UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Ciências Biológicas
(Bioquímica)

PEDRO A. F. GALANTE

Uso de técnicas computacionais no estudo da transcrição e regulação gênica em *Homo sapiens* e *Mus musculus*

Orientadora: BETTINA MALNIC
Co-orientador: SANDRO J. DE SOUZA

São Paulo
07/12/2007
PEDRO A. F. GALANTE

Uso de técnicas computacionais no estudo da transcrição e regulação gênica em *Homo sapiens* e *Mus musculus*

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

Orientadora: BETTINA MALNIC
Co-orientador: SANDRO J. DE SOUZA

São Paulo

07/12/2007
Pedro A. F. Galante

Uso de técnicas computacionais no estudo da transcrição e regulação gênica em Homo sapiens e Mus musculus.

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

Aprovado em: ____________

Banca Examinadora

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________

Prof. Dr. _________________________________________________________
Instituição: _______________________________________________________
Assinatura: _______________________________________________________
Aos meus pais, Pedro e Sonia.
“Não se aprende, senhor, na fantasia, sonhando, imaginando ou estudando, senão vendo, tratando e pelejando”

Luis Vaz de Camões em “Os Lusíadas”
AGRADECIMENTOS

Quando deixei a minha cidade de origem, São Pedro-SP, para iniciar a graduação aqui na USP, sabia que a minha vida mudaria rapidamente. Primeiro, veio o BCC (bacharel em ciência da computação) e seus desafios. Eram muitos, mas eu queria algo maior. Então, veio o convite para prestar o Curso de Ciências Moleculares (CCM). Vi no CCM um ambiente extremamente propício para eu aprender muitas coisas novas de diversas áreas científicas. O CCM me deu uma ótima formação, sem dúvida a melhor que eu poderia ter conseguido. Na hora de escolher uma área para me iniciar cientificamente, vi na Biologia Computacional e na Bioinformática uma extensão natural do CCM, uma maneira de eu continuar envolvido com muitas áreas do conhecimento, além do fato destas áreas serem extremamente interessantes e com muito a ser feito.

Bem, vamos ao objetivo desta seção, os agradecimentos. Primeiro, agradeço à FAPESP pela bolsa concedida e à minha orientadora, a Bettina Malnic, e ao meu co-orientador, o Sandro José de Souza, por me darem todo o suporte e ambiente necessário para o meu crescimento técnico e científico. Vocês me ensinaram o que é e como é pensar cientificamente, vocês também me instruíram no que deve e no que não deve ser feito na ciência, obrigado! Também agradeço à Anamaria Camargo, à Helena Paula Brentani, à Lucila Ohno-Machado, ao Gregory Riggins, à Janete Cerutti e ao Winston Patrick Kuo pelas orientações pontuais e inúmeras colaborações. Todos vocês também foram muito importantes para a minha formação científica.

Ainda na parte científica, agradeço aos meus colegas da turma 9 do CCM – uma turma para ficar na história – e aos meus colegas dos laboratórios por onde passei. Em especial eu gostaria de agradecer ao Noboru Jô Sakabe pela muitas (boas) conversas sobre ciência e por me ajudar no...
desenvolvimento do meu senso critico; ao Luiz Ricardo Hanai pelas ótimas conversas e amizade durante o ciclo básico do CCM; ao Jorge Estefano de Souza pela amizade e pelas muitas colaborações científicas; ao Robson Francisco de Souza pela amizade e pelas discussões técnicas e científicas; ao Elisson Campus Osório pelas dicas técnicas no inicio da minha carreira e pelas inúmeras dicas de empreendedorismo; ao Arthur de Souza Ramos pelo suporte; à Carina Victoria Manzini Baldi pelas horas de estudo durante a prova de ingresso na Pós-Graduação, pela nossa ótima amizade e por ler cuidadosamente esta tese; à Maria Dulcetti Vibranovski pela amizade e pelas nossas boas conversas; à Ana Claudia Pereira pelo suporte; à Natanja Kirschbaum-Slager pela amizade; à Elza Helena Barbosa pela convivência no laboratório. Com todos vocês eu aprendi muitas coisas e passei inesquecíveis e bons momentos.

Por fim, eu agradeço à minha (pequena) família, os meus pais, Pedro e Sonia, duas pessoas extraordinárias, que não tiveram chance de estudar, mas desde cedo perceberam a importância do estudo e permitiram que eu chegasse até aqui. Sem vocês dois nada disso seria possível! Também agradeço à minha, digamos, futura família: a Daniela, a Bete, o Carlos, a Bianca, o Elisson, a Débora e o Lucas. Vocês participaram de uma boa parte deste processo. Em especial, eu gostaria de agradecer à Daniela, uma artista completa, extremamente dedicada e competente, com quem foi muito fácil e prazeroso passar os últimos cinco anos. Dani, eu te admiro muito, você é uma pessoa muito especial.
RESUMO

Galante, P.A.F., Uso de técnicas computacionais no estudo da transcrição e regulação gênica em *Homo sapiens e Mus musculus*. 2007. (134 pág.) Tese (Doutorado) – Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

O gene, uma sequência de nucleotídeos necessária para a síntese de moléculas funcionais, é transcrito e regulado por um conjunto de processos e fatores extremamente complexos. Entender o momento e o tecido em que os genes são expressos, as isoformas funcionais, as regiões controladoras e os fatores envolvidos na regulação da expressão de cada gene é um dos grandes desafios da biologia molecular moderna. Hoje, com a enorme quantidade de informações de sequências genômicas e de transcriptomas, aliado ao desenvolvimento de métodos computacionais para agrupar e analisar estes dados em larga escala, o estudo dos fenômenos relacionados à transcrição e regulação gênica está passando por uma revolução. Por exemplo, é possível medir, concomitantemente, a expressão gênica de milhares de genes em diferentes tecidos, assim como identificar diversos fenômenos que atuam nestes genes. Neste trabalho nós desenvolvemos e aplicamos métodos computacionais no estudo de quatro temas envolvendo aspectos chave da transcrição e regulação gênica. No primeiro trabalho, nós abordamos a expressão gênica tecido-específica através do estudo dos genes expressos no cérebro e em dez regiões cerebrais de camundongo. No segundo trabalho, nós identificamos sequências potencialmente envolvidas no controle da transcrição gênica através do estudo de motivos sobre representados na região promotora dos genes de receptores olfativos. No terceiro trabalho, analisamos o transcriptoma humano quanto a presença de eventos de retenção de intron, um tipo de splicing.
alternativo. No quarto trabalho, nós abordamos a complexidade do transcriptoma e a regulação da expressão gênica através do estudo de pares de genes senso-antisenso em humanos e camundongos. Em todos os trabalhos, obtivemos resultados que nos permitiram tirar conclusões específicas sobre cada fenômeno estudado e nos mostraram a importância de estudá-los através de uma abordagem em larga escala. Adicionalmente, verificamos que os nossos métodos computacionais foram eficientes e adequados para o estudo da transcrição e regulação gênica em *Homo sapiens* e *Mus musculus*.

**Palavras-chave:** expressão gênica, bioinformática, cérebro, receptores olfativos, *splicing* alternativo, transcritos antisenso.
ABSTRACT

Galante, P.A.F., *Use of computational methods to study the transcription and gene regulation in Homo sapiens and Mus musculus*. 2007. (134 pag) PhD Thesis – Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Genes, nucleotide sequences necessary for the synthesis of functional molecules, are transcribed and regulated by extremely complex cellular and molecular processes. To understand when and in which tissues the genes are expressed, their functional isoforms, control regions and the factors involved in gene regulation is one of major challenges of modern molecular biology. Today, the availability of complete genome sequences and transcriptomes, together with the development of new computational methods allows the study of phenomena related to the transcription and gene regulation in a large scale. For example, it is possible to quantify, concomitantly, gene expression of thousands of genes in different tissues and analyze different aspects of their regulation. In this work we developed and applied computational methods to the study of four key aspects of gene transcription and regulation. In the first study, we addressed tissue specific gene expression through the study of genes that are preferentially expressed in the brain and ten different mouse brain regions. In the second study, we identified sequences that are potentially involved in the control of gene transcription through the study of motifs that are over represented in the promoter region of olfactory receptor genes. In the third study, we browsed the human for the presence of intron retention, a type of alternative splicing. In the fourth study, we addressed the transcriptoma complexity and gene expression regulation through the study of pair of sense-antisense genes in human and mouse. In all studies, our results allowed us to
make specific conclusions about each phenomenon analyzed which showed us the importance of a large scale approach. In addition, we verified that our computational methods can be efficiently applied to the study of transcription and gene regulation in Homo sapiens and Mus musculus.

**Keywords:** gene expression, bioinformatics, brain, olfactory receptors, alternative splicing, antisense transcripts.
SUMÁRIO

1 INTRODUÇÃO .......................................................................................................................... 13

1.1 O TRANSCRIPTOMA DOS EUCARIOTOS SUPERIORES .............................................................. 13
1.2 GENES EXPRESSOS NO CÉREBRO DE CAMUNDONGO ......................................................... 16
1.3 ANÁLISE DA REGIÃO PROMOTORA DOS ORs ................................................................. 18
1.4 SPICING ALTERNATIVO TIPO RETENÇÃO DE INTRON ..................................................... 20
1.5 TRANSCRITOS ANTISENSO NATURAIS (NAT) ................................................................. 21
1.6 O ESTUDO EM LARGA ESCALA DA TRANSCRIÇÃO GÊNICA ............................................. 24

2 OBJETIVOS ................................................................................................................................. 25

2.1 GERAIS .................................................................................................................................. 25
2.2 ESPECÍFICOS .......................................................................................................................... 25
2.2.1 Estudo dos genes expressos no cérebro de camundongo ...................................................... 25
2.2.2 Estudo da região promotora dos ORs .................................................................................. 25
2.2.3 Estudo dos genes apresentando retenção de intron .............................................................. 26
2.2.4 Estudo dos transcritos antisenso naturais ........................................................................... 26

3 MATERIAIS E MÉTODOS ............................................................................................................ 27

3.1 DADOS PRIMÁRIOS .............................................................................................................. 27
3.1.1 Genomas .......................................................................................................................... 27
3.1.2 cDNAs ............................................................................................................................ 27
3.1.3 Dados de SAGE e de MPSS ............................................................................................. 28
3.1.4 Dados de hibridização in-situ .......................................................................................... 28
3.2 ALINHAMENTO DOS cDNAS NO GENOMA ....................................................................... 29
3.3 AGRUPAMENTO DOS cDNAS ............................................................................................... 31
3.4 MAPEAMENTO DAS TAGS DE SAGE E DE MPSS ............................................................. 32
3.5 IDENTIFICAÇÃO DOS TRANSCRITOS PREFERENCIALMENTE EXPRESSOS NO CÉREBRO ......................................................................................................................... 34
3.5.1 Quantificação da expressão gênica ...................................................................................... 34
3.5.2 Agrupamento hierárquico dos genes expressos no cérebro ................................................ 35
3.5.3 Ontologia dos genes expressos no cérebro .......................................................................... 35
3.5.4 Seleção dos genes preferencialmente expressos nas regiões do cérebro ................................ 36
3.5.5 Validação experimental da expressão gênica por hibridização in-situ ................................. 36
3.6 ANÁLISE DA REGIÃO PROMOTORA DOS RECEPTORES OLFATIVOS ............................ 37
3.6.1 Amplificação completa da região 5’ dos genes ORs ............................................................. 37
3.6.2 Definição da região promotora dos genes ORs ..................................................................... 38
3.6.3 Mapeamento das ORFs dos ORs e agrupamento das sequências ....................................... 38
3.6.4 Splicing alternativo ........................................................................................................... 39
3.6.5 Busca de motivos conservados na região promotora dos genes ORs .................................. 39
3.7 IDENTIFICAÇÃO DOS EVENTOS DE RETENÇÃO DE INTRON .......................................... 40
3.7.1 Identificação dos eventos de splicing alternativo ................................................................. 40
3.7.2 Análise dos eventos de retenção de intron .......................................................................... 41
3.7.3 Análise dos IRs conservados no genoma de camundongo .................................................. 43
3.8 ESTUDO DOS TRANSCRITOS ANTISENSO NATURAIS ...................................................... 43
3.8.1 Identificação dos pares senso (S) antisenso (AS) ................................................................. 43
3.8.2 Análise dos pares S-AS ........................................................................................................ 44
3.8.3 Identificação de novos NAT através das tags de MPSS ....................................................... 46
3.8.4 Conservação do par S-AS entre homem e camundongo .................................................... 47
3.8.5 Identificação dos variantes de splicing alternativo que contém um NAT ............................. 48
3.8.6 Validação experimental dos NATs ...................................................................................... 49

4 RESULTADOS ........................................................................................................................... 50

4.1 IDENTIFICAÇÃO E ANÁLISE DOS GENES EXPRESSOS PREFERENCIALMENTE NO CÉREBRO DE CAMUNDONGO .......... 50
4.1.1 Os genes expressos no cérebro de camundongo .............................................. 50
4.1.2 A expressão gênica nas regiões cerebrais ......................................................... 58
4.1.3 Ontologia dos genes expressos no cérebro ......................................................... 61
4.1.4 Seleção dos genes expressos preferencialmente no cérebro e nas regiões cerebrais ............................................................................................................ 63
4.1.5 Os genes expressos preferencialmente no bulbo olfativo .................................. 64
4.1.6 Confirmação experimental da expressão gênica no OB ...................................... 67
4.2 IDENTIFICAÇÃO DE SEQUÊNCIAS REGULADORAS NOS PROMOTORES DOS GENES ORs .......................................................... 70
4.2.1 Amplificação da região 5' completa dos genes ORs .............................................. 70
4.2.2 Agrupamento dos cDNAs e definição dos genes ORs ........................................... 71
4.2.3 A região 5' dos genes OR ...................................................................................... 73
4.2.4 Identificação dos motivos na região promotora dos ORs ..................................... 74
4.3 DETECÇÃO E ANÁLISE DOS EVENTOS DE RETENÇÃO DE INTRON ........................................... 78
4.3.1 Identificação dos eventos de retenção de intron ................................................. 79
4.3.2 Distribuição dos eventos de IR ao longo do transcrito ........................................... 80
4.3.3 Dominios protêicos codificados por introns retidos .......................................... 81
4.3.4 Pressão seletiva nos introns retidos ...................................................................... 83
4.3.5 Conservação dos introns retidos em camundongo ............................................. 85
4.4 ANÁLISE FUNCIONAL E EVOLUTIVA DOS PARES DE GENES SENSE-ANTISENSE EM HOMEM E CAMUNDONGO .......... 86
4.4.1 Identificação dos pares S-AS ............................................................................... 86
4.4.2 Organização genômica dos pares S-AS ............................................................. 88
4.4.3 Conservação dos pares S-AS entre humanos e camundongos ............................. 90
4.4.4 Padrão de expressão dos pares S-AS ............................................................... 92
4.4.5 Influência dos NATs no processo de splicing alternativo ................................... 93
4.4.6 Utilizando MPSS na identificação de novos NAT .............................................. 96
4.4.7 A poliadenilação alternativa na definição de novos pares S-AS ......................... 98
4.4.8 Os NAT originados de primer interno ............................................................. 99

5 DISCUSSÃO .................................................................................................................. 102
5.1 ESTUDO DOS GENES EXPRESSOS NAS REGIÕES DO CÉRE布O DE CAMUNDONGO .......................................................... 102
5.1.1 A complexidade da expressão gênica no cérebro ............................................. 102
5.1.2 Genes expressos preferencialmente nas regiões do cérebro ................................ 103
5.1.3 Genes expressos no bulbo olfativo ................................................................. 104
5.2 IDENTIFICAÇÃO DE SEQUÊNCIAS REGULADORAS NOS PROMOTORES DOS GENES ORs .......................................................... 106
5.2.1 Motivos presentes nas regiões promotoras dos genes ORs ............................... 106
5.2.2 A região promotora e a regulação da expressão dos genes ORs ......................... 108
5.3 IDENTIFICAÇÃO DOS EVENTOS DE RETENÇÃO DE INTRONS ........................................... 109
5.3.1 Características dos introns retidos ...................................................................... 109
5.3.2 Considerações evolutivas dos introns retidos .................................................... 112
5.4 ANÁLISE FUNCIONAL E EVOLUTIVA DOS PARES DE GENES SENSE-ANTISENSE EM HOMEM E CAMUNDONGO .......... 113
5.4.1 Os pares S-AS ................................................................................................. 114
5.4.2 A organização genômica dos pares S-AS ........................................................ 114
5.4.3 A evolução dos pares S-AS ............................................................................... 115
5.4.4 A conservação dos pares S-AS entre humanos e camundongos ....................... 116
5.5 O ESTUDO EM LARGA ESCALA DA TRANSCRIÇÃO GÊNICA .................................................. 117

6 CONCLUSÃO .................................................................................................................. 119
6.1 ESTUDO DOS GENES EXPRESSOS NO CÉREBRO DE CAMUNDONGO .......................................................... 119
6.2 ESTUDO DAS SEQUÊNCIAS REGULADORAS NOS PROMOTORES DOS GENES ORs .......................................................... 119
6.3 ESTUDO DOS EVENTOS DE RETENÇÃO DE INTRONS ........................................... 120
6.4 ESTUDO DOS PARES S-AS EM HOMEM E CAMUNDONGO ........................................... 120
6.5 ESTUDO DA TRANSCRIÇÃO GÊNICA ATRAVÉS DE TÉCNICAS COMPUTACIONAIS .......................................................... 121

7 REFERÊNCIAS ............................................................................................................... 122

LISTA DE ANEXOS ......................................................................................................... 132
1 Introdução

1.1 O transcriptoma dos eucariotos superiores

Em termos moleculares, o gene é definido como uma sequência de nucleotídeos necessária para a síntese de uma molécula funcional, a qual pode ser uma cadeia de polipeptídio ou uma molécula de RNA (Snyder e Gerstein, 2003). Entender o momento e o tecido em que os genes são expressos, as isoformas funcionais, as regiões controladoras e os fatores envolvidos na regulação de cada gene, é o grande desafio da biologia molecular moderna. Se no passado se estudava um ou pouquíssimos genes ao mesmo tempo, hoje, com a combinação de técnicas e experimentos em larga escala, dados públicos e análises computacionais, é possível estudar concomitantemente muitos ou todos os genes de um organismo (Rockman e Kruglyak, 2006).

Basicamente, os genes podem ser classificados em dois tipos: i) ubiquamente expressos; ii) com expressão diferencial e restrita a determinados tecidos (Alberts, 2007). Enquanto o primeiro conjunto é responsável por processos comuns à todos os tecidos, tais como os processos metabólicos e fisiológicos básicos, o segundo desempenha funções celulares e moleculares específicas ao tecido onde são preferencialmente expressos (Su, Cooke et al., 2002), por exemplo, o gene GRP (gastrin releasing peptide), expresso apenas no núcleo lateral da amígdala, está evolvido na formação de memórias que desencadeiam o medo (Shumyatsky, Tsvetkov et al., 2002). Apesar da sua importância, muitos genes com expressão diferencial ainda não foram identificados, ou não apresentam uma função conhecida. O estudo do transcriptoma e, em especial, dos genes com expressão restrita deve contribuir para um ganho significativo no entendimento das funções dos órgãos e tecidos onde estes genes são expressos.
O primeiro passo na expressão gênica é a transcrição de uma região genômica, um processo complexo que se inicia com a ligação de fatores de transcrição (fatores em *trans*) aos elementos reguladores (seqüências em *cis*) presentes nas regiões promotoras dos genes (Uptain, Kane *et al.*, 1997). Estas interações de fatores em *trans* e seqüências em *cis* podem determinar o momento e a intensidade em que um gene é transcrito. Apesar da dificuldade de se identificar fatores em *trans*, a detecção dos fatores em *cis* é mais direta e depende de métodos computacionais aliados a validações experimentais (Wasserman e Sandelin, 2004). O estudo de seqüências reguladoras em *cis* contribuirá para a decodificação da complexidade da regulação da expressão gênica em diversos tecidos.

Inesperadamente, não se confirmou uma relação direta entre o número de genes codificadores e a complexidade do organismo (Claverie, 2001). Por exemplo, o homem apresenta ~30k genes codificadores (Lander, Linton *et al.*, 2001; Venter, Adams *et al.*, 2001), o camundongo ~30k (Waterston, Lindblad-Toh *et al.*, 2002), a drosófila ~15k (Adams, Celniker *et al.*, 2000) e o verme *C. elegans* ~19k genes (Waterston, 1998). Uma possível explicação para esta falta de correlação está nos eventos pós transcripcionais: a edição de RNA, a edição de cadeias polipeptídicas, a poliadenilação alternativa e o *splicing* alternativo (Pennisi, 2005). Destes fenômenos, o *splicing* alternativo é o maior provedor de proteínas distintas por gene, afetando, por exemplo, mais da metade dos genes humanos (Modrek e Lee, 2002). A identificação de todas as isoformas de um gene é um passo essencial para se desvendar a complexidade do transcriptoma e se entender o papel funcional de cada gene.

Apesar de muitos genes serem transcritos em moléculas de RNAs que serão traduzidas em sequências polipeptídicas, outros geram seqüências não codificadoras (ncRNAs). Os exemplos mais conhecidos são os rRNA, tRNA (Eddy, 2001). Contudo, análises em larga escala do transcriptoma vem revelando uma abundância surpreendente dos RNAs não codificadores (Storz, 2002). Hoje, são conhecidos os miRNA (Lee, Feinbaum *et al.*, 1993), snRNAs (Will e Luhrmann, 2001), snoRNAs
Entre os ncRNA, muitos são mRNAs e, no geral, transcritos a partir da fita oposta de uma região genômica que contém um outro gene (Kumar e Carmichael, 1998). Estes genes são conhecidos como transcritos naturais antisenso (NATs) (vale destacar que nem todos os NAT são ncRNAs, estima-se que 20% sejam codificadores (Kiyosawa, Yamanaka et al., 2003)). Os NATs foram primeiramente identificados em procariotos (Wagner e Simons, 1994), mas hoje se sabe que eles também são muito comuns nos eucariotos (Zhang, Liu et al., 2006), incluindo o homem (Chen, Sun et al., 2004; Katayama, Tomaru et al., 2005; Zhang, Liu et al., 2006), e estão envolvidos essencialmente na regulação da expressão do gene definido como senso (Reis, Nakaya et al., 2004; Yan, Hong et al., 2005; Nakaya, Amaral et al., 2007; Osato, Suzuki et al., 2007). Identificar e estudar estes transcritos reguladores da expressão gênica é uma tarefa essencial para se entender a complexidade do controle e da expressão gênica.

Hoje a biologia molecular está passando por uma revolução. Os diversos projetos de sequenciamento de genomas e do transcriptoma produziram uma quantidade extraordinária de informação (Carninci, 2007). Assim como, a utilização da computação na exploração, em larga escala, dos resultados produzidos e no direcionamento dos experimentos a serem realizados (Butler, 2001). A junção destas duas áreas deu origem a uma nova maneira de se estudar muitos dos processos biológicos, ajudou a confirmar a complexidade de vários fenômenos celulares e nos mostrou o quanto ainda precisa ser estudado (Kitano, 2002). Sem dúvida, a exploração completa deste mundo molecular só será possível através da junção de técnicas computacionais poderosas e de métodos e experimentos bem direccionados e conclusivos.

Nesta tese estão descritos quatro estudos relacionados à expressão gênica em Homo sapiens e Mus musculus. No primeiro, estudamos os genes expressos preferencialmente no cérebro e nas regiões cerebrais de camundongo. No segundo, analisamos a presença de elementos reguladores nas regiões
promotoras dos genes de receptores olfativos de camundongo. No terceiro, estudamos a ocorrência de splicing alternativo (retenção de intron) nos genes humanos. Por fim, estudamos a presença de genes antisenso no genoma humano e de camundongo. Os principais experimentos em todos os estudos foram realizados através de técnicas computacionais de análise em larga escala.

1.2 Genes expressos no cérebro de camundongo

Algumas das funções específicas e parte da morfologia celular dos diversos órgãos do corpo são conseqüência de genes com expressão diferencial, restrita e localizada (Zhang, Zhou et al., 1997). Por exemplo, sem uma expressão restrita a determinados períodos não seria possível o desenvolvimento correto da maioria dos tecidos do nosso corpo (Bruhn e Cepko, 1996; Chen, Yu et al., 2004; Al-Bader e Al-Sarraf, 2005). Sem uma expressão génica diferencial e localizada não teríamos as diferentes proteínas nos diferentes tipos celulares e, consequentemente, os órgãos e tecidos não apresentariam toda a complexidade funcional que hoje conhecemos (Storch, Lipan et al., 2002).

Entre os órgãos, o cérebro é o que apresenta a maior diversidade funcional e molecular (Hatten e Heintz, 2005). Por exemplo, cada região do cérebro apresenta uma função específica: o hipocampo está envolvido com o resgate de memória declarativa e espacial; o hipotálamo com o controle homeotáxico do corpo; o bulbo olfativo com a discriminação de odores e feromônios (Albright, Jessell et al., 2000). Do ponto de vista molecular, trabalhos da década de 80 já mostravam que o cérebro é o órgão que apresenta a maior diversidade de transcritos (Chaudhari e Hahn, 1983; Milner e Sutcliffe, 1983; Hatten e Heintz, 2005).
Hoje, alguns trabalhos vêm demonstrando a existência de genes expressos especificamente ou preferencialmente no cérebro e em suas regiões. Por exemplo, Piao e colaboradores (Piao, Hill et al., 2004) mostraram que a expressão do gene Gpr56 em células neurais progenitoras é essencial para o desenvolvimento do córtex humano. Sakurai e colaboradores (Sakurai, 2007) mostraram que o gene orexina é expresso apenas no hipotálamo e que está envolvido no controle da fome e do ciclo sono/vigília. É lógico imaginar que a identificação destes genes pode produzir insights sobre o funcionamento de cada região do cérebro e pode contribuir para o entendimento do sistema nervoso como um todo. Entretanto, nenhum dos poucos trabalhos existentes, em larga escala, sobre a expressão gênica nas diferentes regiões cerebrais abordou seis ou mais regiões distintas ao mesmo tempo.

Entre as regiões do cérebro, o bulbo olfativo (OB) é uma das mais interessantes. O OB desempenha uma função central na detecção dos odores, recebendo informações dos neurônios que detectam o odor (neurônios contendo os receptores olfativos) e enviando-as para o córtex olfativo e áreas límbicas do cérebro (Buonviso e Chaput, 1990). Além desta importante tarefa, o OB é muito estudado porque: i) na maioria dos mamíferos está posicionado numa região cerebral de fácil acesso, na parte frontal do cérebro (Kandel, Schwartz et al., 2000); ii) apresenta regiões bem definidas, com células distintas (Shepherd, 2003); iii) junto com o hipocampo, é a única região que apresenta neurogênese nos mamíferos adultos (Gould, 2007); iv) está envolvido diretamente em questões importantes e ainda não solucionadas, tal como o mecanismo responsável por guiar os axônios do epitélio olfativo à alvos específicos no OB (Lin, Yang et al., 2004).

Um dos estudos descritos nesta tese é a expressão gênica no cérebro de camundongos, com ênfase no OB (ver seções relacionadas a “Expressão gênica no cérebro”). Neste trabalho, identificamos, em larga escala, os genes expressos em 10 regiões do cérebro de camundongo, selecionamos aqueles com expressão preferencial em cada região e analisamos diversas suas
características. Entre os genes com expressão preferencial, estudamos em detalhes aqueles expressos no OB.

1.3 Análise da região promotora dos ORs

A discriminação dos odores é realizada por receptores protéicos (receptores olfativos – ORs) presentes nos cílios dos neurônios olfativos da mucosa nasal (Buck e Axel, 1991). Cada neurônio olfativo expressa apenas um entre muitos genes que codificam para estes receptores (Chess, Simon et al., 1994; Malnic, Hirono et al., 1999). Em roedores, por exemplo, há mais de mil escolhas possíveis (Godfrey, Malnic et al., 2004); no homem este repertório é de aproximadamente 350 (Malnic, Godfrey et al., 2004). Além disso, o receptor que cada neurônio expressa não determina apenas o reconhecimento do odor, mas também a convergência do axônio neuronal à uma região específica do bulbo olfativo (Mombaerts, 2006). Entretanto, apesar da escolha do OR ser essencial para a organização funcional do sistema olfativo e uma questão interessantíssima sob o ponto de vista de regulação da expressão gênica, ainda não se conhece o mecanismo que a determina.

Nos últimos anos surgiram algumas hipóteses para explicar essa expressão de apenas um receptor olfativo por neurônio, mas algumas já foram descartadas e outras ainda necessitam ser melhor exploradas. A modificação irreversível do DNA através de rearranjos gênicos foi a primeira hipótese a ser descartada. Eggan e colaboradores (Eggan, Baldwin et al., 2004) e Li e colaboradores (Li, Ishii et al., 2004) mostraram que camundongos gerados a partir da transferência do núcleo de neurônios do epitélio olfativo apresentam um sistema olfativo normal, com o repertório completo de receptores. Recentemente também foi proposta a existência de uma LCR (região controladora de expressão),
denominada região H (de homóloga), a qual poderia conduzir a expressão de um único OR por célula (Lomvardas, Barnea et al., 2006). Entretanto, está hipótese também foi descartada (Fuss, Omura et al., 2007).

Por outro lado, vários estudos com camundongos transgênicos definiram regiões reguladoras envolvidas na expressão dos genes ORs. Vassalli e colaboradores (Vassalli, Rothman et al., 2002) e Rothman e colaboradores (Rothman, Feinstein et al., 2005) definiram pequenas regiões (de 161 e 395 nt upstream ao sítio de início da transcrição (TSS)) necessárias para a expressão dos ORs M71 e MOR23, respectivamente. Nestas regiões foram detectados sítios de ligação em cis para os fatores de transcrição homeodomain e fatores similares ao O/E (O/E-similar). Também foi mostrado que mutações nestes sítios alteram a expressão destes ORs (Rothman, Feinstein et al., 2005).

Apesar destes estudos, pouco se conhece sobre a influência de seqüências reguladoras em cis no controle da expressão dos ORs. Uma dificuldade inerente a este tipo de estudo é a incapacidade de se definir corretamente a região promotora dos ORs. Nos bancos de dados públicos, a grande maioria das seqüências de ORs corresponde à região codificadora e quase não existem cDNAs com a região 5’ completa (apenas para ~30 genes). Outra dificuldade é construir uma biblioteca de cDNAs full length de um conjunto representativo de ORs.

Um dos estudos descrito nesta tese é a análise da região promotora de 198 genes ORs de camundongo (ver as seções relacionadas a “Receptores olfativos” e o artigo “Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences” no anexo). Neste trabalho, definimos experimentalmente o início da região promotora dos genes ORs, determinamos a estrutura exon-intron da região 5’ UTR destes genes, identificamos algumas seqüências (motivos) conservadas entre as regiões promotoras de 198 genes e, por fim, mostramos que alguns dos motivos identificados se ligam especificamente às proteínas do núcleo de neurônios do epitélio olfativo.
1.4 Splicing alternativo tipo retenção de intron

O splicing alternativo é o maior responsável pela diversidade existente entre os transcritos de um mesmo gene (Black, 2003). O exemplo mais conhecido é o gene dscam de drosófila, o qual pode, em potencial, dar origem a 38,016 transcritos codificadores (Schmucker, Clemens et al., 2000). Quanto ao significado biológico do splicing alternativo, existem alguns trabalhos, feitos gene a gene, mostrando que este fenômeno é importante na formação e manutenção de diversos organismos. Por exemplo, são determinados por variantes de splicing: i) as formas solúveis e insolúveis da proteína fibronecitinga; ii) o sexo das drosófilas; iii) a detecção de som nas galinhas (Black, 2003).

Atualmente, o estudo de splicing alternativo está passando por uma revolução causada pelo sequenciamento de genomas completos, pelo acumulo de sequências de transcritos (especialmente as ESTs) e pelo desenvolvimento de técnicas computacionais para a integração e análise em larga escala deste enorme conjunto de dados. Por exemplo, hoje se sabe que as leveduras apresentam pouquíssimos exemplos ou não apresentam splicing alternativo e que os humanos e camundongos apresentam 60-80% e 40-50% de genes com splicing alternativo, respectivamente (Modrek e Lee, 2003).

Existem quatro tipos principais de splicing alternativo: uso alternativo de exon, sítio alternativo 3’, sítio alternativo 5’ e retenção de intron. O uso alternativo de exons é o tipo mais estudado e mais freqüente. Por outro lado, a retenção de intron é, sem dúvida, o tipo menos freqüente e menos estudado de splicing alternativo (Modrek e Lee, 2002). Entre os fatores que levam a esta discrepância, pode-se destacar: i) o fato de que as isoformas de retenção de intron podem ser confundidas com pré-mRNAs não processados ou processados parcialmente; ii) muitos dos cDNA contendo introns retidos são descartados (não submetidos aos bancos públicos) por apresentarem um códon de parada prematuro.
Mas, apesar disto, este tipo de splicing alternativo é extremamente importante e existem alguns exemplos funcionais de seus variantes (Rio, 1991; Le Hir, Charlet-Berguerand et al., 2002).

Um dos estudos descrito nesta tese é a análise de retenção de intron nos genes humanos (seções relacionadas a “Retenção de intron” e artigo “Detection and evaluation of intron retention events in the human transcriptome” no anexo). Neste trabalho buscamos, em larga escala, isoformas de retenção de intron em 21,106 genes humanos e estudamos algumas características funcionais e evolutivas destes variantes, tais como a capacidade de se tornarem regiões codificadoras e a conservação em camundongo.

1.5 Transcritos antisenso naturais (NAT)

Os transcritos naturais antisenso (NATs) são transcritos endógenos com regiões complementares a outros transcritos. Existem dois tipos de NATs: cis, onde o transcrito antisenso localiza-se na mesma região genômica que o gene senso; e o trans, onde o antisenso localiza-se numa região genômica distinta do gene senso (Kumar e Carmichael, 1998). Em geral, o tipo cis apresenta uma longa região de sobreposição e está relacionado a um ou poucos genes senso. O tipo trans, por outro lado, apresenta uma região de sobreposição curta e imperfeita e pode estar relacionado a vários genes senso. O tipo NAT-cis é o mais frequente em mamíferos (Vanhee-Brossollet e Vaquero, 1998; Lavorgna, Dahary et al., 2004).

