados de *tags* alelo específicas de SAGE. A presença de *tags* alelo específicas para esses genes permite inferir a expressão de cada um dos alelos, a partir da frequência de cada uma das *tags* alelo específicas nas diferentes bibliotecas.

Os dados públicos de SAGE foram curados com a finalidade de excluir bibliotecas geradas a partir de uma mistura de RNA de diferentes indivíduos e bibliotecas provenientes de tecidos de origem embrionária. Ressaltamos também que as *tags* alelo específicas de SAGE geradas quando o SNP destrói o sítio de restrição da enzima *Nla III* na *tag* original de SAGE (Figura 9B) não foram utilizadas neste trabalho, uma vez que *tags* alelo específicas de SAGE localizadas mais 5' da *tag* original de SAGE podem também ser associadas com eventos de poliadenilação alternativa.

Assim, no nosso banco de dados, de um total de 20.034 genes (*UniGene clusters)*, 1.295 (6,46%) apresentaram *tags* alelo específicas de SAGE. Desses 1.295 genes, de acordo com o padrão de expressão dessas *tags*, 481 (37,2%) foram classificados com ADE. Interessantemente, identificamos 442 (34,1%) genes para os quais as *tags* alelo específicas de SAGE nunca foram observadas concomitantemente na mesma biblioteca, sugerindo a ocorrência de expressão monoalélica.

Após essa análise inicial *in silico*, passamos a etapa de validação experimental na qual 20 genes com evidências de ADE foram selecionados para o sequenciamento direto do gDNA e cDNA de indivíduos heterozigotos, seguido pela comparação entre a razão de intensidade dos picos das bases polimórficas no cromatograma das amostras de gDNA e cDNA. 13 desses 20 genes, em que foram encontrados mais do que 5 indivíduos heterozigotos, puderam ser avaliados. Desses 13 genes, 10 (76,9%) apresentaram ADE em pelo menos 20% dos indivíduos heterozigotos analisados. Os resultados serão apresentados no manuscrito a seguir.

2.3.2 Manuscrito

Analysis of allelic differential expression in the human genome using allele-specific SAGE tags.

Daniel O Vidal, Jorge E. de Souza, Lilian C. Pires, Cibele Masotti, Anna Christina M. Salim, Maria Cristina F. Costa, Pedro A. Galante, Sandro J. de Souza, Anamaria A. Camargo*.

Ludwig Institute for Cancer Research, São Paulo, SP, Brazil.

Abstract

Recent reports have demonstrated that a significant proportion of human genes display allelic differential expression (ADE). ADE is associated with phenotypic variability and may contribute to complex genetic diseases. Here, we present a computational analysis of ADE using allele-specific SAGE tags representing 1,295 human genes. We identified 481 genes for which unequal representation (>3-fold) of allele-specific SAGE tags was observed in at least one SAGE library, suggesting the occurrence of ADE. Moreover, for 242 genes, represented in more than 10 SAGE libraries, both allele-specific SAGE tags were never concomitantly observed in these libraries, suggesting that these genes might display monoallelic expression. Thirteen genes were subjected to experimental validation and ADE was confirmed for 10 out of these 13 genes. Our results suggest that ~43% of the human genes display ADE and allele-specific SAGE tags can be efficiently used for the identification of such genes.

Keywords: allelic differential expression, allele-specific SAGE tags, monoallelic expression, SAGE.

Introduction

Until recently, it was generally assumed that in diploid eukaryotic organisms both alleles of each gene were expressed at the same level and that allele-specific differences in expression levels were restricted to imprinted, X chromosome-inactivated genes and a few autosomal genes. However, recent reports have demonstrated that a significant proportion of non-imprinted autosomal human genes display allelic differential expression (ADE) (1-6) and that allele-specific differences in expression levels are heritable, involving both genetic and epigenetic mechanisms (1,7).

ADE has been associated with phenotypic variability between individuals and may contribute to both Mendelian and complex genetic diseases (8-11). Because of their implications for human health, several high-throughput methods measuring the relative expression level of different alleles using intragenic polymorphisms have been applied to identify genes displaying ADE (1-6). Together these studies demonstrate that ~20-65% of the human genes display ADE (1-6).

Serial Analysis of Gene Expression (SAGE) is a powerful technique for genome-wide analysis of gene expression that is capable of measuring expression levels irrespective of mRNA abundance and without a priori knowledge of the transcript sequence. In the SAGE technique, a short sequence tag with a variable length (10 or 17 nucleotides) adjacent to the 3´ most NlaIII restriction site is extracted from each transcript (12). The extracted tags are then concatenated for high-throughput sequencing and tag counts are used to measure the relative abundance of their corresponding transcripts.

We have previously analyzed the impact of Single Nucleotide Polymorphisms (SNPs) on the generation of allele-specific SAGE tags (13). The identification of allele-specific SAGE tags was achieved through the construction of a reference database in which the analysis of mRNA sequences from UniGene was combined with information available from the NCBI SNP database (14) and SAGE Genie (15). Allele-specific SAGE tags were identified by analyzing the presence of SNPs within the original SAGE tag sequence or SNPs creating or disrupting NlaIII sites used for SAGE library construction (13).

In the present work, we developed a computational method to identify genes that display ADE using our database of allele-specific SAGE tags and publicly available SAGE expression data. We first identified SAGE libraries in which both allele-specific SAGE tags were present and measured allelic variation in gene expression based on the SAGE tag counts for each allele. We identified 481 genes for which allele-specific SAGE tags were concomitantly expressed with a frequency difference higher than 3-fold in at least one SAGE library, suggesting the occurrence of ADE. Moreover, for 242 genes represented in at least 10 SAGE libraries, both allele-specific SAGE tags were never concomitantly observed in these libraries, suggesting that a fraction of these genes might be subjected to monoallelic gene expression. cDNA sequencing of heterozygotes was then used to validate a subset of genes displaying ADE. Our results suggest that ~43% of the human genes display ADE and allele-specific SAGE tags can be efficiently used for the identification of such genes.

Materials and Methods

Identification of allele-specific SAGE tags

Identification of allele-specific SAGE tags was carried out as previously described (13) except by the fact that updated versions of UniGene and NCBI-SNP database were used in the present analysis. A total of 74,561 mRNA sequences containing a poly-A tail and corresponding to 20,034 human genes according to UniGene (Build #198) were mapped onto the publicly available human genome sequence (Built #35.1) using BLAT and Sim4. Spurious and multiple alignments were eliminated by using an additional set of alignment criteria. These included a minimum identity of 93% and coverage (percentage of sequence length aligned) greater than 55%. Sequences mapping to more than one location on the genome were given a score for alignment quality. A higher score was associated with a higher identity and coverage. Only the alignments with the highest scores were kept in the database. Poly-A containing mRNA sequences were then scanned for the presence of NlaIII restriction sites and virtual SAGE tags downstream the 3' most restriction site were extracted and denominated original tags. A total of 10,054,521 SNPs from the

NCBI-SNP database (Build #124) were also mapped to the human genome sequence. Mapping was achieved through the alignment of sequences flanking the SNPs according to the NCBI criteria for SNP mapping. We have restricted our analysis to SNPs that mapped only once to the human genome sequence. A MySQL database was loaded with mapping information for all mRNAs, SNPs, and original SAGE tags that shared an overlap in genomic coordinates. Alternative allele-specific SAGE tags were then identified by crossing mapping information stored in the relational database and by analyzing the presence of SNPs within the original SAGE tag sequence or SNPs affecting (creating or disrupting) the restriction enzyme site used for SAGE library construction. The complete list of allele-specific SAGE tags and related information are available upon request. Alternative allele-specific SAGE tags were then compared to a list of experimentally obtained tags and were used to measure ADE using SAGE expression data available in public databases.

Analysis of ADE using SAGE data

SAGE expression data was obtained from SAGE Genie (http://cgap.nci.nih.gov/SAGE). SAGE library descriptions were manually curated to exclude libraries that were made with pooled RNA from different individuals or libraries that were derived from a same individual. ADE was measured using the relative frequency (tags per 200.000) of each allele-specific SAGE tag in 180 SAGE libraries (163 short SAGE libraries and 17 long SAGE libraries), assuming that curated libraries represent different single individuals. A complete list of curated SAGE libraries is available as Supplemental Table1.

Biological samples

Blood samples were obtained from the Blood Bank of the Hospital A.C.Camargo and were processed immediately after collection. Samples were collected after informed consent and the study was approved by institution’s ethic committee.

