UNIVERSIDADE DE SÃO PAULO
Programa de Pós-Graduação em Bioinformática

JORGE ESTEFANO SANTANA DE SOUZA

**Identificação *in-silico* de genes humanos submetidos à
expressão alélica diferencial**

São Paulo

Data do Depósito:
07/10/2008

JORGE ESTEFANO SANTANA DE SOUZA

**Identificação *in-silico* de genes humanos submetidos à
expressão alélica diferencial**

Tese apresentada ao Programa Interunidades em
Bioinformática da Universidade de São Paulo para
obtenção do Título de Doutor em Bioinformática.

Área de concentração: Bioinformática.

Orientador: Prof. Dr. Sandro José de Souza

Co-orientador: Prof. Dr. Junior Barrera

Durante a elaboração deste trabalho o autor recebeu apoio financeiro da CAPES.

- São Paulo, outubro de 2008 -

Jorge Estefano Santana de Souza

Identificação *in-silico* de genes humanos submetidos à expressão alélica diferencial.

Este exemplar corresponde à redação final da tese de doutorado devidamente corrigida e defendida por Jorge Estefano Santana de Souza e aprovada pela comissão julgadora.

São Paulo, 02 de dezembro de 2008.

Banca Examinadora

Prof. Dr. Sandro José de Souza

Instituição: Instituto Ludwig de Pesquisa sobre o Câncer

Prof. Dr. Diogo Meyer

Instituição: Universidade de São Paulo, Instituto de Biociências.

Profa. Dra. Aline Maria da Silva

Instituição: Universidade de São Paulo, Instituto de Química.

Profa. Dra. Dirce Maria Carraro

Instituição: Fundação Antônio Prudente.

Dra. Ariane Machado Lima

Instituição: Universidade de São Paulo, Instituto de Matemática e Estatística.

À minha Mãe, Esposa e Filho.

AGRADECIMENTOS

Esta seção é para agradecer aqueles que se tornaram fundamentais para este passo de minha formação acadêmica.

Primeiro, agradeço a CAPES pela bolsa concedida, ao meu orientador, Sandro José de Souza, e ao meu co-orientador, Junior Barrera, pela orientação e todo o suporte necessário para a conclusão deste trabalho, assim como a oportunidade para o desenvolvimento da minha formação. Ainda entre os grandes pesquisadores dos quais tive a oportunidade de trabalhar, agradeço à Anamaria Aranha Camargo pelas inúmeras colaborações, amizade e ensinamentos, à Helena Paula Brentani pelas broncas, ensinamentos, amizade e por me mostrar que pontos de vistas devem ser discutidos sem que isso implique em rompimentos.

Dentre aqueles que quero agradecer, existe uma lista de pessoas que se tornaram fundamentais para a minha formação; pessoas fantásticas a quem devo muito e a quem nunca irei esquecer. Estas pessoas me ensinaram que, independente das circunstâncias, devo sempre procurar o melhor, que não devo me influenciar por maus exemplos e que, independente de uma vitória ou um fracasso, sempre temos algo a aprender. Agradeço ao Pedro Alexander Favoretto Galante pela amizade, pelas inúmeras colaborações, pelo suporte na área biológica, pelas dicas de trabalho, por corrigir todos os meus pôsteres, relatórios, trabalhos, resumos, traduções, por me ajudar a se tornar uma pessoa melhor, por me ajudar a entender melhor os processos da vida, por me ajudar nas dúvidas de estatística e nas inúmeras ferramentas de bioinformática que utilizamos. Ao Noboru Jo Sokabe pela amizade, pelo suporte na área de bioquímica e biologia molecular, pelas infundáveis discussões sobre ética, cidadania, vida, pelas colaborações, pelas repreensões, pelas correções de trabalhos e por ajudar a solucionar as minhas dúvidas mesmo sem ter muito tempo para isto. À Maria Dulcetti Vibranovski pela amizade, pelas viagens ao Rio de Janeiro, pelos suportes na área de evolução, pelas idas aos cinemas, teatros e barzinhos, pelas discussões de arte e cultura, pelas brigas horrendas por

pontos de vistas diferentes, pelos abraços depois destas brigas, pelos conselhos e pelo ombro amigo nos momentos difíceis. Ao Robson Francisco de Souza pela amizade, pelo suporte em todas as áreas do saber, pelas infinitas discussões sobre os mais variados assuntos, por estar sempre disposto a gastar horas e mais horas me explicando relevâncias inerentes aos nossos estudos, das quais eu nem havia imaginado existirem. Sei que estas pessoas são muito mais extraordinárias do que eu consegui descrever, assim como foram muito mais importante do que eu consigo imaginar. Sou grato e espero ter retribuído à altura tanta dedicação.

Agradeço aos amigos de laboratório Natanja Kirschbaum-Slager, Elza Helena Barbosa, Patrícia Marra Carvalho e Daniel T. Ohara pela convivência. Também agradeço ao Daniel Onofre Vidal, à Ana Paula Medeiros Silva e à Lílían Lopes pelas validações experimentais e por toda parte de bancada de muitos dos meus trabalhos. Aqui cabe um agradecimento especial à Ana Claudia Pereira pela ajuda na parte burocrática, pela pronta dedicação e pela amizade que tem me dedicado.

Por fim, agradeço à minha família, à minha mãe, Jane Rita Sampaio Santana, aos meus irmãos, John Robson Santana de Souza e Anderson Santana de Souza, aos meus primos, Janderson Santana Ferreira e Ricardo Santana de Souza por torcerem e acreditarem em mim, por me socorrerem todas as vezes que necessitei e por me cobrarem todas as vezes que lhes faltei. Em especial eu quero agradecer à minha esposa, Maria Joelma de Santana Souza, e ao meu filho, Guilherme de Santana Souza, por tanta dedicação, amor e companheirismo.

“Que ninguém se engane, só consigo a simplicidade
através de muito trabalho.”

Clarice Lispector em “A Hora da Estrela”

RESUMO

Souza, J.E.S. **Identificação *in-silico* de gene humanos submetidos à expressão alélica diferencial**. 2008. (73 pág.). Tese (Doutorado) - Programa de Pós-Graduação em Bioinformática. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Estudos recentes demonstraram que a variação de expressão alelo-específica é mais comum do que se imaginou, podendo chegar, em humanos, a 50% dos genes. Identificar os genes submetidos ao controle de expressão alelo-específica é muito importante para o entendimento de várias doenças, incluindo o câncer. A identificação dos alvos desse tipo de regulação diferencial é difícil, principalmente devido à dificuldade de se avaliar a expressão de cada alelo individualmente. Neste trabalho, abordamos este problema com uma estratégia de análise *in-silico*, fundamentada na integração de dados públicos do genoma humano, dados de expressão (como cDNAs, SAGE e MPSS) e dados sobre polimorfismos (SNPs). Desenvolvemos um banco de dados de polimorfismos de base única (*Single-Nucleotide Polymorphism* - SNPs) associados a etiquetas alternativas de SAGE (*Serial Analysis of Gene Expression*) e MPSS (*massively parallel signature sequencing*). SAGE e MPSS são técnicas desenvolvidas para análise da expressão de genes em larga escala. Ambas as técnicas têm como princípio a produção de pequenas seqüências marcadoras (etiquetas), adjacentes aos sítios de enzimas de restrição que estiverem mais próximo da cauda poli-A do RNA mensageiro. Tais etiquetas são seqüenciadas em grande escala e a quantidade de etiquetas é usada para medir a abundância relativa dos RNAs mensageiros correspondentes. A presença de SNPs nos sítios de restrição ou nas seqüências das etiquetas pode gerar etiquetas distintas para alelos do mesmo gene, que denominamos etiquetas alternativas. Neste trabalho, empregamos o banco de dados de etiquetas alternativas associadas a SNPs para identificar

genes com expressão alélica diferencial. Usando esta estratégia, identificamos 812 genes com expressão monoalélica. Estudos anteriores comprovaram que, dentre os 812 genes identificados, cinco estão sujeitos ao fenômeno de *imprinting* genômico. Durante o decorrer deste estudo, trabalhos realizados por outros grupos apontaram outros 73 genes do nosso repertório como genes que apresentam variação no nível de expressão dos alelos em heterozigotos. Com objetivo de confirmar a expressão alélica diferencial dos nossos candidatos, selecionamos 29 genes para validação experimental. Para 12 destes genes não achamos indivíduos heterozigotos, impossibilitando a análise da expressão dos alelos. Dentre os outros 17 genes, três apresentaram expressão bialélica e 14 apresentaram expressão alélica diferencial nos indivíduos heterozigotos, sendo que 3 deles apresentaram expressão monoalélica. Estes resultados sugerem que nossa estratégia pode contribuir significativamente na identificação de genes com expressão alélica diferencial.

Palavras-chave: expressão alélica diferencial, genes *imprinted*, bioinformática, SAGE, MPSS, SNP.

ABSTRACT

Souza, J.E.S. ***In-silico* identification of human genes submitted to allelic differential expression**. 2008. (73 pag.). PhD Thesis - Programa de Pós-Graduação em Bioinformática. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Recent studies have shown that variation of allelic-specific gene expression is more common than previously thought, reaching up to 50% of human genes. To identify genes displaying differential expression among alleles it is important for the understanding of several diseases, including the cancer. Identification of genes submitted to allelic-specific differential expression is hard, mostly due to the difficulty in evaluating the expression levels of each allele independently. In this work, we developed an *in-silico* approach, based on the integration of public data about the human genome, gene expression data (such as cDNAs, SNPs, SAGE and MPSS) and data on polymorphisms (SNPs). We developed a database of Single Nucleotide Polymorphisms (SNPs) associated to alternative SAGE (Serial Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing) tags. SAGE and MPSS are genome-wide techniques developed for analysis of gene expression. Both techniques rely on the production of short marker sequences (known as tags), adjacent to restriction sites closer to the poly-A tail of messenger RNAs. Such tags are sequenced in a large scale and tag counts are used to measure the relative abundance of their corresponding transcripts. The presence of SNPs in the restriction sites or in the tag sequences might generate allelic-specific tags for the same gene, which we call alternative tags. In this work, we used the database of SNPs and associated alternative tags to identify genes submitted to allelic-specific differential gene expression. Using this approach, we identified 812 genes showing allelic-specific differential gene expression. Previous studies have shown that,

among the 812 candidates, five genes are targets for genomic imprinting. While this study was being performed, work done by other groups suggested other 73 genes in our candidates list to have different expression levels for alleles in heterozygous. Aiming to verify whether variations in the expression levels of alleles existed among our candidate genes, we submitted 29 genes for experimental validation. For 12 genes, we couldn't find heterozygous individuals, thus rendering it impossible to ascertain whether the supposed expression variation was true. Among the other 17 genes analyzed, three genes presented bi-allelic expression and 14 genes have shown clear differential expression among alleles, three of the last ones displaying strict mono-allelic expression. These results suggest that our approach may contribute significantly to the identification of genes with allelic-specific differential expression.

Keywords: allelic-specific differential expression, imprinted genes, bioinformatics, SAGE, MPSS, SNP.

SUMÁRIO

1. INTRODUÇÃO.	13
1.1. CONTRIBUIÇÃO GÊNICA.	13
1.2. EXPRESSÃO ALÉLICA.	13
1.3. EXPRESSÃO MONOALÉLICA ALEATÓRIA.	14
1.4. <i>IMPRINTING</i> GENÔMICO.	15
1.5. ESTUDOS E IDENTIFICAÇÃO DE GENES SUBMETIDOS A EXPRESSÃO MONOALÉLICA.	16
2. OBJETIVOS.	20
2.1. OBJETIVOS ESPECÍFICOS.	20
2.1.1. AVALIAR O IMPACTO DE SNPs NOS DADOS DE SAGE.	20
2.1.2. BUSCAR GENES COM PADRÃO DE EXPRESSÃO ALÉLICA DIFERENCIAL.	20
2.1.3. VALIDAÇÃO EXPERIMENTAL.	20
2.1.4. ANÁLISE DOS NÍVEIS DE EXPRESSÃO GÊNICA AMOSTRADOS	21
3. MATERIAIS E MÉTODOS.	22
3.1. DADOS PRIMÁRIOS.	22
3.1.1. GENOMA.	22
3.1.2. cDNAs.	22
3.1.3. BANCO DE POLIMORFISMOS (dbSNP).	22
3.1.4. BANCO EXPERIMENTAL DE SAGE E MPSS.	23
3.1.5. MAPEAMENTO E AGRUPAMENTO DOS cDNAs.	24
3.1.6. MAPEAMENTO DOS POLIMORFISMOS NOS cDNAs.	24
3.1.7. ESCOLHAS DOS cDNAs COM CAUDA POLI-A.	26
3.1.8. BANCO DE ETIQUETAS “VIRTUAIS”.	27
3.2. QUANTIFICANDO A EXPRESSÃO DE UM GENE.	27
3.3. REMOVENDO AS ETIQUETAS AMBÍGUAS.	28
3.4. IDENTIFICANDO SNPs QUE INTERFEREM NA TÉCNICA DE SAGE E MPSS.	29
3.5. BANCO DE ETIQUETAS ALTERNATIVAS DE MPSS E SAGE.	31
3.6. BIBLIOTECAS EXPERIMENTAIS PROVENIENTES DE APENAS UM ÚNICO INDIVÍDUO.	31
3.7. ETIQUETAS ALTERNATIVAS NA IDENTIFICAÇÃO DE GENES COM EXPRESSÃO MONOALÉLICA.	32
3.8. VALIDAÇÃO DA NOSSA METODOLOGIA PELA LITERATURA.	37
3.9. CANDIDATOS PARA VALIDAÇÃO EXPERIMENTAL DE EXPRESSÃO ALÉLICA DIFERENCIAL.	38
3.10. AMOSTRAS DE SANGUE PROVENIENTES DO BANCO DE SANGUE DO DEPARTAMENTO DE HEMATOLOGIA E HEMOTERAPIA DO HOSPITAL A. C. CAMARGO.	40
3.11. ESTRATÉGIA EXPERIMENTAL NA VALIDAÇÃO DE EXPRESSÃO MONOALÉLICA.	40