Estudos pioneiros em vários eucariotos mostraram três mecanismos pelos quais os NATs podem regular a expressão dos genes senso, a interferência transcricional, o mascaramento de RNA e a formação citoplasmática de RNA dupla fita (Lavorgna, Dahary et al., 2004). A interferência
transcripcional ocorre porque a mesma região genômica não comporta dois complexos de transcrição (RNA polimerase II e proteínas de enovelamento do DNA) caminhando ao mesmo tempo e em direções opostas nas fitas de DNA. A ocorrência desta regulação foi mostrada primeiramente em levedura pelo rearranjo dos genes GAL10 e GAL7, onde se mediu a expressão destes genes quando eles estavam separados e, em seguida, numa região sobreposta (depois do rearranjo). Na segunda arquitetura, o nível de expressão dos transcritos foi reduzido drasticamente (Prescott, Osheim et al., 2004).

O segundo mecanismo é o mascaramento de RNA. Neste mecanismo os RNAs senso e antisenso se ligam, formando um RNA dupla fita e impedindo a ligação dos fatores em trans aos elementos reguladores de ambas as sequências. O resultado é o comprometimento dos processos de splicing, poliadenilação, transporte, tradução e estabilidade do RNA mensageiro. Um exemplo deste mascaramento ocorre com os genes ErbAα (senso) e Rev-ErbAα (antisenso). Quando o gene antisenso Rev-ErbAα não está expresso, aparecem as duas formas, ErbAα1 e ErbAα2. Mas, quanto o Rev-ErbAα está expresso, a isoforma ErbAα2 deixa de ser encontrada (Munroe, 1988).

O terceiro mecanismo é a interferência de RNA (RNAi), a qual é desencadeada pela formação de um RNA dupla fita no citoplasma de células eucariotas. Neste processo, a enzima Dicer cliva, em sequências curtas de 21-23 pb, as moléculas de RNA dupla fita. Estes fragmentos, por sua vez, podem auto alimentar esse processo de degradação, ligando-se a outras moléculas de RNAs, formando novamente RNA dupla fita e causando a degradação de ambas as moléculas (Lavorgna, Dahary et al., 2004).

Os NATs foram inicialmente identificados através de estudos gene-a-gene. Contudo, com o sequenciamento de genomas e com o acúmulo de dados de transcritos (os cDNAs) em bancos públicos, aliado ao desenvolvimento de métodos computacionais, hoje é possível identificar e estudar uma
enorme quantidade de transcritos antisenso. Por exemplo, alguns estudos mostraram que, no mínimo, 20% e 15% dos genes humanos (Chen, Sun et al., 2004) e de camundongo (Kiyosawa, Yamanaka et al., 2003) apresentam transcritos antisenso, respectivamente. Estudos mais recentes elevaram estes números para ~50% em humanos (Engstrom, Suzuki et al., 2006) e em camundongo (Katayama, Tomaru et al., 2005).

O maior obstáculo na identificação em larga escala dos NATs é definir corretamente a orientação das seqüências públicas de cDNAs, especialmente as ESTs. Muitas ESTs não são clonadas direcionalmente ou apresentam uma anotação incorreta de orientação. O método usualmente empregado para definir a orientação destes cDNAs leva em conta sinais intrínsecos ao transcrito, como a cauda de poli(A) ou a cabeça de poli(T) (quando a seqüência está invertida), e sinais presentes na seqüência de DNA, como a presença dos sítios canônicos de splicing, GT-AG e CT-AC (orientação inversa). Usando estes dois conjuntos de sinais é possível definir a orientação de muitos dos cDNAs dos bancos públicos.

Um dos estudos descritos nesta tese é sobre a ocorrência dos NAT em homem e camundongo (ver seções relacionadas aos “Transcritos Antisenso Naturais” e artigo em anexo “Sense-antisense pairs in mammals: functional and evolutionary considerations”). Nós identificamos novos NAT-cis, selecionamos os conhecidos e analisamos algumas características destes genes, em homem e camundongo, sob um aspecto funcional e evolutivo.
1.6 O estudo em larga escala da transcrição gênica

A enorme quantidade de informações de expressão gênica, seqüências de transcritos e seqüências genômicas direcionou o desenvolvimento de uma nova maneira para se estudar os problemas biológicos, a análise em larga escala (Eisen, Spellman et al., 1998). Se há alguns anos os biólogos moleculares tentavam entender o funcionamento de um único gene, ou a ocorrência de um determinado fenômeno em poucos genes, hoje é possível estudar o funcionamento de milhares de genes e medir a ocorrência dos mais diversos fenômenos em todos eles.

Se mal planejadas, as análises em larga escala podem ser pouco informativas, simples ou sem um objetivo claro. Entretanto, quando bem planejadas e direcionadas por uma pergunta biológica específica ela nos permite estudar de maneira profunda os problemas (fenômenos) biológicos. Entre os passos essenciais para uma boa análise em larga escala estão: i) o catálogo de todos os elementos que compõe o conjunto de interesse; ii) o estudo da origem e da evolução dos elementos e do conjunto como um todo; iii) o descobrimento de maneiras pelas quais cada elemento e o conjunto total é controlado; iv) a identificação das alterações nos elementos e a conseqüência para o conjunto total (Kitano, 2002). Com estes passos, os estudos em larga escala são sempre profundos e geram resultados conclusivos.

Nesta tese nós desenvolvemos e aplicamos métodos computacionais no estudo em larga escala de quatro aspectos envolvendo a transcrição gênica em Homo sapiens e Mus musculus (como descrito acima). Em todos os quatro trabalhos, nós tentamos extrair resultados profundos e conclusivos através de análise em larga escala feitas de uma maneira sistemática, planejada e buscando responder perguntas com um objetivo claro.
2 Objetivos

2.1 Gerais

Desenvolver e aplicar técnicas e métodos computacionais para se estudar aspectos da transcrição gênica em *Homo sapiens* e em *Mus musculus*, abordando suas características mecanísticas, funcionais e evolutivas.

2.2 Específicos

2.2.1 Estudo dos genes expressos no cérebro de camundongo

Estudar a expressão gênica no cérebro e nas regiões cerebrais de *Mus musculus*, buscando os genes expressos preferencialmente em cada uma das regiões cerebrais e, em especial, no bulbo olfativo.

2.2.2 Estudo da região promotora dos ORs

Estudar comparativamente as regiões promotoras e as regiões 5’ UTR dos genes OR de *Mus musculus*, buscando sequências conservadas e envolvidas, em potencial, na regulação da expressão dos OR.
2.2.3 Estudo dos genes apresentando retenção de intron

Identificar os genes humanos que apresentam retenção de intron e estudar características funcionais e evolutivas deste tipo pouco comum de splicing alternativo.

2.2.4 Estudo dos transcritos antisenso naturais

Identificar novos pares de genes senso-antisenso no genoma humano e de camundongo, catalogar os pares senso-antisenso conhecidos e estudar características funcionais e evolutivas destes genes.
3 Materiais e Métodos

3.1 Dados primários

3.1.1 Genomas

As sequências dos genomas de *Homo sapiens* e *Mus musculus* foram obtidas de dois repositórios: do NCBI (ftp://ftp.ncbi.nih.gov) e do Genome Browser da Universidade da Califórnia em Santa Cruz (ftp://hgdownload.cse.ucsc.edu/). No trabalho em que estudamos os casos de IR (Galante, Sakabe *et al*., 2004) foi utilizada a versão 29 (NCBI *build* #29) do genoma humano. No trabalho em que estudamos a região promotora dos gene ORs (Michaloski, Galante *et al*., 2006) foi utilizada a versão 33 do genoma de camundongo. No trabalho em que estudamos os transcritos senso-antisenso (Galante, Vidal *et al*., 2007) foi utilizada a versão 35 do genoma humano e a versão 33 do genoma de camundongo. No trabalho em que estudamos os genes expressos no cérebro foi utilizada a versão 36 do genoma de camundongo.

3.1.2 cDNAs

As sequências de cDNAs também foram obtidas do NCBI e do UCSC Genome Browser. No trabalho em que estudamos os casos de IR (Galante, Sakabe *et al*., 2004) as ESTs foram obtidas do
banco dbEST (Boguski, Lowe et al., 1993) e os mRNAs foram obtidos do UniGene (http://www.ncbi.nlm.nih.gov/UniGene). No trabalho em que estudamos a região promotora dos genes ORs (Michaloski, Galante et al., 2006) as seqüências das regiões codificadoras dos ORs foram obtidas diretamente no site do NCBI (identificadores AY072961-AY074256 e AY317244-AY318733 como descrito em (Zhang e Firestein, 2002; Young, Shykind et al., 2003)). No trabalho em que estudamos os genes S-AS (Galante, Vidal et al., 2007) e no trabalho em que estudamos os genes expressos no cérebro de camundongo, as ESTs e os mRNAs foram obtidas no UCSC Genome Browser (ftp://hgdownload.cse.ucsc.edu/), arquivos: est.fa (ESTs), mrna.fa e refMrna.fa (mRNAs).

**3.1.3 Dados de SAGE e de MPSS**

As bibliotecas de SAGE foram obtidas em http://cgap.nci.nih.gov/SAGE/ (Boon, Osorio et al., 2002). As bibliotecas de MPSS de *Homo sapiens* foram obtidas em http://mpss.licr.org. As bibliotecas de MPSS de camundongo foram obtidas no *Mouse Transcriptome Project* do NCBI (http://www.ncbi.nlm.nih.gov/projects/geo/info/mouse-trans.html).

**3.1.4 Dados de hibridização *in-situ***

Os dados de hibridização *in-situ* foram obtidos do Allen Brain Institute. A expressão de cada gene foi analisada manualmente com base nas imagens de cortes coronais e sagitais do cérebro completo de camundongos C57BL/6J adultos.
O Allen Brain Institute foi fundado em 2003 com o objetivo de descrever a expressão de todos os genes no cérebro de camundongo (Lein, Hawrylycz et al., 2007) (http://www.brainatlas.org). Para isto, utilizaram a técnica de hibridização in-situ semi-automática em larga escala (Visel, Thaller et al., 2004). Foram feitas in-situ com mais de 21.000 genes no cérebro de camundongos (Lein, Hawrylycz et al., 2007) e todos estes dados estão disponíveis (http://www.brainatlas.org).

3.2 Alinhamento dos cDNAs no genoma

Atualmente, fazer o alinhamento de uma seqüência de cDNA (ou algumas dezenas) no genoma humano é um processo rápido e simples, basta submeter às seqüências de interesse a um dos muitos portais contendo programas de alinhamento. Entretanto, alinhar alguns milhões de cDNAs num genoma complexo, como o dos mamíferos, requer computadores robustos e ferramentas eficientes de alinhamento. Por exemplo, são necessários 10 computadores (PC-AMD64 2,7G Hertz e 2GB de memória RAM) trabalhando ininterruptamente por ~7 dias para se alinhar ~5 milhões de cDNAs no genoma humano ( dado não apresentados).

Entre as ferramentas de alinhamento, somente o BLAT (Kent, 2002), BLAST (Altschul, Madden et al., 1997) e seus derivados (por exemplo o WU-BLAST (Lopez, Silventoinen et al., 2003) e o MEGABLAST (Zhang, Schwartz et al., 2000)) conseguem realizar em um tempo aceitável o alinhamento de milhões de seqüências de cDNA em genomas complexos. Entretanto, apesar de velozes, estes programas não apresentam alinhamentos com a precisão necessária para se estudar, por exemplo, a estrutura exon-intron dos genes. Tal precisão pode ser obtida, por exemplo, com o programa sim4 (Florea, Hartzell et al., 1998) e com o programa est2genome (Mott, 1997).
Como em todos os nossos trabalhos foi necessária a definição precisa da estrutura exon-intron, desenvolvemos um método, que alia a velocidade do BLAST ou do BLAT e a precisão do sim4, para alinhar milhões de cDNAs no genoma humano e de camundongo. Resumidamente, o processo se inicia com um BLAT ou BLAST das ESTs e dos mRNAs contra o genoma. No segundo passo, estes alinhamentos são processados e as coordenadas de alinhamento de cada cDNA são identificadas. No terceiro passo, cada cromossomo é dividido continuamente em pedaços de 500k bases, com uma sobreposição de 200k bases entre eles, e um programa identifica os cDNAs que melhor (mais equidistantes das bordas) se encaixam nestes fragmentos genômicos. Em seguida, outro programa recebe o arquivo FASTA do cDNA e do fragmento genômico e os alinha usando o sim4. No último passo, um conjunto de programas seleciona apenas os alinhamentos com a identidade e a cobertura pré-definidas, e gera um arquivo para ser carregado no banco de dados.

Para mostrar a eficiência do nosso protocolo, nós o comparamos com o alinhamento direto dos cDNAs no genoma. Utilizamos 1000 sequências de mRNAs humanos. Com o nosso método, o alinhamento ficou pronto em ~2 min; alinhando diretamente (como o sim4) no genoma, o tempo foi de ~54 min. Utilizando o nosso método e 10 computadores PCs, em ~7 dias nós conseguimos alinhar com precisão ~8 milhões de cDNAs no genoma humano e ~4 milhões de cDNAs no genoma de camundongo.

Para finalizar todo o processo, armazenamos o resultado do alinhamento num banco de dados relacional. Neste banco, como um único comando é possível descobrir a posição exata de qualquer gene conhecido no genoma humano, o número de exons deste gene, a posição das regiões codificadoras e UTR, por exemplo. Além das posições de mapeamento, também armazenamos algumas informações sobre as sequências de cDNA, tais como: o tipo da sequência (mRNA ou EST), o tamanho, o tecido de origem, informações patológica, etc.


3.3 Agrupamento dos cDNAs

O agrupamento (*clustering*) é um processo que analisa um conjunto de elementos e forma subconjuntos com aqueles que compartilham certas características. Aplicando este método aos cDNAs alinhados no genoma, agrupamos aqueles que compartilham, essencialmente, a mesma região genômica.

Atualmente, três projetos fazem o agrupamento dos cDNAs públicos e disponibilizam o resultado de maneira organizada e acessível: o UniGene (Boguski e Schuler, 1995), o TIGR Gene Index (Quackenbush, Liang *et al.*, 2000) e o Strack Pack (Burke, Wang *et al.*, 1998). Dos três, o UniGene é o mais utilizado pela comunidade. Entretanto, analisando cuidadosamente os clusters do UniGene, percebemos que muitos são falso-positivos ou falso-negativos (Galante, PAF, Osório, EC e de Souza, SJ; manuscrito não publicado). Para evitar problemas com clusters e podermos definir e controlar todos os parâmetros deste método, nós resolvemos criar o nosso próprio protocolo de agrupamento de cDNAs.

Basicamente, o nosso protocolo utiliza o genoma como referência e agrupa as sequências alinhadas na mesma região genômica. São agrupadas as sequências que: i) compartilham ao menos uma borda exon-intron; ii) sequências com sobreposição exclusiva dos exons terminais ou iniciais; iii) sequências sem introns e com sobreposição. É impedido o agrupamento de sequências transcritas de fitas opostas do DNA e de sequência que compartilham a mesma região genômica, mas não compartilham ao menos uma borda exon/intron.

Escolhemos estes critérios depois de estudar outros protocolos de agrupamento (UniGene; TIGR Gene Index) e de analisar manualmente muitos *clusters* gerados pelo nosso e por outros protocolos.
3.4 Mapeamento das tags de SAGE e de MPSS

SAGE (Velculescu, Zhang et al., 1995) e MPSS (Brenner, Johnson et al., 2000) determinam o nível de expressão gênica através de seqüências curtas, tags, extraídas de uma posição específica do mRNA, o sítio mais 3’ da enzima *NlaIII* (SAGE) e *DpnII* (MPSS). Nestas metodologias, o nível de expressão dos genes é proporcional à freqüência das tags (Velculescu, Zhang et al., 1995; Brenner, Johnson et al., 2000).

Uma biblioteca de MPSS, em média, apresenta 1,1 milhão de tags de 17 ou 20 nt. Já as bibliotecas de SAGE, em média, contém 120 mil tags. Portanto, através do enorme conjunto de tags, MPSS, em potencial, identifica todos os transcritos de uma célula (estimado em 300 mil (Zhang, Zhou et al., 1997)) e mostra a expressão de genes não identificados por SAGE.

Um experimento de SAGE e de MPSS gera uma biblioteca que contém um conjunto de tags e a freqüência de cada uma. Então, o primeiro passo para se analisar estes dados é identificar (mapear) o gene que cada tag representa.

Basicamente, o mapeamento das tags pode ser realizado de duas maneiras: i) através de uma anotação, *tag a tag*, usando websites como o SAGEGenie (Boon, Osorio et al., 2002) e o SAGEmap (Lash, Tolstoshev et al., 2000); ii) através de uma anotação, em larga escala, baseada em banco de dados e ferramentas computacionais construídas localmente (no laboratório do pesquisador).

Aproveitando parte do *know-how* adquirido anteriormente (ver Boon e colaboradores (Boon, Osorio et al., 2002)), e a necessidade de utilizarmos freqüentemente os dados de SAGE e, em especial, de MPSS, nós construímos um protocolo (método) de mapeamento das tags. Este método, chamado por nós de ACTG, foi publicado na revista *Bioinformatics* (Galante, Trimarchi et al., 2007).
Resumidamente, o ACTG (Automatic Correspondence of Tags and Genes) é um conjunto de programas na linguagem Perl + CGI, shell scripts, códigos HTMLs e vários bancos de dados de tags virtuais (uma predição, feita no computador, das tags potencialmente produzidas num experimento de SAGE e de MPSS). Todos podem utilizá-lo no endereço http://retina.med.harvard.edu/ACTG. Para
mais detalhes sobre o método do ACTG, ver (Galante, Trimarchi et al., 2007) e a seção help do website http://retina.med.harvard.edu/ACTG.

Em todos os trabalhos aqui descritos, as tags de SAGE e de MPSS foram mapeadas com o método do ACTG diretamente no website ou localmente no nosso laboratório.

3.5 Identificação dos transcritos preferencialmente expressos no cérebro

Abaixo estão descritos os métodos utilizados no trabalho em que estudamos os genes expressos no cérebro de camundongo.

3.5.1 Quantificação da expressão gênica

A expressão gênica foi medida através das tags de MPSS de 81 bibliotecas correspondendo a 45 tecidos de camundongos C57BL/6J adultos (ver a Tabela 1). Para o mapeamento das tags foi utilizado o protocolo do ACTG (ver seção “3.4 – Mapeamento da tags de SAGE e de MPSS”).
3.5.2 Agrupamento hierárquico dos genes expressos no cérebro

O método de agrupamento hierárquico junta elementos com características similares através de dois passos principais: i) medir as distâncias (correlação) entre os elementos (no nosso caso, os genes) e formar grupos com os pares mais próximos (pseudo-grupos); ii) agrupar (aglomerar) os pseudo-grupos mais próximos, formando grupos maiores. Estes dois passos se repetem até que todos os grupos sejam comparados entre si e que sejam formados todos os grupos (Eisen, Spellman et al., 1998).

Nós aplicamos este método a todos os genes expressos nos 45 tecidos de camundongos e utilizamos a distância Euclidiana como medida de correlação e o Average Linkage como parâmetro de aglomeração. Os cálculos e a apresentação gráfica dos resultados foram feitos com a biblioteca Heatplus do pacote estatístico R (http://www.r-project.org/).

3.5.3 Ontologia dos genes expressos no cérebro

A classificação de ontologia dos genes de camundongo foi baseada nos dados produzidos pelo consórcio Gene Ontology (GO: http://www.geneontology.org). Os dados foram obtidos diretamente do website do projeto (http://www.geneontology.org) e armazenados num banco de dados local. Construímos programas para manipulá-los e extraímos a anotação para os genes de interesse.
3.5.4 Seleção dos genes preferencialmente expressos nas regiões do cérebro

A seleção dos genes expressos preferencialmente nas regiões do cérebro foi feita através de uma simulação de Monte Carlo aplicada ao dado de expressão por MPSS. Basicamente, este processo consiste em: i) definir a expressão de cada gene nas 9 outras regiões do cérebro e nos outros 35 tecidos/órgãos não cerebrais (nos casos onde há duas bibliotecas para o mesmo tecido, utilizamos a expressão média); ii) comparar através de simulação a expressão dos genes de cada biblioteca cerebral contra a expressão deste gene nas outras regiões do cérebro e também em todos os outros tecidos não cerebrais; iii) selecionar os genes sobre expressos (valor \( p < 0.05 \)) em cada uma das 10 regiões do cérebro.

Para selecionar os genes preferencialmente expressos no cérebro como um todo, independente da região, nós comparamos a expressão de cada biblioteca cerebral contra a expressão nos outros tecidos e evitamos a comparação direta entre cada uma das 10 regiões cerebrais.

3.5.5 Validação experimental da expressão gênica por hibridização \textit{in-situ}

Os dados de hibridização \textit{in-situ} foram obtidos do Allen Brain Institute, como descrito na seção “3.1.4 – Dados de hibridização \textit{in-situ}”. Todos os genes expressos preferencialmente no bulbo olfativo foram analisados manualmente e, entre todas as imagens, armazenamos apenas aquelas onde as expressões do gene estavam mais intensas.
3.6 Análise da região promotora dos receptores olfativos

Abaixo estão descritos os métodos utilizados no trabalho em que estudamos os motivos conservados nas regiões promotoras dos genes ORs.

3.6.1 Amplificação completa da região 5’ dos genes ORs

O método descrito nesta seção foi feito pela aluna Jussara Michaloski. Achei necessário descrevê-lo, pois este é o primeiro passo no processo de análise da região promotora dos genes ORs. O objetivo deste método é gerar ESTs contendo a região completa da extremidade 5’ UTR dos genes ORs (EST 5’ completas).

Sucintamente, esse método se inicia com uma 5’ RLM-RACE do RNA total de epitélio olfativo de camundongos C57BL/6J adultos. Em seguida, usando primers degenerados, os quais reconhecem a sexta ou sétima hélice transmembranar de muitos ORs, amplifica-se a região 5’ UTR e uma parte da região codificadora. Por fim, para melhorar a especificidade da seleção dos ORs, realiza-se uma nested-PCR. Na Figura 12 há um esquema geral do método. Além disso, maiores detalhes podem ser obtidos em (Michaloski, Galante et al., 2006). Com este protocolo, obtivemos 434 sequências de 198 genes ORs.
3.6.2 Definição da região promotora dos genes ORs

Por definição, a região promotora fica *upstream* do início de transcrição (TSS), podendo se estender de centenas a milhares de nucleotídeos (Kozak, 1999). Para determinar a região promotora dos genes ORs, nós alinhamos as ESTs 5’ completas no genoma de camundongo, agrupamos as seqüências (ver seção “3.3 – Agrupamento dos cDNAs”) e definimos o TSS do gene como sendo a primeira base do exon mais 5’ do *cluster*. As seqüências foram alinhadas como descrito na seção “3.2 – Mapeamento dos cDNAs no Genoma”.

3.6.3 Mapeamento das ORFs dos ORs e agrupamento das seqüências

Além de alinhar a ESTs, também alinhamos as ORFs preditas de todos os genes ORs de camundongo (ver seção “3.1.2 cDNAs”). Agrupando (ver seção “3.3 – Agrupamento dos cDNAs”) as ORFs e as ESTs 5’ completa, conseguimos definir, além da região promotora, a região codificadora dos ORs.

3.6.3.1 Análise dos genes ORs
Depois do mapeamento e agrupamento dos cDNAs, nós construímos programas para analisar algumas características dos genes ORs. Por exemplo, identificamos a presença dos genes OR por cromossomos, a extensão da região 5’ UTR, o número de exons e introns na região 5’ UTR, o tamanho dos exons e introns, a ocorrência de inícios de transcrição alternativos, a ocorrência de ATGs upstream ao originalmente predito e a presença de uORFs (ORFs pequenas e upstream à ORF original).

3.6.4 Splicing alternativo

Nós também analisamos splicing alternativo nos genes ORs. Para isso, utilizamos o método descrito na seção “3.6 – Identificação dos eventos de splicing alternativo”.

3.6.5 Busca de motivos conservados na região promotora dos genes ORs

Após definir o TSS, extraímos uma seqüência de 600 nt upstream ao início da transcrição de cada um dos ORs (a região promotora). Em seguida, estas seqüências foram analisadas por três programas, o Gibbs Recursive Sampler (Thompson, Rouchka et al., 2003), o Consensus (Hertz e Stormo, 1999) e o Weeder (Pavesi, Mereghetti et al., 2004), os quais identificam seqüências curtas, motivos, sobre representadas no conjunto de regiões promotoras.
3.7 Identificação dos eventos de retenção de intron

Abaixo estão descritos os métodos utilizados no trabalho em que estudamos a retenção de intron no genoma humano.

3.7.1 Identificação dos eventos de splicing alternativo

Primeiramente, todos os cDNAs humanos (4.533.159 ESTs do dbEST e 89.300 mRNAs) foram alinhados no genoma segundo o método descrito na seção “3.2 – Alinhamento de cDNAs no Genoma”. Em seguida, estes cDNAs foram agrupados segundo o método descrito na seção “3.3 – Agrupamento dos cDNAs”. O banco de dados resultante destes dois passos foi utilizado na busca dos eventos de splicing alternativo, em especial de evento de retenção de intron.

Os casos de retenção de intron foram identificados por programas que analisaram todos os transcritos de todos os clusters e identificaram aqueles que contêm ao menos uma região definida como intron segundo os outros cDNAs (ver Figura 16). Apenas eventos de IR contendo ao menos uma sequência de mRNA (sequência de melhor qualidade, quando comparadas com as ESTs) foram considerados válidos. Para eliminar sequências provenientes de bibliotecas de cDNAs contaminadas por DNA genômico, foi exigido que todas as sequências apresentassem múltiplos exons, sobretudo as sequências contendo os introns retidos.
3.7.2 Análise dos eventos de retenção de intron

Depois de identificados os eventos de retenção de intron, foram construídos programas para analisar as características deste tipo de splicing alternativo (ver abaixo).

3.7.2.1 Distribuição dos eventos ao longo do transcrito

Para todos os genes com uma ORF definida, foi analisada a presença de eventos de retenção de introns nas regiões 5’ UTR, 3’ UTR e na própria região codificadora. A presença do intron foi confirmada quando a sua posição mais 5’ estava contida na região determinada. Para evitar qualquer possível viés introduzido pelas ESTs (sequências preferencialmente geradas a partir das regiões 5’ UTR e 3’ UTR dos genes), nesta análise foram utilizados apenas eventos de retenção de intron definidos por mRNAs.

3.7.2.2 Análise da seqüência dos introns retidos.
Foram construídos programas para avaliar o conteúdo de GC e a presença de códons de parada nas sequências dos introns retido e também nos introns e exons não envolvidos com eventos de splicing alternativo.

3.7.2.3 Uso de códons nos exons e introns retidos

Em colaboração com o aluno Noboru Jô Sakabe, construímos programas para analisar o uso de códons nos exons e introns retidos na região codificadora dos genes. Como controle, também analisamos o uso de códons nos introns não envolvidos em retenção.

3.7.2.4 Análise dos domínios protéicos codificados pelos introns retidos

Também em colaboração com o aluno Noboru Jô Sakabe, analisamos a presença de domínios protéicos contidos nos introns retidos. Foram utilizados os dados do Pfam 9.0. Para evitar qualquer viés introduzido pelo tamanho do intron retido, nós excluímos introns menores do que 80 nt e maiores do que 990 nt. Um domínio foi considerado presente no intron retido quando, ao menos, 30% da sua sequência apresentou alinhamento (e-value <10^{-1}) com a sequência intrônica.
3.7.3 Análise dos IRs conservados no genoma de camundongo

Assim como em *Homo sapiens*, nós identificamos os eventos de splicing alternativo em *Mus musculus* (ver seção 3.6.1 – Identificação dos eventos de splicing alternativo) e, em seguida, selecionamos os eventos de retenção de intron conservados entre as duas espécies. Um evento de retenção de introns foi considerado conservado quando o gene ortólogo em camundongo também apresentou ao menos um cDNA mostrando a retenção do mesmo intron retido em humano.

3.8 Estudo dos transcritos antisenso naturais

Abaixo estão descritos os métodos utilizados no trabalho em que estudamos os pares de genes senso-antisenso nos genomas de camundongo e humano.

3.8.1 Identificação dos pares senso (S) antisenso (AS)

O banco de dados contendo os genes senso-antisenso foi construído através de quatro passos principais. Primeiramente, utilizamos os métodos descritos na seção “3.2 – Mapeamento de cDNAs no genoma” para mapear no genoma todos os cDNAs disponíveis para *Homo sapiens* e *Mus musculus*. Para humano, foram mapeadas 5.992.459 ESTs e 186.358 mRNAs na versão 35 do genoma. Para camundongo, foram mapeadas 4.246.824 ESTs e 120.058 mRNAs na versão 33 do genoma.
Em seguida, definimos a orientação de todos os cDNAs mapeados. Para as ESTs, utilizamos como evidência de orientação dois tipos de sinais: i) a cauda de poli(A) ou a cabeça de poli(T), respectivamente, um conjunto de, no mínimo, oito adenosinas na região mais três’ UTR, ou oito timinas na região 5’ UTR da sequência de cDNA; ii) a presença exclusiva dos sítios canônicos de splicing, GT-AG ou CT-AC (o inverso do GT-AG), nos alinhamentos das sequências contendo múltiplos exons. Todas as ESTs sem ao menos um destes sinais foram descartadas. Para os mRNAs, nós simplesmente corrigimos a orientação daqueles que apresentaram os sítios de splicing CT-AC. Nenhum mRNA foi descartado.

No terceiro passo, agrupamos estes cDNAs resultantes. Utilizamos o método como descrito na seção “3.3 – Agrupamento de cDNAs”, com a adição de mais um passo: um programa adicional para evitar o agrupamento de sequências com orientações opostas.

No quarto e último passo, selecionamos os genes com sobreposição, segundo o mapeamento genômico, e posicionados em fitas opostas do DNA (os pares senso-antisenso em cis). Essa tarefa foi realizada através de um conjunto de programas que selecionou e comparou a orientação e as coordenadas de mapeamento de todos os genes presentes no banco de dados de *Homo sapiens* e de *Mus musculus*.

### 3.8.2 Análise dos pares S-AS

Depois de identificados os pares senso-antisenso, nós construímos diversos programas para analisar a organização genômica, a região de sobreposição, a estrutura exon/intron e expressão destes genes (ver abaixo).
3.8.2.1 Análise da organização genômica dos pares S-AS

Os genes senso-antisenso foram classificados de acordo com a posição dos genes no genoma. Quando encontramos apenas um gene senso e outro antisenso, o par foi classificado como bidirecional simples (aqui chamado bidirecional). Quando encontramos dois ou mais genes antisenso a outro(s) gene(s), nós classificamos estes transcritos como pares bidirecionais múltiplos.

Os pares bidirecionais foram classificados segundo a orientação dos dois genes: 5’-5’, 3’-3’ e totalmente incluídos (ver a Figura 19).

3.8.2.2 A estrutura exon-intron e a região de complementaridade dos pares S-AS

Os pares bidirecionais também foram classificados de acordo com a posição da região complementar: exônica, quando a região complementar está contida exclusivamente no exon; intrônica, quando a região complementar está exclusivamente contida no intron; exônica-intrônica, quando a região complementar passa pela borda exon-intron de um dos genes (ver Figura 19).

Para avaliar se o resultado desta classificação poderia ter sido gerado ao acaso, nós fizemos uma simulação de Monte Carlo, onde foi mantida a estrutura de um dos genes e incluído aleatoriamente na sua região o outro transcrito. Esse experimento foi repetido 1000 vezes para cada par
e em todas às vezes foi identificada à região de sobreposição. No final, obtivemos um $\chi^2$ e calculamos o valor $p$.

Também classificamos os pares bidirecionais segundo a presença de introns (definidos pelo alinhamento do cDNA no genoma) nos genes senso-antisenso: ambos com intron, quando o senso e o antisenso apresentam introns; intron-sem-intron, quando apenas um dos genes apresenta intron; ambos sem-intron, quando nenhum dos genes apresenta intron (ver Figura 19).

3.8.2.3 Expressão dos pares S-AS

A expressão dos genes senso-antisenso foi medida através das tags de MPSS. Foram utilizados os dados de MPSS descritos na seção “3.1.3 – Dados de SAGE e de MPSS” e as tags foram mapeadas com o método descrito na seção “3.2 – Mapeamento da tags de SAGE e de MPSS”. Com este dado de expressão, nós classificamos os pares senso-antisenso em co-expressos (quando ambos os genes estavam expressos em mais de quatro bibliotecas de MPSS) ou não co-expressos (quando ambos os genes estavam expressos em menos do que três bibliotecas de MPSS ou quando apenas um dos genes estava expresso).

3.8.3 Identificação de novos NAT através das tags de MPSS
A identificação de novos transcritos antisenso através das tags de MPSS foi feita em três etapas. Primeiro, mapeamos no genoma humano e de camundongo as tags de MPSS (são 340.820 tags distintas de *Homo sapiens* e 56.061 tags distintas de *Mus musculus*). O mapeamento das tags no genoma é feito através da seleção, *in silico*, de todos os sítios genômicos da enzima de restrição DpnII e de uma comparação exata da tag adjacente a cada sítio com as tags das bibliotecas experimentais de MPSS.

Na segunda etapa, nós cruzamos as coordenadas do mapeamento genômico das tags de MPSS com as coordenadas do mapeamento genômico dos genes que não apresentaram um transkrito antisenso e selecionamos apenas as tags de MPSS mapeadas na orientação inversa. Esse cenário indica que a tag de MPSS (em potencial) vem de um transkrito antisenso ao gene no qual ela está mapeada.

Finalizando, 96 destas tags foram submetidas à validação experimental para confirmarmos a sua presença, orientação e identificarmos o transkrito ao qual ela pertence (ver logo abaixo).