DNA extraction and genotyping

Genomic DNA was isolated by digestion with Proteinase K (Invitrogen) followed by phenol/ chloroform extractions. PCR primers used for genotyping (Supplemental Table 2) were designed flanking the SNP region associated with the allele-specific SAGE tag and avoiding other known SNPs within the primer sequences. Sequences corresponding to the universal sequencing primer M13 (5' GTAAAACGACGGCCAGT 3') were appended to the forward or reverse primers used for genotyping in order to apply the same sequencing conditions to all PCR fragments. PCR was carried out in a final volume of 25 µl, containing 50ng of gDNA, 1X Taq Platinum DNA polymerase buffer (Invitrogen), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each primer and 1 U of Taq Platinum DNA polymerase (Invitrogen). PCR conditions were 95ºC for 2 min, followed by 35 cycles at 95ºC for 35 sec, 60ºC for 35 sec, and 72ºC for 40 sec. Reactions were kept at 72ºC for 6 min after the last cycle. The amplified products were treated with 10U Exonuclease and 1U Shrimp Alkaline Phosphatase (USB) according to manufacturer's instructions and used for direct sequencing using Big-Dye Terminator (Applied Biosystems) and an ABI3130 sequencer (Applied Biosystems). Sequence traces were manually inspected to identify heterozygotes.

RNA extraction and RT-PCR

RNA was isolated using Trizol (Invitrogen) according to the manufacturer's instructions. The RNA quality was verified on agarose gels and 50µg of total RNA were treated with 8U of DNAse I (Ambion) for 40 min. at 37°C, extracted with phenol/chloroform and re-preciptated. cDNA synthesis was performed using 1µg DNA-free RNA, oligo (dT) 12-18 primers and Superscript II Reverse Transcriptase (Invitrogen) according to manufacturer 's instruction. Primers used for RT-PCR were the same used for gDNA amplification and genotyping. RT-PCRs were carried out in a final volume of 25 µl, containing 1µl of cDNA, 1X Taq Platinum DNA polymerase buffer (Invitrogen), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each primer and 1 U of Taq Platinum DNA polymerase (Invitrogen). PCR conditions were the same used for genotyping. RT-PCR products were purified and sequenced as described above.

Allele-specific expression analysis using high sensitivity sequencing

Allele expression levels were determined using the PeakPicker software (3) specifically developed to quantify the relative amount of two alleles by measuring peak intensity of the two polymorphic bases from sequence traces. Subsets of at least 5 informative heterozygotes for each SNP associated with the allele-specific SAGE tag were initially identified and their gDNA was amplified in identical conditions to establish the sequence peak intensity ratio between polymorphic bases that would correspond to a 50:50 representation of both alleles. Because peak heights vary depending on sample, base type, and their position within the sequence, the PeakPicker software carries out a normalization step in which the SNP allele peak intensity is compared to the intensity of reference peaks in the flanking sequence. We limited our analysis to sequence traces in which 80% of the bases within a 21 base window flanking the SNP presented Phred quality score>15. A text file with SNP allele peak intensity ratios normalized by the reference peaks is generated as an output. Peak intensity ratios were calculated for DNA and cDNA samples from all informative heterozygotes. Ratios with values above 1 were transformed to 1/(ratio) to set all ratios in a 0-1 scale and then adjusted to the mean of the peak intensity ratios from DNA samples. The adjusted peak intensity ratios (AR) of DNA samples from heterozygotes were used to estimate the methodological variability and establish a 99% confidence interval (CI) for equal representation of both alleles. The 99% CI was calculated assuming that normalized peak height ratios of DNA samples are normally distributed according to the Anderson-Darling test. A similar approach was used to identify samples displaying monoallelic expression. However, in this case, normalized ratios of homozygote DNA samples were used to establish a 99% CI.

Results

ADE analysis using allele-specific SAGE expression data

We have previously demonstrated that allele-specific SAGE tags can be generated due to the presence of SNPs creating or disrupting NlaIII sites used for

SAGE library construction or within the original tag sequence (Figure 1). The identification of allele-specific SAGE tags was achieved through the construction of a reference database in which the analysis of mRNA sequences from UniGene was combined with information available from the NCBI SNP database and SAGE Genie (13).

After updating our reference database, a total of 2,738 allele-specific short SAGE tags and 3,415 allele-specific long SAGE tags, representing 2,892 known human genes, were identified (Figure 2). Alternative allele-specific SAGE tags generated by SNPs that disrupted the NlaIII restriction site (640 short SAGE tags and 640 long SAGE tags) were not used for ADE analysis, since allele-specific SAGE tags located upstream of the original SAGE tag could also be associated with other biological phenomena like alternative polyadenylation or alternative splicing.

Of the remaining allele-specific SAGE tags, 1,584 short SAGE tags and 2,549 long SAGE tags could be unambiguously assigned to a known human transcript. Of those unambiguous tags, 1,226 allele-specific short SAGE tags and 1,098 allele-specific long SAGE tags were represented in SAGE libraries derived from single individuals (Figure 2). These tags were used for allele-specific expression analysis of 1,295 genes for which both allele-specific SAGE tags were represented in curated SAGE libraries. The complete list of genes under analysis is available as Supplemental Table 3. Allele-specific expression was measured using the relative frequency of each allele-specific SAGE tag in these libraries.

We then searched for genes displaying ADE and selected from our database 481 genes that showed at least 3-fold difference in the frequencies of both allele-specific tags in at least one SAGE library. For 96 out of these 481 genes, unequal representation of allele-specific SAGE tags was observed in more than 5 SAGE libraries presenting both allele-specific SAGE tags. The complete list of genes displaying ADE and their allele-specific SAGE expression data is available as Supplemental Table 4.

A representative example of a gene displaying ADE is presented in Table 1. The *NDUFA4* gene (Hs.50098) presents two allele-specific SAGE tags that differ by the presence of a SNP (rs1804855) within the tag sequence. *NDUFA4* expression was observed in 17 SAGE libraries derived mainly from mammary gland and white

blood cells. The presence of both allele-specific SAGE tags was detected in 6 out of these 17 libraries and differential expression between the two alleles (>3-fold) was observed in all of these 6 libraries.

Interestingly, we noticed that for 442 genes out of the 1,295 genes under analysis, both allele-specific SAGE tags were never concomitantly observed in any SAGE library, a finding that is compatible with genes displaying monoallelic expression. However, this number might include false positives, since the number of SAGE libraries in which expression of these genes was detected as well as a low frequency of one of the alleles in the population would directly influence the probability of detecting the simultaneous expression of both allele-specific SAGE tag. To overcome these limitations, we restricted our analysis to genes that were represented in at least 10 SAGE libraries and for which expression of one of the allele-specific tags was detected in at least 1 library, excluding to some extent tags associated with rare alleles and increasing the chances of having an heterozygote among individuals represented by different SAGE libraries. After restricting our initial analysis, 242 genes were classified as displaying putative monoallelic expression. This new subset contains 23 genes already described as displaying monoallelic expression, including the *ZNF597* imprinted gene. The complete list of genes displaying putative monoallelic expression and their allele-specific SAGE expression data is available as Supplemental Table 5.

A representative example of a gene displaying putative monoallelic expression is presented in Table 2. The *FRMD4B* gene (Hs.371681) presents two allele specific SAGE tags due to the presence of a SNP (rs6765309) that creates a new 3´ most NlaIII restriction site. *FRMD4B* expression was observed in 29 SAGE libraries derived from different tissues. The concomitant expression of both allele-specific SAGE tags was not detected in none of these 29 libraries. Expression of the original SAGE tag was detected in 25 libraries and expression of the SNP associated tag was detected in 4 libraries.

For the remaining 372 genes under analysis both allele-specific SAGE tags were equally represented (<3 fold difference) in all SAGE libraries presenting both allele-specific SAGE tags and were therefore classified as displaying biallelic expression.

Allele-specific gene expression analysis using high sensitivity sequencing

Twenty genes displaying ADE were randomly selected for allele-specific gene expression analysis in blood samples. For 13 of these 20 candidates, we were able to identify at least 5 heterozygotes to carry out the analysis. These 13 genes were represented on average in 77 different SAGE libraries (min.22 – max 158 libraries) and the presence of both tags in a same SAGE library was detected on average in 6 libraries (min. 2 – max. 11). ADE for these 13 genes was detected on average in 2 different libraries (min. 1- max 6).

Experimental validation was carried out by direct sequencing of gDNA and cDNA fragments from heterozygous individuals, followed by comparison between peak intensity ratios corresponding to the polymorphic bases in the sequencing traces from gDNA and cDNA samples (Figure 3). The peak intensity ratios from DNA samples of heterozygotes were used to estimate the methodological variability and to establish a 99% confidence interval (CI) for equal representation of both alleles (50:50 ratio). If normalized peak height ratios in cDNA samples showed significant deviation beyond the 99% CI established from genomic DNA data, the sample was classified as displaying allelic differential expression. A similar approach was used to identify samples displaying monoallelic expression. However, in this case, normalized ratios of homozygote DNA samples were used to establish the 99% CI and cDNA samples showing ratios within the 99% CI were considered as displaying monoallelic expression.