3.12. IDENTIFICAÇÃO DE EXPRESSÃO ALÉLICA DIFERENCIAL E BIALÉLICA NOS DADOS EXPERIMENTAIS.	43
4. RESULTADOS E DISCUSSÃO.	47
4.1. O IMPACTO DOS SNPs NA INTERPRETAÇÃO DOS DADOS DE SAGE E DE MPSS.....	47
4.1.1. IDENTIFICAÇÃO DE SÍTIOS ALTERNATIVOS DE RESTRIÇÃO.....	48
4.1.2. PERDA DO SÍTIO DE RESTRIÇÃO ORIGINAL	49
4.1.3. NOVAS ETIQUETAS PRODUZIDAS PELA SUBSTITUIÇÃO DE BASE NAS ETIQUETAS ORIGINAIS.....	50
4.1.4. INTEGRAÇÃO DO BANCO DE ETIQUETAS ALTERNATIVAS COM O SAGE GENIE.	51
4.1.5. CONSIDERAÇÕES SOBRE AS ETIQUETAS ALTERNATIVAS.....	53
4.2. ETIQUETAS ALTERNATIVAS NA BUSCA DE GENES COM EXPRESSÃO ALÉLICA DIFERENCIAL.....	56
4.2.1. A CATEGORIA DOS GENES COM EXPRESSÃO MONOALÉLICA.....	57
4.2.2. A CATEGORIA DOS GENES COM ETIQUETAS NÃO INFORMATIVAS.....	58
4.2.3. A CATEGORIA DOS GENES COM EXPRESSÃO BIALÉLICA.....	59
4.2.4. GENES COM PERDA DE EXPRESSÃO MONOALÉLICA EM TUMOR.....	59
4.2.5. GENES JÁ IDENTIFICADOS COM EXPRESSÃO ALÉLICA DIFERENCIAL NA LITERATURA. (BANCO DE GENES COM EXPRESSÃO ALÉLICA DIFERENCIAL).	60
4.2.6. GENES CANDIDATOS SUBMETIDOS À VALIDAÇÃO EXPERIMENTAL.	63
4.2.7. EXPRESSÃO ALÉLICA DIFERENCIAL E BIALÉLICA NOS DADOS EXPERIMENTAIS.	65
5. CONCLUSÕES.....	68
5.1. O IMPACTO DOS SNPs NOS DADOS EXPERIMENTAIS DE SAGE E MPSS.....	68
5.2. A INTEGRAÇÃO DO BANCO DE ETIQUETAS ALTERNATIVAS COM O SAGE GENIE.	68
5.3. ETIQUETAS ALTERNATIVAS NA BUSCA DE GENES COM EXPRESSÃO MONOALÉLICA.	68
5.4. BANCO DE GENES COM EXPRESSÃO ALÉLICA DIFERENCIAL.	69
5.5. VALIDAÇÃO EXPERIMENTAL DA EXPRESSÃO ALÉLICA DIFERENCIAL.	69
5.6. OBSERVAÇÕES FINAIS.....	69
6. REFERÊNCIAS	71
LISTA DE ANEXOS.....	74

1. Introdução.

1.1. Contribuição Gênica.

Por muito tempo acreditou-se que ambos os alelos de todos os genes eram expressos simultaneamente e eqüitativamente nos organismos eucariontes diplóides. Entretanto, com as descobertas, em mamíferos, da inativação do cromossomo X (Lyon, 1961), da expressão monoalélica aleatória (Pernis, Chiappino *et al.*, 1965; Chess, Simon *et al.*, 1994; Rajewsky, 1996; Bix e Locksley, 1998; Hollander, Zuklys *et al.*, 1998) e do *imprinting* genômico (Reik e Walter, 2001), ficou claro que alguns genes não seguem essa expressão eqüitativa e simultânea. Também ficou claro que estudos no sentido de identificar diferenças no padrão de expressão dos alelos e estudos para modelar os mecanismos moleculares pelos quais isto ocorre são muito importantes e necessários para aprimorarmos o nosso conhecimento sobre este processo tão complexo que é a expressão gênica nos eucariotos.

1.2. Expressão Alélica.

Acredita-se que pequenas variações de expressão entre alelos de um mesmo gene seja algo natural (Knight, 2004) e que a inativação do cromossomo X em mamíferos é apenas um mecanismo de compensação de dose para deixar mais igualitária a expressão gênica entre macho e fêmea de uma mesma espécie (Lyon, 1961). Porém, quando focamos o padrão de expressão dos alelos dos genes dos cromossomos autossômicos, existem duas classes de genes que apresentam em comum um padrão de expressão completamente diferente entre os seus alelos, ou seja, um alelo silenciado (falta ou diminuição drástica no nível de expressão de um alelo) e um alelo funcional (expressão ou aumento significativo no nível de expressão de um único alelo) (Cowles, Hirschhorn *et al.*, 2002; Yan, Yuan *et al.*, 2002; Bray, Buckland *et al.*, 2003; Lo, Wang *et al.*, 2003). Estas classes gênicas estão associadas com os fenômenos de expressão monoalélica aleatória e *imprinting* genômico. Os genes destas duas classes têm

um padrão diferencial de expressão muito semelhante. O que separa essas duas classes é que, no caso de *imprinting* genômico, a diferença de expressão alélica é determinada pela sua origem parental. A assinatura de *imprinting* genômico (expressão alelo-específica) pode ser vista nos indivíduos de uma espécie, e a perda deste padrão está associada a diversas doenças genéticas (Wrzeska e Rejduch, 2004). Na expressão monoalélica aleatória o padrão diferencial de expressão não se dá devido a origem parental. Além disto, enquanto um indivíduo de uma espécie apresenta para um determinado gene expressão monoalélica outro indivíduo pode apresentar para esse mesmo gene expressão bialélica diferencial, e essa diferença de dose de expressão geralmente não é comparável entre indivíduos. Assim, por apresentar esse padrão aleatório de expressão, estudos recentes acreditam que esse mecanismo seja importante para estabelecer a diversidade entre os indivíduos (Gimelbrant, Hutchinson *et al.*, 2007; Bjornsson, Albert *et al.*, 2008).

1.3. Expressão Monoalélica Aleatória.

Os genes classificados como expressão monoalélica aleatória podem apresentar, para uma mesma linhagem de clones celular, um conjunto de células com expressão para o alelo paterno, enquanto que para outro conjunto de células expressão do alelo materno, e ainda para um terceiro conjunto a expressão de ambos alelos (Bix e Locksley, 1998; Hollander, Zuklys *et al.*, 1998; Gimelbrant, Hutchinson *et al.*, 2007). Pouco se sabe sobre os mecanismos que levam à expressão monoalélica aleatória, em alguns casos específicos, como nos genes de receptores de superfície em linfócitos B, o silenciamento de um dos alelos está relacionado com rearranjos cromossômicos (Chess, 1998).

1.4. *Imprinting* Genômico.

Imprinting genômico é o fenômeno no qual dois alelos de um mesmo gene são expressos diferencialmente, dependendo de sua origem parental (Knight, 2004). Para estes genes apenas um dos alelos, paterno ou materno, é expresso. Para alguns genes submetidos a fenômenos de *imprinting* genômico (genes *imprinted*), a falta de expressão de um dos alelos pode ser observada em todos os tecidos de um indivíduo, enquanto que para outros genes *imprinted* o silenciamento de um dos alelos é observado apenas em um tecido ou em um particular estágio de desenvolvimento (Morison, Ramsay *et al.*, 2005; Monk, Arnaud *et al.*, 2006).

O *imprinting* genômico foi descrito na década de 80 em estudos com camundongos. Estes animais foram manipulados geneticamente para apresentar contribuição genética exclusiva da mãe ou do pai (Reik e Walter, 2001). Porém, nenhum destes embriões uniparentais era capaz de se desenvolver normalmente, levando à conclusão da importância da expressão gênica desigual entre os alelos maternos e paternos (Mcgrath e Solter, 1984; Surani, Barton *et al.*, 1984). Posteriormente, experimentos com camundongos apresentando dissomias uniparentais parciais ou completas sugeriram mais uma vez que ambos os genomas parentais são necessários para um desenvolvimento normal (Cattanach e Kirk, 1985).

Acredita-se que o mecanismo pelo qual o *imprinting* genômico é transmitido dos progenitores à sua prole é a metilação do DNA (Li, Beard *et al.*, 1993). Estudos sobre o mecanismo molecular de *imprinting* genômico têm demonstrado que a metilação do DNA possui um papel fundamental durante a aquisição e manutenção do *imprinting* genômico (Reik e Walter, 2001). Entretanto, ainda não foi esclarecido como o padrão de metilação diferencial entre os alelos é determinado nas células germinativas e, principalmente, como é mantido nas células somáticas ao longo do desenvolvimento.

1.5. Estudos e identificação de genes submetidos a expressão monoalélica.

Como já descrito, desde a década de 80 as dissomias uniparentais parciais (envolvendo apenas algumas regiões cromossômicas) ou completas (envolvendo cromossomos individuais inteiros) foram utilizadas com sucesso na identificação de regiões cromossômicas que sofrem *imprinting* genômico. No entanto, as duas técnicas descritas em Leighton, Saam *et al.*, 1996; Oakey e Beechey, 2002 são laboriosas, de difícil análise e não permitem a identificação de genes específicos, apenas da região que sofre *imprinting* genômico. Atualmente, com a disponibilidade de diversos genomas, aliado às técnicas de estudo da expressão gênica em larga escala e ao crescente poder computacional, a identificação de genes com padrão de expressão monoalélica se tornou mais precisa e sua análise mais quantitativa. Por exemplo, a identificação de expressão monoalélica foi feita através de cDNA *microarrays* (Lo, Wang *et al.*, 2003), de ESTs (*Expressed sequence Tags*) (Yang, Hu *et al.*, 2003) e pela utilização de regiões homólogas entre várias espécies (Prawitt, Enklaar *et al.*, 2000; Wang, Fan *et al.*, 2004). Entre outras vantagens destas novas abordagens, podemos destacar a possibilidade de se fazer análises em larga escala e a capacidade de monitoramento da expressão alélica diferencial de genes individuais. Também devemos ressaltar que ainda não foram aplicadas a este tipo de estudo técnicas como SAGE (*Serial Analysis of Gene Expression*) (Velculescu, Zhang *et al.*, 1995) e MPSS (*Massively Parallel Signature Sequencing*) (Brenner, Johnson *et al.*, 2000), as quais são especializadas na quantificação da expressão gênica, na identificação de genes com baixa expressão e, sobretudo, não necessitam de nenhum conhecimento prévio sobre a sequência do gene de interesse.

Um bom exemplo da aplicação de comparações entre genomas na busca por genes com expressão monoalélica e genes *imprinted* pode ser visto em Prawitt, Enklaar *et al.*, (2000). Neste trabalho os autores buscaram novos genes ligados à síndrome de *Beckwith-Wiedemann* (BWS) e submetidos à *imprinting* genômico através da análise de regiões

conservadas entre humano, *Caenorhabditis elegans* e camundongo. Outra vantagem destas comparações entre regiões homólogas é a identificação de elementos associados à regulação do *imprinting* genômico (Wang, Fan *et al.*, 2004). Esta metodologia é importante, mas vale salientar que mesmo que alguns dos genes submetidos a *imprinting* genômico em uma espécie tenham homólogos em outra, isso não significa que o padrão de expressão monoalélica seja conservado (Reik e Walter, 2001).

Outra abordagem promissora no estudo de genes com expressão monoalélica é a análise de seqüências transcritas aliada ao conjunto de dados de SNPs (*single nucleotide polymorphism*). A essência desta análise está na identificação de polimorfismos que acontecem em regiões transcritas, o que, em potencial, poderia ser utilizado para diferenciar a expressão de alelos de um mesmo transcrito. Um exemplo desta análise pode ser visto em Lo, Wang *et al.*, (2003). Neste trabalho, o estudo da expressão alélica foi feito com o *microarray Affymetrix HuSNP* que é capaz de discriminar 1.063 SNPs localizados em regiões de transcritos conhecidos. Como resultado, os autores encontraram 170 genes com diferença de expressão superior a quatro vezes entre os dois alelos. Destes 170 genes, 4 já eram confirmados como genes *imprinted* por outros trabalhos.

Também analisando seqüências expressas e seus SNPs correspondentes, Yang, Hu *et al.*, (2003) avaliaram o padrão de expressão alelo-específico de diversos genes. Estes autores utilizaram ESTs de diferentes bibliotecas de cDNAs provenientes de tecidos normais e tumorais, considerando que cada biblioteca de cDNA representava um indivíduo e que o conjunto destas bibliotecas representavam uma população. Eles foram capazes de analisar 50 SNPs mapeados em genes que sofrem *imprinting* genômico. Ao analisar tecidos normais eles encontraram 4 SNPs com expressão monoalélica, mas quando eles acrescentavam bibliotecas provenientes de tecidos tumorais apenas 1 SNP continuou classificado como expressão

monoalélica. Este resultado é importante e consistente com a hipótese de perda de *imprinting* genômico (LOI – *loss of imprinting*) em tecidos tumorais.

Também utilizando dados de ESTs associados a SNPs e com a finalidade de identificar genes que apresentavam expressão alélica diferencial, Ge, Gurd *et al.*, (2005) validaram experimentalmente a expressão diferencial de 14 genes. Neste trabalho também foi desenvolvido um *software*, chamado *Peak Picker*, para facilitar as análises dos dados de expressão diferencial das sequências. Este *software* foi utilizado nesta presente tese como parte importante para conseguirmos quantificar os resultados obtidos experimentalmente (ver materiais e métodos).

Recentemente, Gimelbrant, Hutchinson *et al.*, (2007) fizeram um estudo em larga escala também com a finalidade de identificar genes submetidos à expressão monoalélica. Neste trabalho, o estudo da expressão alélica foi feito com um *array Affymetrix Human Mapping 500k*, e foram analisados aproximadamente 4.000 genes humanos. Destes 4000 genes, 371 (9,5%) foram classificados como apresentando expressão monoalélica. Este trabalho foi importante, pois mostrou que 80% dos 371 genes classificados como expressão monoalélica também apresentam expressão bialélica em alguns clones celulares, assim como previsto no modelo de expressão monoalélica aleatória.

Atualmente, o interesse no estudo de genes com padrão de expressão monoalélica aleatória e *imprinting* genômico vem crescendo, principalmente porque alguns trabalhos comprovaram que a perda de *imprinting* genômico (*loss of imprinting* - LOI) está associada a diversas doenças genéticas, tais como as síndromes de *Angelman*, de *Prader-Willi*, de *Beckwith-Wiedemann*, de *Silver-Russel* e ao câncer (Wrzeska e Rejduch, 2004). Além destes estudos, recentemente diversos trabalhos estimaram que aproximadamente 50% dos genes dos cromossomos autossômicos podem apresentar uma diferença significativa no nível de expressão entre os seus alelos (Lo, Wang *et al.*, 2003; Ge, Gurd *et al.*, 2005; Pant, Tao *et al.*,

2006; Gimelbrant, Hutchinson *et al.*, 2007) e que esse nível de expressão diferencial entre alelos é importante para estabelecer a diversidade entre os indivíduos e células. Estes estudos ainda argumentaram que um desbalanço da expressão entre os alelos possa levar ao desenvolvimento de uma série de doenças e ao câncer.

Após uma análise da literatura, constatamos que foram identificados e comprovados apenas 96 genes que sofrem *imprinting* genômico (aproximadamente 55 em humanos) (Morison, Paton *et al.*, 2001) e aproximadamente 1.186 genes com expressão alélica diferencial. Acreditamos que este número relativamente baixo é reflexo da uma grande dificuldade de analisar a expressão entre os alelos de um gene. Considerando todas estas informações, acreditamos que a utilização de metodologias e estratégias de bioinformática utilizando bancos de dados públicos de seqüências e técnicas como SAGE, aliadas aos dados de SNPs, possam contribuir significativamente para a identificação de genes submetidos a expressão alélica diferencial.