### 3.8.4 Conservação do par S-AS entre homem e camundongo

A identificação dos pares S-AS conservados entre homem e camundongo foi feita de três maneiras. Na primeira, identificamos os pares S-AS conservados nas duas espécies através da identificação dos genes ortólogos entre *Homo sapiens* e *Mus musculus*. A identificação dos genes ortólogos foi feita através do HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene). Como o HomoloGene utiliza a seqüência protéica para a identificação dos ortólogos, este método não nos permitiu identificar a conservação dos pares S-AS onde apenas um os genes é codificador.
No segundo método, nós partimos de todos os pares S-AS de *Homo sapiens*, identificamos aqueles onde um dos genes é conservado (novamente, utilizando o HomoloGene) e testamos localmente a conservação do outro gene do par. Consideramos uma conservação real quando a identidade do alinhamento global entre as seqüências foi maior do que 30%. Desta maneira, também conseguimos identificar a conservação dos pares S-AS onde temos um gene codificador e outro não codificador.

No terceiro método, utilizamos o dado de homologia do *UCSC Genome Browser database*. Neste método a seleção das regiões homólogas é feita através do alinhamento de cromossomos inteiros, um método custoso do ponto de vista computacional, mas que produz resultados robustos (ver *UCSC Genome Browser database*). O nosso método consiste em verificar se os loci gênicos homólogos entre *Homo sapiens* e *Mus musculus* (definidos pelos dados da UCSC) apresentam pares S-AS em ambas as espécies. Se isto ocorrer, estes pares S-AS são considerados ortólogos.

### 3.8.5 Identificação dos variantes de *splicing* alternativo que contém um NAT

Utilizando o método descrito na seção “3.6 – Identificação dos eventos de *splicing* alternativo”, nós selecionamos os transcritos gerados por eventos de *splicing* alternativo. Essa busca foi realizada em todos os genes contendo seqüências com orientação definida. Para analisar melhor a co-ocorrência do NAT e de *splicing* alternativo, nós dividimos as bordas exon-intron dos eventos de *splicing* alternativo em duas categorias: i) bordas que apresentam ao menos um transcrito antisenso completar; ii) bordas que não apresentam um transcrito antisenso completar.
3.8.6 Validação experimental dos NATs

O experimento (validação experimental) descrito nessa seção foi feito pelo aluno Daniel Onofre Vidal. Foram dois os objetivos dessa validação experimental: i) confirmar alguns dos transcritos antisenso identificados apenas pelas tag de MPSS; ii) confirmar se alguns NAT foram gerados por primers genómico.

Para a confirmação dos transcritos antisenso identificados pelas tag de MPSS foi utilizada a técnica de GLGI (Silva, De Souza et al., 2004). Essencialmente, está técnica expande a tag de MPSS e gera uma seqüência que vai da própria tag até a cauda de poli(A) do gene que ela representa. Foram utilizadas amostras de linhagem celular HB4A de glândula mamária normal. Para mais detalhes, ver (Silva, De Souza et al., 2004) e (Galante, Vidal et al., 2007)

Na segunda etapa, foi testada a influência da contaminação de primer genómico na geração de NATs. Foram feitas RT-PCRs em amostras de figado fetal, colon e pulmão. Resumidamente, nessa análise foi testada a veracidade de 11 NATs. Primeiro, foi feita uma RT-PCR com RNA tratado com DNAse e com RNA não tratado com DNAse. Como controle, foi repetido o experimento com e sem a presença da enzima transcriptase reversa e, também como controle, foi adicionado DNA genómico ao RNA de algumas das reações. Para mais detalhes, ver (Galante, Vidal et al., 2007).
4 Resultados

4.1 Identificação e análise dos genes expressos preferencialmente no cérebro de camundongo

Neste trabalho analisamos o transcriptoma (conjunto dos genes expressos) no cérebro e em dez regiões cerebrais de camundongos C57BL/6J adultos. Para isto, utilizamos dados de expressão gênica em larga escala, métodos computacionais e dados de validação experimental através de hibridização in situ.

4.1.1 Os genes expressos no cérebro de camundongo

Para estudar o transcriptoma de camundongo e selecionar os genes expressos no cérebro e nas regiões cerebrais, partimos de 81 bibliotecas de MPSS de 45 tecidos diferentes, incluindo 10 regiões cerebrais (Tabela 1). Todas essas bibliotecas foram construídas a partir de RNAs poli(A)$^+$ de tecidos de camundongos C57BL/6J adultos. Apenas para avaliar a comparabilidade entre as bibliotecas, quantificamos o total de tags em cada uma: esse número varia de 1.096.689 (coração: ventrículo e septo) a 1.384.689 tags (músculo esquelético). Na média o número é similar (1.232.628 tags para cada
biblioteca), o que nos permite uma comparação igualitária entre as 81 bibliotecas. Mas, para garantir uma precisão maior, todas as bibliotecas foram normalizadas por 1 milhão de *tags*.

### Tabela 1 | Conjunto total de bibliotecas de MPSS de camundongos.
As bibliotecas de cérebro estão destacadas. Os símbolos “|” e “||” indicam bibliotecas de fêmea e de macho, respectivamente. A coluna “Total” representa o número de *tags* em cada biblioteca. A coluna “Distintas” representa o número de *tags* diferentes em cada biblioteca.

| Tecidos                  | Tags  |                  | Tags  |                  | Tags  |                  |
|--------------------------|-------|------------------|-------|------------------|-------|------------------|
|                          | Distintas | Total          | Distintas | Total          | Distintas | Total          |
| Adrenal||              | 10226     | 1.281.909   | Gordura marrom||              | 8917     | 1.293.330   | Glândula mamária||              | 10.389    | 1.230.414   |
| Adrenal|              | 12187     | 1.225.954   | Gordura marrom||              | 7883     | 1.272.802   | Ovário||              | 13.332    | 1.229.036   |
| Bexiga||              | 12554     | 1.239.178   | Cartilagem||              | 9646     | 1.245.771   | Pituitária||              | 7291      | 1.256.845   |
| Bexiga|              | 9149      | 1.279.893   | Cartilagem||              | 9014     | 1.186.482   | Pituitária||              | 6308      | 1.203.656   |
| Osso - Fêmur|      | 7714      | 1.325.633   | Cérvice e vagina||              | 7142     | 1.210.864   | Placenta - E18||              | 11.187    | 1.145.153   |
| Osso - Fêmur|      | 7135      | 1.279.948   | Embrionário E18||              | 11.343    | 1.282.965   | Músculo esq.||              | 5929      | 1.380.576   |
| Próstata||              | 5480      | 1218.104    | Esófago||              | 7432     | 1.293.521   | Músculo esq.||              | 6247      | 1.384.689   |
| Amígdala||              | 9044      | 1.208.674   | Esófago||              | 8437     | 1.200.215   | Pele||              | 10.723    | 1.239.942   |
| Amígdala|              | 14090     | 1.141.545   | Coração: Aorta||              | 9024     | 1.287.839   | Pele||              | 9636      | 1.282.550   |
| Caudado, Putamem, Acumbem|      | 13052     | 1.159.331   | Coração: ventrículo e septo||              | 8434     | 1.130.810   | Intestino grosso||              | 9853      | 1.205.494   |
| Caudado, Putamem, Acumbem|      | 13163     | 1.164.042   | Coração: ventrículo e septo||              | 7210     | 1.096.689   | Intestino grosso||              | 10.805    | 1.275.969   |
| Cerebelo||              | 13941     | 1.173.558   | Coração: Aorta||              | 9515     | 1.285.438   | Coluna espinhal||              | 14.890    | 1.103.143   |
| Cerebelo|              | 14004     | 1.136.630   | Coração: átrio||              | 8229     | 1.215.658   | Coluna espinhal||              | 12.631    | 1.261.833   |
| Córtejo||              | 13963     | 1.149.255   | Coração: Atrio||              | 4101     | 1.236.632   | Baço||              | 9804      | 1.293.676   |
| Córtejo|              | 13375     | 1.196.385   | Rim: córtex||              | 11.278   | 1.215.278   | Baço||              | 7165      | 1.304.431   |
| Hipocampo||              | 14139     | 1.153.051   | Rim: córtex||              | 8611     | 1.255.852   | Estômago||              | 8722      | 1.223.530   |
| Hipocampo|              | 12488     | 1.174.856   | Rim: medula||              | 9202     | 1.257.968   | Testículo||              | 9087      | 1.204.994   |
| Hipotálamo||              | 10980     | 1.213.070   | Rim: medula||              | 9396     | 1.243.408   | Timo||              | 9021      | 1.232.729   |
| Hipotálamo|              | 9614      | 1.214.811   | Intestino delgado||              | 4860     | 1.243.121   | Timo||              | 12.278    | 1.216.292   |
| Meso encéfalo|        | 12017     | 1.174.679   | Intestino delgado||              | 8156     | 1.191.328   | Tireóide/paratireo.||              | 9548      | 1.308.330   |
| Meso encéfalo|        | 12633     | 1.208.107   | Figado: lobo esq.||              | 6457     | 1.301.768   | Tireóide/paratireo.||              | 9627      | 1.285.960   |
| Bulbo olfatório||          | 12359     | 1.155.868   | Figado: lobo dir.||              | 4439     | 1.329.135   | Útero; grávida E18||              | 7464      | 1.244.848   |
| Bulbo olfatório||          | 15864     | 1.126.432   | Figado: lobo dir.||              | 5761     | 1.267.624   | Útero: grávida E18||              | 10.784    | 1.280.175   |
| Tubérculo olfator.|      | 11431     | 1.215.404   | Pulmão||              | 10.669    | 1.229.909   | Gordura branca||              | 10.347    | 1.302.910   |
| Tubérculo olfator.|      | 10197     | 1.128.574   | Pulmão||              | 10.175    | 1.340.241   | Gordura branca||              | 10.875    | 1.218.397   |
| Tálamo||              | 15252     | 1.207.035   | Linho nodo||              | 8187     | 1.169.244   |                      |           |              |
| Tálamo|              | 11547     | 1.264.767   | Linho nodo||              | 8499     | 1.230.414   |                      |           |              |
Além do número total, também podemos obter o número de *tags* distintas por biblioteca de MPSS. Este número varia de 4774 (próstata) a 15.912 *tags* (cerebelo), com a média de 10.805 *tags* (Tabela 1). Aqui já podemos ver o primeiro indício de que as regiões cerebrais apresentam muitos transcritos distintos: enquanto as regiões do cérebro, na média, apresentam 12.587 *tags* distintas, as regiões não cerebrais apresentam 10.220 (Tabela 1).

Depois de organizar todas as 81 bibliotecas de MPSS, iniciamos o processo de análise das *tags*. O primeiro passo foi identificar (mapear) o gene que cada *tag* representa. Este passo foi realizado através do ACTG, uma ferramenta que disponibiliza um bom repertório de bancos de dados para o mapeamento de *tags* de MPSS (ver seção “3.4 – Mapeamento das *tags* de SAGE e de MPSS”). Em seguida, dividimos as *tags* em dois conjuntos, *tags* de bibliotecas de cérebro e *tags* não cerebrais, e fizemos uma comparação entre eles. Novamente vemos uma enorme variedade de *tags* nas regiões cerebrais: apesar do conjunto de *tags* não cerebrais apresentar três vezes mais bibliotecas (61 bibliotecas versos 20), o número de *tags* distintas é apenas ~1/3 maior (33.394 versos 45.524) (Tabela 2). Esse mesmo padrão se repete para o número de genes: temos 14.945 genes expressos nos tecidos cerebrais e 18.310 genes expressos nos tecidos não cerebrais (um número sutilmente maior no segundo conjunto, 26%) (Tabela 2). Quando comparamos o mapeamento das *tags* dos dois conjuntos, temos um número similar (97% das *tags* das bibliotecas cerebrais e 95% das *tags* das bibliotecas não cerebrais foram mapeadas (Tabela 2)), mostrando que ambos os conjuntos apresentam *tags* com ótima qualidade.
O próximo passo foi medir as interseções entre os dois conjuntos, cerebrais e não cerebrais. Para isto, quantificamos o número de genes compartilhados e específicos nos dois conjuntos. No total, temos a expressão de 19.153 genes, sendo que 14.103 são comuns às bibliotecas cerebrais e aos tecidos não cerebrais (Figura 2), 843 só aparecem no cérebro e 4207 genes só aparecem nos outros tecidos. Também verificamos a expressão média destes genes. Os genes comuns aos dois conjuntos apresentam-se de 3 a 5 vezes mais expressos que os genes só presentes no cérebro ou nos outros tecidos, respectivamente. Uma lista completa dos genes expressos em todos os 45 tecidos de camundongo pode ser obtida em: http://www.compbio.ludwig.org.br/mpssmbt/.

Uma organização natural dos dados de expressão gênica é agrupar aqueles genes e tecidos que apresentam padrões similares de expressão. Para isto, aplicamos um método conhecido com

![Diagrama de Venn mostrando as interseções entre genes cerebrais e não cerebrais.](image)

Tabela 2 | Comparação entre as tags de tecidos cerebrais e não cerebrais. O número de tags corresponde ao número de tags distintas.

| Bibliotecas       | Tags de MPSS |           |       |
|-------------------|--------------|-----------|-------|
|                   | Total        | Mapeadas  | No. de Genes |
| Cerebrais         | 33.394       | 32.395 (97%) | 14.946 |
| Não cerebrais     | 45.524       | 43.316 (95%) | 18.310 |

Figura 2 | Interseccão entre os genes expressos no cérebro (cerebrais) e nos tecidos não cerebrais (não cerebrais).
agrupamento hierárquico nos 19.153 genes expressos nos 45 tecidos de camundongo. Neste método, a relação entre os objetos (genes e tecidos) é mostrada através de um dendrograma (uma árvore), onde o comprimento dos ramos reflete o grau de similaridade entre os objetos.

Primeiro, temos o agrupamento hierárquico dos 41 tecidos de camundongo (Figura 3). Rapidamente, podemos ver a existência de 3 grandes grupos de tecidos, com destaque para o grupo formado pelas bibliotecas das regiões do sistema nervoso (ver no destaque). Além disso, dentro deste grupo, notamos a proximidade de algumas regiões com características funcionais e anatômicas similares (ver o agrupamento e o comprimento dos ramos). Por exemplo, o bulbo olfativo e o tubérculo olfativo (regiões envolvidas com a detecção do olfato); o tálamo e o mesoencéfalo (anatomicamente, regiões muito próximas). Além do agrupamento, também notamos a separação de outras regiões. Por exemplo, o cerebelo e a coluna espinhal, anatomicamente regiões muito diferentes das demais, não estão agrupadas com nenhuma outra região cerebral. Aqui vale ressaltar que, apesar da coluna espinhal fazer parte do sistema nervoso central, nós não a analisamos junto com as demais regiões cerebrais, pois demos preferências para as regiões contidas na caixa craniana.
Podemos incrementar o agrupamento hierárquico e mostrar a figura completa contendo, não só o agrupamento dos tecidos, mas também o agrupamento dos genes (Figura 4). O agrupamento dos genes está representado na lateral esquerda da Figura 4. O nível de expressão de cada gene está representado por uma célula colorida e que vai de verde (pouco ou não expresso) a vermelho (muito expresso), passando por preto (expressão intermediária).
Como estamos visualizando a expressão de 19.153 genes em 45 tecidos, inevitavelmente o resultado se torna complexo e não nos permite uma análise detalhada de cada gene. Entretanto, podemos observar alguns padrões de expressão: i) existem ‘blocos horizontais’ contendo genes expressos em praticamente todos os 45 tecidos (duas regiões destacadas em azul); ii) existe um grande ‘bloco vertical’ (destacado em vermelho) contendo muitos genes expressos numa intensidade alta; iii) pequenos ‘blocos verticais’ isolados (regiões destacadas com flechas brancas).

Pelo padrão de expressão, os genes do primeiro conjunto (destacado em azul) parecem ser responsáveis por processos celulares comuns a todos os tecidos. O conjunto dois (destacado em vermelho) é formado pelas bibliotecas cerebrais, indicando que os genes destacados são expressos no cérebro com uma intensidade alta e, muitas vezes, podem ser de uma maneira preferencial (como destacado pelas flechas vermelhas). O terceiro conjunto (destacado pelas flechas brancas) contém genes com expressão preferencial nos tecidos destacados.
Figura 4 | Agrupamento hierárquico dos tecidos de camundongo. Foram utilizados 19.153 genes de 45 tecidos diferentes. As barras azuis destacam os genes com expressão alta em quase todos os tecidos. O retângulo vermelho destaca a expressão gênica no conjunto das 10 regiões cerebrais. As flechas vermelhas destacam alguns genes com expressão preferencial no cérebro (regiões mais escuras). As flechas brancas destacam os genes com expressão restrita a determinados tecidos.

4.1.2 A expressão gênica nas regiões cerebrais

Depois desta comparação geral dos genes expressos no cérebro e nos outros tecidos, partimos para um estudo mais detalhado da expressão gênica nas 10 regiões cerebrais. Primeiro, submetemos, individualmente, cada biblioteca cerebral ao protocolo de mapeamento do ACTG. Na Tabela 3 podemos ver que a porcentagem de mapeamento é similar entre as regiões cerebrais, mas o número de tags distintas e o número de genes variam consideravelmente. Por exemplo, o cerebelo e o bulbo olfativo são as regiões que mais apresentam genes expressos, 10.211, seguidos pelo córtex (10.199) e pelo tálamo (10.173). No outro extremo, com o menor número de genes expressos, está a amígdala (8765 genes) e o hipotálamo (8923 genes).
Tabela 3 | Comparação entre as *tags* das dez regiões cerebrais. O número de *tags* representa o número de *tags* distintas.

| Bibliotecas                      | Total   | Tags de MPSS | No. De Genes |
|----------------------------------|---------|--------------|--------------|
| Amígdala                         | 12.245  | 12.128       | 8765         |
| Caudado / putamem / acubens      | 14.649  | 14.504       | 10.023       |
| Cerebelo                         | 15.912  | 15.681       | 10.211       |
| Córtex                           | 15.152  | 15.026       | 10.199       |
| Hipocampo                        | 14.826  | 14.515       | 10.065       |
| Hipotálamo                       | 12.559  | 12.368       | 8923         |
| Mesoencéfalo                     | 14.455  | 14.253       | 9985         |
| Bulbo olfativo                   | 15.680  | 15.382       | 10.211       |
| Tubérculo olfativo               | 12.932  | 12.723       | 9178         |
| Tálamo                           | 15.010  | 14.766       | 10.173       |

Apesar da diferença, também existe um compartilhamento dos genes expressos entre as regiões cerebrais? Para responder esta pergunta, dividimos os genes da Tabela 3 em três categorias: i) com presença restrita (1 ou 2 regiões cerebrais); ii) com presença mediana (3 a 8 regiões cerebrais); iii) muito presentes (9 ou 10 regiões cerebrais). Como podemos ver na Figura 5 temos ~7600 genes que são compartilhados por oito a dez regiões. Também temos ~3200 genes presentes apenas em uma ou duas regiões cerebrais (presença restrita).
Figura 5 | Presença dos genes nas regiões cerebrais. Os genes estão divididos em três categorias de acordo com o número de regiões cerebrais em que são expressos: i) presença restrita (1 a 2 regiões); presença mediana (3 a 7 regiões); muito presentes (8 a 10 regiões).

Aproveitando essa divisão em categorias, também correlacionamos a presença do gene com o seu nível de expressão médio (medido pela tags de MPSS). Como esperado, vemos uma correlação positiva entre a presença do gene nas regiões cerebrais e o seu nível de expressão gênica: quanto mais presente, mais expresso é o gene (Figura 6).
4.1.3 Ontologia dos genes expressos no cérebro.

O Gene Ontology (GO) (http://www.geneontology.org/) é um projeto que provê um vocabulário controlado para a classificação dos genes em três categorias, função molecular (MF), processo biológico (BP) e componente celular (CC). Usamos os dados do GO para organizar a anotação e comparar os genes expressos nas regiões do cérebro e nos outros órgãos.

Primeiro, classificamos todos os genes expressos nas regiões do cérebro e nos órgãos não cerebrais (Figura 7). Em seguida, classificamos os genes expressos preferencialmente nas diferentes regiões cerebrais ou em alguns órgãos não cerebrais (Figura 8). Observando estes resultados, notamos que não existem grandes diferenças entre o conjunto total de genes (Figura 7). Contudo, quando observamos os genes expressos preferencialmente em 7 regiões cerebrais, as diferenças (reflexos da
especificidade das funções de cada tecido) aparecem (Figura 8). Por exemplo, o OB (quando comparado com as outras regiões) apresenta muitos genes nas categorias: desenvolvimento e transdução de sinal e sinalização célula-célula; o córtex nas categorias: processos metabólicos e transporte; o tubérculo olfativo nas categorias adesão célula-célula, sinalização célula-célula e ciclo celular; o músculo esquelético na categoria: processos metabólicos (Figura 8).

Figura 7 | G.O. classificação dos genes expressos no cérebro (cerebral) e dos genes expressos nos outros tecidos (não cerebral). Os genes foram classificados em dez categorias do tipo processo biológico.
Seleção dos genes expressos preferencialmente no cérebro e nas regiões cerebrais

Apesar de já termos selecionado as seções anteriores alguns genes com expressão restrita às regiões cerebrais, resolvemos aplicar um método que dá uma confiabilidade estatística (um valor $p$) a expressão diferencial dos genes nas regiões cerebrais. O método consiste em avaliar a diferença da expressão gênica através da simulação de Monte Carlo (este método foi anteriormente aplicado à bibliotecas de SAGE (Zhang, Zhou et al., 1997), nós simplesmente o programamos para os dados de MPSS). Foram selecionados os genes com expressão preferencial suportada por um valor de $p < 0,05$ (ver seção “3.5.4 - Seleção dos genes preferencialmente expressos nas regiões do cérebro”).

Figura 8 | G.O. classificação dos genes expressos preferencialmente na dez regiões cerebrais e em dois órgão não cerebrais. Os genes foram classificados em dez categorias do tipo processo biológico. Bulbo olfativo (OB), amígdala (AM), cerebelo (CB), tálamo (TL), tubérculo olfativo (TO), putamem-caudado-acumbens (PCA), córtex (CO), hipocampo (HI), hipotálamo (HY), mesoencéfalo (MB), pulmão (LU) e músculo esquelético (SM).
Utilizando este método, identificamos 1257 genes expressos preferencialmente nas regiões cerebrais (Tabela 4). Por exemplo, encontramos 167 genes expressos preferencialmente no cerebelo, 162 no bulbo olfativo e 102 no mesoencéfalo. Na Tabela 4 temos o número de genes com expressão preferencialmente nas 10 regiões do cérebro. Analisando a anotação destes genes, notamos que muitos são anotados como hipotéticos, sem uma função conhecida e sem uma nomenclatura oficial. Uma lista completa com estes genes está acessível em: http://www.compbio.ludwig.org.br/mpssmtp/.

| Regiões cerebrais          | No. de genes |
|----------------------------|--------------|
| Amigdala                   | 111          |
| Caudado / putamem / acubens| 136          |
| Cerebelo                   | 167          |
| Córtex                     | 126          |
| Hipocampo                  | 135          |
| Hipotálamo                 | 110          |
| Mesoencéfalo               | 104          |
| Bulbo olfativo             | 162          |
| Tubérculo olfativo         | 97           |
| Tálamo                     | 109          |

4.1.5 Os genes expressos preferencialmente no bulbo olfativo.

Apesar de selecionarmos apenas os genes expressos preferencialmente nas regiões do cérebro, ainda temos uma lista extensa de candidatos, 1257 genes, o que dificulta uma análise mais detalhada.
gene a gene. Então, para podermos realizar esta análise, nos concentramos em uma única região do cérebro, o bulbo olfativo (OB).

O bulbo olfativo (OB) está localizado na parte frontal do cérebro da maioria dos mamíferos (Figura 9: OB de camundongo) e desempenha uma função central no mecanismo de detecção dos odores (Leon e Johnson, 2003). O OB recebe informações dos neurônios olfativos que detectam o odor e as envia para o córtex olfativo e áreas límbicas do cérebro, onde a informação olfativa é reconhecida e interpretada (Shepherd, 2005).

Na Tabela 4 vemos que o OB apresenta 162 genes com expressão preferencial. Destes, depois de uma anotação semi-automática, identificamos 78 genes contendo um nome oficial ou, ao menos, uma seqüência de mRNA. Os outros 84 genes são de clusters que contém apenas ESTs (UniGene build #151). Analisando a literatura para alguns destes genes, vemos que muitos já foram descritos como expressos no sistema nervoso. Por exemplo: Trim9, Epha7, Neto2, Arhgef6, Gria2, Hes5, Lrrn3 e Dlx1. Olhando a função destes genes, podemos ver que temos fatores de transcrição (Sp9, Arx e...
Rbpsuhl, receptores (Stk32c e Rgsl2), canais (Cacnb3 e Kcnf1), entre outros. Na Tabela 5 está representada a lista completa dos 78 genes.

| Gene Name | Gene Name | Gene Name |
|-----------|-----------|-----------|
| Gria2     | Trim9     | Neto2     |
| Parp11    | Tera      | Scn10a    |
| A530045M11| Fxyd5     | Dmc1h     |
| Vipr2     | Tgm4      | Epha7     |
| Klhl15    | Trnt1     | Arhgef6   |
| Trh       | S100a5    | Gpsm1     |
| Tiam2     | Rutbc2    | Doc2g     |
| Hes5      | Eomes     | AW121567  |
| Sh3gl2    | Il20      | Gad1      |
| Sdk1      | Nmb       | Pnma1     |
| AB112350  | Stoml3    | Nrip3     |
| Ppfia2    | Olfr1408  | Arx       |
| Akap7     | Shhrs     | Ncoa6     |
| Dazl      | Zcwpw2    | Mdga1     |
| Cacna1f   | Snx7      | Fto       |
| Fmip      | Palm      | Dlx1      |
| Rasl11b   | Cacnb3    | Foxd3     |
| Gm1568    | D7Rp2e    | Gad2      |
| Rbpsuhl   | 4921513^04| Htr5a     |
| Mcm3ap    | Slc25a16  | Gria1     |
| Gtdc1     | Slc26a2   | Maml1     |
| Chrnb4    | Dgkh      | Sp8       |
| Rgsl2     | Za20d2    | Stk32c    |
| Spnb3     | Spata1    | Prok2     |
| Sp9       | Lrrn3     | Cacng5    |
| Apobec1   | Kcnf1     | Gucy1a2   |
4.1.6 Confirmação experimental da expressão gênica no OB

A hibridização *in-situ* (ISH) é um método que localiza sequências específicas de DNA ou RNA em uma secção de tecido. Usando os dados de ISH produzidos pelo Allen Brain Institute, nós testamos a expressão dos 162 genes candidatos a serem expressos preferencialmente no bulbo olfativo.

Na Figura 10 vemos a ISH de 3 destes genes. Na parte esquerda está mostrado o resultado da ISH; na parte direita (imagens escuras), estamos usando um recurso do Allen Brain Atlas que remove todo o *background* e mostra somente as regiões onde o gene está expresso (a escala vai do azul, pouco expresso, ao vermelho, muito expresso). Analisando este resultado, podemos ver que estes três genes apresentam expressão preferencial no OB. O *Csdc2* apresenta-se expresso em todo o OB e também no cerebelo. O *Doc2g* também está preferencialmente expresso no OB e, sobretudo, nas bordas da camada glomerular. O *Sp8* também está expresso preferencialmente no OB, em especial na região granular.

Quando olhamos a imagem direta da ISH parece que o cerebelo também está marcado para o *Doc2g* e *Sp8*, mas quando olhamos a imagem sem o *background* (parte direita da Figura 10) vemos que esta marcação é muito fraca.
Na Figura 11 temos hibridizações *in-situ* de outros 12 genes. Fazendo uma análise geral, vemos que todos estão muito expressos no OB e, para alguns casos, no cerebelo. Entretanto, assim como acontece com o *Doc2g* e *SP8* (ver acima), as imagens sem o *background* confirmam a expressão no OB e descartam a expressão no cerebelo (dado não mostrado).
Figura 11. Híbridização *in-situ* dos genes candidatos a serem expressos preferencialmente no OB. Todas as imagens são cortes sagitais do cérebro de camundongos C57BL/6 adultos.
4.2 Identificação de sequências reguladoras nos promotores dos genes ORs

Neste trabalho nós geramos cDNAs da região 5’ dos ORs, mapeamos estas sequências no genoma, definimos e analisamos a região 5’ UTR destes genes e, na região promotora, buscamos sequências curtas (motivos) sobre representadas. Para finalizar, verificamos se estes motivos se ligam a proteínas presentes no núcleo dos neurônios do epitélio olfativo.

4.2.1 Amplificação da região 5’ completa dos genes ORs

O primeiro passo do trabalho foi obter a região 5’ completa dos genes ORs. Esse método foi feito através de um RLM-RACE (uma técnica que permite a amplificação de mRNAs contendo o cap 5’, ou seja, transcritos com a região 5’ completa) usando RNA total de epitélio olfativo de camundongos adultos (Figura 12). Para selecionar os genes ORs do RNA total de epitélio olfativo, foram utilizados primers degenerados para regiões conservadas entre os ORs. Foram feitas duas reações de PCR, uma direta utilizando primers 5’ e os primers P8 ou P27 e outra nested-PCR com o produto da primeira, utilizando primers 5’ nested e o primer P26R. Os fragmentos de PCR foram purificados, clonados e sequenciados.
4.2.2 Agrupamento dos cDNAs e definição dos genes ORs

Dos 1012 clones sequenciados, 80% foram de genes ORs e 96% dos cDNAs apresentaram sequências *upstream* ao ATG inicialmente predito. Alinhando e agrupando essas sequências, verificamos que elas representam 198 ORs (9 são pseudogenes), correspondendo a 17% do repertório total destes genes em camundongo. Na Tabela 6 temos uma comparação, cromossomo a cromossomo, do total de genes ORs preditos e dos genes para os quais identificamos sequências. Os 198 ORs estão distribuídos de maneira semelhante ao repertório completo de ORs (Godfrey, Malnic et al., 2004): muitos ORs nos cromossomos 2 e 7; poucos ORs nos cromossomos X, 15 e 8. Identificamos
seqüências para os ORs de todos os cromossomos, com exceção aos dois ORs do cromossomo 3 (Tabela 6).

| Tabela 6 | Distribuição dos genes ORs nos cromossomos. |
|----------|----------------------------------------|
|           | Cromossomo | Número de genes | Genes ORs (Godfrey, Malnic et al., 2004) | Genes ORs com ESTs 5’ |
| 1         | 23 (7)     | 6               |
| 2         | 275 (81)   | 44 (5)          |
| 3         | 2          | -               |
| 4         | 19 (9)     | 4               |
| 5         | -          | -               |
| 6         | 22 (9)     | 5               |
| 7         | 198 (51)   | 32              |
| 8         | 4          | 2               |
| 9         | 118 (37)   | 20 (1)          |
| 10        | 47 (13)    | 9               |
| 11        | 35 (20)    | 21 (2)          |
| 12        | -          | -               |
| 13        | 12 (2)     | 4               |
| 14        | 32 (4)     | 6               |
| 15        | 6 (3)      | 3 (1)           |
| 16        | 29 (8)     | 18              |
| 17        | 36 (17)    | 5               |
| 18        | -          | -               |
| 19        | 52 (19)    | 9               |
| X         | 2          | 1               |
| Y         | -          | -               |
| Total     | 1190       | 198             |

Continuando a comparação entre o repertório total de ORs e os 198 com seqüências 5’ completa, analisamos a classificação dos ORs em famílias gênicas e classes filogenéticas. Os ORs podem ser subdivididos em famílias gênicas (onde os membros apresentam ao menos 60% de identidade entre si) e também em duas classes filogenéticas, classe I (similares a receptores de peixes e mais envolvidos no reconhecimento de odores diluídos em água) e classe II (específicos dos animais
terrestres) (Glusman, Sosinsky et al., 2000). Para camundongos, os 1190 ORs estão classificados em 248 famílias gênicas (Godfrey, Malnic et al., 2004), 88% são de classe II e 12% de classe I. Para os 198 ORs que identificamos sequências, temos 102 famílias gênicas representas (41% do total), onde 4,5% dos genes são de classe I e 95,5% de classe II.

4.2.3 A região 5’ dos genes OR

Dos 198 genes, todos apresentaram ao menos uma sequência definindo uma região 5’ UTR. Estas regiões apresentam comprimento médio de 189 pb (estendendo-se de 32 a 659 pb), 72% contém um exon completo, 26% têm dois ou mais exons; o tamanho destes exons varia de 9 a 403 pb. Já os introns apresentam uma grande variação de tamanho, de 91 bases a 22.5 kb, onde 46% e 34% dos genes apresentam introns de 2-4kb e maiores do que 4kb, respectivamente.

Em geral, nos eucariotos a tradução inicia-se no primeiro ATG e termina no códon de parada localizado na parte terminal do transcrito (Kozak, 1999). Contudo, nós encontramos upstream ORFs (uORFs), regiões de até 10 codons, em 108 dos genes ORs. Pesole e colaboradores (Pesole, Mignone et al., 2001) sugeriram que estas uORFs estão envolvidas com o controle da regulação da tradução. Somente uma exploração mais detalhada poderá avaliar o papel dessas uORFs na regulação dos ORs.

Nós também identificamos 35 ORs com um ATG, em fase, upstream ao ATG originalmente predito, indicando que estes genes ORs apresentam uma proteína maior do que a depositada nos bancos de dados públicos.
4.2.4 Identificação dos motivos na região promotora dos ORs

Em teoria, espera-se que genes regulados pelos mesmos fatores de transcrição também apresentem elementos (sequências curtas) conservedos entre as suas regiões promotoras. Apoiando-se nesta hipótese, nós decidimos buscar sequências compartilhadas entre as regiões promotoras dos 198 ORs.

Primeiro, utilizamos as ESTs 5’ completa para determinar com precisão o início da transcrição (TSS) dos 198 genes ORs. Em seguida, extraímos sequências de 600pb upstream ao TSS, as submetemos ao RepeatMasker (http://www.repeatmasker.org/) e retiramos as regiões repetitivas e/ou de baixa complexidade (apenas 7% das bases foram filtrados). Por fim, buscamos os motivos compartilhados entre as 198 sequências.

Antes de buscar os novos elementos, avaliamos a presença do TATA box nas 198 regiões promotoras. Identificamos o TATA em 35% das regiões promotoras. Apesar deste número parecer baixo, ele está de acordo com o descrito previamente por (Sosinsky, Glusman et al., 2000; Lane, Cutforth et al., 2001; Hoppe, Frank et al., 2003).