Of the 13 genes subjected to allele-specific gene expression analysis, 10 presented ADE in at least 20% of the heterozygotes (Figure 4). Moreover, ADE was detected in more than 75% of the heterozygotes analyzed for the *SQSTM1*, *CCDC74B* and *DSC1* genes. Interestingly, monoallelic expression was also detected in some heterozygotes for *CCDC74B*, *SQSTM1*, *ATP5F1* and *DSC1* genes and in all heterozygotes for the *PHC1* gene (Figure 4). Together, these results demonstrate that allele-specific SAGE tags can be efficiently used to measure allele-specific gene expression to identify genes displaying ADE.

Discussion

ADE has been observed in several species and can result from cis-acting sequence polymorphisms that affect the rate of transcription or from epigenetic modifications that cause complete or partial suppression of one allele (7). Expression level differences between alleles are directly associated with phenotypic variability and have been linked to the development of common genetic disorders in humans (8-11).

Several genome-wide computational and experimental studies have been carried out to identify genes displaying ADE (1-6). Collectively, these studies revealed that around 20-60% of all non-imprinted autosomal human genes display ADE. However, it has already been shown that ADE can be context specific with regard to cell type and developmental stage and that for some genes even small differences in expression level between alleles are physiologically relevant (8). In this context, the development of tools for identification of genes displaying ADE and, most importantly, for measuring differences in allele expression levels will allow us to dissect the genetic and epigenetic mechanisms underlying ADE and will contribute to a better understanding of phenotypic variability in humans.

In this study, we developed a computational method to identify genes displaying ADE and to measure allele-specific expression using publicly available SAGE data. Unambiguous allele-specific SAGE tags were identified for 1,295 (6.5%) human genes and allele-specific gene expression was measured by determining the relative frequency of allele-specific SAGE tags in 180 SAGE libraries derived form different single individuals. Using this approach, we were able to identify 481 genes displaying ADE and 242 genes with putative monoallelic expression. The applicability of our approach was experimentally confirmed by cDNA sequencing of heterozygotes. Ten out of 13 genes (77%) selected for experimental validation displayed ADE in at least 20% of the heterozygotes. If we consider that the 1,295 genes under analysis is a representative subset of all human genes and taking into account our experimental validation efficiency (77%), our analysis suggests that ~43% of all human genes (481+242 x 0.77/1,295) display

ADE and confirm previous estimates made with other computational and experimental methods (2,4,6).

Unexpectedly, monoallelic expression was detected in 100% of the heterozygotes for the *PHC1* gene and, interestingly, all heterozygotes expressed the same allele, suggesting that regulation of *PHC1* expression is influenced by cis-acting elements in strong linkage disequilibrium with the assayed exonic SNP. In contrast to the pattern observed in the validation analysis, concomitant expression of both allele-specific SAGE tags for *PHC1* was observed in 8 out of 103 SAGE libraries. Interestingly, all these libraries were derived from tumors, mostly from the central nervous system (CNS). A possible explanation for the unexpected results obtained for *PHC1* would be the occurrence of loss of the monoallelic expression pattern in tumors derived from the CNS. Loss of monoallelic expression has been described for a few imprinted genes in different tumor types, a mechanism called loss of imprinting or LOI (16,17). *PHC1* is a homolog of the Drosophila polyhomeotic gene, which is a member of the Polycomb group (PcG) of proteins (18). PcGs proteins are involved in chromatin remodeling and maintenance of gene expression patterns during development and differentiation (19, 20). Recent data suggest that PcG proteins are over-expressed in tumors and may play an important role in tumorigenesis (21, 22) however the mechanism associated with over-expression remains unknown.

In conclusion, we have demonstrated that allele-specific SAGE tags can be efficiently used to measure allele-specific gene expression and that SAGE data provide a valuable resource for studying phenotypic variation and complex diseases. To our knowledge, this is the first time that SAGE data is used for allele-specific expression analysis. One major limitation of our approach is the yet relatively small number of SAGE libraries available in public databases and the low expression level of a significant fraction of genes under analysis, which are represented by a small number of SAGE tags. These issues preclude the use of statistical methods for addressing differences in expression levels and might lead to an underestimation of the number of genes displaying ADE and to an overestimation of genes displaying monoallelic expression. Another important limitation is the yet relatively small number of human genes with allele-specific SAGE tags, which in turn is related to

limited information on the distribution of SNPs in the human genome. All these limitations will nevertheless be surpassed in the near future by the disseminated use of next generation sequencers for the generation of expression data and for identification of new SNPs.

The SAGE protocol has been recently adapted to be used in combination with next-generation sequencing platforms such as Illumina GA and SOLiD. These adapted protocols allow for the highest level of expression profile sensitivity and quantification with minimal sequencing requirements and will remain the choice protocols for gene expression quantification, especially when large number of samples is to be analyzed. Moreover, although more recent techniques such as RNA-Seq will allow broader gene coverage, and improve SNP detection, the application of this type of data for ADE studies is not straightforward. First, SNP calling in next generation sequencing data is still complex and one of the major advantages of our approach is that SNPs under analysis are experimentally confirmed in an independent fashion by sensitivity to restriction enzyme digestion. Secondly, events of alternative splicing (which are less common in the 3' end of the transcripts) will directly interfere in allele representation, increasing the complexity of the analysis. Finally, since sequence coverage is not equally distributed along the transcript, the use of RNA-Seq data will introduce further complexity to the analysis if more than one SNP per transcript is considered. In this context, we anticipate that the applicability of our approach will certainly increase in the near future with the disseminated use of next generation sequencing technologies.

Funding

This work was supported by Ludwig Institute for Cancer Research; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [141025/2005-0]; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Footnotes

To whom correspondence should be addressed: Tel: +55 11 33883248; Fax: +55 11 35490475; Email: anamaria@ludwig.org.br

"The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors".

References

1. Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B., Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.

2. Lo,H.S., Wang,Z., Hu,Y., Yang,H.H., Gere,S., Buetow,K.H., Lee,M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.*, **13**, 1855-62.

3. Ge,B., Gurd,S., Gaudin,T., Dore,C., Lepage,P., Harmsen,E., Hudson,T.J., Pastinen,T. (2005) Survey of allelic expression using EST mining. *Genome Res.*, **15**, 1584-91.

4. Pant,P.V., Tao,H., Beilharz,E.J., Ballinger,D.G., Cox,D.R., Frazer,K.A. (2006) Analysis of allelic differential expression in human white blood cells. *Genome Res.*, **16**, 331-9.

5. Serre,D., Gurd,S., Ge,B., SlADEk,R., Sinnett,D., Harmsen,E., Bibikova,M., Chudin,E., Barker,D.L., Dickinson,T., et al. (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.*, **4**, e1000006.

6. Palacios,R., Gazave,E., Goñi,J., Piedrafita,G., Fernando,O., Navarro,A., Villoslada,P. (2009) Allele-specific gene expression is widespread across the genome and biological processes. *PLoS One*, **4,** e4150.

7. Pastinen,T., SlADEk,R., Gurd,S., Sammak,A., Ge,B., Lepage,P., Lavergne,K., Villeneuve,A., Gaudin,T., Brändström,H., et al. (2004) A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics*, **16**, 184-93.

8. Wilkins,J.M., Southam,L., Price,A.J., Mustafa,Z., Carr,A., Loughlin,J. (2007) Extreme context specificity in differential allelic expression. *Hum Mol Genet.*,**16**, 537-46.

9. Milani,L., Gupta,M., Andersen,M., Dhar,S., Fryknäs,M., Isaksson,A., Larsson,R., Syvänen,A.C. (2007) Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic Acids Res.*, **35**, e34.

10. Jordheim,L.P., Nguyen-Dumont,T., Thomas,X., Dumontet,C., Tavtigian,S.V. (2008) Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of CDA, DCK, NT5C2, NT5C3, and TP53. *Drug Metab Dispos.*, **36**, 2419-23.

11. Milani,L., Lundmark,A., Nordlund,J., Kiialainen,A., Flaegstad,T., Jonmundsson,G., Kanerva,J., Schmiegelow,K., Gunderson,K.L., Lönnerholm,G., et al. (2009) Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res.*, **19**, 1-11.

12. Velculescu,V.E., Zhang,L., Vogelstein,B., Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484-7.

13. Silva,A.P., De Souza,J.E., Galante,P.A., Riggins,G.J., De Souza,S.J., Camargo,A.A. (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res.*, **32**, 6104-10.

14. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M., Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308-11.

15. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J., et al. (2002) An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A*, **99**, 11287-92.

16. Reik,W., Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nat Rev Genet.*, **2**, 21-32.

17. Murrell,A. (2006) Genomic imprinting and cancer: from primordial germ cells to somatic cells. *ScientificWorldJournal*, **6**, 1888-910.

18. Levine,S.S., Weiss,A., Erdjument-Bromage,H., Shao,Z., Tempst,P., Kingston,R.E. (2002) The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. *Mol Cell Biol.*, **22**, 6070-6078.

19. Lee,T.I., Jenner,R.G., Boyer,L.A., Guenther,M.G., Levine,S.S., Kumar,R.M., Chevalier,B., Johnstone,S.E., Cole,M.F., Isono,K., et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, **125**, 301-13.

20. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K., et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349-53.

21. Bruggeman,S.W., Hulsman,D., Tanger,E., Buckle,T., Blom,M., Zevenhoven,J., van Tellingen,O., van Lohuizen,M. (2007) Bmi1 controls tumor development in an Ink4a/Arf-independent manner in a mouse model for glioma. *Cancer Cell*, **12**, 328-41.

22. Kondo,Y., Shen,L., Cheng,A.S., Ahmed,S., Boumber,Y., Charo,C., Yamochi,T., Urano,T., Furukawa,K., Kwabi-Addo,B., et al. (2008) Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet.*, **40**, 741-50.

Figure Legends

Figure 1. Allele-specific SAGE tags. **(A)** SNPs that generate a new restriction enzyme site downstream to the original tag; **(B)** SNPs that disrupted the 3'-most restriction site associated with the original tag; **(C)** SNPs that did not affect the restriction sites, but occurred within the adjacent tag sequence. Restriction enzyme sites (NlaIII) are represented by black boxes, original tags by hatched boxes and Allele-specific SAGE tags by open boxes. The locations of the SNPs within mRNA sequences are indicated by arrows.

Figure 2. Schematic representation of the computational approach used for the identification of allele-specific SAGE tags and analysis of ADE.

Figure 3. Experimental validation of ADE using gDNA and cDNA direct sequencing. Representative sequencing results for genes presenting bialleic expression (*GLUL*), ADE (*SQSTM1*) and monoallelic expression (*PHC1*). Sequencing traces from gDNA and cDNA samples are represented in the left and right panels, respectively. SNPs are indicated by the arrows; AR means adjusted peak ratio.

Figure 4. ADE analysis in blood samples**.** Graphic representation of adjusted peak ratios (AR) for DNA (left) and cDNA (right) samples from heterozygotes of the 13 genes assayed for ADE. Upper and lower black lines represent the 99% confidence intervals (CI) for biallelic and monoallelic expression calculated using the AR from heterozygotes and homozygotes, respectively. Grey and black squares represent the AR from heterozygotes and homozygotes, respectively.

Table 1. Allele-specific SAGE expression data for a gene displaying ADE.

Gene Name:	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa (NDUFA4)
UniGene	Hs.50098
original tag	(long) TTGGAGATCTCTATTGT
alternative tag	(long) TTGGAGATCTCTATTGC
SNP ID	rs1804855

SAGE	library number	library origin	tissue	original tag frequency	alternative tag frequency
long	655	N	vascular	28.2	0.00
long	1563	N	white blood cells	33.0	0.00
long	**1564**	**N**	**white blood cells**	**25.5**	**1.96**
long	1565	N	white blood cells	7.90	0.00
long	**1566**	**N**	**white blood cells**	**36.0**	**2.00**
long	1567	N	white blood cells	16.1	0.00
long	**645**	**T**	**mammary gland**	**35.4**	**2.72**
long	649	T	mammary gland	36.63	0.00
long	653	T	colon	22.3	0.00
long	657	T	mammary gland	31.5	0.00
long	**673**	**T**	**mammary gland**	**84.1**	**12.0**
long	675	T	mammary gland	72.79	0.00
long	683	T	mammary gland	16.7	0.00
long	703	T	mammary gland	17.6	0.00
long	**723**	**T**	**mammary gland**	**32.3**	**3.59**
long	963	T	lung	88.2	0.00
long	**1645**	**T**	**brain**	**21.3**	**7.10**

Bold: presence of both allele-specific SAGE tags in SAGE libraries. tag frequency: count of tags normalized by 200.000. *: SAGE libraries in which allele-specific SAGE tags displayed ADE ($\geq$ 3). N: normal, T: tumoral.

Table2. Allele-specific SAGE expression data for a gene displaying monoallelic expression.

Gene Name:	FERM domain containing 4B (FRMD4B)	
UniGene	Hs.371681	
original tag	(short) TGAAGCAGGT	(long) TGAAGCAGGTCGCAGTG
alternative tag	(short) GTTCATCTTT	(long) GTTCATCTTTTTATTGA
SNP ID	rs6765309	

SAGE	library number	library origin	tissue	original tag frequency	alternative tag frequency
short	20	N	ovary	4.19	0.00
short	31	N	mammary gland	4.10	0.00
short	99	N	lung	2.25	0.00
short	133	N	brain	2.57	0.00
short	163	N	retina	1.89	0.00
short	181	N	white blood cells	6.25	0.00
short	389	N	thyroid	1.72	0.00
short	523	N	brain	2.75	0.00
long	1567	N	white blood cells	0.00	2.02
short	18	T	cerebellum	0.00	4.17
short	19	T	cerebellum	4.55	0.00
short	97	T	stomach	3.08	0.00
short	146	T	brain	2.25	0.00
short	153	T	mammary gland	2.74	0.00
short	170	T	stomach	2.86	0.00
short	171	T	mammary gland	6.60	0.00
short	172	T	mammary gland	4.64	0.00
short	185	T	vascular	0.00	2.64
short	343	T	brain	1.99	0.00
short	365	T	thyroid	0.00	1.60
short	383	T	cerebellum	3.30	0.00
short	412	T	cerebellum	3.36	0.00
short	524	T	brain	3.34	0.00
short	526	T	brain	2.88	0.00
short	528	T	brain	4.10	0.00
short	604	T	cartilage	4.26	0.00
short	608	T	cartilage	1.82	0.00
long	673	T	mammary gland	3.00	0.00
long	703	T	mammary gland	2.94	0.00

tag frequency: count of tags normalized by 200.000. N: normal, T: tumoral.