2. Objetivos.

Desenvolver uma nova estratégia de análise *in-silico*, que visa combinar os dados de expressão gênica derivados por bibliotecas de SAGE e MPSS com dados de SNPs e cDNAs objetivando a identificação de genes com expressão alélica diferencial.

2.1.Objetivos Específicos.

2.1.1. Avaliar o impacto de SNPs nos dados de SAGE.

Elaborar uma estratégia para identificar e avaliar o impacto de SNPs nos dados de SAGE. Tal objetivo requer a identificação dos genes humanos que apresentam etiquetas com polimorfismo em sua sequência ou no sítio das enzimas de restrição utilizadas na obtenção das etiquetas. Estas etiquetas (aqui chamadas de etiquetas alternativas) foram mais tarde utilizadas para discriminar a expressão de um alelo específico.

2.1.2. Buscar genes com padrão de expressão alélica diferencial.

A partir das etiquetas alternativas identificadas no objetivo 2.1.1, elaboramos filtros e estratégias para identificar genes que apresentavam um padrão de expressão monoalélica, isto é, genes que nunca apresentavam etiquetas de ambos os alelos na mesma biblioteca de SAGE.

2.1.3. Validação experimental.

Confirmar experimentalmente a expressão diferencial postulada para alguns genes candidatos após a obtenção dos resultados de 2.1.2. Esta validação foi feita em colaboração com o laboratório da Dra. Anamaria A. Camargo, do Instituto Ludwig de Pesquisa sobre o Câncer, sendo o aluno Daniel Onofre Vidal o responsável pela coleta de material biológico, extração e quantificação do RNA e pela genotipagem do DNA. As amostras de sangue

utilizadas foram provenientes de 30 doadores normais do Banco de Sangue do departamento de Hematologia e Hemoterapia do Hospital A. C. Camargo.

2.1.4. Análise dos níveis de expressão gênica amostrados

Durante a etapa de validação experimental, tornou-se necessário o desenvolvimento de um novo método de análise para os níveis de expressão alélica diferencial. Tal método deveria levar em conta os erros de incorporação de bases durante o sequenciamento dos genes estudados experimentalmente.

3. Materiais e Métodos.

3.1. Dados primários.

3.1.1. Genoma.

O genoma de *Homo sapiens* foi obtido do *Genome Browser* da *University of California Santa Cruz* (UCSC) (<ftp://hgdownload.cse.ucsc.edu/>). No trabalho em que estudamos o impacto de SNPs na interpretação nos dados experimentais de SAGE e de MPSS (Silva, De Souza *et al.*, 2004) foi utilizada a versão 34.2 do genoma humano. No trabalho em que identificamos os genes submetidos à expressão alélica diferencial em humanos (De Souza, Vidal *et al.* em submissão) foi utilizada a versão 35.1 do genoma humano.

3.1.2. cDNAs.

As seqüências de cDNAs foram obtidas do *National Center for Biotechnology Information* (NCBI) (<ftp://ftp.ncbi.nih.gov>) e do UCSC *Genome Browser* (<ftp://hgdownload.cse.ucsc.edu/>), arquivos: mrna.fa e refMrna.fa (cDNAs). No trabalho em que estudamos o impacto de SNPs na interpretação nos dados experimentais de SAGE e MPSS (Silva, De Souza *et al.*, 2004) utilizamos um total de 130.148 seqüências. No trabalho de identificação de genes submetidos à expressão alélica diferencial em humanos (De Souza, Vidal *et al.* em submissão) utilizamos um total de 229.557 seqüências.

3.1.3. Banco de polimorfismos (dbSNP).

Foi criado um banco de dados local com os polimorfismos de único nucleotídeo (*Single-Nucleotide Polymorphism* - SNPs) obtidos no dbSNP do NCBI (<ftp://ftp.ncbi.nih.gov/SNP/>). No trabalho em que estudamos o impacto de SNPs na interpretação nos dados experimentais de SAGE e MPSS (Silva, De Souza *et al.*, 2004) utilizamos um total de 5.789.183 SNPs (dbSNP versão 118). No trabalho em que

identificamos os genes submetidos à expressão alélica diferencial em humanos (De Souza, Vidal *et al.* em submissão) utilizamos um total de 10.054.521 SNPs (versão 124 do dbSNP).

3.1.4. Banco experimental de SAGE e MPSS.

SAGE e MPSS são técnicas que permitem avaliar, de forma quantitativa, o perfil de expressão gênica de um determinado tecido sem o conhecimento a priori da sequência dos genes que estão sendo expressos. Estas técnicas são bastante similares e quantificam os transcritos através de sequências de 10 a 21 bases (etiquetas), extraídas a partir do sítio de restrição (NlaIII para SAGE e DpnII para MPSS) mais 3' de cada RNA mensageiro (Velculescu, Zhang *et al.*, 1995; Brenner, Johnson *et al.*, 2000). Os dados experimentais de SAGE foram obtidos do “SAGE Genie” (<http://cgap.nci.nih.gov/SAGE>). No trabalho em que estudamos o impacto de SNPs na interpretação nos dados experimentais de SAGE e MPSS (Silva, De Souza *et al.*, 2004) utilizamos 586.144 etiquetas únicas geradas de 260 bibliotecas de *short* SAGE derivadas de 25 tecidos humanos. Os dados experimentais de MPSS foram obtidos do “Ludwig Institute for Cancer Research and National Cancer Institute MPSS database” (<http://mpss.licr.org>) (Jongeneel, Delorenzi *et al.*, 2005). Foram utilizadas 84.555 etiquetas distintas geradas de seis bibliotecas de cólon e de mama. Para o trabalho de identificação de genes submetidos à expressão alélica diferencial em humanos (De Souza, Vidal *et al.* em submissão) foi feita uma atualização dos dados primários e resolvemos utilizar também as bibliotecas de *long* SAGE. A diferença entre *long* e *short* SAGE é que *long* SAGE apresenta etiquetas de 17 pares de base, enquanto *short* SAGE apresenta etiquetas de 10 pares de base. Esta nova versão do SAGE Genie continha 713.492 etiquetas (distintas) de 305 bibliotecas de *short* SAGE derivadas de 25 tecidos humanos e 1.087.047 etiquetas (distintas) de 66 bibliotecas de *long* SAGE derivadas de 6 tecidos humanos.

3.1.5. Mapeamento e agrupamento dos cDNAs.

O mapeamento das seqüências de cDNAs no genoma humano foi realizado através de um abordagem que utiliza BLAT (www.genome.ucsc.edu/blat) e sim4 (Florea, Hartzell *et al.*, 1998). O agrupamento dos cDNAs foi realizado por uma metodologia desenvolvida localmente. Estas abordagens já foram previamente descritas nos nossos trabalhos (Sakabe, De Souza *et al.*, 2003; Galante, Sakabe *et al.*, 2004; Galante, Vidal *et al.*, 2007) e foram implementada por Pedro AF Galante (para detalhes, ver em www.teses.usp.br). Brevemente, todas as seqüências de cDNAs foram alinhadas no genoma humano e somente o melhor alinhamento (aquele com maior identidade e com o maior comprimento alinhado) para cada seqüência foi mantido. Em seguida, todos os alinhamentos foram agrupados, onde cada grupo de seqüências representa um gene humano conhecido. Por fim, foi construído um banco de dados MySQL contendo todas as posições genômicas dos cDNAs (através deste banco é possível descobrir a posição exata de qualquer *exon* ou *intron* de um gene conhecido, o número de exons deste gene, a posição das regiões codificadoras (CDS) e UTRs, o tecido de origem, o tamanho e informações patológicas de suas seqüências).

3.1.6. Mapeamento dos polimorfismos nos cDNAs.

Para mapear os SNPs nos cDNAs nós cruzamos o banco de alinhamentos com o banco de SNPs (ambos os bancos já foram citados anteriormente). Este foi o método: i) sabendo as posições genômicas dos SNPs (PosSNPdna), como pode ser visto na tabela 3.3.1, escrevemos um programa em *Perl* para contar quantos *dels* (deleções nas seqüências) existem entre o início do *exon* (StartExon) e o SNP, tanto no cDNA (delcDNA) quanto no genoma (delDNA). ii) identificamos a posição dos SNPs nos cDNAs usando as fórmulas:

➔ Para alinhamentos de cDNAs na fita positiva:

$$\text{PosSNPcDNA} = (((\text{PosSNPdna} - \text{StartExon} + 1) - \text{delcDNA}) + \text{delDNA})$$

➔ Para alinhamentos de cDNAs na fita negativa:

$$\text{PosSNPcDNA} = (((\text{EndExon} - \text{PosSNPdna} + 1) - \text{delcDNA}) + \text{delDNA})$$

Onde EndExon é a posição genômica que representa o final do *exon* onde o SNP foi identificado.

Na figura 3.3.1 podemos verificar a veracidade da fórmula citada acima. Usando o sim4, a sequência AL834172 foi alinhada no cromossomo 1; as posições dos SNPs foram identificadas na sequência genômica e na sequência de cDNA (ver as setas em vermelho). A tabela 3.3.2 mostra as coordenadas dos polimorfismos no genoma e no cDNA.

Tabela 3.3.1 – Posições genômicas de alguns SNPs do dbSNP.

SNP	Posição Cromossômica	Cromossomo	Alelos
rs3817921	6423420	chr1	G/C
rs3817920	6423426	chr1	T/C
rs3817914	6423619	chr1	T/C
rs2311045	6423631	chr1	G/C
rs3817912	6423649	chr1	T/C
rs3817911	6423687	chr1	G/A
rs3817910	6423690	chr1	G/A

Tabela 3.3.2 – Exemplos de SNPs identificados.

SNP	Alelos	Posição Cromossômica	cDNA	Posição no cDNA	Alelo cDNA
rs3817921	G/C	6423420	AL834172	98	C
rs3817920	T/C	6423426	AL834172	104	C
rs3817914	T/C	6423619	AL834172	171	C
rs2311045	G/C	6423631	AL834172	183	G
rs3817912	T/C	6423649	AL834172	201	C
rs3817911	G/A	6423687	AL834172	239	G
rs3817910	G/A	6423690	AL834172	242	A

```

seq1 = chr1.AL834172.fasta, 676 bp
seq2 = chr1

>AL834172
>CHR1

1-133      (6423320-6423455)      97% ->
134-264    (6423581-6423712)      98% ->
265-466    (6427128-6427329)      100% ->

      0      .      :      .      :      .      :      .      :
      1 CACCGCCATGCACTT  CG CGAGCCGCGGCCACACCAAGGTGCTCAGCT
      |||||-----|||-----|||-----|||-----|||-----|||
6423320 CACCGCCATGCACTTCGCGGCGAGCCGCGGCCACACCAAGGTGCTCAGCT

      50      .      :      .      :      .      :      .      :
      48 GGCTGCTGCTGCACGGCGGGGAGATCTCGGCTGACCTGTGGGGCGGGACC
      |||||-----|||-----|||-----|||-----|||-----|||
6423370 GGCTGCTGCTGCACGGCGGGGAGATCTCGGCTGACCTGTGGGGCGGGACC

      100      .      :      .      :      .      :      .      :
      98 CCGCTGCACGACGCGCCGCGGAGAACGGGGAGCTAGAG      TGCTG
      |||||-----|||-----|||-----|||-----|||-----|||
6423420 CCGCTGCACGACGCGCCGCGGAGAACGGGGAGCTAGAGGTC...CAGTGCTG
      ↑      ↑
      150      .      :      .      :      .      :      .      :
      139 CCAGATCCTGGTAGTGAACGGCG GGAGCTGGACGTCCGCGACCGGGACG
      |||||-----|||-----|||-----|||-----|||-----|||
6423586 CCAGATCCTGGTAGTGAACGGCGCGGAGCTGGACGTCCGCGACCGGACG
      ↑      ↑
      200      .      :      .      :      .      :      .      :
      188 GGTACACGGCCGCGGACCTGTTCGGAACGGCCACAGCCACTGCACC
      |||||-----|||-----|||-----|||-----|||-----|||
6423636 GGTACACGGCCGCGGACCTGTTCGGAACGGCCACAGCCACTGCACC
      ↑
      250      .      :      .      :      .      :      .      :
      238 CGCTACCTGCGCACGGTGGAGAACCTG      AGCGTGGAGCACCG
      |||||-----|||-----|||-----|||-----|||-----|||
6423686 CGCTACCTGCGCACGGTGGAGAACCTGGTA...CAGAGCGTGGAGCACCG
      ↑      ↑

```

FIGURA 3.3.1 - Alinhamento da sequência AL834172 contra o cromossomo 1 usando o programa sim4. As setas em vermelho representam SNPs identificados através do banco dbSNP. Na maioria dos casos a sequência AL834172 e a sequência genômica apresentam o mesmo alelo.

3.1.7. Escolhas dos cDNAs com cauda poli-A.

Experimentalmente a técnica de SAGE e de MPSS obtém a etiqueta adjacente ao sítio de restrição mais 3' para as enzimas NlaIII (SAGE) e DpnII (MPSS). Então, quando vamos extrair as etiquetas virtuais das sequências de cDNAs (predições usadas para anotar as etiquetas reais - para detalhes, ver abaixo) precisamos garantir que estas sequências

apresentam a região 3' completamente seqüenciadas. Uma maneira de identificar se a seqüências de cDNA apresentam esta característica é confirmar a presença de, ao menos, 5 bases 'A' no sua região mais 3' (cauda de poli-A). Nas nossas análises foram utilizadas apenas as seqüências de cDNA com uma cauda de poli-A (54.645 seqüências na construção #163 do UniGene e 74.561 seqüências na construção #198 do UniGene, representando aproximadamente 20.000 genes).

3.1.8. Banco de etiquetas “virtuais”.

No sentido de anotar os dados obtidos experimentalmente por SAGE e MPSS, em um segundo momento, são feitas as associações entre as etiquetas experimentais e as etiquetas virtuais. Resumidamente, este processo consiste em mimetizar, *in silico*, o experimento de SAGE e de MPSS feito *in vitro*: i) é extraída a etiqueta adjacente ao sítio de restrição (CATG para NlaIII e GATC para DpnII) mais 3' de todos os cDNAs conhecidos; ii) estas etiquetas (aqui chamadas de etiquetas virtuais) são armazenadas em um banco MySQL de modo que para cada cDNA (gene) temos etiquetas virtuais; iii) as etiquetas virtuais são comparadas com as etiquetas experimentais e, quando iguais, é associada a etiqueta ao gene correspondente.