Para buscar os elementos conservados entre as sequências, utilizamos três programas, o Gibbs Recursive Sampling, Consensus e o Weeder. Escolhemos estes programas, pois cada um apresenta um método diferente para a identificação dos motivos. Utilizamos os parâmetros default para os três programas, apenas alteramos o tamanho mínimo dos motivos a serem procurados. O mínimo foi de 6 nt e o máximo de 14 nt. Depois de organizar e processar todos os resultados, encontramos 2 grupos de sequências (Figuras 13 e 14). O primeiro grupo, presente em 87% das regiões promotoras (Tabela 7), mostrou-se similar ao sítio de ligação do fator de transcrição Olf-1 (O/E) (Wang, Tsai et al., 1993).
Chamaremos estes sítios de O/E-similar. O segundo grupo, presente em 95% das regiões promotoras (Tabela 7), mostrou-se similar ao sítio de ligação do fator *homeodomain*.

**Tabela 7. Distribuição das sequências (motivos) nas regiões promotoras dos ORs.** A porcentagem é referente aos 198 ORs. O O/E-similar indica a soma, sem redundância, dos genes que apresentam os motivos (M1-M4). O *homeodomain*-s indica a soma dos genes que apresentam os motivos similares ao *homeodomain*.

| Motivos      | Número de Genes ORs | Número total de sítios |
|--------------|---------------------|------------------------|
| M1           | 85 (42%)            | 110                    |
| M2           | 99 (50%)            | 129                    |
| M3           | 71 (36%)            | 79                     |
| M4           | 51 (26%)            | 64                     |
| O/E-similar  | 173 (87%)           | 382                    |
| Homeodomain-s| 188 (95%)           | 1029                   |

Os fatores de transcrição O/E-similares foram identificados originalmente em linfócitos B (Wang et al 2003), mas depois foram encontrados no epitélio olfativo. É interessante notar que motivos de ligação para estes fatores estão presentes na região promotora de mais de 80% dos ORs analisados (Tabela 7).
Figura 13 | Motivos encontrados na região promotora dos ORs. As sequências foram agrupadas de acordo com a similaridade com a sequência do motivo O/E e com as sequências do Homeodomain. As sequências destacadas, motivos 1, 2, 3 e 4 foram selecionadas e analisadas em
Depois de identificados, analisamos a distribuição dos motivos ao longo da região promotora.

Na Figura 15 temos o número de motivos nas 600 pb das regiões promotoras dos 198 ORs. Na Figura 15A temos a distribuição dos motivos M1-M4. Na Figura 15B temos distribuição dos motivos O/E-like, homeodomain-similar e do homeodomain (TAATTG – sequência previamente identificada em promotores de ORs (Vassalli, Rothman et al., 2002)). Podemos ver que os motivos O/E-like estão concentrados mais próximos ao TSS. Já os motivos homeodomain e homeodomain-similar estão espalhados por toda a região promotora, com uma concentração não muito expressiva próxima ao TSS.
Detecção e análise dos eventos de retenção de intron

Neste trabalho mapeamos no genoma todo o conjunto público de cDNA humano, agrupamos estas seqüências e desenvolvemos um protocolo para identificar e estudar os eventos de retenção de intron, um dos tipos de splicing alternativo.

Figura 15. Localização dos motivos em relação ao início de transcrição. (A) indica o número de motivos 1-4 posicionados nas primeiras 600 bases da região promotora dos 198 ORs. (B) indica o número total de motivos O/E-similar (soma dos motivos 1-4), homeodomain-similar e homeodomain (TAATTG) nas primeiras 600 bases da região promotora dos 198 ORs.

4.3 Detecção e análise dos eventos de retenção de intron
4.3.1 Identificação dos eventos de retenção de intron

A retenção de intron é o *splicing* alternativo no qual ocorre, no mRNA maduro, a presença de uma sequência definida previamente como um intron (Figura 16). Para identificar os eventos de retenção de intron, primeiro nós mapeamos as EST e os mRNAs no genoma humano, agrupamos os cDNA pertencentes ao mesmo gene (cluster) e construímos programas para comparar a estrutura exon-intron de todos os transcritos do cluster. Estratégias similares a nossa também foram utilizadas por outros grupos (Mironov, Fickett *et al*., 1999; Modrek, Resch *et al*., 2001; Holste, Huo *et al*., 2006).

Utilizando este método, identificamos ao menos um intron retido em 3127 genes, totalizando 14,8% dos 21.106 genes analisados. Com o objetivo de eliminar falsos IR, eliminamos as mensagens contendo um único exon, pois elas podem ser seqüências não processadas ou derivadas de contaminantes genômicos; comparamos apenas as seqüências que compartilham ao menos uma borda exon-intron; analisamos apenas os eventos de retenção de intron envolvendo ao menos uma sequência de mRNA (Figura 16). Também geramos um controle do nosso banco de *splicing* alternativo, através do cálculo da ocorrência de exons alternativos, o tipo mais freqüente e mais estudado de *splicing* alternativo. Identificamos estes eventos em 52% dos genes, o que está de acordo com o encontrado por outros grupos (Modrek, Resch *et al*., 2001; Kan, States *et al*., 2002).
Na Tabela 8 podemos ver o número de retenção de intron entre os pares mRNAs-mRNAs e mRNAs-ESTs. É interessante notar que em 89% (2793/3127) dos casos, são os mRNAs que apresentam os introns retidos. Por outro lado, são as ESTs que representam a maioria das sequências classificadas como protótipo, 83% (2594/3127).

| Protótipo | Variante | Total |
|-----------|----------|-------|
| mRNA      | EST      | Total |
| 640       | 691 (385)| 1120  |
| 2594      | Não analisado | 2594 |
| 2793      | 691      | 3127  |

4.3.2 Distribuição dos eventos de IR ao longo do transcrito

Os introns retidos estão localizados preferencialmente em alguma região do gene: 5’ UTR, CDS ou 3’ UTR? Para responder esta questão, selecionamos os casos de introns retidos envolvendo
mRNAs, defínimos a região codificadora e região UTR destas sequências, e quantificamos a presença dos eventos de retenção de introns em cada região. Para evitar qualquer viés, normalizamos o número de eventos de IR pelo número de introns de cada região. Na Tabela 9 podemos ver o resultado: há um excesso de introns retidos nas regiões 3’ UTR e 5’ UTR quando comparadas com a sequência codificadora (CDS) ($\chi^2 = 229$, 1 grau de liberdade; valor-$p < 10^{-49}$).

Tabela 9. Número de eventos de IRs nas regiões 5’ UTR, CDS e 3’ UTR. O conjunto “Esperado” foi obtido através do cálculo da densidade de introns ao longo do transcrito. Utilizamos a seguinte fórmula: $(IR + nI) / tI$, onde IR é o número total de retenção de intron; nI é o número total de introns nas regiões 5’ UTR, CDS ou 3’ UTR do transcrito, respectivamente; tI é o número total de introns do transcrito. O grupo mRNA é formado por todos os mRNAs envolvidos na retenção de intron. O grupo MGC contém apenas os mRNAs vindos do Mammalian Gene Collection (Strausberg, Feingold et al., 2002).

| Eventos na | mRNAs | MGC |
|------------|--------|-----|
|            | Observado | Esperado | Observado | Esperado |
| 5’ UTR     | 84 (15%) | 27 (5%)  | 15 (9%)   | 8 (5%)   |
| CDS        | 287 (53%) | 502 (93%) | 87 (52%) | 155 (93%) |
| 3’ UTR     | 170 (32%) | 12 (2%)   | 65 (39%)  | 4 (2%)   |

4.3.3 Domínios protéicos codificados por introns retidos

Uma maneira de verificar o significado biológico do intron retido é avaliar a sua contribuição na codificação dos domínios protéicos. Para isto, buscamos os domínios protéicos do Pfam nos mRNAs contendo introns retidos na região codificadora. Na Figura 17 temos alguns exemplos da contribuição dos introns retidos na formação dos domínios protéicos das proteínas. 145 introns retidos codificam parte dos domínios, mas apenas dois codificam um domínio protéico completo. Comparamos os IR com os introns não retidos e com os exons: os introns retidos são mais eficientes para a codificação de domínios protéicos do que os introns não retidos dos mesmos genes (teste de
Fisher, \( p < 0.21 \); mas, como esperado, codificam menos domínios que os exons, 30 no total (teste de Fisher, \( p < 0.0026 \)).

Nós também analisamos a presença de domínios protéicos nos introns retidos em conjunto com seus dois exons flanqueadores. Primeiro, verificamos que estes trios de seqüências codificam 13 domínios protéicos completos (consideramos apenas os casos onde o intron retido contém, no mínimo, 30% do domínio protéico). Em seguida, excluímos os introns retidos e buscamos os domínios protéicos. Seus 6 domínios não foram detectados. Outros 6 domínios foram detectados com um \( e\)-value sem significado estatístico. Um domínio foi detectado com um \( e\)-value similar ao encontrado quando o intron retido estava presente.

Além dos domínios inteiros, também encontramos 24 casos onde parte do domínio protéico é codificado pelo trio “exon/intronRetido/exon” (novamente exigimos que o intron retido codificasse ao menos 30% do domínio protéico). Juntando todos os resultados, verificamos que 39 dos 147 introns retidos na CDS (26%) estão envolvidos na codificação de domínios protéicos.

Para finalizar essa análise, como controle positivo, testamos se os introns retidos nas regiões UTR também poderiam participar da codificação de domínios protéicos. Como esperado, nenhum dos 254 introns retidos nas regiões UTR, traduzidos nas três fases de leitura, apresentou identidade com os domínios protéicos do Pfam (teste de Fisher, \( p < 10^{-17} \), quando comparado com os introns da região codificadora).
4.3.4 Pressão seletiva nos introns retidos

Olhando os resultados anteriores pode surgir a questão: os introns retidos poderiam estar sob seleção para se tornarem seqüências codificadoras? Para responder esta questão fizemos alguns experimentos com os introns retidos na região codificadora (CDS). Primeiro, sabendo que as regiões codificadoras no homem apresentam um conteúdo de GC maior que as não codificadoras (Lander, Linton et al., 2001), buscamos a frequência destes di-nucleotídeos nos introns retidos e comparamos os exons e introns não retidos (dividimos em categorias de tamanho, pois a composição de GC varia com o tamanho dos introns, (Lander, Linton et al., 2001)). Verificamos que os introns retidos apresentam
conteúdo de GC diferente dos introns não retidos \( (\chi^2 = 2276; 6 \text{ g.l.; valor } p < 10^{-50}) \), mas similar ao dos exons \( (\chi^2 = 7,9; 6 \text{ g.l.; valor } p = 0,23) \) (Figura 18).

No próximo passo, verificamos se a presença de códon de parada no intron retido era igual à esperada por chance. Baseando-se na composição de nucleotídeos dos introns retidos na CDS, calculamos o número esperado de códon de parada: 1064. Entretanto, o número de códons de parada encontrados nestes introns foi de 651, significativamente menor do que os esperado \( (\chi^2 = 103; 1 \text{ g.l.; valor } p < 10^{-25}) \).

Adicionalmente, nós também analisamos em detalhes o conteúdo de GC de 88 cDNAs onde o intron retido adiciona um códon de parada prematuro e trunc a proteína original. Verificamos que, upstream ao códon de parada, o conteúdo de GC é maior do que downstream, 58% e 49% respectivamente, indicando que há um relaxamento da seleção após (downstream) o sinal de parada.

**Figura 18 | Conteúdo de GC nos introns retidos, introns não retidos e nos exons.** Todas as sequências estão dentro das regiões codificadoras do gene. Separamos em categorias de tamanho, pois o conteúdo de GC varia com o tamanho dos introns.
4.3.5 Conservação dos introns retidos em camundongo

Espera-se que eventos de splicing alternativo conservados entre espécies distantes sejam funcionalmente importantes. Para avaliar esta questão, buscamos os IR conservados entre *Homo sapiens* e *Mus musculus*. Primeiro, alinhamos os cDNAs humanos, contendo o intron retido, contra todos os cDNAs de camundongo. Desta maneira, encontramos seqüências ortólogas para 57% dos introns retidos na CDS e para 38% ou 43% dos introns retidos nas regiões UTR (Tabela 10). A identidade dos IR ortólogos foi de 84% em média, próxima aos 87% dos genes ortólogos descritos por Waterston e colaboradores (Waterston, Lindblad-Toh *et al.*, 2002). Como controle, calculamos a identidade, em camundongo, dos introns não retidos: 60%. Além de identificarmos as seqüências representando os introns retidos em camundongo, também buscamos os cDNAs protótipo, o qual definem aquela região como sendo um intron. Para 22% dos casos de IR foi possível identificar também essas seqüências protótipo (última linha da Tabela 10).

| Tabela 10 | Eventos de IR conservados entre homem (Hs) e camundongo (Mm). As análises foram feitas separadamente para as regiões UTR e CDS. |
|-----------|-------------------------------------------------------------|
| **IR em Hs com alinhamento nos cDNAs de Mm** | **5’ UTR** | **CDS** | **3’UTR** |
| Genes de Mm com mais de um cDNA na região | 1/9 | 40/74 | 5/15 |
| Genes de Mm com no mínimo uma sequência protótipo | 0/1 | 9/40 | 1/5 |
4.4 Análise funcional e evolutiva dos pares de genes sense-antisense em homem e camundongo

Neste trabalho nós desenvolvemos um protocolo para a identificação de novos NATs através das tags de MPSS, selecionamos os pares de genes senso-antisenso determinados pelos cDNAs públicos e estudamos estes genes sob um aspecto funcional e evolutivo no genoma de homem e camundongo.

4.4.1 Identificação dos pares S-AS

Para identificarmos os genes transcritos de fitas opostas do mesmo locus genômico, primeiro mapeamos todos os cDNAs de Homo sapiens e de Mus musculus nos seus respectivos genomas (ver a seção “3.2 – Alinhamento dos cDNAs no genoma”), removemos as seqüências sem evidência de orientação (ver seção “3.8.1 – Identificação dos pares S-AS”) e agrupamos todas as seqüências em clusters (ver seção “3.3 – Agrupamento de cDNAs”). Em seguida, construímos programas para analisar todos os clusters formados e identificamos aqueles com orientação de transcrição oposta e com sobreposição exônica e/ou intrônica. Apenas os clusters contendo ao menos um mRNA foram utilizados.

Tabela 11. Distribuição dos pares S-AS no genoma humano e de camundongo. Transcrição bidirecional simples corresponde aos loci contendo apenas um par S-AS.
Transcrição bidirecional múltipla corresponde aos loci contendo dois ou mais pares S-AS. Os conjuntos for subdivididos de acordo com a presença de mRNAs ou de ESTs no clusters.

| Tipo de cDNA | Transcrição bidirecional simples | Transcrição bidirecional múltipla |
|--------------|---------------------------------|----------------------------------|
|              | Humano                          | Camundongo                       |
| mRNA-mRNA    | 2109                            | 1879                             |
|              | 1004                            | 720                              |
| mRNA-EST     | 3299                            | 3265                             |
|              | 3665                            | 2227                             |
| Total        | 5408                            | 5144                             |
|              | 4669                            | 2947                             |

Como resultado, identificamos 10.077 e 8091 pares S-AS em humano e camundongo, respectivamente (Tabela 11). Para explorar melhor estes genes, classificamos os pares S-AS em simples, contendo apenas um gene sense e outro antisense, e múltiplos, contendo dois ou mais pares S-AS. Também classificamos os pares de acordo com a presença de mRNAs ou ESTs nos clusters. Os pares S-AS simples são os mais frequentes em humanos e em camundongos (5408 e 5144, respectivamente; Tabela 11). Também são mais frequentes os pares S-AS definidos por mRNA-EST (Tabela 11).

Como identificamos um número elevado de pares S-AS, resolvemos testar se o nosso método de agrupamento (clustering) de cDNAs não poderia estar gerando casos de falsos positivos. Para isso, alteramos o mínimo de sobreposição necessária para o agrupamento das seqüências de cDNA. Originalmente fizemos o agrupamento com 30 nt de sobreposição, mas alteramos este número para 1 nt. Como esperado, a diminuição para 1 nt fez com que seqüências, até então separadas, se agrupassem, reduzindo o número total de clusters em 2% em humano e 1% em camundongo. Entretanto, o número de pares S-AS diminuiu somente 0.3% nas duas espécies, descartando qualquer problema no processo de clustering.
4.4.2 Organização genômica dos pares S-AS

Depois de identificarmos o repertório de S-AS em *Homo sapiens* e *Mus musculus*, começamos a estudar algumas características destes genes. Por exemplo, exploramos a organização genômica e o padrão de sobreposição dos pares. Para simplificar, descartamos os pares S-AS múltiplos e utilizamos apenas os simples.

Primeiro, os pares S-AS foram classificados de acordo com a orientação no genoma (5’-5’ (sobreposição entre as regiões 5’-5’ dos genes), contidos (incluídos totalmente dentro do gene da fita oposta), 3’-3’ (sobreposição entre as regiões 3’-3’ dos genes)) e com a região de sobreposição dos exons e introns (completamente exônica, exônica-intrônica, completamente intrônica). Na Figura 19 podemos ver todos os tipos organização genômica e de sobreposição dos pares. É interessante notar que o tipo contido é o mais freqüente em ambas as espécies, 47.8% (homem) e 42.5% (camundongo) (Tabela 12). Quando olhamos a sobreposição dos pares, o tipo exônico-intrônico predomina (Tabela 12). Também é importante notar que 3/4 das sobreposições completamente intrônicas são originados de pares S-AS contidos.

Para testar se estas organizações ocorrem simplesmente ao acaso ou se são funcionais, nós desenvolvemos uma simulação onde sorteamos aleatoriamente o início da sobreposição do gene antisense em relação ao sense e re-classificamos os pares em 5’-5’, contido, 3’-3’. Repetimos este processo 1000 vezes para cada par S-AS. No final, o resultado mostrou que os três tipos de sobreposição não ocorrem ao acaso ($p = 10^{-11}$ para 5’-5’; $p = 0,003$ para contido; $p = 10^{-29}$ para 3’-3’).
Tabela 12 | Distribuição dos NATs em relação à estrutura do transcrito senso.

| Localização         | Humano | Camundongo |
|---------------------|--------|------------|
|                     | 5'-5'  | 3'-3'      | 5'-5'  | 3'-3'      |
| Completamente       |        |            |        |            |
| exônico             | 112    | 213        | 156    | 227        |
| (20%)               | (3%)   | (40%)      | (27%)  | (45%)      |
| Exônico / intrônico | 362    | 259        | 360    | 338        |
| (64%)               | (37%)  | (48%)      | (62%)  | (42%)      |
| intrônico           | 92     | 61         | 61     | 448        |
| (16%)               | (60%)  | (12%)      | (11%)  | (56%)      |
| Total               | 566    | 5333       | 577    | 800        |
|                     | 1010   | 502        | 502    |            |

Além de classificar os pares S-AS segundo a sua sobreposição, fizemos uma classificação baseada na presença de exons e de introns destes genes. Quando ambos os genes do par S-AS
apresentaram múltiplos exons, eles foram classificados como “ambos com introns”; “intron-sem-intron” quando apenas um dos genes apresentou múltiplos exons”; “sem intron”, quando nenhum gene do par apresentou múltiplos exons. Aproximadamente metade dos pares S-AS de humanos (47%) apresentou ao menos um gene sem múltiplos exons (“sem introns”). Em camundongo este número foi de 44%. É interessante notar que os genes sem introns estão enriquecidos na categoria contidos ($p < 1,2 \times 10^{-22}$ para humano; $p < 0,045$ para camundongo), 66% para humanos e 49% para camundongo (Tabela 13).

| Pares S-AS | Humano | Camundongo |
|------------|--------|------------|
| 5’-5’ | 351 | 394 |
| 3’-3’ | 417 | 390 |
| Total | 556 | 800 |

| Protocols for the identification of conserved S-AS pairs (5,6,25). We implemented the latter. First, we used the homology data of the HomoloGene and identified 190 pairs conserved between human and mouse.

4.4.3 Conservação dos pares S-AS entre humanos e camundongos

Depois de identificarmos o repertório de pares S-AS em homem e camundongo, comparamos estes dois conjuntos e identificamos os pares conservados. Essencialmente, existem três protocolos para a identificação de pares S-AS conservados (5,6,25). Nós implementamos os três. Primeiro, utilizamos os dados de homologia do HomoloGene e identificamos 190 pares conservados entre homem e camundongo.
Juntando o fato de que o Homologene define os genes ortólogos através da seqüência protéica dos genes (http://www.ncbi.nlm.nih.gov/HomoloGene) e que muitos dos genes antisense não são codificadores (Kiyosawa, Yamanaka et al., 2003), nós resolvemos utilizar um segundo método onde os dados do Homologene são utilizados apenas para identificar o gene ortólogo a um dos genes do par S-AS. A ortologia do outro gene (muitas vezes um gene não codificador) foi testada localmente. Com este método, identificamos outros 546 pares S-AS conservados entre as duas espécies, totalizando 736 (190 + 546).

O terceiro método, mais robusto que os dois primeiros, é baseado em regiões ortólogas definidas a partir do alinhamento de cromossomos inteiros. Neste método, os pares S-AS de homem e de camundongo são definidos como ortólogos quanto ambos estão em loci gênico considerados homólogos e correspondentes entre as duas espécies. Com este método, nós conseguimos encontrar 1136 e 1144 pares S-AS conservados entre humano-camundongo e camundongo-humano, respectivamente. Vale destacar que estes números não são iguais porque um par S-AS em humano pode corresponder a mais de um par em camundongo, e vice versa.

Nós também classificamos os genes conservados segundo a sua organização genômica, 5’-5’, contido e 3’-3’. A organização 3’-3’ é a mais freqüente nos dois conjuntos. Nos dois conjuntos obtidos a partir do HomoloGene, temos 14% (5’-5’), 19% (contido) e 67% (3’-3’); no outro conjunto, temos 38% (5’-5’), 14% (contido) e 48% (3’-3’).
4.4.4 Padrão de expressão dos pares S-AS

Quantificar o padrão de expressão dos genes senso e antisenso é um passo fundamental para se entender a maneira de atuação destes genes. Para isto, utilizamos bibliotecas de MPSS para avaliar a expressão do gene senso e do antisenso. O mapeamento das tags foi feito com o ACTG (ver seção “3.4 – Mapeamento das tags de SAGE e de MPSS”), mas tomamos o cuidado de remover as tags ambiguas.

Na Figura 20 temos o padrão de expressão dos pares S-AS em todas as bibliotecas de MPSS (humanos). Novamente, nós dividimos o conjunto de dados em três categorias, 3’-3’, contido e 5’-5’. A maioria dos pares S-AS apresenta ao menos um dos genes expressos (Figura 20A). Também podemos notar que a co-expressão é significativamente maior nos pares 3’-3’ quando comparado com os pares contido (50,3%; $\chi^2 = 134$; g.l. = 1; $p = 5.4 \times 10^{-31}$); o mesmo acontece com os pares 5’-5’ quando comparados com os pares contido (50; $\chi^2 = 23,5$; g.l. = 1; $p = 1.2 \times 10^{-6}$). Quando selecionamos apenas os pares co-expressos e calculamos a razão da expressão (Figura 20B), vemos que muitos pares apresentam diferença de expressão $> 3x$ (categorias 3-5 e $> 5$). Não notamos diferenças significativas este os casos 5’-5’, contido e 3’-3’.
4.4.5 Influência dos NATs no processo de *splicing* alternativo

Está claro que muitos dos genes humanos apresentam *splicing* alternativo (Modrek e Lee, 2002). Também se sabe desde o final da década de 80 que os NATs podem alterar o *splicing* do transcrito senso, causando *splicing* alternativo (Munroe, 1988). Baseados nisto, nós resolvem testar a co-ocorrência de NAT e *splicing* alternativo.
Nós testamos se a ocorrência de splicing alternativo no transcrito senso poderia ser afetada pela presença do antisenso através da análise das bordas onde ocorre a sobreposição senso-antisenso. Para tornar a análise mais informativa, dividimos as bordas em quatro categorias (sitio doador terminal, sitio doador interno, sitio acceptor terminal e sitio acceptor interno). Para os sitios doadores e acceptores internos, a presença do transcrito antisenso aumentou sutilmente a razao de splicing alternativo em 4% e 3%, respectivamente (Tabela 14). Esse aumento é devido, principalmente, a retenção de intron. Para os sitios terminais, a presença dos NAT teve o efeito contrario (diminuição de 5% e 6% para os sitios doadores e acceptores, respectivamente). Por outro lado, os NATs localizados exclusivamente dentro dos exons e introns não têm efeitos significativos nas bordas destas sequencias (dado não mostrado). Calculando a taxa de splicing alternativo nas bordas com e sem NATs, observamos que há uma diferença significativa para os sitios doadores ($\chi^2 = 31$; g.l. = 1; $p = 2,3 \times 10^{-8}$) e para os sitios acceptores ($\chi^2 = 23$; g.l. = 1; $p = 1,6 \times 10^{-6}$).

**Tabela 14 | Frequência dos diferentes tipos de splicing alternativo nas bordas exon-intron com e sem transcritos antisenso.**

| Bordas           | Total | Bordas alternativas | Retenção de intron | Uso alternativo de exon | Sítio alternativo 3' e 5' |
|------------------|-------|---------------------|--------------------|-------------------------|---------------------------|
| **Com NAT**      |       |                     |                    |                         |                           |
| Doador terminal  | 2578  | 553                 | 130                | 7                       | 416                       |
| Doador interno   | 7632  | 3100                | 535                | 1616                    | 949                       |
| Acceptor terminal| 7749  | 3145                | 493                | 1642                    | 1010                      |
| Acceptor interno | 2763  | 688                 | 208                | 7                       | 473                       |
| **Sem NAT**      |       |                     |                    |                         |                           |
| Doador terminal  | 2200  | 579                 | 101                | 32                      | 446                       |
| Doador interno   | 23.414| 8674                | 1080               | 4997                    | 2597                      |
| Acceptor terminal| 23.447| 8787                | 1022               | 5007                    | 2758                      |
| Acceptor interno | 1732  | 545                 | 154                | 16                      | 375                       |
Analisando os loci com múltiplos pares S-AS, notamos que alguns transcritos antisenso podem formar uma ponte de ligação entre dois genes, e permitir a formação de transcritos quiméricos, contendo exons de dois genes. Encontramos 5 casos onde há ao menos um cDNA que contém exons de dois genes que compartilham o mesmo NAT. Na Figura 21 podemos ver um destes casos. Nele, os transcritos AK095876 e AK000438 contêm exons dos genes SERF2 e HYPK (ambos compartilham o antisenso AK097682).

**Figura 21 | Transcritos contendo exons de dois genes, SERF2 and HYPK, ligados por um NAT.**
4.4.6 Utilizando MPSS na identificação de novos NAT

Assim como SAGE, MPSS quantifica a expressão gênica sem o conhecimento prévio da seqüência dos genes a serem analisados (Brenner, Johnson et al., 2000). Uma vantagem de MPSS em relação a SAGE é a quantidade de tags. Enquanto uma biblioteca de SAGE contém ~150 mil tags, uma biblioteca de MPSS contém ~1 milhão, permitindo a identificação de transcritos raros ou pouco expressos. Para identificar novos NATs, a nossa estratégia foi mapear as tags de MPSS no genoma e verificar se elas encontravam-se na fita oposta de algum gene conhecido (vale destacar que a orientação da tag é dada pela orientação do sítio DpnII). Foram utilizadas 41 bibliotecas de humanos e 81 bibliotecas de camundongo.

Como controle positivo, antes de buscar os novos NATs com as tags de MPSS, nós utilizamos a nossa estratégia para confirmar os pares S-AS previamente identificados com os cDNAs públicos. Isto é, quantos dos pares S-AS definidos por mRNAs e/ou ESTs também são definidos pelas tags de MPSS. Com as tags, confirmamos a transcrição senso-antisenso para 84% e 51% dos pares S-AS em humanos e camundongos, respectivamente.

Então, em seguida, buscamos os novos NATs com as tags de MPSS. Encontramos tags antisenso a 4308 e a 216 genes de humano e de camundongo, respectivamente. Como certos genes apresentam múltiplas tags devido a poliadenilação alternativa e splicing alternativo, alguns genes senso podem apresentar duas ou mais tags antisenso, mapeadas em exons ou em introns. Para tornar a nossa identificação mais detalhada, dividimos os NATs de acordo com o número e a região de mapeamento das tags: única tag exônica; única tag intrônica; tags exônicas e intrônicas; múltiplas tags exônicas; múltiplas tags intrônicas (Tabela 15).
Para validar experimentalmente a existência destes novos NAT, utilizamos a técnica de GLGI aplicada a MPSS (Silva, De Souza et al., 2004) (este experimento foi feito em colaboração, ver métodos). Resumidamente, esta técnica expande a tag de MPSS até o final da região 3’ UTR do transcrito por ela representado. Testamos 96 tags e, para a maioria, vemos a identificação de uma banda (Figura 22). Todos os fragmentos amplificados foram purificados, clonados, seqüenciados e alinhados no genoma. Fomos capazes de gerar um fragmento 3’ específico para 46 das 96 tags testadas. Apesar da orientação da tag (e do fragmento de GLGI) estar bem definida, foi feita uma RT-PCR fita específica para três destes NATs. Confirmamos a orientação dos três. Entre os 52% (96 – 46 = 50) candidatos considerados inválidos, 25 apresentaram orientação antisenso, mas uma organização exon/intron idêntica a do gene senso. Também fizemos RT-PCR fita específica para 2 destes genes, e um teve a orientação confirmada.

|                         | Número de clusters |       |       |
|-------------------------|--------------------|-------|-------|
|                         | Humano             | Camundongo |
| Única tag exônica       | 2121 (51,3%)       | 124 (57,3%) |
| Única tag intrônica     | 875 (20,3%)        | 90 (41,7%)  |
| Tags exônicas e intrônicas | 707 (16,4%)  | 2 (1%)     |
| Múltiplas tags exônicas | 318 (7,4%)        | 0     |
| Múltiplas tags intrônicas | 196 (4,6%) | 0     |
| Total                   | 4308               | 216   |
Como sugerido por Dahary e colaboradores (Dahary, Elroy-Stein et al., 2005), a sobreposição de alguns pares S-AS envolve transcritos gerados por poliadenilação alternativa. Então, decidimos quantificar a presença deste fenômeno na definição dos nossos pares S-AS. Para os pares S-AS 3’-3’, verificamos que a poliadenilação alternativa é o principal mecanismo responsável pelo surgimento da sobreposição entre dois genes. Em 51% (274 / 533) dos casos a sobreposição só ocorre porque existe um transcrito mais longo gerado por poliadenilação alternativa.

**Figura 22 | Amplificação de GLGI-MPSS para 96 tags de MPSS.** Estão sendo mostrados os 96 fragmentos marcados com brometo de etídeo num gel de agarose. Enquanto algumas colunas apresentam uma simples banda, outras contem múltiplas bandas ou mesmo um *smear*. Foi utilizado um marcador molecular de 100 pb.

### 4.4.7 A poliadenilação alternativa na definição de novos pares S-AS
Num outro experimento, verificamos que a proporção dos pares S-AS é similar em todos os
cromossomos, com exceção do cromossomo X (dado não mostrado). Então, para tentar explicar esta
menor frequência, analisamos a presença de poliadenilação alternativa nos cromossomos humanos.
Verificamos que a taxa de poliadenilação alternativa é 27,5%, em média, nos cromossomos
autossônicos e, apenas, 20% no cromossomo X, uma diferença significativa ($\chi^2 = 34.91; gl = 1; p < 0.0001$).

4.4.8 Os NAT originados de primer interno

Durante a validação experimental dos novos NATs identificados por MPSS, verificamos que
uma fração significativa dos transcritos apresentou a cauda de poli(A) alinhada diretamente no genoma
(19 dos 46 fragmentos validados). Isso nos motivou a procurar casos com essas características no
conjunto de pares S-AS definidos por cDNAs. Em 18% (humano) e 26% (camundongo) dos pares S-
AS, ao menos um dos genes apresentou a cauda de poli(A) alinhada diretamente no genoma. Quando
consideramos apenas os pares mRNA-mRNA, estes valores se reduzem para 11,7% (homem) e 12,6%
(camundongo).

Duas hipóteses podem explicar o alinhamento da cauda poli(A) no genoma: i) um simples
artefato metodológico, a ocorrência de primer genômico gerados por contaminação de DNA na
bibliotecas de cDNA; ii) um evento biológico, a ocorrência de retrotransposição gênica (Long, Betran
et al., 2003), onde inevitavelmente a cauda de poli(A) é integrada na sequência genômica.

Para testar a primeira hipótese, desenvolvemos uma estratégia experimental que avalia a
influência de primer genômico no conjunto de transcritos S-AS. Começamos com reações de RT-PCR
em bibliotecas de cDNA (de fígado fetal, colon e pulmão) tratadas e não tratadas com DNAse (estes experimentos foram realizados em colaboração, ver métodos). Nas amostras não tratadas com DNAse, dos 11 pares S-AS testados, todos os foram amplificados (Figura 23, linhas 3 e 4). Entretanto, nas amostras tratadas com DNAse, não observamos amplificação específica para 7 (63,6%) candidatos (Figura 23, linhas 1 e 2). Sugerindo que parte dos NATs são artefatos, mas que parte é real.