Figure 1

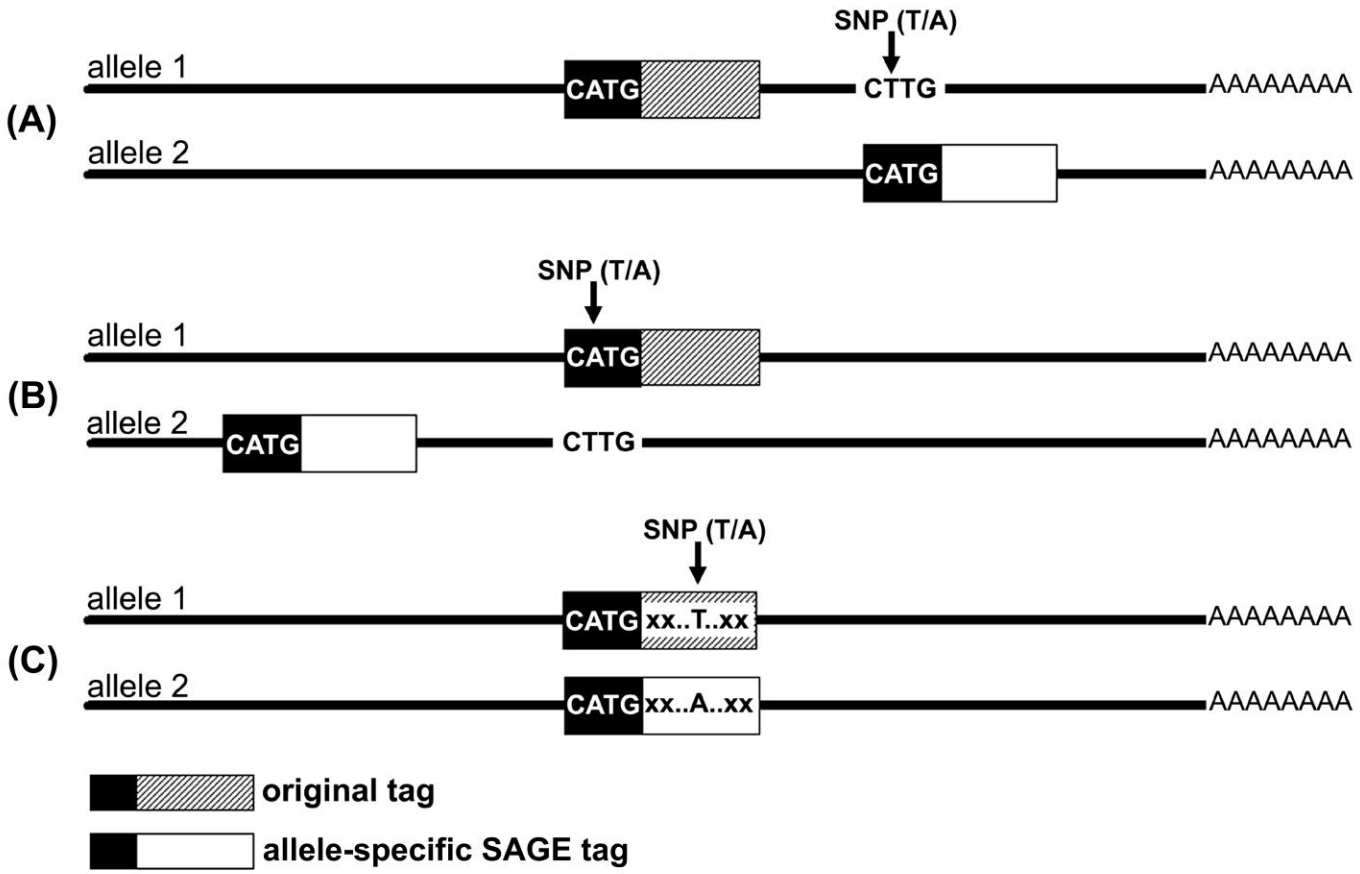

SNP (T/A)
allele 1
CATG
CTTG
AAAAAAAA
(A)
allele 2
CATG
AAAAAAAA
SNP (T/A)
allele 1
CATG
AAAAAAAA
(B)
allele 2
CATG
CTTG
AAAAAAAA
SNP (T/A)
allele 1
CATG xx..T..xx
AAAAAAAA
(C)
allele 2
CATG xx..A..xx
AAAAAAAA
original tag
allele-specific SAGE tag

Figure 2

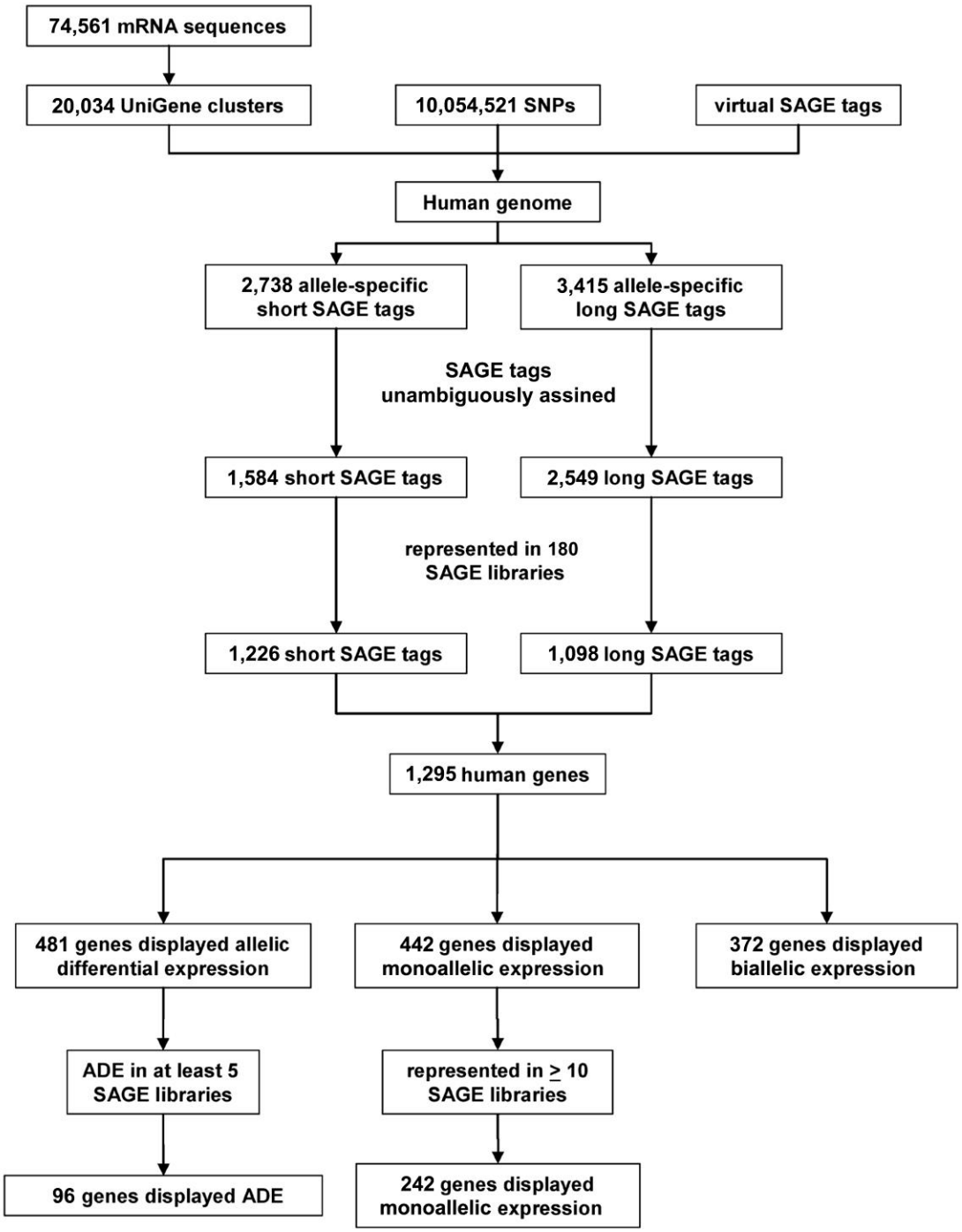

74,561 mRNA sequences
20,034 UniGene clusters
10,054,521 SNPs
virtual SAGE tags
Human genome
2,738 allele-specific short SAGE tags
3,415 allele-specific long SAGE tags
SAGE tags unambiguously assined
1,584 short SAGE tags
2,549 long SAGE tags
represented in 180 SAGE libraries
1,226 short SAGE tags
1,098 long SAGE tags
1,295 human genes
481 genes displayed allelic differential expression
442 genes displayed monoallelic expression
372 genes displayed biallelic expression
ADE in at least 5 SAGE libraries
represented in ≥ 10 SAGE libraries
96 genes displayed ADE
242 genes displayed monoallelic expression