3.2. Quantificando a expressão de um gene.

Para quantificar a expressão de um gene por SAGE ou MPSS usamos a associação descrita acima (etiqueta-gene) e o dado de freqüência das etiquetas obtidas experimentalmente. Gene mais expressos apresentam uma freqüência maior de etiquetas. Os dados experimentais são agrupados por bibliotecas, onde cada biblioteca representa um único experimento e, no geral, um único tecido. Utilizando estes dados podemos quantificar a expressão gênica em diversos tecidos humanos. Por exemplo, o transcrito da figura 3.8.2 apresenta a etiqueta virtual ATGAAACCCC (ver região em destaque de 10 bases após o sítio

mais 3' de NlaIII, a sequência CATG). Buscando essa etiqueta no banco experimental obteremos um resultado apresentado na tabela 3.6.1, aonde podemos saber a frequência das etiquetas e inclusive informações patológicas (como exemplo se o tecido é tumoral ou normal). Para que a comparação da expressão entre as diferentes bibliotecas seja factível, é necessário uma normalização. Nossos dados foram normalizados pela seguinte fórmula:

$$(\text{Freq} / \text{ttTAGs}) * 200.000$$

Onde Freq é a frequência de ocorrência de uma determinada etiqueta em uma biblioteca e ttTAGs é o total de etiquetas seqüenciadas nesta mesma biblioteca.

Tabela 3.6.1 - Banco de etiquetas experimentais de *Short* SAGE.

Biblioteca	Nome descritivo da biblioteca	Etiqueta	Frequência absoluta	Frequência normalizada
22	Prostate_adenocarcinoma_CL_LNCAp	ATGAAACCCC	7	62,656
23	Ovary_adenocarcinoma_B_OVT-6	ATGAAACCCC	11	53,085
25	Ovary_adenocarcinoma_B_OVT-7	ATGAAACCCC	20	74,746
29	Vascular_normal_CS_control	ATGAAACCCC	20	77,456
30	Vascular_normal_CS_VEGF+	ATGAAACCCC	22	76,767
31	Breast_normal_epithelium_AP_1	ATGAAACCCC	15	61,565
39	Ovary_carcinoma_CL_A2780	ATGAAACCCC	7	65,515
40	Ovary_cystadenoma_CL_ML10-10	ATGAAACCCC	17	61,656
41	Brain_glioblastoma_control_CL_H247	ATGAAACCCC	15	49,645
42	Brain_glioblastoma_hypoxia_CL_H247	ATGAAACCCC	16	44,590
43	Brain_oligodendroglioma_B_H988	ATGAAACCCC	11	78,954
50	Brain_astrocytoma_grade_I_B_H1043	ATGAAACCCC	44	115,908
51	Brain_medulloblastoma_B_1273	ATGAAACCCC	9	46,615
53	Brain_normal_thalamus_B_1	ATGAAACCCC	2	16,656
54	Brain_astrocytoma_grade_III_B_H1020	ATGAAACCCC	34	131,852
55	Brain_normal_cerebellum_B_BB542	ATGAAACCCC	4	19,753
56	Kidney_embryonic_CL_293-control	ATGAAACCCC	3	14,301
57	Prostate_carcinoma_CL_LNCAp	ATGAAACCCC	8	26,556

3.3. Removendo as Etiquetas Ambíguas.

Com o banco de etiquetas virtuais temos como identificar genes diferentes que compartilham a mesma etiqueta (etiqueta ambígua). Para estes genes não temos com precisar qual foi a expressão destes genes, pois não sabemos se a frequência encontrada no banco experimental é uma somatória da expressão destes genes, ou a expressão de um e ausência do outro. Por este motivo, todas as etiquetas ambíguas foram excluídas das nossas análises.

3.4. Identificando SNPs que interferem na técnica de SAGE e MPSS.

Como mostramos em nosso trabalho Silva, De Souza *et al.*, (2004), os SNPs podem afetar as etiquetas produzidas por SAGE e MPSS de três maneiras: (1) o SNP pode gerar um novo sítio de restrição *downstream* ao sítio originalmente considerado como mais 3' (Figura 3.8.1-A). (2) o SNP pode ocorrer no sítio de restrição da enzima de restrição, destruindo-o e 'criando' um novo sítio de restrição mais 3', o sítio imediatamente *upstream* ao sítio mais 3' original (Figura 3.8.1-B)(Obs. A 'destruição' do sítio original pode ocorrer devido a outros fatores como *splicing* alternativo, por essa razão resolvemos retirar estas etiquetas de nossas análises de expressão alélica diferencial). (3) o SNP pode ocorrer na própria sequência da etiqueta (Figura 3.8.1-C). Nestes três casos, sempre temos uma etiqueta diferente da etiqueta que seria obtida quando o polimorfismo não ocorre. Chamaremos as etiquetas afetadas por SNP de etiquetas alternativas e as não afetadas de etiquetas originais.

Figura 3.8.1 – Formação de etiquetas alternativas decorrentes da presença de SNPs. Em A, o SNP gera um novo sítio de restrição entre a etiqueta original e final do transcrito. Em B, o SNP destrói o sítio de restrição original, 'criando' um novo sítio de restrição mais 3'. Em C, o SNP altera diretamente a sequência da etiqueta.

Para identificar SNPs com potencial para gerar etiquetas alternativas, primeiro cruzamos os dados de mapeamento dos SNPs nos cDNAs com os bancos de etiquetas virtuais. Como isso, identificamos SNPs que ocorrem entre o último sítio CATG (SAGE) ou GATC (MPSS) e o final do transcrito (Figura 3.8.2). Em seguida testamos se o SNP está no sítio da enzima, na etiqueta ou se foi criado um novo sítio de NlaIII ou DpnII. Metodologicamente, a identificação do último tipo é a mais complexa, pois tivemos que: i) analisar 7 nucleotídeos, três upstream e três downstream ao SNP; ii) buscar novos sítios de NlaIII e DpnII nestas seqüências de 7 nucleotídeos; iii) recortar as etiquetas adjacente aos ‘novos’ sítios. Por exemplo, na Figura 3.8.2-B o SNP rs1142895 ‘criou’ um novo sítio para a enzima de restrição NlaIII, gerando a etiqueta alternativa TCTACTAAAA.

FIGURA 3.8.2 – Busca de SNPs com potencial de gerar etiquetas alternativas. Em A seqüência fasta do transcrito BC008600 com os sítios da enzima NlaIII (em vermelho), com os SNPs (destacado em verde) e com a etiqueta original de SAGE (10 nt em negrito e sublinhado). Três bases *upstream* e três bases *downstream* ao SNP e os dos alelos esperados para cada SNP. No destaque em verde, janelas de 4 bases ao redor dos SNP usadas na busca de novos sítios CATG potencialmente gerados pelos SNPs.

Tabela 3.8.1 – SNPs identificados no transcrito BC008600.

SNP	cDNA	Posição no cDNA	Alelos
rs35359791	BC008600	162	C/T
rs11554344	BC008600	485	C/G
rs3817656	BC008600	519	C/T
rs1142895	BC008600	568	C/G
rs1142896	BC008600	570	C/G
rs9915517	BC008600	732	T/C

3.5. Banco de etiquetas alternativas de MPSS e SAGE.

Após termos feito um varredura em todas as seqüências de cDNAs e identificarmos os SNPs com potencial para gerar etiquetas alternativas, criamos um banco de dados disponível no SAGE *Genie* (<http://cgap.nci.nih.gov/SAGE>). Uma versão mais completa incluindo as etiquetas alternativas de MPSS está disponível em (<http://www.compbio.ludwig.org.br/~jorge/monoallelic/>). A tabela 3.9.1 é um pequeno exemplo deste banco. Nele podemos obter de maneira fácil a seqüência da etiqueta original, a etiqueta alternativa e o SNP associado, assim como as suas posições no transcrito.

Tabela 3.9.1 – Banco de etiquetas Alternativas associadas com SNPs.

cDNA	Posição do Sítio original	Etiqueta Original	Posição do Sítio alternativo	Etiqueta Alternativa	Posição do SNP	SNP	Alelos
AB032929	1532	GGCCAGCAAGTCCTGGA	1547	GATAATTTCTTTTCGGTC	1548	rs3219496	A/C
AB036429	2807	AAGATCCAGCTGCTCTG	3264	GCTGAGACAGGAAGTGC	3265	rs6843860	G/A
AB040450	954	ACTGTTGGAATTGCTCT	1029	TCGGCACCTAGTAATGG	1032	rs7092831	C/G
AB047004	3701	CTGAACCTCCCCAACAA	3881	GTCCTGTAGGACGGCA	3882	rs34418000	G/A
AB049211	1623	ATCTGAGGCCAGCTCCC	1816	GACACTAGCTGCCCCAG	1818	rs28382696	C/T
AB056722	2052	AACAGCAAGGAGTGTTT	2518	GAAAAGTATCTGTAATT	2519	rs10468616	A/G

3.6. Bibliotecas experimentais provenientes de apenas um único indivíduo.

Para evitar falsos negativos, tivemos o cuidado de selecionar manualmente apenas bibliotecas feitas com tecido de um único indivíduo. Essa seleção foi baseada na descrição da biblioteca encontrada no banco público. Quando havia dúvidas sobre a origem da biblioteca a

mesma era classificada como proveniente de múltiplos indivíduos. Além disso, bibliotecas geradas a partir de tecido de origem embrionária também foram excluídas, pois este tecido pode apresentar expressão diferente dos tecidos adultos. Como as bibliotecas de MPSS eram construídas com amostras provenientes de mais de um indivíduo, excluímos os dados MPSS de nossas análises de expressão monoalélica. Aplicando estes filtros, das 305 bibliotecas de *short* SAGE ficamos com 233 bibliotecas derivadas de 20 tecidos humanos, contendo 646.077 etiquetas únicas; das 66 bibliotecas de *long* SAGE ficamos com 30 bibliotecas de seis tecidos humanos contendo 443.735 etiquetas únicas.

3.7. Etiquetas alternativas na identificação de genes com expressão monoalélica.

Após a identificação do conjunto de etiquetas afetadas por SNPs e seus genes correspondentes, partimos para a comparação com os dados experimentais. Pela teoria, levando em consideração que neste trabalho cada biblioteca experimental representa um indivíduo, se a expressão monoalélica ocorrer, encontraremos frequência de um determinado gene em uma determinada biblioteca experimental para apenas uma das etiquetas. Além disso, este padrão deve-se manter na maioria das demais bibliotecas que apresentam a expressão destes genes. Neste sentido separamos os genes que possuem etiquetas alternativas em quatro categorias diferentes como descritas a seguir:

Na primeira categoria, chamada de expressão monoalélica (EM), estão os genes que apresentam apenas uma das etiquetas, original ou alternativa, mas ambas estão presentes não concomitantemente no banco experimental. Na tabela 3.11.1 podemos ver um gene que tem etiquetas alternativas tanto para *short* SAGE quanto para *long* SAGE em 20 bibliotecas, 3 *long* SAGE e 17 *short* SAGE. Vale ressaltar que em nenhuma biblioteca vemos a presença das duas etiquetas original e alternativa simultaneamente.

Tabela 3.11.1 – Exemplo de Expressão Monoalélica.

Gene:	Phosphatidylinositol-4-phosphate 5-kinase, type II, alpha				
Seqüência:	NM_005028.3				
Etiqueta Original:	(Long) AACATTGGAGGGACAGA	(Short) AACATTGGAG			
Etiqueta Alternativa:	(Long) AACATTGGATGGACAGA	(Short) AACATTGGAT			
SNP:	rs1053454	Alelo1: G (0,636)	Alelo 2: T (0,363)		
Tipo	Biblioteca	tumor	Etiqueta Original	Etiqueta Alternativa	Tecido
Long Sage	654	S	2,53	0,00	Colon
Long Sage	647	N	2,90	0,00	mammary gland
Long Sage	1568	N	0,00	3,87	white blood cells
Short Sage	431	S	3,94	0,00	mammary gland
Short Sage	40	S	0,00	3,63	ovary
Short Sage	41	S	0,00	3,31	brain
Short Sage	2	S	3,30	0,00	colon
Short Sage	8	N	2,12	0,00	brain
Short Sage	384	S	1,68	0,00	brain
Short Sage	609	S	5,30	0,00	cartilage
Short Sage	79	S	0,00	3,29	mammary gland
Short Sage	1575	N	0,00	3,87	white blood cells
Short Sage	1576	N	0,00	3,72	white blood cells
Short Sage	610	S	2,26	0,00	cartilage
Short Sage	1368	S	2,94	0,00	cartilage
Short Sage	1445	S	3,40	0,00	stomach
Short Sage	99	N	0,00	2,25	lung
Short Sage	420	N	2,90	0,00	mammary gland
Short Sage	351	S	0,00	1,86	brain
Short Sage	428	S	2,51	0,00	colon

Na segunda categoria, chamada de não informativa (NI), estão os genes que apresentam apenas uma das etiquetas, original ou alternativa, presentes no banco experimental. Uma etiqueta não estar presente em nenhuma das bibliotecas sugere a ausência do SNPs em todas as amostras utilizadas para a construção das bibliotecas ou que há um erro metodológico com estes SNPs. Portanto, para evitar qualquer problema, excluímos estes genes de nossas análises. Na tabela 3.11.2 podemos ver que a etiqueta alternativa não está presente em nenhuma das 8 bibliotecas de short SAGE.

Tabela 3.11.2 – Casos não informativos.

Gene:		POU domain, class 5, transcription factor 1			
Seqüência:		AY484516.1			
Etiqueta Original:		(Short) CCAGCCGCCA			
Etiqueta Alternativa:		(Short) CCAGCCGCCT			
SNP:		rs2269711	Alelo1: A (0,621)	Alelo 2: T (0,377)	
Tipo	Biblioteca	tumor	Etiqueta Original	Etiqueta Alternativa	Tecido
Short Sage	585	N	4,74	0,00	Lung
Short Sage	154	S	4,91	0,00	Prostate
Short Sage	157	S	2,24	0,00	mammary gland
Short Sage	669	N	11,57	0,00	mammary gland
Short Sage	185	S	2,64	0,00	vascular
Short Sage	1366	S	2,02	0,00	lung
Short Sage	1443	S	2,13	0,00	stomach
Short Sage	97	S	6,16	0,00	stomach

Na terceira categoria, chamada de expressão bialélica (EB), estão os genes que ambas as etiquetas, original e alternativa, estão presentes concomitantemente em pelo menos uma biblioteca experimental. Na tabela 3.11.3 podemos ver um exemplo. Das 23 bibliotecas de *long* SAGE, 6 (destacadas de vermelho) apresentam ambas as etiquetas.

Tabela 3.11.3 – Exemplo de Expressão Bialélica.