A retrotransposição ocorre quando uma sequência de mRNA é retrotranscrita a cDNA e integrada no genoma. Portanto, todos os genes gerados por retrotransposição devem apresentar a cauda de poli(A) na sua extremidade 3’. Para testar se os NAT contendo a cauda de poli(A) alinhada no genoma são casos de retrotransposição, cruzamos estes NAT contra 4000 genes retrotransposto identificados por Marques e colaboradores (Marques, Dupanloup et al., 2005). Dos 4000 genes retrotranspostos, 413 estão no nosso banco de dados. Destes, 138 são NAT (70 dão origem aos pares S-AS tipo mRNA-mRNA e 68 do tipo mRNAs-ESTs). Dos 70 pares mRNA-mRNA, 78% dos NAT retrotranspostos se localizam integralmente no intron de um outro gene, coerente com o fato desta categoria ‘contido’ apresentar muitos pares e, principalmente, pares introns/sem introns.
Figura 23  RT-PCR de pares S-AS onde um dos genes apresentam a cauda de poli(A) alinhada no genoma. Foram utilizados RNA total de amostras de figado fetal, colon e pulmão. A reação de RT-PCR foi feita com o tratamento com DNAses (linhas 1 e 2) e sem o tratamento com DNAses (linhas 3 e 4). Além disso, as amostras das reações 3 e 4 foram contaminadas propositadamente com DNA genômico. Como controle, as reações de RT-PCR foram conduzidas com (linhas 1 e 3) e sem (linhas 2 e 4) a presença de uma enzima transcriptase. O DNA genômico foi o controle positivo para essas reações de PCR, linha 5. Reações sem o template, linha 6, foram o controle negativo.
5 Discussão

5.1 Estudo dos genes expressos nas regiões do cérebro de camundongo

O cérebro localiza-se no topo da hierarquia funcional dos órgãos dos mamíferos. Este órgão é responsável por diversas funções essenciais, tais como o controle homeostático, as funções cognitivas e a coordenação das atividades motoras. Outra característica do cérebro são suas regiões com identidades anatômicas e funcionais distintas. Acredita-se que grande parte das funções e características do cérebro, e de suas regiões cerebrais, seja consequência de uma expressão gênica diferenciada e restrita. Como o objetivo de estudar a expressão gênica no cérebro, nós analisamos, através de dado de MPSS, o transcriptoma de 10 regiões cerebrais, selecionamos os genes preferencialmente expressos em cada região e utilizamos dados públicos de hibridização in situ para confirmar a expressão preferencial de diversos genes no bulbo olfativo.

5.1.1 A complexidade da expressão gênica no cérebro

Nos mamíferos, o cérebro é o órgão que apresenta o maior conjunto de transcritos diferentes (Chaudhari e Hahn, 1983; Milner e Sutcliffe, 1983; Hatten e Heintz, 2005). Os nossos resultados confirmaram essa diversidade transcricional em camundongo. Primeiro, observamos que o número de
tags de MPSS para as regiões cerebrais supera a média dos números de tags para os outros órgãos. Como é razoável imaginar que o número de tags é proporcional ao número de transcritos, podemos sugerir que os nossos resultados confirmam a grande diversidade de transcritos no cérebro (Tabelas 1 e 2). Já no resultado do agrupamento hierárquico (Figura 4), observamos uma enorme diferença entre o conjunto de tecidos cerebrais e os não cerebrais. O cérebro apresentou muitos genes exclusivos ou com uma intensidade de expressão maior do que a maioria dos outros órgãos.

5.1.2 Genes expressos preferencialmente nas regiões do cérebro

Todos os órgãos do corpo devem apresentar um transcriptoma característico. Entre eles, o cérebro, por sua complexidade, deve apresentar transcritomas diferentes para as diferentes regiões cerebrais. No nosso resultado de agrupamento hierárquico podemos ver essa identidade transcripcional entre os órgãos e, em especial, entre as regiões cerebrais (Figura 4). Alguns genes são sempre expressos, mas em níveis diferentes; outros são exclusivos. Num experimento mais quantitativo, identificamos mais de mil genes expressos de maneira preferencial em 10 regiões cerebrais analisadas. Sem dúvida, estudar estes genes exclusivos é uma ótima via para se explorar funções desconhecidas e patologias dos órgãos ou regiões cerebrais onde eles são expressos preferencialmente.

Entre as regiões distintas do cérebro de camundongo, algumas apresentam uma maior complexidade funcional e anatômica. Por exemplo, o cerebelo é uma região grande e independente, localizada na parte posterior do cérebro de camundongo. Entre suas funções, podemos destacar o controle e parte do planejamento motor (Kandel, Schwartz et al., 2000). O bulbo olfativo (especialmente nos roedores) é uma região enorme, com vários tipos celulares e está localizada na
parte anterior do cérebro. Funcionalmente, esta região recebe informações dos neurônios do epitélio olfativo e as transmite para o córtex olfativo e regiões límbicas (Kandel, Schwartz et al., 2000). A amígdala é uma pequena região localizada na parte inferior do cérebro de camundongo e está envolvida em processos comportamentais, tais como o desencadeamento do medo (Kandel, Schwartz et al., 2000). Parece razoável imaginar que as diferença anatômicas e funcionais estejam intimamente relacionadas às diferenças na expressão gênica. Avaliamos esta questão através da comparação, entre as regiões cerebrais, do número total de genes expressos e de genes expressos preferencialmente. Nossos resultados mostraram que há mais genes expressos nas regiões mais complexas. No extremo inferior está a amígdala e o superior o cerebelo ou o bulbo olfativo, com uma diferença de 16%. Contudo, a maior variação (proporcional) ocorre com os genes expressos de maneira preferencial. Para estes genes, no limite inferior temos o tubérculo olfativo e no superior, novamente, o cerebelo, com uma diferença de 72%. É razoável imaginar que a complexidade transcricional do órgão seja mais evidente quando estamos tratando os genes com expressão preferencial, pois, como discutido anteriormente, são eles que determinam parte das funções específicas destes órgãos.

5.1.3 Genes expressos no bulbo olfativo

Previamente, alguns trabalhos já haviam estudado a expressão gênica no bulbo olfativo (OB). Por exemplo, Sandberg e colaboradores (Sandberg, Yasuda et al., 2000) analisaram a expressão gênica em múltiplas regiões do cérebro, incluído o OB. Zirlinger e colaboradores (Zirlinger, Kreiman et al., 2001) fizeram um estudo comparativo da expressão gênica em 5 regiões do cérebro, entre as regiões estava o OB. Lin e colaboradores (Lin, Yang et al., 2004) fizeram um estudo espacial (em três
dimensões) da expressão génica em regiões do OB. Entretanto, em nenhum destes trabalhos foi feito um estudo comparativo entre a expressão génica no OB e tantas outras regiões do cérebro ou órgãos não cerebrais. Essa comparação é importante, pois quanto mais completo o repertório de tecidos, menor a chance de não se identificar o tecido no qual o gene realmente tem uma expressão diferencial, isto é, evitar falso positivo. Também nunca foi feito um estudo detalhado dos genes expressos no OB utilizando uma técnica em larga escala, que mede a expressão génica *ab initio* e quantifica, inclusive, genes com expressão muito baixa (como devem ser os genes preferenciais). Vale ressaltar que na maioria dos casos, o método usado para medir a expressão génica foi *microarray*, um método que tem suas vantagens, mas que não é eficiente para identificar novos genes ou genes sem muito conhecimento prévio.

Nossos resultados mostraram candidatos a serem expressos preferencialmente no OB, e a expressão de muitos foi validada por ISH. Entre os genes, temos fatores de transcrição, receptores e muitos genes sem uma função conhecida. Como discutido anteriormente, o estudo de genes com expressão restrita num órgão pode ser uma ótima via para se explorar funções desconhecidas ou pouco estudadas destes órgãos. Estes genes podem ajudar na resposta de questões relacionadas à neurogênese no OB (Gould, 2007) e aos mecanismos que guiam os axônios dos neurônios do epitélio olfativo a regiões específicas do OB (Lin, Yang *et al.*, 2004), por exemplo.
5.2 Identificação de seqüências reguladoras nos promotores dos genes ORs

Em camundongos existem mais de 1000 genes de receptores olfativos (OR). Estes genes estão presentes em quase todos os cromossomos e são expressos especificamente nos neurônios do epitélio olfativo, onde cada neurônio expressa um único OR, de um único alelo. Entretanto, o mecanismo responsável por este controle fino é pouco conhecido. Neste trabalho nós realizamos um estudo comparativo da região promotora de 198 ORs e identificamos elementos reguladores em cis que podem estar envolvidos na regulação da expressão destes genes.

5.2.1 Motivos presentes nas regiões promotoras dos genes ORs.

Através da comparação entre as regiões promotoras de 198 genes ORs, identificamos elementos (motivos) que podem estar envolvidos no mecanismo de regulação da expressão destes genes responsáveis pela detecção dos odores. Os motivos identificados foram divididos em duas categorias: i) similar ao sítio de ligação do domínio O/E (O/E-similares); ii) similar ao sítio de ligação do domínio homeodomain (homeodomain-similares). Os domínios O/E-similares e homeodomain-similares estão presentes em 87% e 95% das regiões promotoras dos ORs, respectivamente. Apesar de estes elementos serem os mais freqüentes e de termos utilizado três métodos (Hertz e Stormo, 1999; Thompson, Rouchka et al., 2003; Pavesi, Mereghetti et al., 2004) para identificá-los, não está excluída a possibilidade de existirem outros motivos conservados, e menos freqüentes que estes.
Já foi demonstrado que a região promotora mínima do gene OR M71 deve conter um sítio de ligação O/E-similar e homeodomain-similar (Vassalli, Rothman et al., 2002). Este resultado também foi corroborado com estudos de mutação destas regiões em camundongos transgenicos (Rothman, Feinstein et al., 2005). Por outro lado, a destruição do gene Olf-1 não altera a expressão dos genes ORs (Wang, Lin et al., 1993; Lin e Grosschedl, 1995), possivelmente porque há uma redundância funcional entre os genes da família do Olf-1 (esta família gênica é composta de 4 membros, O/E-1, O/E-2, O/E-3 e O/E-4) (Wang, Tsai et al., 1997; Wang, Betz et al., 2002). Por outro lado, foi mostrado que camundongos sem os genes O/E-2 e O/E-3 apresentam defeitos na projeção dos axônios dos neurônios do epitélio nasal para o bulbo olfativo, indicando que a função destes genes O/E-similares não apresenta uma completa redundância.

Foi mostrado que a proteína do gene Lhx2 LIM-homeodomain se liga ao sítio homeodomain na região promotora do gene OR M71 (Hirota e Mombaerts 2004). Camundongos knockout para o Lhx2 não expressam ORs, mas também não contêm os neurônios olfativos maduros, sugerindo que não há uma evidência direta do envolvimento da proteína do homeodomain na expressão dos ORs (Hirota e Mombaerts, 2004; Kolterud, Alenius et al., 2004). É possível que outras proteínas, além do Lhx2, também se liguem aos elementos homeodomain-similar presentes na região promotora dos ORs.

Apesar de estarem presentes na maioria das regiões promotoras dos ORs (87%), os motivos O/E-similar são razoavelmente variáveis (ver Figuras 13 e 14). Foi mostrado anteriormente que diferentes proteínas O/E podem se ligar a motivos similares (Wang, Tsai et al., 1997). Nossos resultados indicam que os motivos M1-M4 interagem diferentemente com as proteínas nucleares do epitélio olfativo. Várias possibilidades podem explicar esta diferença de afinidade. Diferentes proteínas O/E-similares ou variantes de splicing destas proteínas podem estar se ligando, com diferentes
afinidades, aos motivos M1-M4 (Wang, Tsai et al., 1997). Como os motivos M1-M4 não são idênticos, podemos imaginar que alguns têm maior afinidade pelas proteínas O/E-similares.

5.2.2 A região promotora e a regulação da expressão dos genes ORs

Até o momento, existem alguns modelos que tentam explicar parte da regulação da expressão dos genes ORs (Sosinsky, Glusman et al., 2000; Mombaerts, 2004; Serizawa, Miyamichi et al., 2004; Shykind, Rohani et al., 2004). Recentemente foi demonstrado que a expressão monoalélica dos ORs é regulada por um mecanismo de retroalimentação negativo, o qual requer a expressão de um gene OR funcional (Serizawa, Miyamichi et al., 2003; Lewcock e Reed, 2004). Adicionalmente, foi mostrado que se um neurônio olfativo imaturo expressar um OR funcional, a troca deste gene praticamente não ocorre. Por outro lado, se um neurônio olfativo expressar um OR não funcional, a troca deste por um OR funcional é muito rápida (Shykind, Rohani et al., 2004). Baseado nestes resultados, foi proposto um novo modelo para explicar a expressão dos ORs. Neste modelo, primeiramente, fatores em trans presentes em quantidades limitadas escolhem estocasticamente um OR funcional para ser expresso e, em seguida, a proteína gerada a partir da tradução deste OR participa do sinal de retroalimentação negativa, mantendo a expressão deste único OR. As sequências reguladoras em cis identificadas neste trabalho podem ter um papel fundamental, pois elas podem funcionar com ponto de ligação destes fatores em trans e participar diretamente do processo de escolha e manutenção da expressão dos genes ORs. Acreditamos que estudos mais detalhados e com mais análises funcionais devem ajudar na determinação do real papel destes elementos em cis e de elementos em trans no controle da expressão dos ORs.
5.3 Identificação dos eventos de retenção de introns

Nos últimos anos, com a enorme disponibilidade de dados de genomas e transcriptoma, identificou-se uma alta frequência de splicing alternativo em humano (Mironov, Fickett et al., 1999; Croft, Schandorff et al., 2000; Modrek e Lee, 2002). No mínimo, metade dos genes humanos apresenta splicing alternativo (Modrek e Lee, 2002). Entre os quatro tipos de splicing alternativo (uso alternativo de exon, sitio alternativo 3’, sitio alternativo 5’ e retenção de intron), sem dúvida o uso alternativo de exon é o mais freqüente e o mais estudado, enquanto a retenção de intron é o menos freqüente e o menos estudado. Neste trabalho, nós fizemos uma identificação em larga escala dos introns retidos em humanos e abordamos algumas de suas características funcionais e evolutivas.

5.3.1 Características dos introns retidos

É razoável imaginar que o processo de splicing tem uma taxa de erro intrínseca, muito difícil de ser estimada, e que o resultado deste erro pode estar presente nos bancos de dados públicos de cDNAs. Então, um aspecto importante nos estudos de splicing alternativo é a adição de filtros metodológicos para se eliminar estas sequências problemáticas. Quando avaliamos a ocorrência de erros nos quatro tipos de splicing alternativo (uso alternativo de exons, retenção de intron, sitio alternativo 3’ e sitio alternativo 5’), percebemos que a retenção de intron é a mais afetada. Além de apresentar erros que
também ocorrem nos outros tipos de *splicing*, a retenção de intron também é afetada por mRNAs não processados ou processado parcialmente e por contaminantes genômicos nas bibliotecas de cDNAs.

Fazendo os filtros necessários para evitar contaminantes, nós encontramos a presença de retenção de intron em 14.8% dos genes humanos. Entretanto, (Kan, States *et al.*, 2002) encontraram taxas diferentes. Primeiro eles identificaram IR em 35% dos genes humanos, depois eles aplicaram métodos estatísticos e chegaram a uma taxa de apenas 5%. Qual a razão desta discrepância? Acreditamos que (Kan, States *et al.*, 2002), primeiro, podem ter usado critérios pouco precisos (obtendo os 35%) e depois, critérios estatísticos muito estringentes. Nós, em contrapartida, não usamos critérios estatísticos, mas fomos conservadores em relação aos critérios biológicos. Por exemplo, consideramos apenas seqüências com múltiplos exons, exigimos que as seqüências protótipo e variante compartilhassem ao menos uma borda de *splicing* e só selecionamos eventos de retenção de introns como ao menos uma seqüência de mRNA envolvida. É interessante notar que se considerarmos apenas os casos de retenção de intron envolvendo exclusivamente mRNAs, temos uma taxa de 4.6%.

Depois de identificarmos os transcritos afetados por retenção de introns, partimos para um estudo das características deste *splicing* alternativo. Primeiro, verificamos a distribuição dos introns retidos ao longo dos genes. Verificamos que ocorre uma baixa frequência de IR na CDS e uma alta taxa nas regiões 3’ UTR (a mais frequente) e 5’ UTR. Uma possível explicação para este viés é a não publicação de seqüências contendo um códon de parada prematuro. Para avaliar isso, só quantificamos os IR através dos mRNA do Mammalian Gene Collection (MGC) (Strausberg, Feingold *et al.*, 2002), pois neste projeto todas as seqüências produzidas são publicadas, incluindo aquelas com códon de parada prematuro. Entretanto, esta hipótese foi descartada, pois, mesmo com as seqüências do MGC, os introns retidos continuam significativamente mais frequentes na região UTR.
Uma outra explicação para o viés UTR/CDS é a ocorrência de nonsense-mediated decay (NMD) (Gonzalez, Ruiz-Echevarria et al., 2000) (um mecanismo celular que degrada transcritos com CDSs trucadas) causado por códons de parada presente nos introns retidos. Entretanto, o NMD não explica a diferença (de 14 vezes) de introns retidos na região 3’ UTR e 5’ UTR. Uma explicação para este viés pode estar relacionada com a ordem de retirada dos introns, já que o processo de splicing ocorre praticamente junto com a transcrição (Kornblihtt, 2006), a região 3’ UTR deve ter menos tempo para realizar o splicing. Porém, existe uma evidência contrária a está hipótese: 90% dos introns retidos identificados apresentam ao menos um intron não retido downstream. Uma segunda explicação para este viés entre as regiões UTR é a possível ligação do evento de retenção e a estabilidade da molécula de mRNA. Elementos em cis presentes na região 3’ UTR são conhecidos por alterarem a estabilidade do mRNA (Bashirullah, Cooperstock et al., 2001), tornando-o mais ou menos estável, e a presença do IR poderia ter um papel fundamental neste processo. Parece mais razoável imaginar que o viés da distribuição dos IR em relação as regiões CDS e UTR é uma mistura de todos estes elementos aqui destacados.

Stamm e colaboradores (Stamm, Zhu et al., 2000) observaram que os intron retidos são significativamente menores que os introns não retidos. Nós também observamos esta tendência (ver Figura suplementar 1 em http://www.compbio.ludwig.org.br/~pgalante/IR). Contudo, tomamos muito cuidado ao interpretar este resultado, pois existe um limitante técnico: a eficiência de clonagem pode favorecer a amplificação de fragmentos curtos, desfavorecendo a identificação de longos introns retidos.

Um viés em relação ao tamanho da seqüência e a retenção de intron pode ser observado entre as ESTs e os mRNA. Nos nossos dados observamos que a categoria mais frequente é a retenção na seqüência de mRNA e a definição da retenção na seqüência de EST. Qual a razão? Sem dúvida o
tamanho das seqüências. Enquanto as ESTs são seqüências curtas de ~500 pb geradas a partir do sequenciamento parcial de um clone de cDNA, os mRNAs são seqüências longas geradas a partir do sequenciamento completo de um clone de cDNA. Além disso, no geral escolhe-se o maior clone do gene para o seqüenciamento completo (este clone pode ser o afetado pelo evento de retenção).

Também obtivemos resultados importantes quando comparamos os introns retidos e os exons. Verificamos que os introns retidos apresentam alto conteúdo de GC, mais similar ao dos exons do que ao dos introns não retidos. O uso de códons nos introns retidos também é mais similar ao dos exons. Por fim, verificamos que os introns retidos codificam domínios protêicos (algumas vezes apenas uma parte; outras o domínio inteiro), o que não ocorre com os introns não retidos.

5.3.2 Considerações evolutivas dos introns retidos

Nossos resultados mostraram que os introns retidos na CDS estão sob pressão seletiva para se tornar regiões codificadoras em potencial. Além dos introns retidos apresentarem alto conteúdo de GC, uso de códons e domínios protêicos similares aos dos exons, eles também contêm uma taxa menor de códons de parada do que o esperado por chance. Outra evidência de pressão seletiva é a conservação de eventos de IR nos camundongos (vale destacar que a identidade média entre os IR conservados é de 83%, similar a identidade de exons ortólogos entre estas duas espécies).

Apesar da baixa presença de códons de parada nos IRs, nós encontramos 88 casos de retenção de intron contendo um códon prematuro de parada. Para 40% destes casos, apesar de prematuro, o códon de parada estava localizado no exon mais 3’ (o que impede que a seqüência seja degradada por NMD). Quando analisamos o conteúdo de GC upstream e downstream ao códon de parada (sempre dentro do
intron retido), vemos que a concentração de GC é maior na região codificadora (região upstream) (58%) do que na região 3’ UTR (região downstream) (49%). Novamente um resultado que mostra a seleção atuando nestes introns retidos.

Como discutido originalmente por Gilbert (Gilbert, 1978), e reforçado recentemente por Black (Black, 2003), pode ocorrer casos onde os variantes de splicing dão origem a proteínas distintas das codificadas pelos transcritos constitutivos e, se estes novos variantes apresentarem alguma característica vantajosa, espera-se que ocorra a manutenção dos variantes. Colocando esta discussão no contexto dos introns retidos, podemos imaginar que os sinais observados (e compartilhados por quase todos os IR) são resultado de uma pressão seletiva atuante, o que pode indicar um significado biológico (ou o seu surgimento) para muitos dos introns retidos.

5.4 **Análise funcional e evolutiva dos pares de genes sense-antisense em homem e camundongo**

Os transcritos naturais antisenso (NATs) são transcritos endógenos com regiões complementares a outros transcritos. O maior obstáculo na identificação em larga escala dos NATs é definir corretamente a orientação das sequências públicas de cDNAs, especialmente as ESTs. Neste trabalho desenvolvemos um protocolo para a identificação de novos NATs através das tags de MPSS, selecionamos os pares de genes senso-antisenso determinados pelos cDNAs públicos (mRNAs e ESTs) e estudamos estes genes sob um aspecto funcional e evolutivo no genoma de homem e camundongo.
5.4.1 Os pares S-AS

Os pares S-AS são mais freqüentes do que o estimado originalmente para os mamíferos (Lehner, Williams et al., 2002; Shendure e Church, 2002). Nossa análise sugere que, no mínimo, 21.000 dos 35.000 genes humanos e 16.000 dos 32.000 genes de camundongo estão envolvidos em pares S-AS. Apesar de parecerem altos, estes números estão de acordo com os de outros trabalhos (Cheng, Kapranov et al., 2005; Katayama, Tomaru et al., 2005; Engstrom, Suzuki et al., 2006) e são resultado do protocolo em larga escala e da quantidade de dados que utilizamos.

5.4.2 A organização genômica dos pares S-AS

É possível imaginar que a organização genômica dos pares S-AS ocorre simplesmente por chance? Os nossos resultados e os de outros trabalhos (Chen, Sun et al., 2005a; b; Dahary, Elroy-Stein et al., 2005) sugerem que não, e se ocorrer é apenas para alguns pares S-AS. Do ponto de vista dos mecanismos de regulação entre os genes S-AS, a organização genômica tem um papel fundamental. Por exemplo, a regulação dos pares S-AS contido só pode ocorrer no compartimento nuclear, pois a sobreposição entre os pares só é intrônica. Por outro lado, para a maioria dos pares 3’-3’ e 5’-5’, por exemplo, a regulação também pode ocorrer no citoplasma através da formação de RNA dupla fita.

Do ponto de vista evolutivo, a organização genômica dos pares S-AS também é muito importante. Dahary e colaboradores (Dahary, Elroy-Stein et al., 2005) concluíram que os pares S-AS tem um efeito importante na evolução dos genomas dos vertebrados, pois esta organização é mais
conservada do que a organização dos genes em geral. A princípio, qualquer mudança nestas regiões de sobreposição gênica não vai alterar apenas um gene, mas sim dois.

5.4.3 A evolução dos pares S-AS

Neste trabalho, propusemos dois cenários evolutivos, não mutuamente exclusivos, para explicar a origem dos pares S-AS. No primeiro, a sobreposição entre o par S-AS é resultado da existência de transcritos com poliadenilação alternativa. Nossos resultados demonstraram que em 51% dos pares S-AS 3’-3’ é o variante de poliadenilação que apresenta sobreposição com o gene da fita complementar. No segundo cenário, o fenômeno de retrotansposição é que gera os pares S-AS. Estes genes retrotranspostos deram origem principalmente aos pares S-AS classificados como contido. Os genes antisenso produzidos por estes dois cenários evolutivos, poliadenilação alternativa e retrotransposição, devem regular os genes senso de maneiras diferentes. Enquanto os pares S-AS gerados por poliadenilação alternativa apresentam principalmente sobreposição entre os exons (sugerindo uma regulação através de RNA dupla fita no citoplasma), pares S-AS originados de retrotransposição apresentam complementaridade com a região intrônica do gene senso (sugerindo uma regulação a nível nuclear). Outra diferença esta na plasticidade da regulação: enquanto os pares gerados por poliadenilação podem apresentar uma regulação temporal (apenas no momento que o variante de poli(A) mais longo está sendo expresso), o casos de retrotransposição são sequências curtas e sem intron, o que favorece a velocidade de transcrição e consequentemente velocidade de regulação do gene senso (ver abaixo).
Como discutido acima, parece razoável imaginar que quanto mais rápida a transcrição e o processamento do gene antisenso, mais eficiente é a regulação do gene senso. Chen e colaboradores (Chen, Sun et al., 2005b) observaram que os genes antisenso apresentam poucos introns e sugeriram que isto está relacionado com uma transcrição rápida. Entre os três conjuntos de genes S-AS identificados, os pares contidos devem responder mais rapidamente a sua função de regulação do gene senso, pois eles não só apresentam uma transcrição rápida por não apresentarem introns, mas também não necessitam do processo de splicing.

5.4.4 A conservação dos pares S-AS entre humanos e camundongos

Características vantajosas são conservadas na evolução dos organismos (Darwin, Murray et al., 1859). Portanto, os S-AS conservados entre humanos e camundongos devem apresentar funções importantes nestes organismos. Neste trabalho, nós identificarmos ~1150 pares S-AS conservados entre Homo sapiens e Mus musculus. Vale ressaltar que este número deve ser uma sub estimativa do conjunto total de genes conservados entre estas duas espécies (já que eles representam apenas ~12% do total de pares S-AS) e, provavelmente, o número menor de seqüências de cDNAs em camundongo pode estar impedindo a identificação de muitos pares S-AS conservados.

Como foi mostrado, a distribuição 5’-5’, contido, 3’-3’ dos pares S-AS não ocorre por acaso. Então nós classificamos, nestas categorias, os pares S-AS conservado entre camundongo e homem. Verificamos que os casos 3’-3’ são os mais freqüentes e os contidos os menos freqüentes. Duas hipóteses devem explicar esta diferença: i) a alta conservação dos NAT 3’-3’ pode ser conseqüência deste tipo de sobreposição conseguir regular a expressão do gene senso através de múltiplos
mecanismos, tais como a interferência na transcrição, a formação de RNA dupla fita no núcleo e a formação de RNA dupla fita no citoplasma; ii) a abaixa freqüência de conservação dos pares contidos pode ser conseqüência destes genes serem gerados por retroposição, um fenômeno que deve ocorrer de maneira espécie específica em humano e em camundongo.

5.5 O estudo em larga escala da transcrição gênica

As análises em larga escala trouxeram uma nova maneira de se estudar os problemas biológicos. A visão do todo e não de um único componente do sistema está revolução entendimento dos fenômenos biológicos. Por exemplo, hoje nós sabemos a posição genômica exata, a seqüência completa, os polimorfismos e as isoformas de ~20k genes humanos e de camundongo. Tudo isso, graças à combinação precisa entre o acúmulo de dados e a análise destes dados em larga escala.

Nesta tese, nós desenvolvemos e aplicamos métodos computacionais no estudo em larga escala de quatro temas ligados à transcrição gênica em Homo sapiens e Mus musculus. Entretanto, todas as técnicas desenvolvidas empregadas não se limitaram a integrar e catalogar as informações, nós fizemos trabalhos com embasamento teórico e baseado em hipóteses (com uma pergunta biológica bem definida).

Todos os projetos aqui mostrados começaram com um insight (uma pergunta biológica) e com a análise do conjunto de informação depositada nos bancos de dados públicos. Os resultados foram obtidos através dos métodos computacionais que nós desenvolvemos ou a partir da aplicação dos métodos existentes ao nosso conjunto de dados. Em alguns casos aliamos as análises in silico com experimentos (na bancada) feitos por colaboradores. Sem dúvida, a integração entre métodos
computacionais e experimentos na bancada traz um ganho enorme, pois é possível testar exatamente o resultado que foi obtido no computador.

A análise *in silico* em larga escala é uma nova maneira de se estudar fenômenos biológicos e eu acredito que cada vez mais este tipo de estudo vai ser empregado e que, num futuro próximo, as análises em larga escala serão o paradigma dominante nas análises biológicas.
6 Conclusão

6.1 Estudo dos genes expressos no cérebro de camundongo

Utilizando praticamente todos os transcriptomas de camundongo, nós identificamos os genes expressos em 10 regiões do cérebro, incluindo os genes com expressão preferencial em diferentes regiões do cérebro. Nós também observamos a enorme diversidade de transcritos no cérebro e, por fim, confirmamos por hibridização in-situ a expressão gênica preferencial de alguns genes candidatos no bulbo olfativo. Acreditamos que o método desenvolvido foi adequado e que os resultados gerados neste projeto podem contribuir para se entender muitas das funções e características do cérebro como um todo, assim como a complexidade das regiões cerebrais.

6.2 Estudo das sequências reguladoras nos promotores dos genes ORs

Nossos resultados mostram que as regiões promotoras dos genes ORs apresentam sítios de ligação para dois conjuntos de fatores de transcrição, os fatores O/E-similar e os fatores homedomain-similar e que a quantidade de sítios (motivos) de ligação varia de gene para gene. Também mostramos que proteínas extraídas do núcleo de células do epitélio olfativo se ligam especificamente aos motivos identificados na análise in silico. Portanto, acreditamos que a nossa estratégia foi eficiente na
identificação dos motivos presentes nas regiões promotoras e que os motivos identificados são importantes no controle da expressão dos genes ORs.

6.3 Estudo dos eventos de retenção de introns

Primeiro, os nossos resultados mostraram que a retenção de intron é o tipo menos frequente de splicing alternativo, ocorrendo em ~15% dos genes humanos. Em seguida, mostramos que os introns retidos apresentam características similares aos exons e não aos introns não retidos. Por fim, mostramos que a seleção está atuando no introns retidos, deixando-os similares às sequências codificadoras. Agrupando e analisando todos os resultados, concluímos que estes introns retidos não são simplesmente produtos de transcrição espúria ou artefato experimental, mas sim eventos de splicing alternativo com uma função biológica no contexto celular.

6.4 Estudo dos pares S-AS em homem e camundongo

Dos trabalhos de identificação e estudo de características dos genes senso-antisenso, o nosso está entre os mais completos. Nós utilizamos todos os cDNAs públicos e 122 bibliotecas de MPSS na identificação do conjunto de genes S-AS em camundongo e humano. Em humanos, com os cDNAs, identificamos 10.077 pares S-AS; com as bibliotecas de MPSS identificamos mais 4308 pares S-AS. Em camundongo, com os cDNAs, identificamos 8091 pares S-SA; com as bibliotecas de MPSS
identificamos outros 216 pares S-AS. Outra contribuição importante do nosso trabalho foi dada no aspecto evolutivo dos genes S-SA: nós propusemos dois mecanismos para explicar a origem destes pares S-AS. No primeiro, os pares S-AS são gerados a partir de poliadenilação alternativa. No segundo, os pares S-SA são originados a partir de eventos de retrotransposição gênica. Também estudamos a co-ocorrência de NAT e splicing alternativo: com uma análise quantitativa, sugerimos que a presença NAT nas bordas exon/intron causa splicing alternativo e em especial a retenção do intron respectivo. Por fim, observamos que uma pequena fração dos pares S-AS de humano e camundongo são artefatos gerados pelo primer genômico durante a construção das bibliotecas de cDNAs.

### 6.5 Estudo da transcrição gênica através de técnicas computacionais

Como conclusão final, acreditamos que os métodos computacionais construídos e utilizados foram eficientes em todos os quatro estudos que realizamos nesta tese. Com estes métodos nós conseguimos explorar de maneira ampla todos os temas estudados e obtivemos resultados compreensivos, consistentes e conclusivos. Por fim, acreditamos que os resultados aqui descritos geraram importantes contribuições para o entendimento de cada um dos temas abordados e, de uma forma geral, para o entendimento dos processos relacionados à transcrição gênica como um todo.
7 Referências

Adams, M. D., S. E. Celniker, et al. The genome sequence of Drosophila melanogaster. *Science*, v.287, n.5461, Mar 24, p.2185-95. 2000.

Al-Bader, M. D. e H. A. Al-Sarraf. Housekeeping gene expression during fetal brain development in the rat-validation by semi-quantitative RT-PCR. *Brain Res Dev Brain Res*, v.156, n.1, Apr 21, p.38-45. 2005.

Alberts, B. *Molecular Biology of the Cell*. New York: GARLAND. 2007

Albright, T. D., T. M. Jessell, et al. Neural science: a century of progress and the mysteries that remain. *Cell*, v.100 Suppl, Feb 18, p.S1-55. 2000.

Altschul, S. F., T. L. Madden, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, v.25, n.17, Sep 1, p.3389-402. 1997.

Bachellerie, J. P., J. Cavaille, et al. The expanding snoRNA world. *Biochimie*, v.84, n.8, Aug, p.775-90. 2002.

Bashirullah, A., R. L. Cooperstock, et al. Spatial and temporal control of RNA stability. *Proc Natl Acad Sci U S A*, v.98, n.13, Jun 19, p.7025-8. 2001.

Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, v.72, p.291-336. 2003.

Boguski, M. S., T. M. Lowe, et al. dbEST--database for "expressed sequence tags". *Nat Genet*, v.4, n.4, Aug, p.332-3. 1993.

Boguski, M. S. e G. D. Schuler. ESTablishing a human transcript map. *Nat Genet*, v.10, n.4, Aug, p.369-71. 1995.

Boon, K., E. C. Osorio, et al. An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A*, v.99, n.17, Aug 20, p.11287-92. 2002.

Brenner, S., M. Johnson, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, v.18, n.6, Jun, p.630-4. 2000.

Bruhn, S. L. e C. L. Cepko. Development of the pattern of photoreceptors in the chick retina. *J Neurosci*, v.16, n.4, Feb 15, p.1430-9. 1996.

Buck, L. e R. Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, v.65, n.1, Apr 5, p.175-87. 1991.
Buonviso, N. e M. A. Chaput. Response similarity to odors in olfactory bulb output cells presumed to be connected to the same glomerulus: electrophysiological study using simultaneous single-unit recordings. *J Neurophysiol*, v.63, n.3, Mar, p.447-54. 1990.

Burke, J., H. Wang, *et al.* Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res*, v.8, n.3, Mar, p.276-90. 1998.

Butler, D. Genomics. Are you ready for the revolution? *Nature*, v.409, n.6822, Feb 15, p.758-60. 2001.

Carninci, P. Constructing the landscape of the mammalian transcriptome. *J Exp Biol*, v.210, n.Pt 9, May, p.1497-506. 2007.