Figure 3

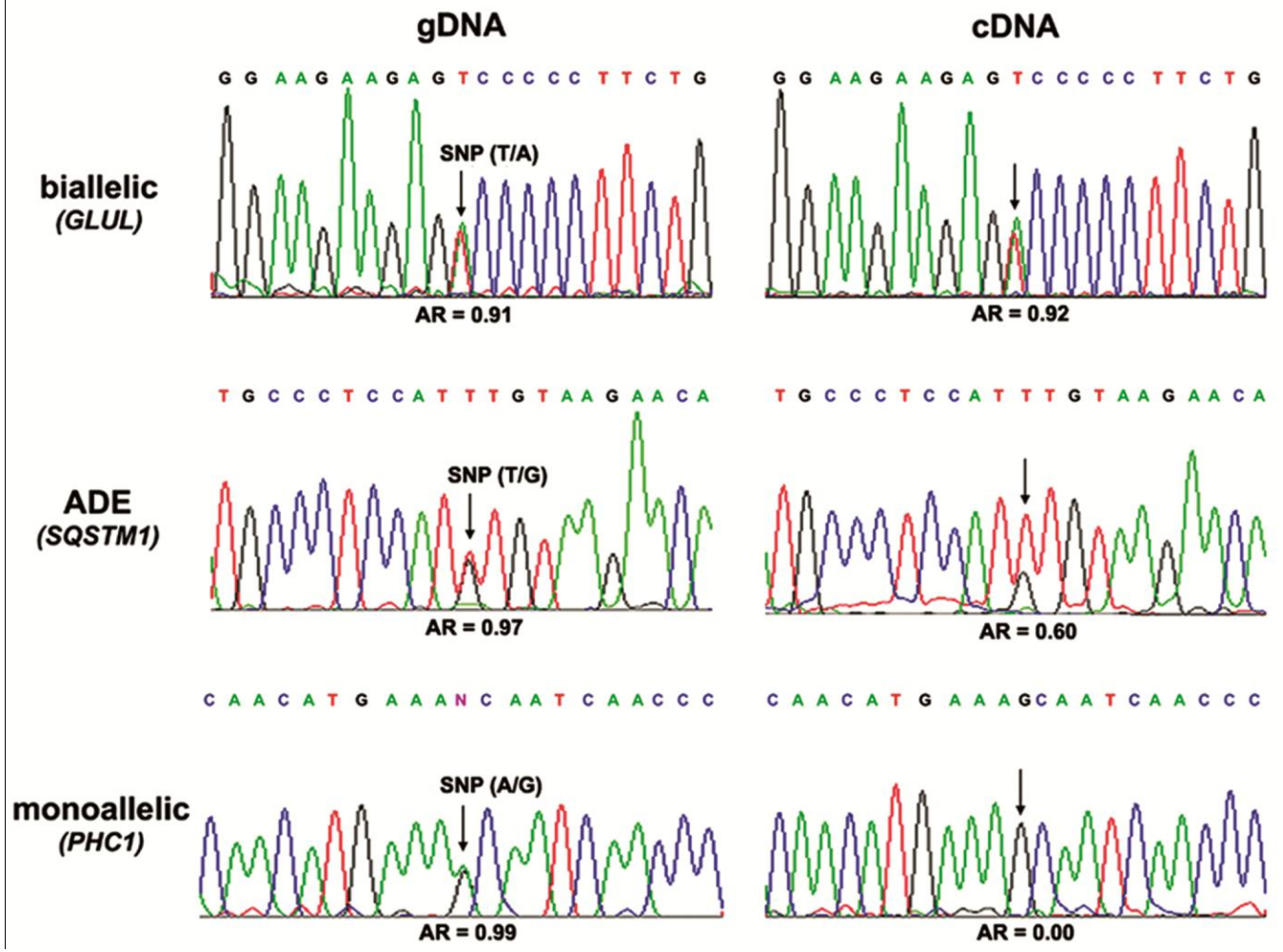
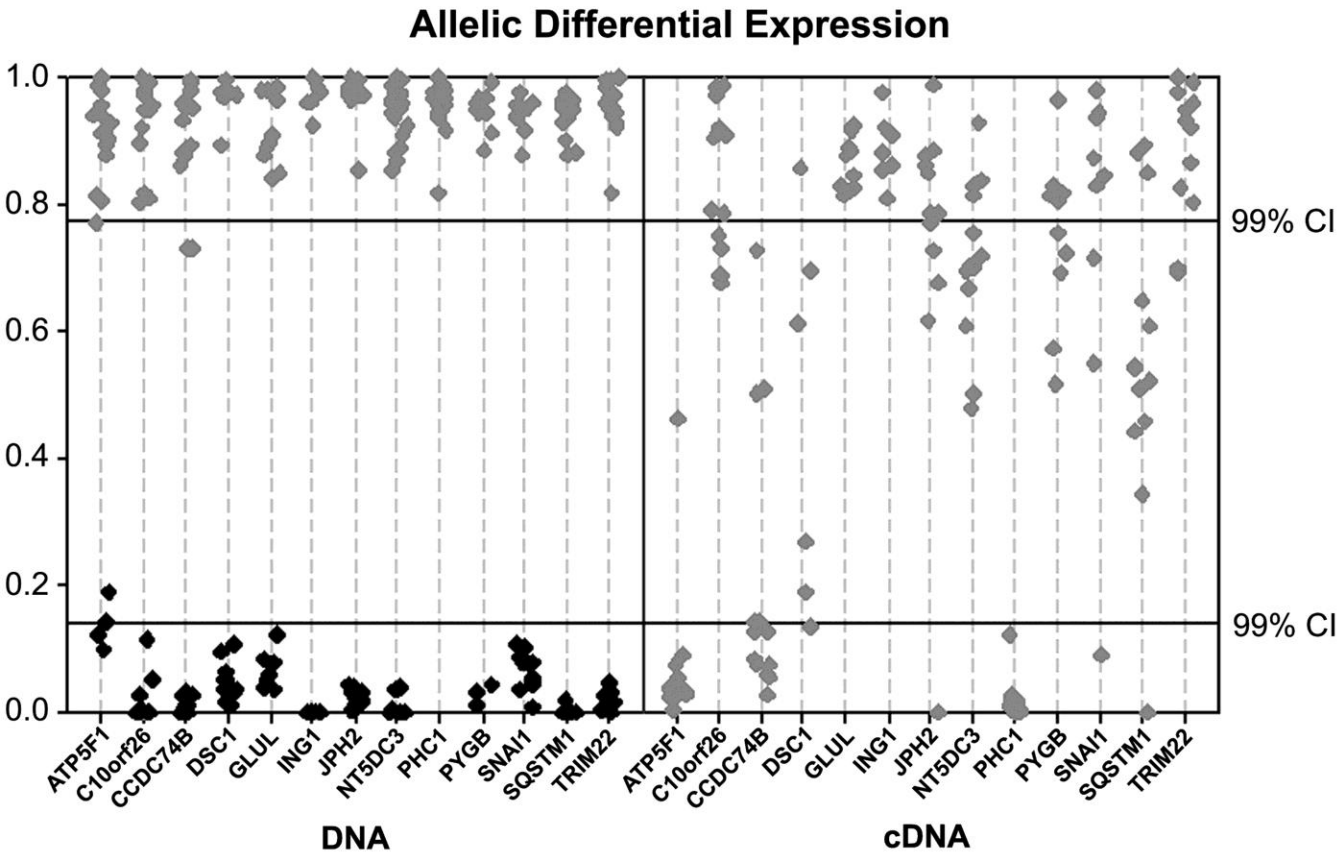

gDNA
cDNA
biallelic
(GLUL)
G G A A G A A G A G T C C C C C T T C T G
SNP (T/A)
AR = 0.91
G G A A G A A G A G T C C C C C T T C T G
AR = 0.92
ADE
(SQSTM1)
T G C C C T C C A T T T G T A A G A A C A
SNP (T/G)
AR = 0.97
T G C C C T C C A T T T G T A A G A A C A
AR = 0.60
monoallelic
(PHC1)
C A A C A T G A A A N C A A T C A A C C C
SNP (A/G)
AR = 0.99
C A A C A T G A A A G C A A T C A A C C C
AR = 0.00

Figure 4


Allelic Differential Expression
1.0
0.8
0.6
0.4
0.2
0.0
99% CI
99% CI
ATP5F1
C10orf26
CCDC74B
DSC1
GLUL
ING1
JPH2
NT5DC3
PHC1
PYGB
SNAI1
SQSTM1
TRIM22
DNA
cDNA

Supplemental Material

Analysis of allelic differential expression in the human genome using allele-specific SAGE tags

2.4 CONCLUSÕES

Em conjunto, os dados apresentados neste projeto nos permitem concluir que:

- *tags* alelo específicas de SAGE podem ser utilizadas na identificação de genes com expressão alélica diferencial;

- 37,2% (481) dos genes avaliados apresentam expressão alélica diferencial;

- 18,7% (242) dos genes avaliados apresentam expressão monoalélica;

- Levando em conta a eficiência da nossa validação experimental, nossos resultados sugerem que 43% de todos os genes humanos apresentam expressão alélica diferencial.

3 REFERÊNCIAS BIBLIOGRÁFICAS

Adams MD, Celniker SE, Holt RA, et al. The genome sequence of Drosophila melanogaster. **Science** 2000; 287:2185-95.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **Nature** 2000; 408:796-815.

Bachellerie JP, Cavaillé J, Hüttenhofer A. The expanding snoRNA world. **Biochimie** 2002; 84:775-90.

Beiter T, Reich E, Williams RW, Simon P. Antisense transcription: a critical look in both directions. **Cell Mol Life Sci** 2009; 66:94-112.

Bignone PA, Lee KY, Liu Y, et al. RPS6KA2, a putative tumour suppressor gene at 6q27 in sporadic epithelial ovarian cancer. **Oncogene** 2007; 26:683-700.

Bjornsson HT, Albert TJ, Ladd-Acosta CM, et al. SNP-specific array-based allele-specific expression analysis. **Genome Res** 2008; 18:771-9.

Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. Cis-acting variation in the expression of a high proportion of genes in human brain. **Hum Genet** 2003; 113:149-53.

Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. **Nat Biotechnol** 2000; 18:630-4.

Capaccioli S, Quattrone A, Schiavone N, et al. A bcl-2/IgH antisense transcript deregulates bcl-2 gene expression in human follicular lymphoma t(14;18) cell lines. **Oncogene** 1996; 13:105-15.

Chen JJ, Rowley JD, Wang SM. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. **Proc Natl Acad Sci U S A** 2000; 97:349-53.

Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD. Over 20% of human transcripts might form sense-antisense pairs. **Nucleic Acids Res** 2004; 32:4812-20.

Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. **Trends Genet** 2005; 21:326-9.

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

Clamp M, Fry B, Kamal M, et al. Distinguishing protein-coding and noncoding genes in the human genome. **Proc Natl Acad Sci U S A** 2007; 104:19428-33.

Engström PG, Suzuki H, Ninomiya N, et al. Complex Loci in human and mouse genomes. **PLoS Genet** 2006; 2:e47.

Faghihi MA, Modarresi F, Khalil AM, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. **Nat Med** 2008; 14:723-30.

Gagneur J, Sinha H, Perocchi F, Bourgon R, Huber W, Steinmetz LM. Genome-wide allele- and strand-specific expression profiling. **Mol Syst Biol** 2009; 5:274.

Ge B, Gurd S, Gaudin T, et al. Survey of allelic expression using EST mining. **Genome Res** 2005; 15:1584-91.

Ge X, Wu Q, Jung YC, Chen J, Wang SM. A large quantity of novel human antisense transcripts detected by LongSAGE. **Bioinformatics** 2006; 22:2475-9.

Ge X, Rubinstein WS, Jung YC, Wu Q. Genome-wide analysis of antisense transcription with Affymetrix exon array. **BMC Genomics** 2008; 9:27.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. **Science** 2007; 318:1136-40.

Gingeras TR. Origin of phenotypes: genes and transcripts. **Genome Res** 2007; 17:682-90.

Hastings ML, Milcarek C, Martincic K, Peterson ML, Munroe SH. Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. **Nucleic Acids Res** 1997; 25:4296-300.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisenso transcriptomes of human cells. **Science** 2008; 322:1855-7.

Jordheim LP, Nguyen-Dumont T, Thomas X, Dumontet C, Tavtigian SV. Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of CDA, DCK, NT5C2, NT5C3, and TP53. **Drug Metab Dispos** 2008; 36:2419-23.

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

Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. **Science** 2002; 296:916-9.

Katayama S, Tomaru Y, Kasukawa T, et al. RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. Antisense transcription in the mammalian transcriptome. **Science** 2005; 309:1564-6.

Kawasaki H, Taira K, Morris KV. siRNA induced transcriptional gene silencing in mammalian cells. **Cell Cycle** 2005; 4:442-8.

Khatib H. Is it genomic imprinting or preferential expression? **Bioessays** 2007; 29:1022-8.

Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y; RIKEN GER Group; GSL Members. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. **Genome Res** 2003; 13:1324-34.

Knight JC. Allele-specific gene expression uncovered. **Trends Genet** 2004; 20:113-6.

Kumar M, Carmichael GG. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. **Microbiol Mol Biol Rev** 1998; 62:1415-34.

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

Lapidot M, Pilpel Y. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. **EMBO Rep** 2006; 7:1216-22.

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, et al. Characterization of the piRNA complex from rat testes. **Science** 2006; 313:363-7.

Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G. In search of antisense. **Trends Biochem Sci** 2004; 29:88-94.

Lehner B, Williams G, Campbell RD, Sanderson CM. Antisense transcripts in the human genome. **Trends Genet** 2002; 18:63-5.

Levanon EY, Eisenberg E, Yelin R, et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. **Nat Biotechnol** 2004; 22:1001-5.

Li Y, Grupe A, Rowland C, Nowotny P, et al. DAPK1 variants are associated with Alzheimer's disease and allele-specific expression. **Hum Mol Genet** 2006; 15:2560-8.

Lin W, Yang HH, Lee MP. Allelic variation in gene expression identified through computational analysis of the dbEST database. **Genomics** 2005; 86:518-27.

Lo HS, Wang Z, Hu Y, et al. Allelic variation in gene expression is common in the human genome. **Genome Res** 2003; 13:1855-62.

Lyon MF. Gene action in the X-chromosome of the mouse (Mus musculus L.). **Nature** 1961; 190:372-3.

Mahr S, Burmester GR, Hilke D, et al. Cis- and trans-acting gene regulation is associated with osteoarthritis. **Am J Hum Genet** 2006; 78:793-803.

Milani L, Gupta M, Andersen M, et al. Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells**.** **Nucleic Acids Res** 2007; 35:e34.

Milani L, Lundmark A, Nordlund J, et al. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. **Genome Res** 2009; 19:1-11.

Monti L, Cinquetti R, Guffanti A, et al. In silico prediction and experimental validation of natural antisense transcripts in two cancer-associated regions of human chromosome 6. **Int J Oncol** 2009; 34:1099-108.

Neeman Y, Dahary D, Levanon EY, Sorek R, Eisenberg E. Is there any sense in antisense editing? **Trends Genet** 2005; 21:544-7.

Palacios R, Gazave E, Goñi J, et al. Allele-specific gene expression is widespread across the genome and biological processes. **PLoS One** 2009; 4:e4150.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. Analysis of allelic differential expression in human white blood cells. **Genome Res** 2006; 16:331-9.

Pasquinelli AE, Hunter S, Bracht J. MicroRNAs: a developing story. **Curr Opin Genet Dev** 2005; 15:200-5.

Pennisi E. Why do humans have so few genes? **Science** 2005; 309:80.

Pollard KS, Serre D, Wang X, et al. A genome-wide approach to identifying novel-imprinted genes. **Hum Genet** 2008; 122:625-34.

Quéré R, Manchon L, Lejeune M, et al. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression**. Nucleic Acids Res** 2004; 32:e163.

Reik W, Walter J. Genomic imprinting: parental influence on the genome. **Nat Rev Genet** 2001; 2:21-32.

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

Rougeulle C, Lalande M. Angelman syndrome: how many genes to remain silent? **Neurogenetics** 1998; 1:229-37.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. **Methods Mol Biol** 2000; 132:365-86.

Sachidanandam R, Weissman D, Schmidt SC, et al. International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature** 2001; 409:928-33.

Saha S, Sparks AB, Rago C, et al. Using the transcriptome to annotate the genome. **Nat Biotechnol** 2002; 20:508-12.

Serre D, Gurd S, Ge B, et al. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. **PLoS Genet** 2008; 4:e1000006.

Shendure J, Church GM. Computational discovery of sense-antisense transcription in the human and mouse genomes. **Genome Biol** 2002; 3:RESEARCH0044.

Silva AP, Chen J, Carraro DM, Wang SM, Camargo AA. Generation of longer 3' cDNA fragments from massively parallel signature sequencing tags. **Nucleic Acids Res** 2004a; 32:e94.

Silva AP, De Souza JE, Galante PA, Riggins GJ, De Souza SJ, Camargo AA. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. **Nucleic Acids Res** 2004b; 32:6104-10.

Silverman TA, Noguchi M, Safer B. Role of sequences within the first intron in the regulation of expression of eukaryotic initiation factor 2 alpha. **J Biol Chem** 1992; 267:9738-42.

Sleutels F, Zwart R, Barlow DP. The non-coding Air RNA is required for silencing autosomal imprinted genes. **Nature** 2002; 415:810-3.

Smilinich NJ, Day CD, Fitzpatrick GV, et al. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. **Proc Natl Acad Sci U S A** 1999; 96:8064-9.

Thrash-Bingham CA, Tartof KD. aHIF: a natural antisense transcript overexpressed in human renal cancer and during hypoxia. **J Natl Cancer Inst** 1999; 91:143-51.

Tonkin LA, Saccomanno L, Morse DP, Brodigan T, Krause M, Bass BL. RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. **EMBO J** 2002; 21:6025-35.

Tufarelli C, Stanley JA, Garrick D, et al. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. **Nat Genet** 2003; 34:157-65.

Vanhée-Brossollet C, Vaquero C. Do natural antisense transcripts make sense in eukaryotes? **Gene** 1998; 211:1-9.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. **Science** 1995; 270:484-7.

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

Wang XJ, Gaasterland T, Chua NH. Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. **Genome Biol** 2005; 6:R30.

Watanabe T, Totoki Y, Toyoda A, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. **Nature** 2008; 453:539-43.

Waterston R. Genome sequence of the nematode C. elegans: a plataform for investigating biology. **Science** 1998; 282:2012-8.

Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. **Nature** 2002; 420:520-62.

Werner A, Schmutzler G, Carlile M, Miles CG, Peters H. Expression profiling of antisense transcripts on DNA arrays. **Physiol Genomics** 2007; 28:294-300.

Werner A, Sayer JA. Naturally occurring antisense RNA: function and mechanisms of action. **Curr Opin Nephrol Hypertens** 2009; 18:343-9.

Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, Loughlin J. Extreme context specificity in differential allelic expression. **Hum Mol Genet** 2007; 16:537-46.

Will CL, Lührmann R. Spliceosomal UsnRNP biogenesis, structure and function. **Curr Opin Cell Biol** 2001; 13:290-301.

Yamamoto T, Manome Y, Nakamura M, Tanigawa N. Downregulation of surviving expression by induction of the effector cell protease receptor-1 reduces tumor growth potential and results in an increased sensitivity to anticancer agents in human colon cancer. **Eur J Cancer** 2002; 38:2316-24.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. **Science** 2002; 297:1143.

Yan H, Zhou W. Allelic variations in gene expression. **Curr Opin Oncol** 2004; 16:39-43.

Yang HH, Hu Y, Edmonson M, Buetow K, Lee MP. Computation method to identify differential allelic gene expression and novel imprinted genes. **Bioinformatics** 2003; 19:952-5.

Yelin R, Dahary D, Sorek R, et al. Widespread occurrence of antisense transcription in the human genome. **Nat Biotechnol** 2003; 21:379-86.

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

Zhang Y, Liu XS, Liu QR, Wei L. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. **Nucleic Acids Res** 2006; 34:3465-75.

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.

ANEXOS

ANEXO 1

Anexo 1 - RT-PCR fita específica

Na reação de RT-PCR fita específica, a orientação do transcrito é definida restringindo-se qual iniciador específico (senso ou antisenso) estará presente durante a síntese da primeira fita do cDNA. Dessa forma foram realizadas quatro reações de RT-PCR para cada candidato. Na primeira reação foi utilizado apenas o iniciador complementar a orientação senso do transcrito (transcrição senso). Na segunda foi utilizado apenas o iniciador complementar a orientação antisenso do transcrito (transcrição antisenso). Na terceira, nenhum dos iniciadores foi utilizado, servindo assim como controle negativo para a formação de *self-priming* do RNA no momento da síntese de cDNA. Na quarta, foram utilizados os dois iniciadores, entretanto na ausência da enzima transcriptase reversa, o que serviu de controle negativo para a contaminação com DNA genômico.

Assim, 25µg de RNA total das linhagens celulares *Hb4a* ou *Hb4a – C5.2* foi tratado com 10U da enzima *RQ1 RNAse-free DNAse* (*Promega*) e incubado a 37^{o}C durante 30 minutos. Posteriormente, o RNA tratado foi purificado com fenol-clorofórmio 1X, centrifugado a 12.000 rpm a 4^{o}C por 10 minutos. O sobrenadante foi recolhido e submetido a precipitação com 1/10 do volume de acetato de sódio (3M, pH 5,2) e 3 vezes o volume de etanol 100%, a -20^{o}C durante 16 horas. Então, o RNA foi lavado com etanol 70% e, após secagem em temperatura ambiente, foi solubilizado em 20 a 30µl de água DEPC. Após esse processo, o RNA foi quantificado em *Nanodrop* (*ND-1000 Spectrophotometer)*. A ausência de contaminação com DNA genômico foi avaliada, utilizando-se o RNA tratado como molde para a amplificação de um fragmento correspondente a uma região intrônica

(flanqueia o exon 12) do gene *hMLH*-1 (Direto 5' TGGTGTCTCTAGTTCTGG 3' e Reverso 5' CATTGTTGTAGTAGCTCTGC 3'). Assim, de 100 a 500ng do RNA tratado foi submetido a síntese de cDNA. A síntese da primeira fita de cDNA foi feita a 50^{o}C durante 1 hora usando 200U de *Superscript II* (*Invitrogen*) e 0,9µM do iniciador complementar ao transcrito senso ou antisenso para cada candidato. Na mesma reação, também adicionamos 0,9µM do iniciador complementar ao transcrito senso do gene *GAPDH*, o que serviu como controle positivo da síntese de cDNA fita específica.

Para a análise de RT-PCR fita específica os iniciadores foram desenhados manualmente com o auxílio do programa *primer 3* (ROZEN e SKALETSKY 2000). Em conjunto a esse programa, para avaliar a formação de heterodímeros ou *hairpins* e calcular a temperatura de ligação dos iniciadores, também foi utilizado o programa *Oligotech* (http://www.oligosetc.com/analysis.php). A sequência de cada um dos iniciadores e o tamanho dos fragmentos estão apresentados na tabela a seguir.

As reações de PCR foram feitas utilizando-se 1µl do cDNA como molde, 1X tampão da enzima, 1,5mM $MgCl_2$, 0,1mM de dNTPs, 0,4µM dos iniciadores específicos para cada candidato e 1U de *Platinum Taq DNA polymerase* (*Invitrogen*) em um volume final de 25µl. As condições de amplificação foram: desnaturação inicial a 95^{o}C durante 2 minutos, seguido de 35 ciclos de 95^{o}C durante 30 segundos, 60 – 62^{o}C durante 30 segundos, 72^{o}C durante 40 segundos, e após o último ciclo a extensão final a 72^{o}C durante 6 minutos. Após a reação, 5µl do produto da PCR foram carregados em gel de acrilamida 8% e corados com nitrato de prata.

Iniciadores utilizados na RT-PCR fita específica. Sequência dos iniciadores específicos utilizados na RT-PCR fita específica para a validação dos transcritos antisenso nas linhagens celulares *Hb4a* e *Hb4a-C5.2* e o tamanho (pb) do fragmento esperado.

Tag	iniciador direto (5 ´ - 3´)	iniciador reverso (5 ´- 3´)	Fragmento (pb)
03	ACT CAA CAA AAT AGC TGC TG	GCT ACC TGA CGG TTG C	105
09	GAT CAG AAA AAA TCA GCC AAT ATA	GGT TCT TGT AGC TGT TTA TGT	63
13	D as - GAT CTT CAT GAT GGA GGC D s - ATC AAG GAG GCC ATT CTC T	GGT TTT CCC TGG TTC CC	68* 62
17	GCA GAT TCC TTA AGC GAC C	TCC AGA GCT TCT TTT CCC TAA	154
19	CAG CTG GAG AGC TCA GAT GGA	TCC ACC CTC ACT CTG CCA TT	60
20	TCA AAA CGT GTC ACA GCT G	GGG GAA TAT TTG TGG GTA TT	61
24	CTG TAG AGG GAA CAT CAC C	GGA GGA TGG TGT GGA AAC	50
25	CGT CAG CGA CGC GAT GT	CAG TTT CTG CTC CCG GTC AT	54
28	CAC TGC AAC ACG CCA CCT TA	AGG CGC AGT TGT GTA GCA GTT	61
34	ACA ACT GTC TCT GAA TAT TAC C	TTC TGG AAA GGA AAG TTC TAT TAA	145
40	AAT CTC CAG GAG TGA GGG T	AAC TTG GTC TCT GCC CTG	146
41	GAT CAA GAG CAG AGG AGG A	AAC TGC CTT CCT GCC TCT	42
43	CGT GCA GAA GGA GGA CC	AGT GCC TTT ATT GGG AGA CTT	96
44	AAT AAT GCA TTG CCT CTA TCA T	TGT ATT TGA GTC TAA TGC ACG	52
49	GAG CAC ACT ATA TAA ATC CTT TG	ATA TTT TGC TAG GGA AGT GAA AC	92
52	TGT TCG TGG TAG GCT TTC	GCA TCC ACT TCC CTC TG	64
53	ATC CAC ATC ACC GCC TG	CAG TCT GTA AAT GGA AAC TTC	118
58	CGG CAA AAC TAA CTG GTT C	CAA TGA ATG TGG AAT AGA AAA TCT	75
64	GGC TGA CTA TAT TGA CAA GAT	CTG TAT TTT GTA TTG TAT GTT TTC	66
65	CTT AGC AGG ACT GTG GAG	GCT ATT ATA GGA AAC ATC AGG G	110
70	GGG TGT GGT TTT GGA ATG AC	CCT GCA CAT ACT CAC TGG A	237
72	CCT TTA GAC AGA ATC TGA GAT	AGG AGC ACC TCT ATA CAG	108
77	GTT GTC AGT TGG GAT GGA C	GCA GCC TCT TTG ACC TAA AC	128
83	TGT AAG ATG TGA GAG GTG TTG A	CTC TTA AAG GCA AGG GTT GAA	178
87	GAT CGA TGG TGT TAC TCA G	TTC ACA AGC TAT TCC CTC AAA	50
94	ATC GTC AGG CGG CTT G	AGA TGG GTT TTT AAT GAT ACT AAG	70
95	CAC AAA AAA AGA TGG AAG TGG	CCT CAA AGG ACC CTT CTT	90

*: Das – direto antisenso, Ds – direto senso.