Gene:		TAF11 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 28kDa			
Seqüência:		AF118094.1			
Etiqueta Original:		(Long) AAGGATGCGGTAATGGC			
Etiqueta Alternativa:		(Long) AAGGATGCGGTGATGGC			
SNP:		rs2985	Alelo1: A (0,870)	Alelo 2: G (0,129)	
Tipo	Biblioteca	tumor	Etiqueta Original	Etiqueta Alternativa	Tecido
Long Sage	723	S	0,00	10,77	mammary gland
Long Sage	651	S	0,00	2,53	colon
Long Sage	652	S	2,52	0,00	colon
Long Sage	644	S	3,07	0,00	mammary gland

Long Sage	653	S	2,48	0,00	colon
Long Sage	654	S	7,58	0,00	colon
Long Sage	655	N	5,14	5,14	vascular
Long Sage	1563	N	11,02	0,00	white blood cells
Long Sage	673	S	3,00	0,00	mammary gland
Long Sage	657	S	15,77	7,88	mammary gland
Long Sage	1645	S	7,10	0,00	brain
Long Sage	1564	N	3,93	3,93	white blood cells
Long Sage	1565	N	7,90	0,00	white blood cells
Long Sage	683	S	6,71	0,00	mammary gland
Long Sage	675	S	12,13	0,00	mammary gland
Long Sage	659	S	17,44	0,00	white blood cells
Long Sage	1566	N	2,00	2,00	white blood cells
Long Sage	676	S	6,29	0,00	mammary gland
Long Sage	1567	N	4,04	8,09	white blood cells
Long Sage	1568	N	11,62	7,74	white blood cells
Long Sage	1569	N	7,43	0,00	white blood cells
Long Sage	963	S	0,00	5,51	lung
Long Sage	703	S	8,84	0,00	mammary gland

A quarta categoria é um subconjunto da terceira e é chamada de perda de expressão monoalélica em tumor (PEMT). Nesta categoria estão os genes em que ambas etiquetas (original e alternativa), assim como na terceira categoria, estão presentes no banco experimental e que em pelo menos uma biblioteca temos a expressão de ambas etiquetas, no entanto, as bibliotecas que apresentam essa expressão bialélica são provenientes de tecido tumoral. Em outras palavras, se consideramos apenas bibliotecas provenientes de tecido normal estes genes seriam classificados na primeira categoria (expressão monoalélica). Na tabela 3.11.4 podemos ver um gene que tem etiqueta alternativa tanto para *long* SAGE quanto para *short* SAGE e das 24 bibliotecas onde elas estão presentes 3 delas (destacadas de vermelho) são provenientes de tumor e apresentam frequência de expressão para ambas etiquetas (original e alternativa).

Tabela 3.11.4 – Exemplo de Perda de Expressão Monoalélica em Tumor.

Gene:		Protein S (alpha)			
Seqüência:		BC015801.1			
Etiqueta Original:		(Long) TCCATCAGTTTGGAAAA		(Short) TCCATCAGTT	
Etiqueta Alternativa:		(Long) TCCGTCAGTTTGGAAAA		(Short) TCCGTCAGTT	
SNP:		rs6123	Alelo1: A (0,657)	Alelo 2: G (0,342)	
Tipo	Biblioteca	tumor	Etiqueta Original	Etiqueta Alternativa	Tecido
Long Sage	644	S	3,07	0,00	mammary gland
Long Sage	647	N	2,89	0,00	mammary gland
Long Sage	676	S	3,14	0,00	mammary gland
Short Sage	358	S	3,90	0,00	brain
Short Sage	359	S	4,03	4,03	brain
Short Sage	3	S	3,57	0,00	brain
Short Sage	155	S	3,02	0,00	white blood cells
Short Sage	156	S	2,45	0,00	mammary gland
Short Sage	158	S	0,00	15,38	mammary gland
Short Sage	303	S	3,00	0,00	cerebellum
Short Sage	523	N	2,75	0,00	brain
Short Sage	525	S	0,00	3,65	brain
Short Sage	526	S	2,88	2,88	brain
Short Sage	385	S	1,85	1,85	brain
Short Sage	610	S	4,52	0,00	cartilage
Short Sage	183	S	0,00	2,52	mammary gland
Short Sage	1366	S	2,02	0,00	lung
Short Sage	416	S	3,07	0,00	mammary gland
Short Sage	344	S	1,94	0,00	brain
Short Sage	1169	S	3,14	0,00	mammary gland
Short Sage	565	S	4,28	0,00	brain
Short Sage	135	N	0,00	3,01	liver
Short Sage	420	N	2,89	0,00	mammary gland
Short Sage	352	S	0,00	2,24	brain

No intuito de tornar nossas análises mais confiáveis, para todas as categorias foram aplicados os seguintes filtros:

- 1) Ambas as etiquetas, original e alternativa, devem estar presentes no banco experimental (com exceção da categoria 2 – “não informativa”);

- 2) As bibliotecas experimentais devem ser provenientes de um único indivíduo.
- 3) As etiquetas virtuais devem ser provenientes de seqüências de cDNAs com uma região 3' completa (identificada pela cauda de poli-A).
- 4) As etiquetas virtuais, tanto as originais quanto as alternativas, devem ser etiquetas não ambíguas.
- 5) Se, para um mesmo gene, dois ou mais SNPs geram etiquetas alternativas diferentes, este gene será classificado como expressão monoalélica se todas as suas etiquetas, originais e alternativas, puderem ser classificadas como tal.

3.8. Validação da nossa metodologia pela Literatura.

Para termos uma idéia do potencial de nossa metodologia, fomos verificar quantos dos genes categorizados em nossas análises já foram identificados por outros estudos como apresentando expressão alélica diferencial ou gene *imprinted*. Para isto, buscamos na literatura os principais trabalhos que identificaram expressão monoalélica em larga escala. Veja abaixo: uma breve descrição destes trabalhos:

➔ O primeiro é *The database of imprinted genes and parent-of-origin effects in animals* (<http://www.otago.ac.nz/IGC>) (Morison, Paton *et al.*, 2001), que atualmente descreve 55 genes humanos categorizado como genes *imprinted* e mais 12 genes humanos que apesar de terem indícios de expressão monoalélica necessitam de estudos adicionais para que de fato sejam categorizados como genes *imprinted*.

➔ O segundo trabalho que contribuiu para nossa base de dados de foi o de Lo, Wang *et al.*, (2003), que fez uma análise em larga escala construindo um *microarray* baseado em SNPs para distinguir as expressões alelo-específica dos genes analisados. Neste trabalho foram identificados 170 genes que apresentaram diferença de expressão alélica superior a quatro vezes em pelo menos uma das amostras utilizadas no trabalho.

➔ O terceiro trabalho que contribuiu com nossa base de dados foi o de Pant, Tao *et al.*, (2006), assim como o trabalho de Lo, Wang *et al.*, (2003). Neste trabalho foi construído um *microarray* baseado em SNPs para distinguir as expressões alelo-específicas dos genes. Neste trabalho foram identificados 731 com expressão diferencial superior a duas vezes.

➔ O quarto trabalho que contribuiu para nossa base de dados foi o de Gimelbrant, Hutchinson *et al.*, (2007). Esse trabalho, assim como os dois anteriores, é baseado na construção de um *microarray* que utiliza SNPs para distinguir as expressões alelo-específicas dos genes. Porém este trabalho foi focado na identificação de genes com padrão de expressão monoalélica e, de um conjunto inicial de aproximadamente 4.000 genes, identificou 330 com padrão de expressão monoalélica.

Após unirmos esses trabalhos conseguimos um catálogo de 1.186 genes humanos com indícios de expressão monoalélica. Este catálogo foi utilizado para quantificarmos a porcentagem de genes de nossas análises que já haviam sido identificados por outros estudos como expressão alélica diferencial. (disponível em: www.compbio.ludwig.org.br/~jorge/monoallelic/).

3.9. Candidatos para validação experimental de expressão alélica diferencial.

A partir dos dados gerados em material e métodos 3.7, escolhemos arbitrariamente 31 genes (ver tabela 3.13) no intuito de comprovar experimentalmente o padrão de expressão monoalélica destes genes. Para realizar a validação experimental foi feita uma colaboração com a Dra. Anamaria Aranha Camargo do laboratório de Biologia Molecular do Instituto Ludwig de Pesquisa sobre o Câncer, ficando a cargo do aluno de Doutorado Daniel Onofre Vidal a coleta de amostras e a realização dos experimentos biológicos aqui mencionados.

Tabela 3.13 – Genes escolhidos para validação experimental.

SNP	Alelos	UniGene	Categoria	Gene
rs2304511	T/G	Hs.529488	EM	Solute carrier family 6 (neurotransmitter transporter, taurine), member 6
rs1065368	C/T	Hs.243678	EM	SRY (sex determining region Y)-box 8
rs470939	G/T	Hs.524812	EM	G protein-coupled receptor 109A
rs1053454	G/T	Hs.588901	EM	Phosphatidylinositol-4-phosphate 5-kinase, type II, alpha
rs2281656	G/A	Hs.596214	EM	Homo sapiens, clone IMAGE:5787583, mRNA
rs1057755	A/T	Hs.165950	EM	Fibroblast growth factor receptor 4
rs2272761	G/A	Hs.458644	EM	R3H domain and coiled-coil containing 1
rs1568918	A/G	Hs.272759	EM	Phosphatidylinositol transfer protein, membrane-associated 2
rs2587512	G/C	Hs.369819	EM	TBC1 domain family, member 16
rs17081950	C/T	Hs.89404	EM	Msh homeobox homolog 2 (Drosophila)
rs1060298	A/G	Hs.594773	EM	CDNA clone IMAGE:5259272
rs2071863	G/A	Hs.178695	PEMT	Mitogen-activated protein kinase 13
rs9132	A/G	Hs.48428	PEMT	5'-nucleotidase domain containing 3
rs3818499	C/G	Hs.508725	PEMT	Inhibitor of growth family, member 1
rs1543442	A/G	Hs.48029	PEMT	Snail homolog 1 (Drosophila)
rs17027704	A/G	Hs.514870	PEMT	ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit B1
rs1558525	A/G	Hs.305985	PEMT	Polyhomeotic-like 1 (Drosophila)
rs1065154	T/G	Hs.437277	PEMT	Sequestosome 1
rs6831	C/T	Hs.482491	PEMT	Mitochondrial ribosomal protein S27
rs12907665	G/A	Hs.292949	PEMT	INO80 complex homolog 1 (S. cerevisiae)
rs2518996	G/A	Hs.133183	PEMT	Hypothetical protein LOC643837
rs12856	T/C	Hs.713563	PEMT	CDNA clone IMAGE:3030163
rs4781	A/G	Hs.132342	PEMT	Lipin 2
rs3193677	C/A	Hs.107153	PEMT	Inhibitor of growth family, member 2
rs474058	A/C	Hs.632702	PEMT	Glioma-associated oncogene homolog 1 (zinc finger protein)
rs2233128	A/G	Hs.653138	PEMT	Lymphocyte antigen 86
rs2241838	A/G	Hs.434993	PEMT	Ras-associated protein Rap1
rs266805	A/G	Hs.515016	PEMT	Melanoma associated antigen (mutated) 1

3.10. Amostras de Sangue provenientes do banco de sangue do departamento de Hematologia e Hemoterapia do Hospital A. C. Camargo.

Para a realização deste projeto foram colhidas 30 amostras de sangue periférico de indivíduos normais provenientes do Banco de Sangue do departamento de Hematologia e Hemoterapia do Hospital A. C. Camargo. A coleta de sangue realizada pelo doutorando Daniel Onofre Vidal foi feita sob a orientação e coordenação da Dra. Mônica Cristovão Poli responsável banco de sangue.

3.11. Estratégia Experimental na validação de expressão monoalélica.

Para a validação experimental dos genes com evidência de expressão monoalélica, primeiramente foi realizada a genotipagem do DNA de todas as amostras normais para cada um dos SNPs associados com as etiquetas alternativas dos genes escolhidos (ver tabela 3.13). Foram realizadas reações de PCR (Sigla em inglês de Reação em Cadeia da Polimerase) para cada gene, utilizando iniciadores (*primers*) flanqueando a região onde está mapeada a etiqueta alternativa associada ao SNP (Figura 3.15.1 A).

Figura 3.15.1 – Genotipagem do SNP associado a etiqueta alternativa. Em A temos os iniciadores (em cinza) flanqueando a região onde está mapeada a etiqueta alternativa. A etiqueta alternativa está em Azul, o sítio de NLAIII em vermelho e o SNP em verde. Em B temos o resultado do PCR. Cada banda representa o fragmento de sequência descrito no item A em uma amostra de sangue.

Após conseguir através de PCR amplificar a região genômica ao redor dos SNPs das etiquetas alternativas (entre 250 e 300 bases) os produtos de PCR foram submetidos ao sequenciamento. Sequências de baixa qualidade foram submetidas ao sequenciamento novamente até conseguirmos uma boa qualidade para cada produto de PCR gerado. A qualidade das sequências foram medidas através de uma estratégia de bioinformática utilizando o programa Phred (Ewing e Green, 1998; Ewing, Hillier *et al.*, 1998), para extrair dos arquivos cromatogramas gerados pelo sequenciador um valor de qualidade para cada base das sequências. Para cada sequência foram avaliadas apenas as 51 bases ao redor do SNP (25 para cada lado mais a base do SNP). Sequências que apresentavam valor de Phred acima de 25 para pelo menos 70% das 51 bases foram consideradas de boa qualidade.

Após o sequenciamento, os fragmentos de DNA foram submetidos ao programa CHROMAS (<http://www.technelysium.com.au/>) para a identificação da região genômica onde estavam representadas as bases dos SNPs. Quando a base representando o SNP apresentou dois picos, a sequência de DNA foi categorizada como heterozigota (ver figura 3.15.2-B). Quando a base apresentou apenas um pico a sequência de DNA foi categorizada como homozigota (ver figura 3.15.2-A).

Para os genes candidatos foram identificadas as sequências de DNA heterozigotas nas 30 amostras de sangue colhidas (ver tabela 3.13). Os candidatos que apresentaram menos de três amostras de DNA heterozigotas foram excluídos das nossas análises.

Figura 3.15.2 – Análise visual dos cromatogramas das seqüências de DNA seqüenciadas. Em A temos o cromatograma para uma seqüência de DNA homozigota para o SNP estudado. Em B temos o cromatograma para uma seqüências de DNA heterozigota para o SNP estudado.

Após identificar genes com amostras de DNA heterozigotas para os SNPs associados com as etiquetas alternativas, partimos para a avaliação de seus respectivos cDNAs, da mesma forma que fizemos para as amostras de DNA e utilizando os mesmos iniciadores. As amostras de cDNA que com um pico no cromatograma (homozigose para os SNPs) sugerem genes com expressão monoalélica. As amostras de cDNAs com dois picos no cromatograma (heterozigose para os SNPs) sugerem genes com expressão bialélica. Surpreendentemente identificamos uma terceira classe em que era possível visualizar dois picos, porém a diferença nas alturas dos picos referentes aos alelos não era compatível com as diferenças visualizadas nas amostras de DNA, dando um indicativo que essa diferença entre os picos era causada não apenas pelo erro implícito na incorporação de bases durante a reação de seqüenciamento, mas também poderia ser devido a um desbalanço da quantidade de cDNAs produzidos ou melhor

expressos durante a reação de PCR. Para avaliar essas sutis diferenças alélicas criamos uma estratégia no intuito de tornar mais quantitativas as nossas análises.