Chaudhari, N. e W. E. Hahn. Genetic expression in the developing brain. *Science*, v.220, n.4600, May 27, p.924-8. 1983.

Chen, H. W., S. L. Yu, *et al.* Dynamic changes of gene expression profiles during postnatal development of the heart in mice. *Heart*, v.90, n.8, Aug, p.927-34. 2004.

Chen, J., M. Sun, *et al.* Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet*, v.21, n.6, Jun, p.326-9. 2005a.

______. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet*, v.21, n.4, Apr, p.203-7. 2005b.

______. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res*, v.32, n.16, p.4812-20. 2004.

Cheng, J., P. Kapranov, *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, v.308, n.5725, May 20, p.1149-54. 2005.

Chess, A., I. Simon, *et al.* Allelic inactivation regulates olfactory receptor gene expression. *Cell*, v.78, n.5, Sep 9, p.823-34. 1994.

Claverie, J. M. Gene number. What if there are only 30,000 human genes? *Science*, v.291, n.5507, Feb 16, p.1255-7. 2001.

Croft, L., S. Schandorff, *et al.* ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet*, v.24, n.4, Apr, p.340-1. 2000.

Dahary, D., O. Elroy-Stein, *et al.* Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res*, v.15, n.3, Mar, p.364-8. 2005.

Darwin, C., J. Murray, *et al.* On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life. London: John Murray ... : Printed by W. Clowes and Sons ... 1859. ix, [1], 502 p., [1] folded leaf of plates p.
Eddy, S. R. Non-coding RNA genes and the modern RNA world. Nat Rev Genet, v.2, n.12, Dec, p.919-29. 2001.

Eggan, K., K. Baldwin, et al. Mice cloned from olfactory sensory neurons. Nature, v.428, n.6978, Mar 4, p.44-9. 2004.

Eisen, M. B., P. T. Spellman, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, v.95, n.25, Dec 8, p.14863-8. 1998.

Engstrom, P. G., H. Suzuki, et al. Complex Loci in human and mouse genomes. PLoS Genet, v.2, n.4, Apr, p.e47. 2006.

Florea, L., G. Hartzell, et al. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res, v.8, n.9, Sep, p.967-74. 1998.

Fuss, S. H., M. Omura, et al. Local and cis effects of the H element on expression of odorant receptor genes in mouse. Cell, v.130, n.2, Jul 27, p.373-84. 2007.

Galante, P. A., N. J. Sakabe, et al. Detection and evaluation of intron retention events in the human transcriptome. RNA, v.10, n.5, May, p.757-65. 2004.

Galante, P. A., J. Trimarchi, et al. Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data. Bioinformatics, v.23, n.7, Apr 1, p.903-5. 2007.

Galante, P. A., D. O. Vidal, et al. Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biol, v.8, n.3, p.R40. 2007.

Gilbert, W. Why genes in pieces? Nature, v.271, n.5645, Feb 9, p.501. 1978.

Glusman, G., A. Sosinsky, et al. Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. Genomics, v.63, n.2, Jan 15, p.227-45. 2000.

Godfrey, P. A., B. Malnic, et al. The mouse olfactory receptor gene family. Proc Natl Acad Sci U S A, v.101, n.7, Feb 17, p.2156-61. 2004.

Gonzalez, C. I., M. J. Ruiz-Echevarria, et al. The yeast hnRNP-like protein Hrp1/Nab4 marks a transcript for nonsense-mediated mRNA decay. Mol Cell, v.5, n.3, Mar, p.489-99. 2000.

Gould, E. How widespread is adult neurogenesis in mammals? Nat Rev Neurosci, v.8, n.6, Jun, p.481-8. 2007.

Hatten, M. E. e N. Heintz. Large-scale genomic approaches to brain development and circuitry. Annu Rev Neurosci, v.28, p.89-108. 2005.

Hertz, G. Z. e G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, v.15, n.7-8, Jul-Aug, p.563-77. 1999.
Hirota, J. e P. Mombaerts. The LIM-homeodomain protein Lhx2 is required for complete development of mouse olfactory sensory neurons. Proc Natl Acad Sci U S A, v.101, n.23, Jun 8, p.8751-5. 2004.

Holste, D., G. Huo, et al. HOLLYWOOD: a comparative relational database of alternative splicing. Nucleic Acids Res, v.34, n.Database issue, Jan 1, p.D56-62. 2006.

Hoppe, R., H. Frank, et al. The clustered olfactory receptor gene family 262: genomic organization, promotor elements, and interacting transcription factors. Genome Res, v.13, n.12, Dec, p.2674-85. 2003.

Kan, Z., D. States, et al. Selecting for functional alternative splices in ESTs. Genome Res, v.12, n.12, Dec, p.1837-45. 2002.

Kandel, E. R., J. H. Schwartz, et al. Principles of neural science. New York: McGraw-Hill, Health Professions Division. 2000. xli, 1414 p. p.

Katayama, S., Y. Tomaru, et al. Antisense transcription in the mammalian transcriptome. Science, v.309, n.5740, Sep 2, p.1564-6. 2005.

Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res, v.12, n.4, Apr, p.656-64. 2002.

Kitano, H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet, v.41, n.1, Apr, p.1-10. 2002.

Kiyosawa, H., I. Yamanaka, et al. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res, v.13, n.6B, Jun, p.1324-34. 2003.

Kolterud, A., M. Alenius, et al. The Lim homeobox gene Lhx2 is required for olfactory sensory neuron identity. Development, v.131, n.21, Nov, p.5319-26. 2004.

Kornbllihtt, A. R. Chromatin, transcript elongation and alternative splicing. Nat Struct Mol Biol, v.13, n.1, Jan, p.5-7. 2006.

Kozak, M. Initiation of translation in prokaryotes and eukaryotes. Gene, v.234, n.2, Jul 8, p.187-208. 1999.

Kumar, M. e G. G. Carmichael. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. Microbiol Mol Biol Rev, v.62, n.4, Dec, p.1415-34. 1998.

Lander, E. S., L. M. Linton, et al. Initial sequencing and analysis of the human genome. Nature, v.409, n.6822, Feb 15, p.860-921. 2001.

Lane, R. P., T. Cutforth, et al. Genomic analysis of orthologous mouse and human olfactory receptor loci. Proc Natl Acad Sci U S A, v.98, n.13, Jun 19, p.7390-5. 2001.
Lash, A. E., C. M. Tolstoshev, et al. SAGEmap: a public gene expression resource. *Genome Res*, v.10, n.7, Jul, p.1051-60. 2000.

Lau, N. C., A. G. Seto, et al. Characterization of the piRNA complex from rat testes. *Science*, v.313, n.5785, Jul 21, p.363-7. 2006.

Lavorgna, G., D. Dahary, et al. In search of antisense. *Trends Biochem Sci*, v.29, n.2, Feb, p.88-94. 2004.

Le Hir, H., N. Charlet-Berguerand, et al. 5'-End RET splicing: absence of variants in normal tissues and intron retention in pheochromocytomas. *Oncology*, v.63, n.1, p.84-91. 2002.

Lee, R. C., R. L. Feinbaum, et al. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, v.75, n.5, Dec 3, p.843-54. 1993.

Lehner, B., G. Williams, et al. Antisense transcripts in the human genome. *Trends Genet*, v.18, n.2, Feb, p.63-5. 2002.

Lein, E. S., M. J. Hawrylycz, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, v.445, n.7124, Jan 11, p.168-76. 2007.

Leon, M. e B. A. Johnson. Olfactory coding in the mammalian olfactory bulb. *Brain Res Brain Res Rev*, v.42, n.1, Apr, p.23-32. 2003.

Lewcock, J. W. e R. R. Reed. A feedback mechanism regulates monoallelic odorant receptor expression. *Proc Natl Acad Sci U S A*, v.101, n.4, Jan 27, p.1069-74. 2004.

Li, J., T. Ishii, et al. Odorant receptor gene choice is reset by nuclear transfer from mouse olfactory sensory neurons. *Nature*, v.428, n.6981, Mar 25, p.393-9. 2004.

Lin, D. M., Y. H. Yang, et al. Spatial patterns of gene expression in the olfactory bulb. *Proc Natl Acad Sci U S A*, v.101, n.34, Aug 24, p.12718-23. 2004.

Lin, H. e R. Grosschedl. Failure of B-cell differentiation in mice lacking the transcription factor EBF. *Nature*, v.376, n.6537, Jul 20, p.263-7. 1995.

Lomvardas, S., G. Barnea, et al. Interchromosomal interactions and olfactory receptor choice. *Cell*, v.126, n.2, Jul 28, p.403-13. 2006.

Long, M., E. Betran, et al. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, v.4, n.11, Nov, p.865-75. 2003.

Lopez, R., V. Silventoinen, et al. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res*, v.31, n.13, Jul 1, p.3795-8. 2003.
Malnic, B., P. A. Godfrey, et al. The human olfactory receptor gene family. Proc Natl Acad Sci U S A, v.101, n.8, Feb 24, p.2584-9. 2004.

Malnic, B., J. Hirono, et al. Combinatorial receptor codes for odors. Cell, v.96, n.5, Mar 5, p.713-23. 1999.

Marques, A. C., I. Dupanloup, et al. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol, v.3, n.11, Nov, p.e357. 2005.

Michaloski, J. S., P. A. Galante, et al. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. Genome Res, v.16, n.9, Sep, p.1091-8. 2006.

Milner, R. J. e J. G. Sutcliffe. Gene expression in rat brain. Nucleic Acids Res, v.11, n.16, Aug 25, p.5497-520. 1983.

Mironov, A. A., J. W. Fickett, et al. Frequent alternative splicing of human genes. Genome Res, v.9, n.12, Dec, p.1288-93. 1999.

Modrek, B. e C. Lee. A genomic view of alternative splicing. Nat Genet, v.30, n.1, Jan, p.13-9. 2002.

Modrek, B. e C. J. Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet, v.34, n.2, Jun, p.177-80. 2003.

Modrek, B., A. Resch, et al. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res, v.29, n.13, Jul 1, p.2850-9. 2001.

Mombaerts, P. Odorant receptor gene choice in olfactory sensory neurons: the one receptor-one neuron hypothesis revisited. Curr Opin Neurobiol, v.14, n.1, Feb, p.31-6. 2004.

______. Axonal wiring in the mouse olfactory system. Annu Rev Cell Dev Biol, v.22, p.713-37. 2006.

Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci, v.13, n.4, Aug, p.477-8. 1997.

Munroe, S. H. Antisense RNA inhibits splicing of pre-mRNA in vitro. Embo J, v.7, n.8, Aug, p.2523-32. 1988.

Nakaya, H. I., P. P. Amaral, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol, v.8, n.3, p.R43. 2007.

Osato, N., Y. Suzuki, et al. Transcriptional interferences in cis natural antisense transcripts of humans and mice. Genetics, v.176, n.2, Jun, p.1299-306. 2007.
Pavesi, G., P. Mereghetti, et al. Weede Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res, v.32, n.Web Server issue, Jul 1, p.W199-203. 2004.

Pennisi, E. Why do humans have so few genes? Science, v.309, n.5731, Jul 1, p.80. 2005.

Pesole, G., F. Mignone, et al. Structural and functional features of eukaryotic mRNA untranslated regions. Gene, v.276, n.1-2, Oct 3, p.73-81. 2001.

Piao, X., R. S. Hill, et al. G protein-coupled receptor-dependent development of human frontal cortex. Science, v.303, n.5666, Mar 26, p.2033-6. 2004.

Prescott, E. M., Y. N. Osheim, et al. Transcriptional termination by RNA polymerase I requires the small subunit Rpa12p. Proc Natl Acad Sci U S A, v.101, n.16, Apr 20, p.6068-73. 2004.

Quackenbush, J., F. Liang, et al. The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res, v.28, n.1, Jan 1, p.141-5. 2000.

Reis, E. M., H. I. Nakaya, et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene, v.23, n.39, Aug 26, p.6684-92. 2004.

Rio, D. C. Regulation of Drosophila P element transposition. Trends Genet, v.7, n.9, Sep, p.282-7. 1991.

Rockman, M. V. e L. Kruglyak. Genetics of global gene expression. Nat Rev Genet, v.7, n.11, Nov, p.862-72. 2006.

Rothman, A., P. Feinstein, et al. The promoter of the mouse odorant receptor gene M71. Mol Cell Neurosci, v.28, n.3, Mar, p.535-46. 2005.

Sakurai, T. The neural circuit of orexin (hypocretin): maintaining sleep and wakefulness. Nat Rev Neurosci, v.8, n.3, Mar, p.171-81. 2007.

Sandberg, R., R. Yasuda, et al. Regional and strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A, v.97, n.20, Sep 26, p.11038-43. 2000.

Schmucker, D., J. C. Clemens, et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell, v.101, n.6, Jun 9, p.671-84. 2000.

Serizawa, S., K. Miyamichi, et al. Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. Science, v.302, n.5653, Dec 19, p.2088-94. 2003.

______. One neuron-one receptor rule in the mouse olfactory system. Trends Genet, v.20, n.12, Dec, p.648-53. 2004.
Shendure, J. e G. M. Church. Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol, v.3, n.9, Aug 22, p.RESEARCH0044. 2002.

Shepherd, G. M. The single capillary and the active brain. Proc Natl Acad Sci U S A, v.100, n.22, Oct 28, p.12535-6. 2003.

______. Outline of a theory of olfactory processing and its relevance to humans. Chem Senses, v.30 Suppl 1, Jan, p.i3-i5. 2005.

Shumyatsky, G. P., E. Tsvetkov, et al. Identification of a signaling network in lateral nucleus of amygdala important for inhibiting memory specifically related to learned fear. Cell, v.111, n.6, Dec 13, p.905-18. 2002.

Shykind, B. M., S. C. Rohani, et al. Gene switching and the stability of odorant receptor gene choice. Cell, v.117, n.6, Jun 11, p.801-15. 2004.

Silva, A. P., J. E. De Souza, et al. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. Nucleic Acids Res, v.32, n.20, p.6104-10. 2004.

Snyder, M. e M. Gerstein. Genomics. Defining genes in the genomics era. Science, v.300, n.5617, Apr 11, p.258-60. 2003.

Sosinsky, A., G. Glusman, et al. The genomic structure of human olfactory receptor genes. Genomics, v.70, n.1, Nov 15, p.49-61. 2000.

Stamm, S., J. Zhu, et al. An alternative-exon database and its statistical analysis. DNA Cell Biol, v.19, n.12, Dec, p.739-56. 2000.

Storch, K. F., O. Lipan, et al. Extensive and divergent circadian gene expression in liver and heart. Nature, v.417, n.6884, May 2, p.78-83. 2002.

Storz, G. An expanding universe of noncoding RNAs. Science, v.296, n.5571, May 17, p.1260-3. 2002.

Strausberg, R. L., E. A. Feingold, et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A, v.99, n.26, Dec 24, p.16899-903. 2002.

Su, A. I., M. P. Cooke, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A, v.99, n.7, Apr 2, p.4465-70. 2002.

Thompson, W., E. C. Rouchka, et al. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res, v.31, n.13, Jul 1, p.3580-5. 2003.

Uptain, S. M., C. M. Kane, et al. Basic mechanisms of transcript elongation and its regulation. Annu Rev Biochem, v.66, p.117-72. 1997.
Vanhee-Brossollet, C. e C. Vaquero. Do natural antisense transcripts make sense in eukaryotes? Gene, v.211, n.1, Apr 28, p.1-9. 1998.

Vassalli, A., A. Rothman, et al. Minigenes impart odorant receptor-specific axon guidance in the olfactory bulb. Neuron, v.35, n.4, Aug 15, p.681-96. 2002.

Velculescu, V. E., L. Zhang, et al. Serial analysis of gene expression. Science, v.270, n.5235, Oct 20, p.484-7. 1995.

Venter, J. C., M. D. Adams, et al. The sequence of the human genome. Science, v.291, n.5507, Feb 16, p.1304-51. 2001.

Visel, A., C. Thaller, et al. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acids Res, v.32, n.Database issue, Jan 1, p.D552-6. 2004.

Wagner, E. G. e R. W. Simons. Antisense RNA control in bacteria, phages, and plasmids. Annu Rev Microbiol, v.48, p.713-42. 1994.

Wang, M. M., R. Y. Tsai, et al. Genes encoding components of the olfactory signal transduction cascade contain a DNA binding site that may direct neuronal expression. Mol Cell Biol, v.13, n.9, Sep, p.5805-13. 1993.

Wang, S. S., A. G. Betz, et al. Cloning of a novel Olf-1/EBF-like gene, O/E-4, by degenerate oligo-based direct selection. Mol Cell Neurosci, v.20, n.3, Jul, p.404-14. 2002.

Wang, S. S., R. Y. Tsai, et al. The characterization of the Olf-1/EBF-like HLH transcription factor family: implications in olfactory gene regulation and neuronal development. J Neurosci, v.17, n.11, Jun 1, p.4149-58. 1997.

Wang, Z. Y., X. H. Lin, et al. Identification of a single-stranded DNA-binding protein that interacts with an S1 nuclease-sensitive region in the platelet-derived growth factor A-chain gene promoter. J Biol Chem, v.268, n.14, May 15, p.10681-5. 1993.

Wasserman, W. W. e A. Sandelin. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet, v.5, n.4, Apr, p.276-87. 2004.

Waterston, R. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science, v.282, n.5396, Dec 11, p.2012-8. 1998.

Waterston, R. H., K. Lindblad-Toh, et al. Initial sequencing and comparative analysis of the mouse genome. Nature, v.420, n.6915, Dec 5, p.520-62. 2002.

Will, C. L. e R. Luhrmann. Spliceosomal UsnRNP biogenesis, structure and function. Curr Opin Cell Biol, v.13, n.3, Jun, p.290-301. 2001.
Yan, M. D., C. C. Hong, et al. Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum Mol Genet*, v.14, n.11, Jun 1, p.1465-74. 2005.

Young, J. M., B. M. Shykind, et al. Odorant receptor expressed sequence tags demonstrate olfactory expression of over 400 genes, extensive alternate splicing and unequal expression levels. *Genome Biol*, v.4, n.11, p.R71. 2003.

Zhang, L., W. Zhou, et al. Gene expression profiles in normal and cancer cells. *Science*, v.276, n.5316, May 23, p.1268-72. 1997.

Zhang, X. E S. Firestein. The olfactory receptor gene superfamily of the mouse. *Nat Neurosci*, v.5, n.2, Feb, p.124-33. 2002.

Zhang, Y., X. S. Liu, et al. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res*, v.34, n.12, p.3465-75. 2006.

Zhang, Z., S. Schwartz, et al. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, v.7, n.1-2, Feb-Apr, p.203-14. 2000.

Zirlinger, M., G. Kreiman, et al. Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. *Proc Natl Acad Sci U S A*, v.98, n.9, Apr 24, p.5270-5. 2001.
LISTA DE ANEXOS

1. Súmula Curricular
2. Artigos incluídos nesta tese:
   a. Michaloski JS, Galante PA, Malnic B. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. *Genome Res.* 2006; 16(9): 1091-8.
   b. Galante PA, Trimarchi J, Cepko CL, de Souza SJ, Ohno-Machado L, Kuo WP. Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data. *Bioinformatics.* 2007; 23(7): 903-5.
   c. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. *RNA.* 2004; 10(5): 757-65.
   d. Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ. Sense-antisense pairs in mammals: functional and evolutionary Considerations. *Genome Biol.* 2007; 8(3): R40.
Anexo 1 – Súmula curricular
SÚMULA CURRICULAR

DADOS PESSOAIS
Nome: Pedro Alexandre Favoretto Galante
Local e data de nascimento: São Pedro - SP, 05/01/1979

EDUCAÇÃO
Julho/2003 – Dezembro/2007
Doutorado Direto em Bioquímica e Biologia Molecular no Departamento de
Bioquímica do Instituto de Química da USP / Laboratório de Biologia Computacional
do Instituto Ludwig de Pesquisa sobre o Câncer.

Agosto/1999 – Julho/2003
Curso de Ciência Moleculares na Universidade de São Paulo.

Fevereiro/1999 – Julho/1999
Bacharelado em Ciências da Computação no Instituto de Matemática e Estatística da
Universidade de São Paulo

FORMAÇÃO COMPLEMENTAR
Março/2003 – Junho/2003
Iniciação científica no exterior
Instituição: Duke University
Supervisor: Dr. Gregory Riggins

Julho/2005 – Agosto/2005
Summer student
Instituição: Harvard University
Supervisor: Dr. Lucila Ohno-Machado, Dr. Winton P. Kuo e Dr. Constance L. Cepko

BOLSAS RECEBIDAS
Doutorado Direto
Fevereiro/2004 - Fevereiro/2008
Instituição: FAPESP

Iniciação Científica.
Novembro/1999 – Agosto/2003
Instituição: CNPq
PUBLICAÇÕES (Artigos Completos)

10: Ferreira EN, Galante PA, Carraro DM, de Souza SJ. Alternative splicing: a bioinformatics perspective. Mol Biosyst. 2007; 3(7): 473-7.

9: Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ. Sense-antisense pairs in mammals: functional and evolutionary Considerations. Genome Biol. 2007; 8(3): R40.

8: Galante PA, Trimarchi J, Cepko CL, de Souza SJ, Ohno-Machado L, Kuo WP. Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data. Bioinformatics. 2007; 23(7): 903-5.

7: Michaloski JS, Galante PA, Malnic B. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. Genome Res. 2006; 16(9): 1091-8.

6: Parmigiani RB, Magalhães GS, Galante PA, Manzini CV, Camargo AA, Malnic B. A novel human G protein-coupled receptor is over-expressed in prostate cancer. Genet Mol Res. 2004; 3(4): 521-31.

5: Kirschbaum-Slager N, Lopes GM, Galante PA, Riggins GJ, de Souza SJ. Splicing factors are differentially expressed in tumors. Genet Mol Res. 2004; 3(4): 512-20.

4: 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. 2004; 32(20): 6104-10.

3: Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. RNA. 2004; 10(5): 757-65.

2: Sakabe NJ, de Souza JE, Galante PA, de Oliveira PS, Passetti F, Brentani H, Osório EC, Zaiats AC, Leerkes MR, Kitajima JP, Brentani RR, Strausberg RL, Simpson AJ, de Souza S, Stevenson BJ, Strausberg RL, Tajara EH, Verjovski-Almeida S, Acencio ML, Bengtson MH, Béton F, Bodmer WF, Briones MR, Camargo LP, Cavenee W, Cerutti JM, Coelho Andrade LE, Costa dos Santos PC, Ramos Costa MC, da Silva IT, Estécio MR, As Ferreira K, Furnari FB, Faria M Jr, Galante PA, Guimaraes GS, Holanda AJ, Kimura ET, Leerkes MR, et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A. 2003; 100(23): 13418-23.
Anexo 2a — Artigo: Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences
Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences

Jussara S. Michaloski, Pedro A.F. Galante and Bettina Malnic

*Genome Res.* 2006 16: 1091-1098; originally published online Aug 10, 2006; Access the most recent version at doi:10.1101/gr.5185406

### Supplementary data

"Supplemental Research Data"

[http://www.genome.org/cgi/content/full/gr.5185406/DC1](http://www.genome.org/cgi/content/full/gr.5185406/DC1)

### References

This article cites 49 articles, 20 of which can be accessed free at:

[http://www.genome.org/cgi/content/full/16/9/1091#References](http://www.genome.org/cgi/content/full/16/9/1091#References)

Article cited in:

[http://www.genome.org/cgi/content/full/16/9/1091#otherarticles](http://www.genome.org/cgi/content/full/16/9/1091#otherarticles)

### Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://www.genome.org/cgi/content/full/16/9/1091#otherarticles)

### Notes

To subscribe to *Genome Research* go to:

[http://www.genome.org/subscriptions/](http://www.genome.org/subscriptions/)

© 2006 Cold Spring Harbor Laboratory Press
Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences

Jussara S. Michaloski,1 Pedro A.F. Galante,1,2 and Bettina Malnic1,3

1Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, C.P.26077 CEP 05513-970, São Paulo, Brazil; 2Ludwig Institute for Cancer Research, São Paulo, 01509-010, SP, Brazil

Mouse odorant receptors (ORs) are encoded by >1000 genes dispersed throughout the genome. Each olfactory neuron expresses one single OR gene, while the rest of the genes remain silent. The mechanisms underlying OR gene expression are poorly understood. Here, we investigated if OR genes share common cis-regulatory sequences in their promoter regions. We carried out a comprehensive analysis in which the upstream regions of a large number of OR genes were compared. First, using RLM-RACE, we generated cDNAs containing the complete 5′-untranslated regions (5′-UTRs) for a total number of 198 mouse OR genes. Then, we aligned these cDNA sequences to the mouse genome so that the 5′ sequence and transcription start sites (TSSs) of the OR genes could be precisely determined. Sequences upstream of the TSSs were retrieved and browsed for common elements. We found DNA sequence motifs that are overrepresented in the promoter regions of the OR genes. Most motifs resemble O/E-like sites and are preferentially localized within 200 bp upstream of the TSSs. Finally, we show that these motifs specifically interact with proteins extracted from nuclei prepared from the olfactory epithelium, but not from brain or liver. Our results show that the OR genes share common promoter elements. The present strategy should provide information on the role played by cis-regulatory sequences in OR gene regulation.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. DR065530–DR065963.]
Results

Amplification of OR cDNA 5’-ends

To obtain 5’-ends sequences from OR cDNAs, we performed RNA ligase-mediated rapid amplification of 5’ cDNA ends (RLM-RACE) using total RNA purified from mouse olfactory epithelium. This method has the advantage that only full-length transcripts (authentic 5’-end capped mRNAs) are amplified (Fig. 1A). In order to obtain 5’ cDNA ends for a large number of OR genes, we used degenerate primers matching to conserved regions in ORs (Fig. 1B) because these primers can amplify the majority of the members of the OR family (Buck and Axel 1991; Malnic et al. 1999). The initial PCR reaction was performed using the 5’ GeneRacer primer together with the reverse P27 degenerate primer. In order to eliminate PCR artifacts, a secondary (nested) PCR was performed using the first PCR product as a template and the 5’ GeneRacer nested primer together with the P26R degenerate primer (Fig. 1A). The primary PCR product contained a heterogeneous mixture of cDNA fragments ranging from 0.8 to 2 kb in length (Fig. 1C, lane 2). This result was expected, since different OR cDNAs must have different 5’-UTR sizes. The secondary PCR reaction produced a similar range of cDNA fragments, except that their sizes were ~350 bp smaller (Fig. 1C, lane 3). This was also expected, since the P26R primer matches to a region in the OR coding sequence that is located ~350 bp upstream to the region matched by P27 (Fig. 1B). Since the region between the first AUG codon and TM-III in OR genes is ~380 bp long, only the nested PCR products ~380 bp long were gel-purified, cloned, and sequenced.

The same procedure was also performed using primers 5’ GeneRacer and P8 degenerate primer for the primary PCR reaction (Fig. 1C, lane 5), and primers 5’ GeneRacer nested and P27 for the secondary PCR reaction (Fig. 1C, lane 6). In this case, since the region between the first AUG codon and TM-VI in ORs is ~750 bp long, only the nested PCR products ≥750 bp were analyzed.

OR gene clusters

We sequenced 1012 clones from their 5’-ends, and 80% of them correspond to ORs, indicating that our strategy preferentially amplifies OR sequences. In addition, 96% of the OR cDNAs contain 5’-UTRs downstream of the predicted initial AUG codon, indicating that full-length mRNAs were amplified. Sequence analysis showed that 5’-RACE products were obtained for a total number of 198 different OR genes, corresponding to ~17% of the complete mouse OR gene repertoire (Table 1). Only nine of the OR genes are pseudogenes.

OR sequences are classified into two phylogenetic classes, referred to as Class I (fish-like) and Class II (terrestrial-specific) ORs (Ngai et al. 1993; Freitag et al. 1995; Glusman et al. 2000a). The Class I ORs constitute ~12% of the mouse OR repertoire (Zhang and Firestein 2002). Nine of the 198 ORs (4.5%) are Class I ORs, indicating that the method we used amplified members of the two OR Classes, although it may favor amplification of Class I ORs.

The OR genes for which cDNA sequences were obtained are distributed among all of the mouse chromosomes previously shown to contain OR genes (Godfrey et al. 2004), except for chromosome 3 (Table 1). The majority of the sequences corre-
Table 1. Chromosomal distribution of OR genes

| Chromosome | Annotated OR genes | OR genes in this study |
|------------|--------------------|------------------------|
| 1          | 23 (7)             | 6                      |
| 2          | 275 (81)           | 44 (5)                 |
| 3          | 2                  | —                      |
| 4          | 19 (7)             | 4                      |
| 5          | —                  | —                      |
| 6          | 22 (9)             | 5                      |
| 7          | 198 (51)           | 32                     |
| 8          | 4                  | 2                      |
| 9          | 118 (37)           | 20 (1)                 |
| 10         | 47 (13)            | 9                      |
| 11         | 35 (20)            | 21 (2)                 |
| 12         | —                  | —                      |
| 13         | 12 (2)             | 4                      |
| 14         | 32 (4)             | 6                      |
| 15         | 6 (3)              | 3 (1)                  |
| 16         | 29 (8)             | 18                     |
| 17         | 36 (17)            | 5                      |
| 18         | —                  | —                      |
| 19         | 52 (19)            | 9                      |
| X          | 2                  | 1                      |
| Y          | —                  | —                      |
| Total      | 1190               | 198                    |

*The number of annotated mouse OR genes from Godfrey et al. (2004) and of mouse OR genes for which cDNA sequences were obtained in the present study. Numbers of pseudogenes are indicated in parentheses.

spond to genes in chromosomes 2 and 7, which had been previously shown to contain the higher numbers of OR genes (Young et al. 2002; Zhang and Firestein 2002; Godfrey et al. 2004). In addition, the 198 OR genes can be subdivided into 102 (41%) out of the total 248 mouse OR subfamilies (where all members of a subfamily are ≥60% identical to all other members in amino acid sequence, as described by Godfrey et al. 2004; see Supplemental Table 1). We therefore believe that our sequences are representative of a random sample of the mouse OR genes. However, it is important to note that the present method may favor the amplification of OR cDNAs that have short 5'-UTRs or that are highly expressed in the olfactory epithelium.

We next used BLAST (Altschul et al. 1990) and Sim4 (Florea et al. 1998) to align all of the cDNA sequences with the mouse genome sequence. The previously annotated mouse OR genes (Young et al. 2002; Zhang and Firestein 2002) were also included in the alignment, to help with the localization of the cDNAs 5'-UTR regions. Each cDNA sequence aligned to one single genomic region, and the cDNAs that aligned to the same genomic region as one of the annotated ORs were considered to correspond to that particular OR gene. In this way, we obtained a total of 198 OR clusters, where each cluster corresponds to one different OR gene. Each one of the clusters contains at least one cDNA sequence, the largest cluster contains 71 sequences, and 54% of the clusters contain more than two sequences (Supplemental Table 1).

The structural organization of each one of the 198 clusters can be visualized using the Olfactory Receptor cDNA Clusters Viewer (http://gbrowser.compbio.ludwig.org.br/or/) by entering the corresponding cluster numbers shown in Supplemental Table 1. In summary, of the 198 OR gene clusters, only two do not have introns in their 5'-ends (Supplemental Tables 1 and 2), and 39 showed alternative splicing (in this case, only clusters containing more than two cDNA sequences were analyzed).

5'-structure of the OR genes

On average, the 5'-UTR is 189 bp long, ranging from 32 to 659 bp. The 5'-UTR exons range from 9 to 403 bp (Supplemental Table 2). Most of the OR genes (72%) have only one 5'-UTR exon, 23% have two, and 3% have three (Supplemental Tables 1 and 2). The 5'-UTR introns show a wide variation in size, ranging from 91 bp to 22.5 kb. A large fraction of the OR genes (46%) have introns with sizes between 2 kb and 4 kb, and 34% of the genes have introns >4 kb.

It is predicted that in most eukaryotic mRNAs translation initiates at the first AUG starting from the 5'-cap (Kozak 1999). Therefore, one would expect not to find AUG codons or upstream ORFs (uORFs) in the 5'-UTRs. However, we found that 108 of the 198 OR genes have uORFs at least 10 codons long. It has been suggested that uORFs could be involved in down-regulation of protein translational efficiency (Pesole et al. 2001). Further analysis should clarify whether these uORFs play a role in OR gene regulation.

We also found that a significant percentage of the OR genes (18%) have an in-frame upstream AUG, indicating that these genes code for OR proteins with a longer N-terminal region than the one originally predicted from their genomic sequences.

OR promoter regions

The generation of 5' full-length cDNAs allowed us to precisely determine the TSSs for the 198 OR genes. On average, the TSSs are located 4.3 kb upstream of the initial AUG codon, the furthest TSSs being located 22.5 kb away and the closest only 18 bp away. Sequences (600 bp) upstream of each TSS were excised from the mouse genomic sequence and analyzed. We first screened the
sequences with RepeatMasker (http://www.repeatmasker.org/) and found that only 7% of the total sequences contain repeats or low-complexity regions. Typical TATA-boxes were found in only a small number of the OR gene promoters (35%), consistent with previous reports (Hoppe et al. 2000; Sosinsky et al. 2000; Lane et al. 2001).

Because most TFBS are usually short, they can occur very frequently in the sequences, making it difficult to identify significant sites. In order to reduce the false-positive predictions, we decided to search for motifs that are common to a large fraction of the promoter sequences and thereby identify elements that are more likely to be functionally important. To do this, we used the Gibbs recursive sampler (Thompson et al. 2003), Consensus (Hertz and Stormo 1999), and Weeder (Pavesi et al. 2004) tools, which were designed to locate common elements in collections of unaligned DNA sequences. We found several motifs that are shared by the OR gene promoter sequences (Fig. 2). A closer inspection of the motifs revealed that although they are diverse, the majority of them resemble Olf-I (O/E) like sites (Figs. 2 and 3; Wang et al. 1993). The O/E-like motifs can be divided into four groups, denominated M1–M4. Motifs in each one of these four groups show different conserved nucleotide sequences (Fig. 3).