3.12. Identificação de expressão alélica diferencial e bialélica nos dados experimentais.

Para medir as razões de heterozigose e homozigose (razões das alturas dos picos nos cromatogramas) dos SNPs que submetemos a validação experimental, elaboramos uma estratégia que se inicia com o programa “*PeakPicker*” (Ge, Gurd *et al.*, 2005). O programa *PeakPicker* (<http://genomequebec.mcgill.ca/ESTHapMap>) foi desenvolvido para analisar de forma quantitativa a razão entre alelos de um polimorfismo. Como dado inicial, ele utiliza os cromatogramas gerados durante o seqüenciamento. O *PeakPicker* pode ser usado para determinar a expressão alélica diferencial em células heterozigóticas através do cálculo e comparação dos picos referentes aos alelos de um SNP. Uma vantagem deste programa é a facilidade na análise de diversos cromatogramas seqüenciados de amostras de DNA e cDNA ao mesmo tempo. As alturas dos picos nos cromatogramas variam e dependem da amostra, tipo básico, e da posição delas dentro do cromatograma. Para tornar as alturas dos picos nos diversos cromatogramas comparáveis, o programa executa um passo de normalização levando em conta a qualidade das seqüências e, como referência, os picos das bases adjacentes ao SNP estudado (picos de referência). No final do processo o programa gera um arquivo com as razões normalizadas dos picos referentes ao SNP estudado (ver Figura 3.16). Chamando os alelos de um polimorfismo arbitrariamente de alelo1 e alelo2, quando temos uma razão, alelo1/alelo2, maior que 1, significa que o alelo1 do SNP tem um pico maior no cromatograma.

#seq	sequence	A/G(N)	A/G(O)	G/A(N)	G/A(O)
1	DNA_cand-8.1.ab1	0,721	0,741	1,386	1,350
2	DNA_cand-8.3-2.ab1	0,656	0,705	1,524	1,419
3	DNA_cand-8.4.ab1	0,815	0,738	1,228	1,355
4	DNA_cand-8.8.ab1	0,735	0,836	1,361	1,196
5	DNA_cand-8.13.ab1	0,829	0,971	1,206	1,030
6	DNA_cand-8.16.ab1	0,803	0,909	1,246	1,100
7	DNA_cand-8.17.ab1	0,694	0,806	1,440	1,241
8	DNA_cand-8.25.ab1	0,768	0,895	1,301	1,118
9	DNA_cand-8.32.ab1	0,718	0,826	1,392	1,211
10	cDNA_cand-8.1.ab1	1,191	1,221	0,840	0,819
11	cDNA_cand-8.3.ab1	0,554	0,556	1,806	1,799
12	cDNA_cand-8.4.ab1	1,008	1,053	0,992	0,949
13	cDNA_cand-8.8.ab1	0,811	0,883	1,234	1,133
14	cDNA_cand-8.13.ab1	1,328	1,212	0,753	0,825
15	cDNA_cand-8.16.ab1	0,968	0,902	1,033	1,109
16	cDNA_cand-8.17.ab1	0,782	0,836	1,278	1,196
17	cDNA_cand-8.25.ab1	0,495	0,297	2,020	3,372
18	cDNA_cand-8.32.ab1	0,976	0,691	1,025	1,448

Figura 3.16 – Exemplo de um resultado do programa *PeakPicker*. Na primeira coluna temos um identificador numérico para a sequência analisada. Na segunda coluna temos o nome da sequência. Na terceira coluna temos as razões normalizadas entre os alelos (alelo1/alelo2) do SNP levando em consideração os picos de referência. Na quarta coluna temos as razões dos alelos (alelo1/alelo2) sem a normalização. Na quinta coluna temos as razões normalizadas entre os alelos (alelo2/alelo1) levando em consideração os picos de referência. Na sexta coluna temos as razões dos alelos (alelo2/alelo1) sem normalização.

Nossa estratégia consiste em submeter ao programa *PeakPicker* todos os cromatogramas das amostras de DNA e cDNA de um determinado gene candidato e extrair, para cada cromatograma, a razão normalizada dos alelos do SNP associado à etiqueta alternativa (ver terceira coluna da Figura 3.16). Para deixar todas as razões na escala entre 0 e 1, quando uma razão é maior do que 1, aplicamos a fórmula: $1 / \text{razão}$. Um problema na quantificação executada pelo *PeakPicker* é que a razão dos picos das bases de uma amostra de DNA heterozigoto deveria ser 1 e por causa da diferença da fluorescência provocado pela erro

de incorporação de bases implícito no processo de sequenciamento isso não ocorre. Para tirar o erro de incorporação foi aplicado nas razões das amostras de DNA a seguinte fórmula:

$$RT_{dna} = 1 - (ABS(R_{dna} - M_{dna})).$$

Onde R_{dna} é a razão entre os alelos de um SNP de um gene candidato para uma determinada amostra de DNA heterozigoto. M_{dna} é a média das R_{dna} de todas as amostras de DNAs heterozigotos de um gene. ABS é uma função que retorna o valor absoluto de um número. RT_{dna} é a razão transformada entre os alelos de um SNP de um gene candidato para uma determinada amostra de DNA heterozigoto.

Diferente do DNA, a divergência encontrada entre os picos dos alelos dos cDNAs não se dão apenas pelo erro de incorporação de bases, mas também pela diferença de expressão entre os alelos. Além disso, para corrigir qualquer erro de incorporação de bases no cDNAs, assumindo que este erro é o mesmo entre o cDNA e o DNA e aplicamos a seguinte fórmula:

$$RT_{cdna} = 1 - (ABS(R_{cdna} - M_{dna}))$$

Onde R_{cdna} é a razão entre os alelos de um SNP de um gene candidato para uma determinada amostra de cDNA, M_{dna} é a média das R_{dna} de todas as amostras de DNA heterozigoto do mesmo gene estudado. ABS é uma função que retorna o valor absoluto de um número. RT_{cdna} é a razão transformada entre os alelos de um SNP de um gene candidato para uma determinada amostra de cDNA.

Para determinar se uma amostra de cDNA apresentava expressão alélica diferencial ou monoalélica, foi construído um intervalo de confiança baseado nas RT_{dna} de todos os genes estudados. O intervalo de confiança (CI) foi calculado de acordo com a seguinte fórmula:

$$CI = (AVG_{dna} - (3 * STD_{dna}); AVG_{dna} + (3 * STD_{dna})),$$

Onde AVG_{dna} é a média dos RT_{dna} das amostras de DNA genômico. STD_{dna} é o desvio padrão dos RT_{dna} das amostras de DNA genômico. O CI foi calculado assumindo que os RT_{dna} para cada amostra de DNA genômico seguem uma distribuição Normal (foi usado o teste de Anderson-Darling para confirmar a Normalidade das amostras). Os RT_{cdna} das amostras de cDNA localizados fora do CI foram considerados candidatos a “expressão alélica diferencial”; os RT_{cdna} das amostras de cDNA localizados dentro do CI foram considerados como tendo “expressão bialélica”.

4. Resultados e Discussão.

4.1. O impacto dos SNPs na interpretação dos dados de SAGE e de MPSS.

Etiquetas alternativas podem ser geradas quando há sobreposição dos SNPs com a região do transcrito que corresponde a uma etiqueta ou quando gera um novo sítio mais 3' do sítio usual ou com a região do transcrito que corresponde aos sítios de restrição das enzimas usadas na construção de bibliotecas de SAGE e MPSS. Neste último caso, a mutação do sítio de restrição leva à clivagem de um sítio mais distante da cauda poli-A, levando ao sequenciamento de uma nova etiqueta para o mesmo gene. Para analisar o impacto de SNPs na geração de etiquetas alternativas, desenvolvemos uma estratégia computacional baseada nas sobreposições de etiquetas e SNPs em todos os genes humanos conhecidos.

Para a construção do nosso banco de dados de etiquetas alternativas utilizamos 54.645 seqüências de cDNAs com cauda de poli-A, representando 20.300 genes conhecidos (agrupamentos feito pelo UniGene; versão #163). Em seguida, identificamos os sítios de restrição das enzimas NlaIII (SAGE) e DpnII (MPSS). Das 54.645 seqüências de cDNAs analisadas, 54.124 (99,0%) continham pelo menos um sítio de restrição da NlaIII e 52.779 (96,6%) continham pelo menos um sítio de restrição da DpnII.

Em seguida, identificamos as seqüências de cDNAs que se sobrepunham a SNPs de acordo com o dbSNP (versão #118) e o alinhamento do cDNA contra o genoma. Das 54.124 seqüências de cDNAs com sítio de restrição NlaIII, 44.033 (81,4%) continham ao menos um SNP. Para DpnII, 43.125 das 52.779 seqüências de cDNAs (81,7%) continham ao menos um SNP.

4.1.1. Identificação de sítios alternativos de restrição.

A identificação de etiquetas alternativas foi realizada em separado para cada uma das três categorias definidas na Figura 3.8.1 (ver também materiais e métodos para mais detalhes). Primeiro, nós identificamos seqüências de cDNAs na qual a presença de um SNP criou um sítio de restrição mais próximo da cauda poli-A. Este sítio está localizado entre o sítio da etiqueta inferido com base no genoma de referência (etiqueta original) e o final do transcrito (Figura 3.8.1-A). Dentre as 44.033 seqüências de cDNA que contêm ao menos um sítio de NlaIII e um SNP, identificamos 573 cDNAs (1,3%) nos quais a presença de SNPs criou um novo sítio de restrição NlaIII após a etiqueta original. Estas 573 seqüências de cDNA correspondem a 294 genes humanos e 305 etiquetas alternativas de SAGE (Tabela 4.1.1). A análise semelhante feita para as 43.125 seqüências de cDNAs com pelo menos um sítio de DpnII e um SNP identificou 393 (0,9%) seqüências de cDNA que correspondem a 205 genes humanos e 217 etiquetas alternativas (Tabela 4.1.1). É importante notar que nesta categoria foram contadas também 56 seqüências de cDNA que correspondem a genes que não apresentam um sítio de restrição NlaIII ou DpnII, mas que adquirem um sítio devido à presença dos SNPs.

TABELA 4.1.1. Etiquetas alternativas produzidas por SNPs que geram um novo sítio de restrição mais próximo da cauda poli-A do transcrito (Figura 3.8.1-A).

	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas
Análise de SAGE	305	275 (90,2%)	38 (12,8%)
Análise de MPSS	217	40 (18,4%)	7 (3,2%)
Total	522	315 (60,3%)	45 (8,6%)

Para validar a existência destas etiquetas alternativas, verificamos sua presença nos bancos de dados de resultados de experimentos de SAGE e MPSS. Estes bancos contêm 586.144 etiquetas únicas de SAGE derivadas de 260 bibliotecas e 84.555 etiquetas únicas de MPSS derivadas de seis bibliotecas. Das 305 etiquetas alternativas de SAGE catalogadas em nosso banco de dados, 275 (90,2%) foram encontradas no banco de dados de SAGE, e das 217 etiquetas alternativas de MPSS, 40 (18,4%) foram encontradas no banco de dados de MPSS (Tabela 4.1.1).

Porém, a presença de uma etiqueta alternativa dentro de um conjunto de etiquetas obtidas experimentalmente não é uma evidência irrefutável para existência da etiqueta alternativa, pois também podem acontecer casos de ambigüidade de etiquetas alternativas (quando um gene passa a ser representado, na presença do SNP, pela mesma etiqueta de outro gene). Para remover as etiquetas alternativas ambíguas, determinamos a porcentagem das etiquetas alternativas que também correspondem à etiqueta original de um outro gene humano. Uma porcentagem pequena (12,8%) das 305 etiquetas alternativas de SAGE correspondeu à etiqueta original de outro gene e esta porcentagem foi ainda menor (3,2%) para as etiquetas alternativas de MPSS, devido ao seu tamanho mais longo (Tabela 4.1.1). Todas as etiquetas alternativas ambíguas identificadas acima foram removidas das análises posteriores. A grande proporção de etiquetas não ambíguas nos bancos de dados de SAGE e MPSS nos permite afirmar que os SNPs constituem, ao lado de mecanismos como *splicing* alternativo e poliadenilação alternativa, uma fonte adicional para a geração de etiquetas alternativas.

4.1.2. Perda do sítio de restrição original

Analizamos as seqüências de cDNA onde existe um SNP em uma das 4 bases do sítio de restrição associado com a etiqueta original. Neste caso, a perda do sítio de restrição

original implica a obtenção de uma etiqueta a partir do uso de um sítio de restrição mais distante da cauda poli-A (Figura 3.8.1-B). Nas 44.033 seqüências de cDNA que contêm ao menos um sítio de *Nla*III e um SNP, identificamos 498 seqüências (1,1%) com um SNP destruindo o sítio de restrição original. Estas 498 seqüências de cDNA correspondem a 236 genes e um total de 235 etiquetas alternativas de SAGE (Tabela 4.1.2). Das 235 etiquetas alternativas de SAGE, 218 (92,8%) estavam presentes no banco de etiquetas experimentais de SAGE, e apenas uma pequena fração (13,2%) destas etiquetas correspondia à etiqueta original de outro gene, o que as valida como etiquetas alternativas genuínas (Tabela 4.1.2).

Para 43.125 seqüências de cDNAs que contêm ao menos um sítio de *Dpn*II e ao menos um SNP, identificamos 422 seqüências (1%) nas quais o sítio original dos transcritos foi desfeito por um SNP. Estas 422 seqüências de cDNA correspondem a 208 genes com 196 etiquetas alternativas de MPSS. Destas, 78 (39,8%) foram encontradas dentro do banco de dados de MPSS, com apenas 40 etiquetas alternativas ambíguas, o que corresponde a uma frequência bastante baixa (4,6%) (Tabela 4.1.2).

TABELA 4.1.2. Etiquetas alternativas devido à presença de SNPs no sítio de restrição original.

	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas
Análise de SAGE	235	218 (92,8%)	31 (13,2%)
Análise de MPSS	196	78 (39,8%)	9 (4,6%)
Total	431	296 (68,7%)	40 (9,3%)

4.1.3. Novas etiquetas produzidas pela substituição de base nas etiquetas originais.

Por último, consideramos as seqüências de cDNA nas quais o SNP não afeta os sítios de restrição, mas ocorre dentro da etiqueta, levando ao aparecimento de etiquetas alternativas

que diferem da original por uma única base (Figura 3.8.1-C). Identificamos 1.136 seqüências (2,6%) com ao menos um sítio de *Nla*III e um SNP na etiqueta original. Estas 1.136 seqüências de cDNAs correspondem a 543 genes humanos 560 etiquetas alternativa de SAGE (Tabela 4.1.3), das quais 512 (91,4%) estavam presentes no banco experimental de SAGE e 92 (16,4%) eram ambíguas (Tabela 4.1.3). Da mesma forma, identificamos 1.009 seqüências de cDNA (2,3%) com ao menos um sítio de *Dpn*II e um SNP na etiqueta. Estas 1.009 seqüências de cDNAs correspondem a 481 genes humanos representando 507 etiquetas alternativas, das quais 127 (25,0%) foram encontradas no banco experimental de MPSS e 33 (6,5%) foram classificadas como ambíguas (Tabela 4.1.3).

TABELA 4.1.3. Etiquetas alternativas geradas pela sobreposição de SNPs com etiquetas de SAGE e MPSS.

	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas
Análise de SAGE	560	512 (91,4%)	92 (16,4%)
Análise de MPSS	507	127 (25,0%)	33 (6,5%)
Total	1.067	639 (59,9%)	125 (11,7%)

4.1.4. Integração do banco de etiquetas alternativas com o SAGE Genie.

Para tornar acessível à comunidade científica os resultados produzidos neste estudo, disponibilizamos o banco de dados das etiquetas alternativas no SAGE *Genie* (<http://cgap.nci.nih.gov/SAGE>). Os dados podem ser acessados de duas maneiras: baixados diretamente como arquivo texto ou visualizados na interface *web* (seguir *link* correspondente ao ‘*Ludwig Transcript Viewer*’). Na Figura 4.1.4 temos um caso onde a ocorrência de um SNP gera um novo sítio de *Nla*III, produzindo uma etiqueta alternativa. Nesta figura, podemos ver a representação esquemática do transcrito, as quatro etiquetas virtuais de SAGE,

a posição do SNP no transcrito, o identificador do SNP e a substituição da base. Usando o *Ludwig Transcript Viewer*, o usuário do *SAGE Genie* pode verificar se uma determinada sequência de cDNA contém uma etiqueta alternativa ou se uma etiqueta de SAGE específica corresponde a uma etiqueta alternativa de um gene humano conhecido.

Figura 4.1.4. - Integração do banco de dados de etiquetas alternativas no *SAGE Genie*. O *Ludwig Transcript Viewer* mostra o transcrito que codifica para o gene MAF1 (BC018714) como uma linha azul. As caixas coloridas representam as quatro últimas etiquetas virtuais e os números abaixo delas informam a posição de cada etiqueta no transcrito. A etiqueta alternativa para o transcrito de MAF1 está representada na tabela indicada pela seta. Essa tabela contém a sequência da etiqueta, a posição dentro do transcrito, a frequência da etiqueta no banco de SAGE, o identificador do SNP no dbSNP, a substituição alélica e a posição do SNP dentro do transcrito. A tabela na parte inferior da figura informa a frequência absoluta para cada uma das quatro etiquetas virtuais no banco de SAGE e suas posições no transcrito.

4.1.5. Considerações sobre as etiquetas alternativas.

Os impactos da poliadenilação alternativa e do *splicing* alternativo na geração de etiquetas alternativas já foram estudados em nosso grupo (Galante, P.A.F., *et al.*, manuscrito em preparação). Porém, etiquetas alternativas também podem ser geradas pela presença de SNPs em seqüências de cDNAs e a identificação destas etiquetas é importante para quantificamos mais precisamente a expressão dos genes por SAGE e MPSS.

Neste trabalho nós mostramos que a presença de SNPs dentro de seqüências de cDNA humano é responsável pela geração de ao menos 2.020 etiquetas alternativas (1.100 etiquetas alternativas de SAGE e 920 etiquetas alternativas de MPSS) e que 8,6% de todos os genes humanos conhecidos apresentam ao menos uma etiqueta alternativa associada à presença de um SNP. Entretanto, vale ressaltar que estes números devem ser uma sub-estimativa, pois a versão do banco de dados de SNPs que usamos contém apenas 40% dos SNPs disponíveis na versão atual do dbSNP. Adicionalmente, o banco de SNPs tende a aumentar muito nos próximos anos devido às novas tecnologias de seqüenciamento (por exemplo: SOLEXA-Illumina, SOLiD-ABI e 454-Roche), que são rápidas e baratas e produzem quantidades sem precedentes de dados genômicos. É razoável, portanto, esperar um grande aumento no número de etiquetas alternativas devido à identificação de novos polimorfismos.

Por outro lado, devemos também estar atentos ao fato de que, dentre os SNPs incluídos em nossas análises, cerca de 54,5% não foram validados (ver estatísticas e critérios de validação no endereço <http://www.ncbi.nlm.nih.gov/projects/SNP/>). Como os dados não validados podem corresponder a polimorfismos que não existem, as etiquetas alternativas previstas e identificadas com base nesses SNPs podem ser falso-positivos. Em nossas análises, tentamos reduzir o impacto desses falso-positivos desconsiderando os casos de expressão diferencial nos quais as etiquetas alternativas estavam ausentes dos bancos de dados de SAGE e MPSS (ver materiais e métodos). Adicionalmente, esperamos que um

aumento da proporção de SNPs validados nas próximas versões do dbSNP permita reduzir o número de etiquetas alternativas falso-positivas, minimizando este problema.

Todas as análises aqui apresentadas foram feitas com seqüências contendo uma região 3' completa (identificada pela presença da cauda poli-A). Porém, sabemos que alguns laboratórios removem a cauda poli-A dos transcritos antes de submetê-los aos bancos de dados públicos. É provável que entre as 75.503 seqüências de cDNA excluídas das nossas análises, que correspondem a 8.677 agrupamentos do UniGene, existam algumas com a região 3' completa. Se nós analisássemos todas as 130.148 seqüências de cDNAs catalogadas no UniGene, independentemente da presença de uma cauda poli-A, um total de 3.520 etiquetas alternativas poderiam ser identificadas (1.950 etiquetas alternativas de SAGE e 1.570 etiquetas alternativas de MPSS). Este número corresponde ao acréscimo de 1.320 etiquetas alternativas (74,3%) ao número identificado na análise de seqüências com cauda de poli-A. Aproximadamente 62% das etiquetas alternativas identificadas em seqüências de cDNAs com cauda poli-A foram identificadas no banco experimental; esta proporção cai para 57,8% quando consideramos todas as seqüências de cDNAs disponíveis. Essa redução é relativamente pequena, sugerindo que muitas das seqüências de cDNAs sem cauda poli-A apresentam uma região 3' completamente seqüenciada e que estamos subestimando o número de etiquetas alternativas em nossas análises.

Apesar do número de etiquetas alternativas identificadas por SAGE e MPSS serem similares, uma fração maior de etiquetas SAGE alternativas foi encontrada no banco de dados experimental (91,4% de SAGE contra 26,6% de MPSS). Acreditamos que esta diferença é consequência do maior número de bibliotecas de SAGE geradas experimentalmente (305 de bibliotecas de SAGE versus 6 de MPSS).

A identificação das etiquetas alternativas é importante para uma quantificação mais precisa da expressão dos genes por SAGE e MPSS e, sobretudo, estas etiquetas também

podem ser usadas no estudo da expressão gênica alelo-específica. Variações na expressão gênica entre alelos foram classicamente associadas com a inativação do cromossomo X (Lyon, 1961) e *imprinting* genômico (Reik e Walter, 2001). Recentemente, estudos mostraram que estas variações são eventos relativamente rotineiros entre os genes não *imprinted*, podendo abranger ~50% de todos os genes humanos, e que a expressão monoalélica aleatória é mais comum do que o esperado (Lo, Wang *et al.*, 2003; Pant, Tao *et al.*, 2006; Gimelbrant, Hutchinson *et al.*, 2007; Bjornsson, Albert *et al.*, 2008). Estes importantes estudos foram eficientes na identificação de genes com variações na expressão de seus alelos, porém as suas limitações estão nos fatos de serem técnicas caras e trabalhosas, envolvendo o *design* e a construção de *arrays*. O uso dos bancos de dados públicos como ferramenta de bioinformática, assim como feitos nos trabalhos de Yang, Hu *et al.*, (2003) e Ge, Gurd *et al.*, (2005), foram pouco explorados e apresentam um enorme potencial no estudo da expressão gênica alelo-específica, especialmente quando empregada a nossa metodologia complementar, precisa e inovadora.

Para dar continuidade ao nosso trabalho e aumentar a utilidade da análise apresentada, antes de iniciarmos as análises de expressão alelo-específica, nós atualizamos o banco de dados de etiquetas alternativas. Os resultados desta atualização podem ser vistos nas Tabelas 4.1.5.1, 4.1.5.2 e 4.1.5.3. Futuras atualizações podem revelar novos candidatos e tornar nossas análises mais interessantes.

TABELA 4.1.5.1 – Etiquetas alternativas produzidas por SNPs que geram um novo sítio de restrição mais próximo da cauda poli-A do transcrito (Atualização)

Análise	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas.
<i>Short SAGE</i>	823	786	212
<i>Long SAGE</i>	823	320	72
MPSS	497	204	50
Total	2.143	1.310	334

TABELA 4.1.5.2 - Etiquetas alternativas devido à presença de SNPs no sítio de restrição original (Atualização).

Análise	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas.
<i>Short SAGE</i>	640	584	165
<i>Long SAGE</i>	640	360	54
<i>MPSS</i>	532	297	48
<i>Todas análises</i>	1.812	1.241	267

TABELA 4.1.5.3 - Etiquetas alternativas geradas pela sobreposição de SNPs com etiquetas de SAGE e MPSS (Atualização).

Análise	Etiquetas alternativas	Confirmadas experimentalmente	Etiquetas ambíguas.
<i>Short SAGE</i>	1.275	1.187	302
<i>Long SAGE</i>	1.952	8.35	154
<i>MPSS</i>	1.516	5.97	198
<i>Todas análises</i>	4.743	2.619	654

4.2. Etiquetas alternativas na busca de genes com expressão alélica diferencial.

Após explorar este conjunto de etiquetas alternativas, descobrimos que podemos utilizá-las para identificar a expressão dos diferentes alelos de um gene. Por exemplo, se um SNP gera uma etiqueta alternativa e se o indivíduo for heterozigoto para esse polimorfismo, este gene deve apresentar duas etiquetas diferentes, uma para cada alelo. A frequência de cada etiqueta dentro de uma dada biblioteca pode então ser utilizada para inferir o nível de expressão de cada um dos alelos.

Analisando nosso banco de dados, identificamos 1.882 genes que apresentam SNPs com potencial de gerar etiquetas alternativas. Geramos um relatório de genes com etiquetas alternativas para as técnicas de *long* SAGE e de *short* SAGE (MPSS foi excluída destas análises, ver material e métodos 3.6). Para cada gene foi estudado o padrão de expressão dos dois alelos nas 263 bibliotecas experimentais (233 de *short* SAGE e 30 de *long* SAGE) e eles foram classificados em quatro categorias diferentes de acordo como o seu padrão de expressão (ver material e métodos 3.7).

4.2.1. A categoria dos genes com expressão monoalélica.

Para 23,38% (440) dos genes estudados, ambas as etiquetas (original e alternativa) estão presentes no banco experimental, porém em bibliotecas diferentes (ver tabela 3.11.1). Estes genes foram classificados como candidatos a apresentarem expressão monoalélica. Aqui vale ressaltar que este padrão de expressão monoalélica só corre em dois casos: i) quando o gene realmente está sendo expresso de maneira monoalélica; ii) quando as bibliotecas de SAGE forem provenientes de indivíduos homozigotos para o SNP, isto é, gerando candidatos falso positivos. Como a nossa metodologia inicial não consegue distinguir entre os dois casos, fizemos diversas análises, incluindo validações experimentais, para estimar a taxa de falso positivos na nossa lista de genes com expressão monoalélica.

No intuito de responder se nossos 440 genes estavam enriquecidos com falsos positivos, medimos a taxa de heterozigose utilizando o princípio de Hardy-Weinberg (Hardy, 1908; Weinberg, 1908). Assumimos que o número de bibliotecas era a nossa população e que a evidência de expressão da etiqueta nas bibliotecas era a frequência alélica para cada alelo. Usando essa estratégia percebemos que 85,22% dos nossos candidatos tinham uma taxa de heterozigose superior a 0,10 (ver Figura 4.2.1-A), mostrando que nosso conjunto de dados não estavam enriquecidos com alelos raros de SNPs e por tanto não estavam enriquecidos

com falsos positivos. Numa segunda estratégia, utilizamos as frequências alélicas dos SNPs disponível no dbSNP, e chegamos a resultados similares, 71,81% dos genes candidatos apresentaram uma taxa de heterozigose superior a 0,10 (ver Figura 4.2.1-B). Um limitante das frequências alélicas dos SNPs é o fato de que apenas 25% (110) dos 440 genes estudados possuem frequência catalogadas no dbSNP e que as bibliotecas de SAGE podem não ser de amostras das mesmas populações que o dado do dbSNP. Entretanto, esta análise nos dá um indicativo de que uma boa porcentagem de nossos genes candidatos são realmente expressos de maneira monoalélica.

Figura 4.2.1. – Frequência de heterozigotos para cada gene. Em A utilizamos expressão da etiqueta alternativa como frequência alélica e número de bibliotecas como população. Em B utilizando as frequências alélicas dos SNPs disponível no dbSNP.

4.2.2. A categoria dos genes com etiquetas não informativas.

Para 27,10% (510) dos genes estudados, apenas uma das duas etiquetas (original ou alternativa) estavam presentes no banco experimental (ver tabela 3.11.2). A falta de expressão de uma das etiquetas pode ser ocasionada por diversos fatores tais como: i) SNPs com alelos raros; ii) genes com expressão baixa; iii) falsos SNPs (só uma fração do banco de SNP (40.7%) foi validado). Por estes motivos, todos os genes com etiquetas alternativas não presentes no banco de SAGE foram excluídos das nossas análises.

4.2.3. A categoria dos genes com expressão bialélica.

Para 49,52% (932) dos genes estudados ambas as etiquetas foram encontradas no banco experimental e, em pelo menos uma biblioteca de um determinado tecido, temos a presença de ambas as etiquetas simultaneamente (ver tabela 3.11.3), sugerindo a expressão bialélica. Segundo a expressão monoalélica aleatória seria possível encontrarmos algumas bibliotecas com expressão bialélica, enquanto outras (a fração predominante) apresentam um padrão de expressão monoalélica (Bix e Locksley, 1998; Hollander, Zuklys *et al.*, 1998; Gimelbrant, Hutchinson *et al.*, 2007). Entretanto, resolvemos ser conservadores com a classificação dos genes, e definimos como expressão bialélica todos os genes que apresentarem expressão de ambas etiquetas simultaneamente em pelo menos uma biblioteca de SAGE.

4.2.4. Genes com perda de expressão monoalélica em tumor.

Essa categoria é na verdade um subconjunto categoria de genes com expressão bialélica. Nesta categoria estão 372 (19,77%) genes que apresentam: i) ambas as etiquetas (original e alternativa) presentes no banco experimental; ii) padrão de expressão monoalélica nas bibliotecas de SAGE de tecido normal; iii) padrão de expressão bialélica em pelo menos uma biblioteca provenientes de tecido tumoral (ver tabela 3.11.4). O padrão de expressão destes genes é consistente com a hipótese de perda de *imprinting* genômico (LOI – *loss of imprinting*) durante a tumorigênese (Yang, Hu *et al.*, 2003) e, como no geral cada gene apresentou poucas bibliotecas com a perda de expressão monoalélica, assim, consideramos esta categoria como genes candidatos à expressão alélica diferencial.