By comparing a variable set of OR genes, we aimed to identify common promoter elements that may be involved in the general mechanisms of OR gene regulation. Strikingly, 87% of the OR gene promoter regions contain O/E-like sites, and 95% contain homeodomain-like sites. No other types of motifs were found to be overrepresented in these sequences, although we cannot ex-
in the projection of olfactory neurons to the olfactory bulb, in-
O/E2, O/E3, and O/E4) (Wang et al. 1997, 2002). Nevertheless, it
possibly because of the functional redundancy of the multiple
gene expression (Lin and Grosschedl 1995; Wang et al. 2003),
are indicated by filled circles.

unlabeled specific oligonucleotide (Vassalli et al. 2006). The
results indicate that the M1–M4 motifs interact differently with
proteins from olfactory epithelium nuclei (Fig. 5). Several possi-
bilities could explain the different DNA–protein complex affini-
ties. The motifs could bind to different O/E proteins, or to alter-
atively spliced versions of these proteins (Wang et al. 1997,
2002). It is also possible that the same O/E protein types could
bind to the M1–M4 motifs, but with different affinities.

Our findings suggest that different OR gene promoters are
bound by different combinations or amounts of O/E-like pro-
teins. The consequences of these differential interactions for OR
gene regulation are unknown. It is known that different ORs are
expressed in different levels (Young et al. 2003). One interesting
possibility is that the types of O/E-like sites in an OR gene pro-
mitter region may determine its probability of being transcribed.

The identification of the proteins that interact with each one of
the motifs and the analysis of the expression patterns of OR
genes that have different motifs should clarify the role of these
O/E-like sites in OR gene regulation.

Promoter DNA elements and OR gene regulation

Different models for OR gene regulation have been considered to
date (for review, see Sosinsky et al. 2000; Mombaerts 2004; Ser-
izawa et al. 2005; Shykoid 2005). It has been recently dem-
strated that the monoallelic expression of an OR gene is regulated
by a negative feedback mechanism that requires a functional OR
protein (Serizawa et al. 2003; Lewcock and Reed 2004). In addi-
tion, it was shown that immature olfactory neurons that express
a given odorant receptor can switch receptor expression at a low
frequency, while neurons expressing a mutant (nonfunctional)
OR can switch expression with a greater probability (Shykoid et
al. 2004). Based on these results, a new model has been proposed
(Serizawa et al. 2004; Shykoid 2005). In this model, after an OR
gene is stochastically selected for expression by a limiting factor,
its corresponding OR protein product mediates a feedback signal
that results in the maintenance of the receptor choice.

Here we show that a collection of random OR genes will
have the same types of cis-regulatory elements, suggesting that
these common promoter elements are likely to play an important
role in OR gene expression. It is possible that enhancers or LCRs
interact with elements in one OR gene promoter to select that
specific OR for expression. Interestingly, it was shown that the H
region, which works as an LCR and is located 75 kb upstream of
the MOR28 gene cluster (Serizawa et al. 2003), also contains at
least one set of homeodomain- and O/E-like sites (Hirota and
Mombaerts 2004). Alternatively, cis-elements and protein factors
that bind to these elements could bring one given OR gene pro-
moter to a single expression site body in the nucleus (Borst 2002;
Voss et al. 2006).

Table 2. Distribution of the sequence motifs among the OR
genes

| Motif No. of OR genes (%) | Total no. of sites |
|--------------------------|-------------------|
| M1 85 (42%)              | 110               |
| M2 99 (50%)              | 129               |
| M3 71 (36%)              | 79                |
| M4 51 (26%)              | 64                |
| O/E-like sites 173 (87%) | 382               |
| Homeodomain sites 188 (95%) | 1029           |

*Number of OR gene promoters (% of 198 promoters) containing motifs M1–M4, total O/E-like sites (sum of M1–M4 motifs) and homeodomain-
like sites (based on the HAATTA consensus sequence).

Total number of motif sites found in the 198 OR gene promoters.
However, it is important to note that other olfactory genes that are expressed in all mature olfactory neurons, such as Omp, Gmα, and AdhI, also have O/E-like sites (Fig. 3). Therefore, the mere presence of O/E-like sites in the promoter regions does not explain the mosaic pattern of OR expression in the olfactory epithelium.

In conclusion, our results indicate that intraspecies comparisons of promoter sequences are likely to be a useful strategy for identifying common regulatory motifs that may be involved in regulation of OR gene expression. A similar strategy can also be applied to other multigene families whose members are coordinately regulated, such as the pheromone receptor families (Dulac and Torello 2003).

Methods

5’ RLM-RACE

Total RNA was purified from C57BL/6j mice (6–8 wk old) olfactory epithelium using TRIzol reagent (Invitrogen), following the manufacturer’s instructions. RLM-RACE was performed using the GeneRacer kit (Invitrogen) and 4 µg of total RNA. Twenty-five-microliter PCR reactions containing 1 µL of RLM-RACE cDNA, 0.2 mM dNTP, 1.5 mM MgCl2, 0.5 µM each forward and reverse primer (or 2 µM degenerate primers), 1.25 U of Platinum Taq DNA polymerase (Invitrogen) were heated to 95°C for 2 min, followed by 40 thermal cycles of 95°C for 1 min, 50°C for 3 min, 72°C for 2 min, and a final incubation at 72°C for 10 min.

Twenty-five-microliter nested PCR reactions were done using 1 µL of a 200-fold dilution of the primary PCR product and 25 cycles as above.

Degenerate oligonucleotide primers

The following degenerate OR primers were used in this study:

Reverse primers
P8 (TM-VII): (GA)TT(TG)IA(GT)I(GC)(TA)(GA)TA IAT(AG)AAIGG(GA)TT
P27 (TM-VI): ACIAICGAIAG(GA)TGIGAI(GC)(GA)CAIGT
P26R (TM-III): CAIATIGCIAC(AG)TAICG(GA)TCAIGT
P26 (TM-III): GCIT(CT)GATGTCITGGAT

Cloning and sequencing

The RACE nested PCR products were gel-purified and cloned into the pCRII vector (Invitrogen). Colonies containing OR cDNAs were selected by colony PCR using the OR degenerate primers P26 and P27, and their orientation was determined also by colony PCR using the pair of primers T7/P27 or SP6/P27. Plasmid DNA was prepared from positive colonies using Filter Plate for high-throughput separations (Multiscreen Millipore). DNA was sequenced with the ABI PRISM Big Dye Terminator V3.1 Cycle sequencing kit using T7 or SP6 primers on an ABI PRISM 3100 Genetic Analyzer (Hitachi). Four percent of the OR sequences were truncated OR RNAs, probably because the CIP reaction during RLM-RACE was not 100% efficient.

Genomic alignment of cDNA sequences

We aligned the cDNA sequences against the mouse genome using BLAST. Only the alignments with percent identities >93% were considered. The position of each alignment was calculated, and the flanking 50-kb genomic sequences were extracted from the corresponding genomic contigs. Each sequence was realigned with its corresponding extracted genomic sequence using the Sim4 program (Florea et al. 1998). Only the Sim4 alignments showing average percent identity >93%, entire sequence alignment >50%, and with the best score (based on the nucleotide identity over the entire alignment) were selected. A MySQL database was loaded with the alignment information.

Clustering of cDNA sequences

The cDNA sequences were clustered based on their genomic coordinates. Sequences that share at least one same exon/intron boundary were included in the same cluster. When no exon/intron boundaries were defined, sequences with at least 30-bp overlap in one same genomic location were included in the same cluster. The Olfactory Receptor cDNA Clusters Viewer site was generated using the Generic Genome Browser (Stein et al. 2002; http://www.gmod.org/ggb/).

Promoter sequence analysis

Promoter sequences were analyzed using the Gibbs Recursive Sampler (Thompson et al. 2003). A FASTA sequence file containing the 198 promoter sequences (600 bp upstream of the TSS) (Supplemental material 1) was analyzed using the eukaryotic default values for all parameters and motif lengths 8, 6, 8, 12, 10, 12, 12; or 14, 14, 12, 14, 14. The parameters used to identify some of the motifs are shown in the Gibbs output files (Supplemental material 2). The promoter sequences were also analyzed using Consensus (Hertz and Stormo 1999) and Weeder (Pavesi et al. 2004) (in both cases, motif widths were set to 6, 8, 10, or 12). Potential TATA-box sequences were predicted using HCTata (http://l25.itba.mi.cnr.it/~webgene/www/HC_Tata.html). The location of the motifs within the promoter regions was determined using SiteSeer (http://rocky.bms.umist.ac.uk/Seiteer). Motifs were searched in both strands of the input sequences.

Preparation of nuclear extracts

Nuclear extracts were prepared using the method described by Kudrycki et al. (1993) from olfactory epithelium dissected from 30 4–7-wk-old C57BL/6j mice. The extract was first concentrated using a Microcon centrifugal filter device (Millipore), and then the buffer was exchanged with binding buffer (10 mM Tris-HCl at pH 7.9, 1 mM EDTA, 5 mM MgCl2, 30 mM KCl, 10% glycerol, 3 mM DTT, 0.3 mM PMSF) using a Micro Bio-Spin P-6 chromatography column (Bio-Rad). Aliquots were stored at −80°C. Protein concentration was determined using the Bradford assay (BioRad).

Gel shift assay

The digoxigenin (DIG) gel shift kit (Roche Applied Science) was used for gel shift assays. Binding reactions contained 2 µg of poly[d(I-C)], 0.1 µg of poly-L-lysine, 1.2 ng of labeled oligonucleotide, and 15 µg of nuclear protein extract. After a 10-min incubation on ice and a 15-min incubation at room temperature, the mixture was added with 5 µL of loading buffer (60% 0.25× TBE buffer, 40% glycerol, 0.2% bromophenol blue) and electrophoresed in 0.5× TBE on a nondenaturing 4% polyacrylamide gel in 0.5× TBE containing a 2-cm 15% acrylamide layer at the bottom to retain the unbound probe in the gel, as described by Bell et al. (1999). The gel was pre-electrophoresed for 1 h at 80 V before the samples were applied. Competition experiments were performed by incubating the binding reaction mixtures with a 100× excess of unlabeled competitor oligonucleotide for 5 min before the addition of the labeled oligonucleotide. Blotting was performed using a Bio-Rad electroblotting system, and chemiluminescence detection of the DIG-labeled DNA–protein complexes was performed using anti-digoxigenin antibody conju-
gated to alkaline phosphatase and the CSPD substrate (Roche Applied Science).

The following pairs of complementary oligonucleotides were used as double-stranded DNA probes for the gel shift reactions (motif sequences are underlined, and the OR genes from which sequences were extracted are indicated):

- M1 motif (from olfr720):
  5′-CTAGCATGTTTCTAGCCAGTGGGAGATGCCAG-3′
  and
  5′-CCCTGAGATGTCTCCAGGAGAAAAGTCTGAG-3′
- M2 motif (olfr165):
  5′-TAAGATGCTAATTTCCTTGGAAGAGTGACTTGA-3′
  and
  5′-TTTACCAATTCCCGGAGAATTGCACTCT-3′
- M3 motif (olfr211):
  5′-CGCTGACATCCCTCAGGGCTTATTTCCG-3′
  and
  5′-AGCAATATAAGGCCACTTGGGAGATGCCAG-3′
- M4 motif (olfr1339):
  5′-CCTTAGCTTCCATCCCTGAGGAGAGAGG-3′
  and
  5′-GCTCCCTGCTTCAAGGAGATGACCTGAA-3′.

**Acknowledgments**

We thank Hiroaki Matsunami for critically reading the manuscript. We are grateful to Sandro José de Souza and Noboro Jo Sakabe for providing bioinformatics assistance. We also thank Erica Bandeira and Luci Navarro for technical assistance. This work was supported by research grants from FAPESP and CNPq (to B.M.) and fellowships from FAPESP (to J.S.M. and P.A.F.G.).

**References**

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

Bell, A., Feng, X., and Reder, A. 1999. Improved band resolution, loading reliability and reduced 22P contamination in mobility shift assays by retention of unbound probe. *Biotechniques* **27**: 1122–1126.

Borst, P. 2002. Antigenic variation and allelic exclusion. *Cell* **109**: 5–8.

Buck, L. and Axel, R. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65**: 175–187.

Bulger, M., Bender, M.A., van Doorninck, J.H., Wertman, B., Farrell, C.M., Felsenfeld, G., Grudine, M., and Hardison, R. 2000. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse 33-globin gene clusters. *Proc. Natl. Acad. Sci.* **97**: 14560–14565.

Chess, A., Simon, I., Cedar, H., and Axel, R. 1994. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**: 823–834.

Dulac, C. and Torello, A.T. 2003. Molecular detection of pheromone receptor rule in the mouse olfactory system. *Proc. Natl. Acad. Sci.* **100**: 1122–1127.

Hertz, G. and Stormo, G. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**: 563–577.

Hirota, J. and Mombaerts, P. 2004. The LIM-homeodomain protein Lhx2 is required for complete development of mouse olfactory neurons. *Proc. Natl. Acad. Sci.* **101**: 2156–2161.

Hoppe, R., Weimer, M., Beck, A., Breer, H., and Strotmann, J. 2000. Sequence analyses of the olfactory receptor gene cluster mOR37 on mouse chromosome 4. *Genomics* **66**: 284–295.

Hoppe, R., Frank, M.E., Breer, H., and Strotmann, J. 2003. The clustered olfactory receptor gene family 262: Genomic organization, promoter elements, and interacting transcription factors. *Genome Res.* **13**: 2674–2685.

Kolterud, A., Alenius, M., Carlsson, L., and Bohn, S. 2004. The Lim homeobox gene Lhx2 is required for olfactory sensory neuron identity. *Development* **131**: 5319–5326.

Kozak, M. 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**: 187–208.

Kudrycki, K., Stein-Iszak, C., Behn, C., Grillo, M., Akeson, R., and Margolis, F. 1993. Olfr-1 binding site: Characterization of an olfactory neuron-specific promoter motif. *Mol. Cell. Biol.* **13**: 3002–3014.

Lane, R., Cutforth, T., Young, J., Athanasiou, M., Friedman, C., Rowe, L., Evans, G., Axel, R., Hood, L., and Trask, B.J. 2001. Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl. Acad. Sci.* **98**: 7390–7395.

Lewcock, J.W. and Reed, R.R. 2004. A feedback mechanism regulates monoallelic odorant receptor expression. *Proc. Natl. Acad. Sci.* **101**: 1069–1074.

Lin, H. and Grosschedl, R. 1995. Failure of B cell differentiation in mice lacking the transcription factor EBF. *Nature* **376**: 263–267.

Malnic, B., Hiro, J., Sato, T., and Buck, L.B. 1999. Combinatorial receptor codes for odors. *Cell* **96**: 713–723.

Mombaerts, P. 2004. Odorant receptor gene choice in olfactory sensory neurons: The one receptor-one neuron hypothesis revisited. *Curr. Opin. Neurobiol.* **14**: 31–36.

Mombaerts, P., Wang, F., Dulac, C., Zhao, S., Nemes, A., Mendelsohn, M., Edmondson, J., and Axel, R. 1996. Visualizing an olfactory sensory map. *Cell* **87**: 675–686.

Ngai, J., Dowling, M.M., Buck, L., Axel, R., and Chess, A. 1993. The family of genes encoding odorant receptors in the channel catfish. *Cell* **72**: 657–660.

Pavesi, G., Meretheotti, P., Mauri, G., and Pessol, G. 2004. Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**: W199–W203.

Pevske, G., Mignone, F., Gissi, C., Grillo, G., Licitiulli, F., and Liuni, S. 2001. Structural and functional features of eukaryotic mRNA untranslated region. *Gene* **276**: 73–81.

Qasba, P. and Reed, R.R. 1998. Tissue and zonal-specific expression of an olfactory receptor transgene. *J. Neurosci.* **18**: 227–236.

Ressler, K.J., Sullivan, S.I., and Buck, L.B. 1993. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73**: 597–609.

Rothman, A., Feinstein, P., Hirota, J., and Mombaerts, P. 2005. The promoter of the mouse odorant receptor gene M71. *Mol. Cell. Neurosci.* **28**: 535–546.

Serizawa, S., Ishi, T., Nakatani, H., Tsuboi, A., Nagawa, F., Asano, M., Sudo, K., Sakagami, J., and Sakano, H. 2000. The clustered Lim homeobox gene Lhx2 is required for olfactory sensory neuron development. *Genome Res.* **10**; 2161–2167.

Shykind, B.M. 2005. Regulation of odorant receptors: One allele at a time. *Hum. Mol. Genet.* **14**: R33–R39.

Shykind, B.M., Rohani, S.C., O'Donnell, S., Nemes, A., Mendelsohn, M., Sudo, K., Sakagami, J., and Sakano, H. 2004. Gene switching and the one neuron hypothesis revisited. *Curr. Opin. Neurobiol.* **14**: 801–815.

Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, S., and Sakano, H. 2003. Negative feedback regulation ensures the one receptor-one olfactory neuron rule in the mouse. *Science* **302**: 2088–2094.

Serizawa, S., Miyamichi, K., and Sakano, H. 2004. One neuron–one receptor rule in the mouse olfactory system. *Trends Genet.* **20**: 648–653.

———. 2005. Negative feedback regulation ensures the one neuron–one receptor rule in the mouse olfactory system. *Chem. Senses* **30**: 1199–1208.

Shykind, B.M. 2005. Regulation of odorant receptors: One allele at a time. *Hum. Mol. Genet.* **14**: R33–R39.

Shykind, B.M., Rohani, S.C., O’Donnell, S., Nemes, A., Mendelsohn, M., Sudo, K., Sakagami, J., and Barnea, G. 2004. Gene switching and the stability of odorant receptor gene choice. *Cell* **117**: 801–815.

Sosinsky, A., Glusman, G., and Lancer, D. 2000. The genomic structure of human olfactory receptor genes. *Genomics* **70**: 49–61.

Stein, L., Mungall, C., Shu, C., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J., Harris, T., Arva, A., et al. 2002. The Generic Genome Browser: A building block for a model organism system database. *Genome Res.* **12**: 1599–1610.

Thompson, W., Rouchka, E., and Lawrence, C. 2003. Gibbs recursive sampler: Finding transcription factor binding sites. *Nucleic Acids Res.* **31**: 3580–3585.
Voss, T.S., Healer, J., Marty, A.J., Duffy, M.F., Thompson, J.K., Beeson, J.G., Reeder, J.C., Crabb, B.S., and Cowman, A.F. 2006. A var gene promoter controls allelic exclusion of virulence genes in Plasmodium falciparum malaria. Nature 439: 1004–1008.

Wang, M.M., Tsai, R.Y.L., Schrader, K.A., and Reed, R.R. 1993. Genes encoding components of the olfactory signal transduction cascade contain a DNA binding site that may direct neuronal expression. Mol. Cell. Biol. 13: 5805–5813.

Wang, S., Tsai, R., and Reed, R. 1997. The characterization of the Olf-1/EBF-like HLH transcription factor family: Implications in olfactory gene regulation and neuronal development. J. Neurosci. 17: 4149–4158.

Wang, F., Nemes, A., Mendelsohn, M., and Axel, R. 1998. Odorant receptors govern the formation of a precise topographic map. Cell 93: 47–60.

Wang, S., Betz, A., and Reed, R. 2002. Cloning of a novel Olf-1/EBF-like gene, O/E-4, by degenerate oligo-based direct selection. Mol. Cell. Neurosci. 20: 404–414.

Wang, S.S., Lewcock, J.W., Feinstein, P., Mombaerts, P., and Reed, R.R. 2003. Genetic disruptions of O/E2 and O/E3 genes reveal involvement in olfactory receptor neuron projection. Development 131: 1377–1388.

Young, J.M., Friedman, C., Williams, E.M., Ross, J.A., Tonnes-Priddy, L., and Trask, B.J. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum. Mol. Genet. 11: 535–546.

Young, J.M., Shykind, B.M., Lane, R.P., Tonnes-Priddy, L., Ross, E.M., Walker, M., Williams, E.M., and Trask, B.J. 2003. Odorant receptor expressed sequence tags demonstrate olfactory expression of over 400 genes, extensive alternate splicing and unequal expression levels. Genome Biol. 4: R71.

Zhang, X. and Firestein, S. 2002. The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. 5: 124–133.

Received February 22, 2006; accepted in revised form July 6, 2006.
Anexo 2b — Artigo: Automatic correspondence of *tags* and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data
Gene expression

Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data

Pedro A. F. Galante¹,²,*, Jeff Trimarchi³, Constance L. Cepko³,⁴, Sandro J. de Souza², Lucila Ohno-Machado⁵,⁶ and Winston P. Kuo⁵,⁷,⁸

¹Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, ²Laboratory of Computational Biology, Ludwig Institute for Cancer Research, São Paulo Branch, Brazil, ³Department of Genetics, Harvard Medical School, ⁴Howard Hughes Medical Institute, Harvard Medical School, ⁵Decision Systems Group, Brigham and Women’s Hospital, ⁶Division of Health Sciences and Technology, Harvard Medical School and MIT, Boston, ⁷Department of Organismic and Evolutionary Biology/FAS, Harvard University, and ⁸Department of Developmental Biology, Harvard School of Dental Medicine, Harvard University, Cambridge, MA, USA

Received on October 23, 2006; revised on December 22, 2006; accepted on January 19, 2007

Advance Access publication February 3, 2007

Associate Editor: John Quackenbush

ABSTRACT

Summary: A critical step in any SAGE, MPSS and SBS data analysis is tag-to-gene assignment. Current available tools are limited by a tag-by-tag annotation process and/or do not provide the dataset that is used to produce a complete tag-to-gene mapping. We developed ACTG, a web-based application that allows a large-scale tag-to-gene mapping using several reference datasets. ACTG can annotate SAGE (14 or 21 bp), MPSS (17 or 20 bp) and SBS (16 bp) data for both human and mouse organisms.

Availability: http://retina.med.harvard.edu/ACTG/

Contact: pga@ludwig.org.br

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Serial analysis of gene expression (SAGE) (Velculescu et al., 1995), Massively parallel signature sequencing (MPSS) (Brenner et al., 2000) and sequencing-by-synthesis (SBS) (www.illumina.com/pages.illum?ID¼201) determine the expression level of genes by measuring the frequency of sequence tags derived from polyadenylated transcripts. SAGE allows the construction of a comprehensive expression profile in which each mRNA is defined by specific 14 or 21 bp sequences adjacent to the 3’ most NlaIII site. MPSS generates 17–20 bp sequences adjacent to the 3’ most DpnII site from millions of mRNA molecules in a sample, providing a quantitative assessment of the transcript abundance. SBS is a new technology that is similar to MPSS (it also generates millions of 20-bp-long tags corresponding to sequences adjacent to the 3’ most DpnII site). SAGE and MPSS have been widely used for the study of gene expression in pathological tissues (Hermeking, 2003) as well as in the study of general aspects of gene expression in different organisms (Hermeking, 2003; Jongeneel et al., 2005; Meyers et al., 2004). An essential step in the analyses of SAGE, MPSS and SBS data is the correct assignment/mapping of a tag to a gene. Basically, there are two strategies to perform this process: an annotation based on data from websites, such as SAGE Genie (Boon et al., 2002) and SAGEmap (Lash et al., 2000) or an annotation based on databases constructed using in-house computer approaches (Blackshaw et al., 2004; Silva et al., 2004). SAGEmap and SAGE Genie are web-based tools that map ‘tags to cDNAs’ and ‘tags to gene/cDNA references’, respectively (these web tools are exclusive to SAGE data). These strategies for tag-to-gene assignments either do not produce a complete annotation or cannot batch process a large number of tags. To address these limitations, we have created a web-based application, called automatic correspondence of tags and genes (ACTG). ACTG is designed to map a large number of tags into several reference datasets. ACTG is user-friendly and generates an output file that is simple to analyze and can be integrated with other applications.

2 OVERVIEW OF ACTG

ACTG is a collection of several Perl scripts, Perl + CGI scripts, shell scripts and HTML codes that performs three main tasks: (i) The assembling of virtual tag databases used in the ACTG, (ii) The uploading and processing of information submitted by the user and (iii) The mapping of the submitted tag list and generation of the output files.

ACTG datasets are composed of virtual tags (a computer prediction of tags produced by a SAGE, a MPSS or a SBS experiment) extracted from three major databases, SAGEGenie, SAGEmap and almost all public cDNAs from GenBank. The following are the main steps for the assembly of each dataset. For SAGEGenie: (i) Download of the files containing the best virtual tags matches to UniGene clusters, (ii) Parsing of the raw data, (iii) Assembly of the data and generation of the final dataset. For SAGEmap, the process is similar: (i) Download of the files containing the reliable virtual tags matches to UniGene clusters, (ii) Parsing of the raw data,
(iii) Selection of the best tag to cDNA assignment (based on the SAGE map tag-to-gene score) and (iv) Assembly of the final dataset. For all public cDNAs sequences the process is more complex: (i) Download of all cDNAs from UniGene (Boguski and Schuler, 1995), RefSeq (Pruitt et al., 2005), MGC (Strausberg et al., 1999) and dbEST (Boguski et al., 1993), (ii) Split of dbEST ESTs, RefSeq, MGC and UniGene mRNAs in subsets presenting either a poly(A) tail (at least five adenosines at the cDNA 3’ end), or a canonical poly(A) signal (AAUAAA or AUUAAA at the 3’ most 50 bp segment), or both, poly(A) tail and signal and (iii) Extraction of the virtual tags and assembly of the final datasets in the ACTG format (for dbEST data, only ESTs containing poly(A) tail and poly(A) tail and signal have been included). These subdivisions of the cDNA datasets, based on the 3’ end information, are important to evaluate the tag to cDNA mapping, from which a well-defined 3’ end produces a more reliable tag-to-gene match (Boon et al., 2002).

Since ACTG presents virtual tags from commonly obtained sources of cDNA data, some users may have difficulty to fully utilize the datasets and interpret the redundancy between each. We therefore created an additional virtual tag dataset, which is a non-redundant tag list that merges and removes the redundancy of every ACTG datasets. In addition, we classified the virtual tags to three categories of tag-to-gene assignment in terms of their reliability (high, medium and low), thus producing a non-redundant and ranked virtual tag dataset (details of the non redundant tag list are described on the ACTG website). These new datasets would help simplify the tag mapping process and the user’s interpretation of results.

We have also created another additional data, a set of putative artifactual tags. This dataset contains tags similar (allowing 1 mismatch) to the sequence of the linker (utilized in the construction of SAGE library) and ambiguous tags, tags mapped to two or more genes. Putative tags are identified by special characters in the output files and should be considered as non-reliable in the tag-to-gene assignments. These specific features are also available in SAGE Genie (Boon et al., 2002).

Figure 1 illustrates the main steps necessary to map a list of tags using ACTG. The first step is the submission of a file with
a list of tags (a plain text file containing the tag sequences and optional columns, for example, the tag frequencies). Second, selecting an organism (human or mouse), tag type (SAGE, MPSS or SBS), and at least one database to map the tags. The final steps include the execution of the program (‘run ACTG’) and downloading of the output files. ACTG produces three output files: (1) A file(s) containing the tag mapping for each selected database, (2) A file that merges all mapping results, and (3) A file containing statistics of the mapping process.

Below is a simplified guide to help best utilize the functions of ACTG and the interpretation of its mapping results. If a user is interested in: (1) Identifying known genes and produce the most ‘reliable’ set of tag-to-gene mapping, use databases containing virtual tags from mRNAs sequences with poly(A) tails and poly(A) tails and poly(A) signals, (2) Identifying new genes and/or new transcripts, use databases containing virtual tags from ESTs with poly(A) tails and poly(A) tails and poly(A) signals, and (3) Performing the most complete mapping, use databases containing virtual tags from SAGEGenie, SAGEmap or virtual tags from the non-redundant tag lists. In reference to the interpretation of the results, a critical aspect to be carefully inspected is the ambiguity of ‘tag-to-gene’ assignments. If a tag is mapped to virtual tags from two or more different cDNAs, but all sequences are in the same UniGene cluster, this ‘redundancy’ is acceptable and will not influence the tag-to-gene match. However, the assignment is ambiguous if a tag is mapped on distinct UniGene clusters.

For further details on the functions of ACTG, the construction of the ACTG datasets, the mapping of a submitted tag list, and the interpretation of results; please visit the ‘HELP & FAQ’ section of the ACTG website.

3 APPLICATION

In order to evaluate our tool, we have selected and submitted to ACTG sets of tags from two recent publications (Blackshaw et al., 2004; Jongeneel et al., 2005). In Blackshaw et al. (2004), they were not able to map 37 813 distinct SAGE tags from 12 retina libraries. By, using ACTG, we were able to map 16 927 of these tags (databases: SAGEGenie, UniGene mRNAs and ESTs with poly(A) tail). In Jongeneel et al. (2005), they constructed an atlas of human gene expression based on tags from 32 MPSS library, where, for each library, the tags were assigned to transcribed regions (genes). For example, tags from cerebellum were mapped to 8183 genes and tags from bone marrow were mapped to 7182 genes. By using ACTG, tags from cerebellum and bone marrow were mapped to 13 773 and 12 689 genes, respectively (database: mRNAs from UniGene, RefSeq, MGC and ESTs with poly(A) tail and poly(A) signal and EST with poly(A) tail). The annotations for these datasets are available in the ‘Publication’ section of ACTG website as supplementary data.

4 SUMMARY

ACTG, a web-based tool, allows the user to quickly annotate a complete SAGE (14 or 21 bp long), MPSS (tags 17 or 20 bp long) or SBS (20 bp long) library using complete datasets for both human and mouse organisms. In addition, ACTG can filter redundant and artifact tags and generates a report of the mapping process. ACTG is a simple publicly available tag-to-gene mapping tool packaged in a user-friendly environment that addresses an essential step in SAGE, MPSS, and SBS data analyses.

ACKNOWLEDGEMENTS

We thank Arthur Ramos for bioinformatics support and Noboru Jo Sakabe for discussions. PA FG was supported by a FAPESp fellowship. This study was supported in part by grant 5D43TW007015-02 from the Fogarty International center, NIH. WPK was supported by HSDM Dean’s Scholar Award. Funding to pay the Open Access publication charges was provided by grant 5D43TW007015-02 from the Fogarty International center, NIH.

Conflict of Interest: none declared.

REFERENCES

Blackshaw,S. et al. (2004) Genomic analysis of mouse retinal development. PLoS Biol., 2, E247.
Boguski,M.S. and Schuler,G.D. (1995) Establishing a human transcript map. Nat. Genet., 10, 369–371.
Boguski,M.S. et al. (1993) dbEST – database for “expressed sequence tags”. Nat. Genet., 4, 332–333.
Boon,K. et al. (2002) An anatomy of normal and malignant gene expression. Proc. Natl Acad. Sci. USA, 99, 11287–11292.
Brenner,S. et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol., 18, 630–634.
Hermeking,H. (2003) Serial analysis of gene expression and cancer. Curr. Opin. Oncol., 15, 44–49.
Jongeneel,C.V. et al. (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc. Natl Acad. Sci. USA, 100, 4702-4705.
Jongeneel,C.V. et al. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res., 15, 1007–1014.
Lash,A.E. et al. (2000) SAGEmap: a public gene expression resource. Genome Res., 10, 1051–1060.
Meyers,B.C. et al. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res., 14, 1641–1653.
Pruitt,K.D. et al. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res., 33, D501–D504.
Silva,A.P. et al. (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. Nucleic Acids Res., 32, 6104–6110.
Strausberg,R.L. et al. (1999) The mammalian gene collection. Science, 286, 455–457.
Velchevski,V.E. et al. (1995) Serial analysis of gene expression. Science, 270, 484–487.

Automatic correspondence of tags and genes

905
Anexo 2c – Artigo: Detection and evaluation of intron retention events in the human transcriptome
Detection and evaluation of intron retention events in the human transcriptome

PEDRO ALEXANDRE FAVORETTO GALANTE, NOBORU JO SAKABE, NATANJA KIRSCHBAUM-SLAGER and SANDRO JOSÉ DE SOUZA

RNA 2004 10: 757-765
Access the most recent version at doi:10.1261/rna.5123504

References
This article cites 28 articles, 10 of which can be accessed free at:
http://www.rnajournal.org/cgi/content/full/10/5/757#References

Article cited in:
http://www.rnajournal.org/cgi/content/full/10/5/757#otherarticles

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

Notes

To subscribe to RNA go to:
http://www.rnajournal.org/subscriptions/

© 2004 RNA Society
Detection and evaluation of intron retention events in the human transcriptome

PEDRO ALEXANDRE FAVORETTO GALANTE,1,2 NOBORU JO SAKABE,1,2 NATANJA KIRSCHBAUM-SLAGER,1 and SANDRO JOSÉ DE SOUZA1

1Ludwig Institute for Cancer Research, Sao Paulo Branch, São Paulo, Brazil
2Ph.D. Program, Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Sao Paulo, Brazil

ABSTRACT
Alternative splicing is a very frequent phenomenon in the human transcriptome. There are four major types of alternative splicing: exon skipping, alternative 3' splice site, alternative 5' splice site, and intron retention. Here we present a large-scale analysis of intron retention in a set of 21,106 known human genes. We observed that 14.8% of these genes showed evidence of at least one intron retention event. Most of the events are located within the untranslated regions (UTRs) of human transcripts. For those retained introns interrupting the coding region, the GC content, codon usage, and the frequency of stop codons suggest that these sequences are under selection for coding potential. Furthermore, 26% of the introns within the coding region participate in the coding of a protein domain. A comparison with mouse shows that at least 22% of all informative examples of retained introns in human are also present in the mouse transcriptome. We discuss that the data we present suggest that a significant fraction of the observed events is not spurious and might reflect biological significance. The analyses also allowed us to generate a reliable set of intron retention events that can be used for the identification of splicing regulatory elements.

Keywords: Alternative splicing; intron retention; transcriptome; EST

INTRODUCTION
The past few years have witnessed the emergence of a picture showing a high frequency of alternative splicing in the human transcriptome (Hanke et al. 1999; Mironov et al. 1999; Croft et al. 2000; Modrek et al. 2001; Modrek and Lee 2002; Sakabe et al. 2003). It seems that at least half of all human genes undergo alternative splicing, but the right number probably will exceed that because the detection of splicing variants of lowly expressed genes is difficult (Kan et al. 2002; Zavolan et al. 2002). The biological significance of alternative splicing is well documented for several models. A classical example is sex determination in Drosophila, in which alternative splicing acts like a genetic trigger (Baker 1989). Another classical example is the gene slo that corresponds to a calcium-activated potassium channel (Xie and Black 2001). Splice variants of this protein form a gradient in the hair cells of the inner ear of the chicken that is involved in sound detection.

Although extremely important due to their biological significance, these examples are isolated cases reflecting particular interests of individual laboratories. Due to the genomics and bioinformatics revolution, the cDNA databases remain growing at an exponential rate. This, coupled to the availability of the human genome sequence, has allowed the use of an integrated approach for the detection of splicing variants (Modrek et al. 2001; Sakabe et al. 2003). It remains doubtful however, whether a significant fraction of the events have biological significance or whether they are spurious products from the splicing machinery (Kan et al. 2002).

There are four major types of alternative splicing: exon skipping, alternative 3' splice site, alternative 5' splice site, and intron retention. Exon skipping seems to be the most frequent type as shown by Modrek et al. (2001). Intron retention is certainly the least studied of all types of alternative splicing, especially because these variants are believed to be largely derived from unspliced or partially spliced pre-mRNAs. There are at least a couple of cases of intron retention events with known biological consequences. The
P-element transcript in *Drosophila* is a transposon active in the *Drosophila* germ line only (Rio 1991). The blocking of the transposition in somatic cells is due to a truncation of the encoded transposase caused by the retention of the third intron in the P-element transcript. The Ret tyrosine kinase also presents a couple of intron retention events, one of them generates a truncated protein that is enriched in familial and sporadic pheochromocytomas (Le Hir et al. 2002).

Here we present a large-scale analysis of intron retention for a set of 21,106 known human genes. We evaluated the distribution of intron retention events in relation to several features such as GC content, codon usage, coding potential, and conservation in mouse among others. The observed data suggest that a significant fraction of the intron retention events is not spurious and probably reflects biological significance.

**RESULTS**

**Identification of intron retention events**

Because of the large-scale nature of their generation, EST data are enriched with artifactual sequences like genomic contamination and unprocessed or partially processed mRNAs. These issues are critical for the detection of true splicing variants. The availability of an almost complete draft of the genome sequence has allowed the integration of both genomic and cDNA data to identify splicing variants (Modrek et al. 2001). We have used a similar strategy by mapping all human cDNAs onto the draft sequence of the human genome (Sakabe et al. 2003). We have also developed a clustering strategy that groups all cDNAs mapping to the same genomic region into a single cluster (see Materials and Methods for more details). A nonredundant set of 21,106 different cDNA clusters containing at least one known human full-insert cDNA (by that we mean a fully sequenced cDNA clone) was used in the analyses described below. To validate our database, we performed an analysis of exon skipping and compared our results to data previously published by different groups. We found, for instance, that 52% of all known human genes undergo exon skipping and that 77% of the events are located within the coding region. Similar numbers were found by different groups (Hanke et al. 1999; Croft et al. 2000; Modrek et al. 2001; Kan et al. 2002).

To be identified as a cDNA containing an intron retention event, a sequence was required to span the entire length of the respective retained intron. Furthermore, the same sequence had to contain at least one exon/exon boundary in order to minimize the chance of the retention being an artifact. This critical step eliminates unprocessed messages (although there is still the possibility of finding partially processed messages). The identification of an intron retention event was based on a pairwise comparison of a transcript to other sequences belonging to the same cDNA cluster. A retention event could be identified in different possible situations (Fig. 1). The prototype used to characterize the retention, spanning the respective exon/exon boundary, could be either an EST or a full-insert cDNA. The sequence containing the retained intron could also be either an EST or a full-insert cDNA. We found 3127 known genes with at least one intron retention event, which corresponds to 14.8% of the total number of known genes analyzed here. Table 1 shows the numbers of genes found for every possible case except the cases where both the prototype and the sequence containing the retention are ESTs. We decided to exclude those cases from our analysis because they involve only ESTs, which are known to be low-quality sequences. The majority of cases correspond to those pairs where the prototype is an EST and the sequence showing the retention is a full-insert cDNA. This is probably due to at least two factors: the average length of EST sequences and a bias for longer clones in the generation of full-insert cDNA sequences. A complete list of all GenBank accessions reporting an event of intron retention is available at http://www.compbio.ludwig.org.br/~pgalante/IR.

Four genes reporting an intron retention event were selected for experimental validation. RT-PCR reactions were performed using primers flanking the intron being retained and primers specific to the retained sequence. Figure 2 shows the products of amplification for one of the four genes in all tissues sampled. We observed the variant containing the retained intron for all genes as well as the corresponding prototypes.

![Figure 1](http://www.compbio.ludwig.org.br/~pgalante/IR)

**FIGURE 1.** Possible cases of intron retention. The prototype sequence and the sequence containing the retained intron can be either a full-insert cDNA or an EST. Cases in which both sequences correspond to ESTs were excluded from our data set. Exons are represented as open bars, introns as lines, and retained introns are shown as black bars (not in scale).
the distribution of events is primarily dependent on the frequency of introns in each part of the sequence, we first obtained the number of introns in the 5’ UTR, CDS, and 3’ UTR. This analysis was done using a set of 16,951 full-insert cDNA sequences that contained annotated 5’ and 3’ UTRs. Based on that, we calculated the number of expected intron retention events in each of the three sections of the transcript

\[ IR \times \frac{li}{I}, \]

where \( IR \) is the total number of retention events, \( li \) is the number of introns in the respective transcript section, and \( I \) is the total number of introns in the data set, and compared that with the observed numbers of events (Table 2). We noticed a significant excess of intron retention events within both UTRs, with a corresponding decrease of observed events within the CDS (\( \chi^2 = 229 \), 1 degree of freedom, \( p < 10^{-49} \)).

It is well known that the presence of premature stop codons in mRNA triggers its degradation through a process of nonsense-mediated decay (NMD; Gonzalez et al. 2000). It is, therefore, possible that the observed biased distribution for retention events in the UTRs is simply a product of NMD. This would actually suggest that the rate of intron retention is higher than observed and a fraction of the events is filtered off by NMD.

Another possibility is underreporting of sequences containing a premature stop codon. To test whether this was a factor affecting the above described distribution, we selected all cases of intron retention reported by the Mammalian Gene Collection (MGC) initiative, which reports sequences to GenBank even when they contain premature stop codons (Strausberg et al. 2002). Intron retention events reported in MGC sequences are still biased for UTRs, showing that underreporting does not seem to affect the above distribution (Table 2).

### TABLE 1. Number of genes containing at least one intron retention event

| Prototype          | Variant       | Full-insert cDNA | EST | Total |
|--------------------|---------------|------------------|-----|-------|
| EST                |               |                  |     |       |
| Total              |               |                  |     |       |

Those cases confirmed by at least two cDNA sequences are shown in parentheses. “Total” corresponds to the nonredundant sum of the values in the corresponding row or column. The reduction is due to the fact that the same gene may have intron retention events reported by both ESTs and full-insert cDNAs (in those cases, the gene was counted only once).

### Distribution of events along transcripts

Because the presence of an intronic sequence is expected to cause a dramatic effect on the corresponding protein when it occurs within the coding sequence (CDS), we evaluated the CDS/untranslated region (UTR) distribution of all events within the data set of 640 genes where both the prototype and the sequence defining the retention event were full-insert cDNAs. For the sake of simplicity, from now on we will call these pairs of sequences the “elite group.” Because

\[ \text{IR} \times \frac{li}{I}, \]

We considered only those cases in which both sequences correspond to full-insert cDNAs. For the statistical analysis we grouped the 5’ and 3’ UTRs. The excess of retention in both UTRs is statistically significant: elite group \( p = 10^{-49} \), MGC (Mammalian Gene Collection) \( p = 10^{-17} \). Expectation was calculated based on the density of introns in the respective transcript section (\( IR \times \frac{li}{I} \), where \( IR \) is the total number of retention events, \( li \) is the number of introns in the respective transcript section and \( I \) is the total number of introns in the dataset).

### TABLE 2. Number of intron retention events in the CDS and 5’ and 3’ UTRs

| Events in | Elite group | MGC |
|-----------|-------------|-----|
| CDS       | 287 (53%)   | 87 (52%) |
| 5’ UTR    | 84 (15%)    | 15 (9%)  |
| 3’ UTR    | 170 (32%)   | 65 (39%) |

We considered only those cases in which both sequences correspond to full-insert cDNAs. For the statistical analysis we grouped the 5’ and 3’ UTRs. The excess of retention in both UTRs is statistically significant: elite group \( p = 10^{-49} \), MGC (Mammalian Gene Collection) \( p = 10^{-17} \). Expectation was calculated based on the density of introns in the respective transcript section (\( IR \times \frac{li}{I} \), where \( IR \) is the total number of retention events, \( li \) is the number of introns in the respective transcript section and \( I \) is the total number of introns in the dataset).
Retained introns code for protein domains

One way to assess the biological significance of an intron retention event is to evaluate its contribution to the protein domains encoded by the respective full-insert cDNA. One hundred forty-seven full-insert cDNAs from the elite group containing at least one retention event totally located in the coding region were searched against Pfam (Bateman et al. 2002). See Materials and Methods for further details. Figure 3 illustrates the different types of contribution by a given retained intron to the coding of a protein domain. Only two domains were encoded entirely by two different retained introns, out of 147. Retained introns are less efficient in coding for protein domains than exons (30 exons, out of 830, encode an entire domain in the same set of full-insert sequences). There is no statistical difference between the numbers obtained for retained introns and exons (Fisher’s exact test, p = 0.21). However, retained introns are more efficient than nonretained introns present in the same set of full-insert cDNAs (no non-retained intron, out of 785, encodes an entire domain). This difference is statistically significant (Fisher’s exact test, p < 0.026).

Furthermore, if the retained intron was joined to the two flanking exons (exon/retained intron/exon), the number of domains entirely encoded by such segments increased to 15 (a hit was considered significant when at least 30% of the domain was encoded by the retained intron). For the 13 cases in which the intron partially encodes a domain, we tested if the exclusion of the sequence corresponding to the retained intron affected the ability to detect the correct domain. We performed a manual inspection of the protein domains encoded by a modified version of the intron retention variant (without the amino acids encoded by the retained intron). In six of these sequences, the domain encoded with the contribution of the retained intron was not found and in six others the domain was found with scores considerably lower and with e-values increased by orders of magnitude. In only one case could we not detect a substantial change in the scores after deleting the intronic sequence. This observation suggests that the amino acids encoded by the retained introns are a relevant part of a structural unit.

In addition, we found 24 other cases where the domain was longer than the “exon/retained intron/exon” unit (i.e., the domain was only partially coded by this unit) with a significant contribution of the intronic sequence (at least 30% of the domain). Thus, 39 out of 147 retained introns (26%) participated in the coding of a protein domain.

As an additional control, we tested if the retained introns located in the UTRs would participate in the coding of protein domains. None of the 762 amino acid sequences (254 UTR-retained introns translated in three phases) showed a complete or partial match against a domain in Pfam (Fisher’s exact test, p < 10^{-17} when compared to 39 cases in 147 retained introns in the CDS).

Selection on intron sequences involved in retention events

To evaluate whether selection for coding potential was acting on intron sequences involved in retention events, we analyzed those introns from our elite group located within the CDS (287 introns as shown in Table 2). We first compared the frequencies of all four bases in the set of introns involved in retention events to the set containing all non-retained introns. Because it has been shown that shorter introns have a higher GC content in humans (Lander et al. 2001), we divided both data sets into length categories. A significant difference ($\chi^2 = 2276$, 6 degrees of freedom, $p < 10^{-30}$) is observed between retained (higher GC content) and nonretained introns (Fig. 4). Figure 4 also shows that the GC distribution for retained introns is similar to the GC distribution for their flanking exons ($p = 0.23$, $p < 10^{-17}$ when compared to 39 cases in 147 retained introns in the CDS).
The distribution observed for both exons and nonretained introns is in accordance with Lander et al. (2001). The frequency of stop codons in retained introns is lower than expected. Based on the nucleotide frequencies observed for retained introns located within the CDS, the expected number of stop codons is 1064 in 26,941 codons (see Materials and Methods). The observed number of stop codons (651) was significantly smaller than expected ($\chi^2 = 103$, 1 degree of freedom, $p < 10^{-25}$). We found 88 full-insert cDNAs, out of 287, in which the retention event generated a putative truncated protein due to the presence of a premature stop codon (available at http://www.compbio.ludwig.org.br/~pgalante/IR). The identification of 88 cases in which the retention generates a putative truncated protein allowed us to examine the GC content within the retained intron for sequences upstream of and downstream from the premature stop codon. It is expected that selection would be relaxed on intronic sequences downstream from the premature stop codon. The GC content for sequences upstream of and downstream from the premature stop codon is 58% and 49%, respectively.

**Codon usage in retained introns, exons, and nonretained introns**

Codon usage in retained introns, exons, and nonretained introns was evaluated as described in Materials and Methods. Two strategies were used to compare the different data sets. The first strategy compares the whole codon usage table (61 codons) of the three data sets (results are labeled A in Fig. 5). The second strategy is based on the calculation of the average number of amino acids using the same codons in the three different data sets (results are labeled B in Fig. 5). Both approaches show that retained introns are more similar to exons than to nonretained introns.

**Conservation of intron retention in mouse cDNA sequences**

To assess if events of intron retention found in the human transcriptome are conserved in the mouse transcriptome we aligned human full-insert cDNAs presenting intron retention (elite group with at least one mouse cDNA hit) to mouse full-insert cDNAs and ESTs using BLAST 2.0 (Altschul et al. 1997) with an e-value threshold of $e^{-10}$. We found that 38%–57% of all retained introns (depending on the UTR/CDS localization) present a mouse hit (Table 3, first line). The sequence identity of orthologous retained introns is 84%, significantly higher than the average identity found for orthologous nonretained introns (60%) and similar to the average degree of identity found for orthologous exons (87%). These numbers are similar to those observed by Waterston et al. (2002). Interestingly, we found only three cases of retained intron in mouse with a premature stop codon, suggesting that the putative retention in mouse preserved the open reading frame.

To verify if the mouse cDNA also corresponds to an intron retention variant, we searched for at least one other cDNA that could define the corresponding intron at the genome level. For many examples, the region containing the retained intron was not covered by any other cDNA sequence, thus not allowing the identification of an intron in that region (second line of Table 3). An intron retention in human is deemed “conserved” in mouse if a mouse hit covers the retained intron in human and if there is also evidence that the retained sequence is also an intron in the mouse genome. We found 10 cases, out of 46, of conserved intron retention. This number is certainly an underestimate.
Galante et al.

| TABLE 3. Intron retention events in human (detected in the elite group with annotated CDS) that also occur in mouse |
|---------------------------------------------------------------|
| Intron retention in human with mouse cDNA hits            | 5′ UTR | CDS | 3′ UTR |
| Mouse clusters with more than one cDNA in the region      | 9/21 (43%) | 74/130 (57%) | 15/39 (38%) |
| Conserved intron retention event                          | 1/9    | 40/74 | 5/15    |
| We performed the analysis separately for CDS and UTR. The first row corresponds to the results of searching the retained introns in human against mouse cDNAs (ESTs and full-insert cDNA). The second row defines the cases that have more than one mouse cDNA in the region of the retained intron (these cases are informative for seeking an intron at the genome level). The third row corresponds to those cases defined in the second row where an intron is observed at the genome level. See text for more details. |

because the prototype may not have been sequenced and all cDNA sequences covering that region may correspond to the variant showing the retention, as also pointed out by Thanaraj et al. (2003).

DISCUSSION

One could expect that the splicing process has an intrinsic error rate that is difficult to estimate. It is reasonable to suppose that at least a fraction of all splicing variants found in the cDNA databases correspond to artifacts caused by such errors. This is even more pronounced nowadays as databases are reaching a critical in-depth sampling. Thus, an important question refers to the amount of artifactual variants identified by the use of cDNA databases. Although valid for all types of alternative splicing, this seems to be more critical for intron retention, because unprocessed or partially processed messages would be easily detected as variants.

Kan et al. (2002) found an intron retention rate of around 35%. They then used some statistics to infer the reliability of a given splicing variant. They found, for instance, that less than 5% of all genes exhibited intron retention at a 95% confidence interval (p < 0.05). We have found that 14.8% of all genes sampled here exhibited putative events of intron retention. This discrepancy may be due to the fact that we have not applied any statistics in our data set and/or to differences in the data sets used. On the other hand, we considered only sequences showing evidence of splicing to exclude totally unprocessed transcripts, and most of our analyses were performed on an elite group. Interestingly, if we consider the elite group only, we find evidence of intron retention in 4.6% of the whole data set used (640 genes with intron retention divided by 13,841 genes with at least two full-insert cDNAs).

An important aspect of our analysis is the apparent nonrandom distribution of intron retention events regarding some specific features. The rate of intron retention is more pronounced in the UTRs as clearly shown in Table 2. There is a threefold and 14-fold excess (above the expectation) of events in the 5′ and 3′ UTRs, respectively. Much of the bias is probably due to NMD that would trigger the degradation of transcripts bearing premature stop codons. Although this would explain the lower number of events within the CDS, it does not explain the different rates of intron retention between the 5′ and 3′ UTRs. This bias toward the 3′ UTR could be explained by the direction of RNA synthesis, because splicing seems to be coupled to transcription (for review, see Black 2003) and therefore the last intron would have less time available to be spliced out. However, for 90% of the retained introns, we were able to find at least one exon/exon boundary downstream from the retained intron, suggesting that the last intron was spliced as efficiently as the first introns in the transcript. One possible explanation implying biological significance is an effect of the retained intron on mRNA stability. Cis-acting elements present in the 3′ UTR are known to alter the stability of mRNAs (Bashirulla et al. 2001), making them either more or less stable. The presence of such elements in the retained introns would affect the stability of that variant transcript.

Stamm et al. (2000) observed that retained introns are significantly shorter than nonretained introns. We observed the same trend in our data set (see Supplemental Fig. 1 at http://www.compbio.ludwig.org.br/~pgalante/IR/). However, the interpretation that this feature is related to biological significance is not straightforward because the retention of very long introns is unlikely to occur. Furthermore, other factors like cloning efficiency could contribute to the biased length distribution observed for retained introns.

Some analyzed features show that the retained introns are considerably more similar to exons than to other introns. The GC content of such sequences is higher than that of nonretained introns in the same size range and is similar to exons. Introns with a higher GC content might have a lower excision rate as demonstrated by Goodall and Filipowics (1991) in dicots. This phenomenon, however, does not seem to exist for other organisms, including vertebrates (Goodall and Filipowics 1989). The codon usage follows the same trend and is overall much more similar to that of exons than to introns. Likewise, retained introns encode more domains than introns but less than exons, showing that the intronic sequence is a relevant part of a larger functional unit, even when it is not coding the entire domain.

These aspects show that the pattern emerging from the data set of variants is not random and therefore can be
Intron retention in the human transcriptome

The definition of introns and exons may become fuzzy when we take into account intron retention. One could establish two requirements for a genomic sequence to be considered an exon: (1) It should be present in the mature mRNA, and (2) it should be flanked by sequences excised at the RNA level. Because the retained introns identified in this study (which seem to be under selection) violate the second requirement, they should be considered introns subjected to regulated alternative splicing. This raises the question of whether retained introns are portions of exons becoming introns or introns being incorporated into exons. The analysis performed here cannot answer this question. Expression data coupled to more complex interorganism comparisons could help to elucidate this issue.

Finally, regarding the mechanistic nature of the observed retention events, it is possible that the flanking exons or the retained introns themselves bear weak splice sites and/or do not contain proper splicing regulatory elements. An analysis of splice sites and regulatory elements currently underway will hopefully shed some light on this subject.

MATERIALS AND METHODS

cDNA mapping and clustering

All human cDNAs available in dbEST (July 2002, Boguski et al. 1993) and mRNA sequences from known human genes from UniGene release 153 (Schuler et al. 1996) were aligned to the masked human genome sequence (build 29, obtained from NCBI) by using pp-Blast (Osorio et al. 2003), an implementation of MEGABLAST (Zhang et al. 2000) for a parallel cluster. The parameters used in MEGABLAST were: -f T -J F -F F -W 24. The MEGABLAST output was parsed and a MySQL database was loaded with the mapping information. Spurious hits were excluded from the mapping database by using an additional set of alignment criteria. These include a minimum degree of identity for a cDNA-genome alignment set to 93% over at least 45% of the total EST length or 55% of the total length of the full-insert sequence. Furthermore, sequences mapping to more than one location on the genome were given a score for alignment quality. Higher score was associated with a higher identity over a longer alignment. Only the sequences with the highest scores were used in the analyses reported here. Clustering of cDNA sequences was based on their genomic coordinates as described by Sakabe et al. (2003). Briefly, if two sequences shared at least partially the same gene structure they were joined into the same cluster. If no exon/intron boundary was defined, a sequence had to have at least a 100-bp overlap with another sequence at the genome level to be added to the respective cluster.
Identification of intron retention events

A set of 21,106 distinct cDNA clusters containing at least one known full-insert cDNA sequence was used in the present analysis. Introns were characterized by the presence of at least one cDNA sequence in which the corresponding intron was not spliced out. We considered valid only those cases in which the sequence characterizing the retention also defined at least one more intron that was not retained. This excludes from our analysis artifacts derived from unprocessed messages.

Distribution of events along transcripts

We selected from the 21,106 full-insert cDNAs those sequences (16,951) containing an annotated CDS. We then mapped all retention events to these sequences. For this analysis, we have only used those cases in which both the prototype and the sequence containing the retained intron were full-insert cDNAs to avoid any positional bias that could be introduced by using 3′ and 5′ ESTs. A retention was considered to be located within a given transcript section if the 5′ end of the intron was mapped to that section. In a few cases in which the presence of the intron generated a stop codon, the 3′ end of the intron would be located in the 3′ UTR whereas the 5′ end would be located in the CDS.

Frequency of stop codons

The number of observed stop codons in the set of retained introns (287 sequences with 26,941 codons from the elite group) was counted considering codons in frame with the upstream exon. The number of expected stop codons is \((f(T) \times f(G) \times f(A) + f(T) \times f(A) \times f(T) \times f(A) \times f(G))\times\text{number of codons in the data set}\) where \(f(T)\) is the frequency of T in the data set and so on.

Codon usage in nonretained introns, exons, and retained introns

Codon usage was determined for those cases belonging to the elite group. The exons and nonretained introns used as a control came also from the elite group to avoid any possible bias in the selection of genes. In the case of both retained and nonretained introns, the frame was defined by the frame of the upstream exon. To evaluate the similarity in codon usage in the three sets we made two distinct measurements. The first was a comparison of the entire codon usage table (61 codons, 60 degrees of freedom) among the three sets. We performed 1000 pairwise comparisons of the random distributions of 8216 codons (the number of codons in the set of retained introns) along the 61 possible codons to calculate the average \(\chi^2\) values (labeled A in Fig. 5) and standard deviation. The second measurement refers to the number of amino acids that presented similar codon usage \(\chi^2\) tests at \(p \leq 0.05\); labeled B in Fig. 5. As methionine and tryptophan have only one codon, they were excluded from this last analysis. Again 1000 random sets were built from each set and pairwise compared, grouped by amino acid. The numbers of amino acids with \(p\) values <0.05 in \(\chi^2\) tests were counted and presented in Figure 5.

Analysis of protein domains encoded by retained introns

Amino acid sequences of full-insert cDNAs, belonging to the elite group, with a retained intron located entirely in the CDS were submitted to domain search in Pfam 9.0 (Bateman et al. 2002) using the program hhm in a GeneMatcher hardware (Paracel Inc). Retained introns smaller than 99 bp and longer than 990 bp were excluded from this analysis. The final number of sequences submitted to Pfam was 147. To be considered a significant hit, the retained intron had to encode at least 30% of the protein domain and have an \(e\)-value <10\(^{-4}\). The coordinates of the domains were then converted to their cDNA positions. The cDNA coordinates of retained introns or exons (control set) entirely in the CDS were then compared to domain coordinates and grouped in those matching a whole domain or part of it. Domains were considered different in a given protein sequence when they presented distinct annotation and coordinates. For the comparison of exons, introns, and retained introns, domain search was performed with each exon or intron (of the full-insert cDNAs with intron retention) being a separate query. Only sequences with similar lengths were analyzed (<330 and >33 amino acids).

Experimental validation

The four genes tested by RT-PCR are: (1) BC015569—similar to SRp25 nuclear protein, (2) BC004239—jumping translocation breakpoint, (3) BC025673—similar to transcription elongation factor B (SII), and (4) BC011903—similar to the small subunit of calpain 4. Reverse transcription was carried out using the SuperScript First Strand Synthesis Kit, according to the manufacturer’s instructions (Life Technologies). The cDNA was synthesized by incubating 2 µg of either brain, lung, prostate, or breast total RNA in 20 µL of 1× reverse transcriptase buffer containing 0.5 µg oligo dT primer, 0.5 mM dNTP, 5 mM MgCl₂, 10 µM DTT, 80 U ribonuclease inhibitor, and 200 U SuperScript II reverse transcriptase at 42°C for 1 h. The quality of total RNA was tested by PCR using MLH1 primers located at intronic sequences flanking exon 12 (forward: 5′-TGGTGTCCTAGTTCTGG-3′, and reverse: 5′-CATTGGTAGCTTGCTGC-3′), as an indicator of possible genomic DNA contamination. Special primers were designed for each intron retention event: two primers annealing to the flanking exons of the retained intron and one additional primer annealing to the retained intron itself. RT-PCR reactions were carried out in a 25-µL reaction mixture containing 1 µL of cDNA, 1× Taq DNA polymerase buffer, 1 µL of dNTP 2.5 mM, 6 pmoles of primers, 0.5 µL of MgCl₂ 50 mM, and 1 unit Taq DNA polymerase (GIBCO/BRL). Standard PCR conditions were: 4 min at 94°C (initial denaturation), 45 sec at 94°C, 45 sec at 58°C, and 1 min at 72°C for 40 cycles and a final extension step of 10 min at 72°C. PCR products were analyzed both on 8% silver-stained polyacrylamide gels and on ethidium bromide agarose gel.

ACKNOWLEDGMENTS

We thank Maria D. Vibranovski, Anamaria A. Camargo, and Ricardo R. Brentani for critical reading of the manuscript. N.J.S. is supported by a Ph.D. fellowship from Fapesp. N.K.S. is supported by a Ph.D. fellowship from Capes. The authors are also indebted
to two anonymous reviewers for excellent comments on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received July 8, 2003; accepted January 26, 2004.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Baker, B.S. 1989. Sex in flies: The splice of life.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Bashirullah, A., Cooperstock, R.L., and Lipshitz, H.D. 2001. Spatial and temporal control of RNA stability. Proc. Natl. Acad. Sci. 98: 7025–7028.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. Nucleic Acids Res. 30: 276–280.

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. 72: 291–336.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. 1993. dbEST—Database for "expressed sequence tags." Nat. Genet. 4: 332–333.

Croft, L., Schandofer, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J.S. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. 24: 340–341.

Gilbert, W. 1978. Why genes in pieces? Nature 271: 501.

Gonzalez, C.I., Ruiz-Echevarria, M.J., Vasudevan, S., Henry, M.F., and Peiltz, S.W. 2000. The yeast hnRNP-like protein Hrp1/Nab4 marks a transcript for nonsense-mediated mRNA decay. Mol. Cell 5: 489–499.

Goodall, G.J. and Filipowics, W. 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell 58: 473–483.

———. 1991. Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J. 10: 2635–2644.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J., and Bork, P. 1999. Alternative splicing of human genes: More than the rule exceptation. Trends Genet. 15: 389–390.

Kan, Z., States, D., and Gish, W. 2002. Selecting for functional alternative splices in ESTs. Genome Res. 12: 1837–1845.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Le Hir, H., Charlet-Berguerand, N., de Franciscis, V., and Thermes, C. 2002. 5' end Ret splicing: Absence of variants in normal tissues and intron retention in pheocromocytomas. Oncology 63: 84–91.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. Genome Res. 9: 1288–1293.

Modrek, B. and Lee, C. 2002. A genomic view of of alternative splicing. Nat. Genet. 30: 13–19.

Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. 29: 2850–2859.

Osorio, E., de Souza, J.E., Zaiats, A.C., de Oliveira, P.S.L., and de Souza, S.J. 2003. pp-Blat: A "pseudo-parallel" Blast. J. Mol. Biol. Res. 36: 463–464.

Rio, D.C. 1991. Regulation of Drosophila P-element transposition. Trends Genet. 7: 282–287.

Sakabe, N.J., de Souza, J.E., Galante, P.F.A., de Oliveira, P.S.L., Passetti, F., Brentani, H., Osorio, E.C., Zaiats, A.C., Leerkes, M.R., Kitajima, J.P., et al. 2003. ORESTES are enriched in rare exon usage variants affecting the encoded protein. C. R. Biol. 326: 979–985.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. Science 274: 540–546.

Stamm, S., Zhu, J., Nakai, K., Stoiolov, P., Stoss, O., and Zhang, M.Q. 2000. An alternative-exon database and its statistical analysis. DNA Cell Biol. 19: 739–756.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shennem, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. 99: 16899–16903.

Thanaraj, T.A., Clark, F., and Mullu, J. 2003. Conservation of human alternative splice events in mouse. Nucleic Acids Res. 31: 2544–2552.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., An, P., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Proc. Natl. Acad. Sci. 100: 10577–10582.

Xie, J. and Black, D.L. 2001. A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. Nature 410: 936–939.

Zavolan, M., van Nimwegen, E., and Gaasterland, T. 2002. Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. Genome Res. 12: 1377–1385.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. J. Comp. Biol. 7: 203–214.
Anexo 2d – Artigo: Sense-antisense pairs in mammals: functional and evolutionary Considerations
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

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
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 contain 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.

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

Table 1

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

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).

Table 2

Human Mouse

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

**Table 2**

| NAT pair | 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.

**Table 3**

| NAT pair | 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.

**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
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-

### Table 4

| Distribution of MPSS tags in an antisense orientation in human and mouse genomes |
|----------------------------------|----------------------------------|
|                                  | 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).
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 (chisquare = 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...
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...
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.

![Figure 3](http://genomebiology.com/2007/8/3/R40)

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.
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 × 10^-9). 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 × 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. Over-expression 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
preponderantly 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 × 10^{-8} for donor sites; and chi-square = 23, df = 1, p = 1.6 × 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 responders" 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 (synthetic) 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 antisense MPSS 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 (CAGGGGA), giving a total of 23 bases for each primer (5’-CAGGGGATCCGCCCCCCT-3’). We also used an antisense primer (ACTATCTAGACGCCGCCGCTT) 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 1x PCR buffer, 2.0 mM MgCl₂, 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 1 (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’ TGGTGTTCTCTAGTCTGGG3’; reverse 5’ CATTGTGTAGTAGCTCGC3’). All PCR amplifications were carried out using 2 µl of cDNA as a template to the final volume of 25 µl and 1x buffer, 1.5 mM MgCl₂, 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 assessed 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 RQ2 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-
cript. 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 specific primers and 1 U of Platinum Taq DNA polymerase (Invitrogen). 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 indebted 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, Vaguer 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. Daly A, 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 QJ, Wei L: Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-AST) in ten species.
8. Katayama S, Tomaru Y, Kasukawa T, Kaki K, Nakashima 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. Brandl S: Antisense-RNA regulation and RNA interference. Biochim Biophys Acta 2002, 1575:15-25.
12. Rouguerre C, Heard E: Antisense RNA in imprinting: spreading silencing through Air. Trends Genet 2002, 18:434-437.
13. Prescott EM, Proudfoot NJ: Translational 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, Zalesiewicz RA, Byrket 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, Williams 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, Khorosrai 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, 3:243-253.
21. Shendure J, Church GM: Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol 2002, 3:R44.
22. Shen J, Sun M, Ken 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, 32:8412-8420.
23. Quere R, Manchon L, Lejeune M, Clement O, Pierrat F, Bonafoux B, Commes T, Piquemal D, Marjat M: 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 PO, Suzuki H, Ninnomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, 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 of intron retention in the human transcriptome. Genome Biol 2004, 5:R14.

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. Veleculescu 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 microarrays. Nat Biotechnol 2000, 18:630-634.

31. Boon K, Osorio EC, Greenhurst SF, Schaefer CF, Shoemaker J, Polvak K, Morin PJ, Buettow KH, Strausberg RL, De Souza SJ, et al.: An anatomy of normal and malignant gene expression. Proc Nat 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, Flatochart AV, Almeida GT, Epedio CM, Paquola AC, Machado AA, Festa F, et al.: Anti-sense intronic non-coding RNA levels correlate to the degree of tumour differentiation in prostate cancer. Oncogene 2004, 23:6684-6692.

34. Kyosawa H, Iwase N, Iwashita Y, Abe K, Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K: Disclosure 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. NCBI Mouse Transcriptome Project [http://www.ncbi.nlm.nih.gov/genome/guide/mouse/MouseTranscriptome.html]

37. Silva AP, Duanloup I, Virkenbosch N, Reymond A, Kaessmann H: Rapid identification and characterization of a novel gene SAF transcribed from the opposite strand of FAS. Hum Mol Gen 2005, 14:1465-1474.

38. Wiemann S, Kolb-Kokocinski A, Poustka A: Alternative pre-mRNA processing regulates cell-type specific expression of the IL4R and NUP62 genes. BMC Biol 2005, 3:16.

39. Takahara T, Kanazu S, Yanagisawa S, Akanuma H: Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic trans-splicing. Trends Genet 2005, 21:326-329.

40. 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, 21:203-207.

41. Nott A, Meislin SH, Moore MJ: A quantitative analysis of intron effects on mammalian gene expression. RNA 2003, 9:607-617.

42. UCSC Genome Browser: Download Page [http://hgdownload.cse.ucsc.edu/]

43. Kent WJ: BLAT - the BLAST-like alignment tool. Genome Res 2002, 12:656-664.

44. 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.

45. Rice P, Longden I, Bleasby A: EMBOSS: The European Molecular Biology open software suite. Trends Genet 2000, 16:276-277.

46. NCBI: Mouse Transcriptome Project [http://www.ncbi.nlm.nih.gov/genome/guide/mouse/MouseTranscriptome.html]

47. U C S C G e n o m e B r o w s e r : D o w n l o a d P a g e [http://hgdownload.ucsc.edu/]

48. BLAT - the BLAST-like alignment tool. Genome Res 2002, 12:656-664.

49. EmboSS: The European Molecular Biology open software suite. Trends Genet 2000, 16:276-277.