Assim nosso estudo categorizou 812 genes candidatos a ter expressão alélica diferencial (440 com padrão de expressão monoalélica e 372 perda de expressão monoalélica em tumor. A lista de genes está disponível em <http://www.compbio.ludwig.org.br/~jorge/>

monoallelic/), representando 43,15% do conjunto inicial de 1.882 genes (figura 4.2.4 resume a nossa categorização). É interessante notar que este resultado é consistente com outros estudos que também identificaram genes com expressão monoalélica em humanos (Lo, Wang *et al.*, 2003; Ge, Gurd *et al.*, 2005; Pant, Tao *et al.*, 2006; Gimelbrant, Hutchinson *et al.*, 2007).

Figura 4.2.4. – Esquema de categorização dos 1.882 genes com etiquetas alternativas identificadas. Caixa com fundo em preto: Total de genes candidatos a expressão monoalélica.

4.2.5. Genes já identificados com expressão alélica diferencial na literatura. (Banco de genes com expressão alélica diferencial).

Para termos uma idéia do potencial de nossa metodologia, nós construímos um banco de dados com os genes com expressão alélica diferencial ou genes *imprinted* presentes na literatura (ver material e métodos 3.12) e cruzamos estes genes com a nossa lista de genes candidatos. Dentro do nosso conjunto de 812 genes, 78 haviam sido descrito previamente como apresentando expressão alélica diferencial (ver figura 4.2.5.1). Entre estes 78 genes, um foi identificado no trabalho de Lo, Wang *et al.*, (2003) (figura 4.2.5.1 Coluna A); 59 foram identificados no trabalho de Pant, Tao *et al.*, (2006) (figura 4.2.5.1 Coluna B); 17 foram identificados no trabalho de Gimelbrant, Hutchinson *et al.*, (2007) (figura 4.2.5.1 Coluna C); 5 foram identificados no trabalho de Morison, Paton *et al.*, (2001) (figura 4.2.5.1 coluna D). Entre os 5 genes candidatos a *imprinted* (INS, ZNF597, DIO3, SLC22A2,

SLC22A3), o DIO3 foi previamente classificado como *imprinted* em camundongo (Morison, Paton *et al.*, 2001), mas seu status de *imprinted* em humano ainda não foi reportado.

UniGene	Gene	A	B	C	D	UniGene	Gene	A	B	C	D
Hs.433060	ACPP					Hs.588289	MAPK14				
Hs.353175	AGPAT4					Hs.592047	MKL2				
Hs.460260	AKR1C2					Hs.101840	MR1				
Hs.156727	ANKH					Hs.530735	MS4A7				
Hs.306764	ANKRD15					Hs.85937	MYBPC2				
Hs.489207	ASNS					Hs.414332	OAS2				
Hs.121028	ASPM					Hs.583792	PARP14				
Hs.468688	C10orf137					Hs.477352	PDIA5				
Hs.500897	C10orf26					Hs.520421	PERP				
Hs.164705	C11orf63					Hs.146100	PIP3-E				
Hs.497253	C14orf101					Hs.498570	PRKCQ				
Hs.148609	C6orf157					Hs.493516	PSIP1				
Hs.100322	CA6					Hs.231850	RIC3				
Hs.490203	CALD1					Hs.154628	RSHL2				
Hs.5353	CASP10					Hs.316931	SCAP1				
Hs.18912	CCDC17					Hs.136309	SH3GLB1				
Hs.424183	CCDC35					Hs.284813	SIGLEC10				
Hs.35433	CDC42BPA					Hs.75231	SLC16A1				
Hs.100914	CEP192					Hs.125482	SLC22A15				
Hs.191164	CHD1L					Hs.436385	SLC22A2				
Hs.434250	CKAP2L					Hs.567337	SLC22A3				
Hs.632335	COQ6					Hs.423163	SLC35A1				
Hs.212140	CWF19L2					Hs.435661	SPTLC2				
Hs.481980	DAB2					Hs.584803	ST3GAL1				
Hs.507755	DCAMKL1					Hs.44385	STON1				
Hs.49322	DIO3					Hs.11615	STYXL1				
Hs.2375	EMR1					Hs.12967	SYNE1				
Hs.167115	ENDOD1					Hs.517155	TMEPAI				
Hs.471933	FKBP1A					Hs.34024	TNIK				
Hs.632389	FPGT					Hs.153504	TOM1L1				
Hs.518525	GLUL					Hs.591961	VMD2				
Hs.163173	IFIH1					Hs.497369	XX				
Hs.524117	IL15RA					Hs.505231	YARS2				
Hs.89832	INS					Hs.400802	ZBTB1				
Hs.65758	ITPR3					Hs.48589	ZNF228				
Hs.189915	KIF13A					Hs.349444	ZNF558				
Hs.476517	LOC92689					Hs.148322	ZNF577				
Hs.187636	LRRK2					Hs.88630	ZNF597				
Hs.178695	MAPK13					Hs.517418	ZNF74				

Figura 4.2.5.1 – 78 Genes do nosso conjunto de dados de expressão alélica diferencial que já haviam sido categorizados por outros estudos. Coluna A representa o trabalho de Lo, Wang *et al.*, (2003), Coluna B representa o trabalho de Pant, Tao *et al.*, (2006), Coluna C representa o trabalho de Gimelbrant, Hutchinson *et al.*, (2007) e Coluna D representa o trabalho de Morison, Paton *et al.*, (2001). “The database of imprinted genes and parent-of-origin effects in animals”. Os Quadrados em vermelho representam em quais trabalhos os genes foram identificados como expressão alélica diferencial.

Também cruzamos nosso conjunto de dados de genes candidatos a apresentarem expressão bialélica contra os dados da literatura. Dos nossos 560 genes, 49 foram categorizados como expressão alélica diferencial pela literatura (ver figura 4.2.5.2), sendo que 6 foram identificados no trabalho de Lo, Wang *et al.*, (2003) (figura 4.2.5.2 Coluna A), 35 foram identificados no trabalho de Pant, Tao *et al.*, (2006) (figura 4.2.5.2 Coluna B), 5 foram identificados no trabalho de Gimelbrant, Hutchinson *et al.*, (2007) (figura 4.2.5.2 Coluna C) e 4 foram identificados no trabalho de Morison, Paton *et al.*, (2001) (figura 4.2.5.2 coluna D). Já sabíamos que isso poderia acontecer, pois como já dito fomos bastante conservadores em nossas categorizações (ver material e métodos 3.11). Além disto, os trabalhos aqui descritos também estão sujeitos a erros de interpretação. Apenas para citar um exemplo, em 2007 Khatib (Khatib, 2007) publicou uma revisão que avaliava a literatura de 50 genes *imprinted* em camundongo. Ele chegou à conclusão que destes 24 genes apresentavam expressão preferencial para um determinado alelo e que muitos destes genes apresentavam expressão bialélica dependendo do tecido estudado. Ele termina o artigo sugerindo que estes 24 genes não deveriam ser classificados como *imprinted* ou que a definição de *imprinting* genômico deveria ser mudada.

UniGene	Gene	A	B	C	D	UniGene	Gene	A	B	C	D
Hs.438823	ABP1					Hs.433442	KIFAP3				
Hs.512033	ACTR2					Hs.2785	KRT17				
Hs.500645	ALDH18A1					Hs.509414	KTN1				
Hs.90280	ATIC					Hs.213289	LDLR				
Hs.489603	ATXN7L1					Hs.148656	LOC91316				
Hs.642611	BCL2L13					Hs.631506	MCM8				
Hs.438064	BCL7B					Hs.297413	MMP9				
Hs.412842	C10orf7					Hs.533499	MNAB				
Hs.10649	C1orf38					Hs.520943	NCF1				
Hs.516853	C20orf194					Hs.73797	NCLN				
Hs.446146	CARD8					Hs.463320	OSBPL7				
Hs.420269	COL6A2					Hs.107410	PEX6				
Hs.191219	CPNE3					Hs.154036	PHLDA2				
Hs.221941	CYBRD1					Hs.102336	PRR5				
Hs.353022	ETAA16					Hs.227777	PTP4A1				
Hs.265018	FAM118A					Hs.512606	SCLY				
Hs.471452	FARSLB					Hs.101307	SLC14A1				
Hs.352642	FCGR2B					Hs.435735	SLC5A6				
Hs.407190	FKBP5					Hs.485195	SORT1				
Hs.40735	FZD3					Hs.8261	SPSB1				
Hs.125898	GNAS					Hs.153026	SWAP70				
Hs.164060	GRB10					Hs.396358	TMEM106B				
Hs.1051	GZMB					Hs.30345	TRAP1				
Hs.533566	H19					Hs.364544	TSPAN13				
Hs.590881	JAG1										

Figura 4.2.5.2 – 49 Genes do nosso conjunto de dados de expressão bialélica que foram categorizados por outros estudos como expressão alélica diferencial. Coluna A representa o trabalho de Lo, Wang *et al.*, (2003), Coluna B representa o trabalho de Pant, Tao *et al.*, (2006), Coluna C representa o trabalho de Gimelbrant, Hutchinson *et al.*, (2007) e Coluna D representa o trabalho de Morison, Paton *et al.*, (2001). “*The database of imprinted genes and parent-of-origin effects in animals* “. Os Quadrados em vermelho representam em quais trabalhos os genes foram identificados como expressão alélica diferencial.

4.2.6. Genes candidatos submetidos à validação experimental.

A partir dos 812 genes candidatos a expressão alélica diferencial, escolhemos arbitrariamente 29 genes (ver tabela 3.13) para testarmos experimentalmente o seu padrão de expressão monoalélica. Primeiramente foi feita uma análise para identificar indivíduos heterozigotos para os SNPs correspondentes às etiquetas alternativas através da genotipagem

do DNA de 30 amostras de sangue. Foram excluídos das nossas análises 12 genes candidatos para os quais não foi possível identificar mais do que quatro indivíduos heterozigotos. Dos 17 genes restantes foram seqüenciadas, em amostras de DNA e cDNA, as regiões onde estavam mapeadas as etiquetas alternativas associadas aos SNPs. O resultado do seqüenciamento também foi submetido à avaliações de qualidade (ver material e métodos 3.11).

Tabela 4.2.6 – número de heterozigotos para os genes candidatos à validação experimental.

Candidato	SNP	Alelos	Gene	# de ind. Heterozigotos
01	rs2304511	T/G	SLC6A6	09
02	rs1065368	C/T	SOX8	11
03	rs470939	G/T	GPR109A	08
04	rs1053454	G/T	PIP5K2A	12
05	rs2587512	G/C	TBC1D16	08
06	rs1060298	A/G	Hs.594773	14
Excluído	rs2272761	G/A	R3HCC1	< 4
Excluído	rs1568918	A/G	PITPNM2	< 4
Excluído	rs2281656	G/A	Hs.596214	< 4
Excluído	rs17081950	C/T	MSX2	< 4
Excluído	rs1057755	A/T	FGFR4	< 4
07	rs2071863	G/A	MAPK13	08
08	rs9132	A/G	NT5DC3	13
09	rs3818499	C/G	ING1	07
10	rs1543442	A/G	SNAI1	09
11	rs17027704	A/G	ATP5F1	14
12	rs1558525	A/G	PHC1	12
13	rs1065154	T/G	SQSTM1	12
14	rs6831	C/T	MRPS27	08
15	rs12907665	G/A	INOC1	11
16	rs266805	A/G	MUM1	11
17	rs14849	T/C	C10orf26	06
Excluído	rs4781	A/G	LPIN2	< 4
Excluído	rs3193677	C/A	ING2	< 4
Excluído	rs474058	A/C	GLI1	< 4
Excluído	rs2233128	A/G	LY86	< 4
Excluído	rs2241838	A/G	RBJ	< 4
Excluído	rs2518996	G/A	LOC643837	< 4
Excluído	rs12856	T/C	Hs.713563	< 4

4.2.7. Expressão alélica diferencial e bialélica nos dados experimentais.

Para analisar a expressão dos genes candidatos elaboramos uma estratégia que utiliza o programa *PeakPicker* e um método estatístico. O *PeakPicker* analisa os SNPs a partir dos cromatogramas das amostras de DNA e cDNA. O método estatístico torna possível analisar a expressão alélica de um determinado gene candidato através da razão normalizada das diferenças entre os alelos dos SNPs associado às etiquetas alternativas nas amostras de DNA e de cDNA (ver material de métodos 3.12).

Após obtermos o padrão de expressão alélica dos nossos 17 genes candidatos, determinamos se uma amostra de cDNA apresentava expressão alélica diferencial através de um intervalo de confiança (IC) baseado na variância da razões encontradas nas amostras de DNA (ver material de métodos 3.12). O intervalo de confiança vai de 1, o qual representa expressão igual e bialélica, a 0,81, o qual representa a expressão de 60% para um alelo e 40% para o outro. A figura 4.2.7 é um painel aonde mostramos as razões encontradas para todas as nossas amostras de DNA e cDNA. A linha azul é a representação do nosso IC e as amostras de cDNAs abaixo desta linha são as amostras que apresentaram expressão diferencial.

Figura 4.2.7 – Panel de genes submetidos a validação experimental. Reta em azul é a representação do IC99%. Em vermelho os pontos representam as razões das diferenças de expressão entre os alelos de todas as amostras (cDNA à esquerda e DNA à direita). No eixo X painel de candidatos e no eixo Y a razão alélica 1,0 representa expressão 50:50% entre os alelos e 0,0 representa 100:0%.

Para três genes (SLC6A6 candidato 1, ING1 candidato 9 e PIP5K2A candidato 4) não encontramos em nenhuma de suas amostras de cDNAs um padrão de expressão diferencial (ver figura 4.2.7). Dois destes genes (SLC6A6 e PIP5K2A) já tinham sido descritos previamente com expressão bialélica (Gimelbrant, Hutchinson *et al.*, 2007). Para os outros 14 genes, ao menos uma amostra de cDNA apresentou expressão alélica diferencial (ver figura 4.2.7). Destes, 3 genes foram classificados como monoalélicos (ATP5F1 candidato 11, PHC1 candidato 12 e INOC1 candidato 15), pois apresentaram expressão diferencial para todas as amostras de cDNAs (em 95% das amostras não era possível ver a expressão de um segundo

alelo através de uma análise visual nos cromatogramas). Em dois destes três genes, ATP5F1 candidato 11 (gene que codifica o complexo F0 na formação ATP sintase mitocondrial), PHC1 candidato 12 (complexo protéico associado à repressão de expressão gênica), houve uma expressão preferencial entre os alelos, sugerindo um fenômeno de exclusão alélica (fenômeno incluso dentro da categoria dos genes com expressão monoalélica aleatória). Já o terceiro gene, INOC1 candidato 15 (uma subunidade do complexo de remodelamento de cromatina), seria um forte candidato a estar sob *imprinting* genômico, pois para esse candidato não houve expressão preferencial para nenhum dos seus alelos.

Dentre os 11 genes restantes, o MAPK13 e o C10orf26 já haviam sido descritos como expressão diferencial por Pant, Tao *et al.*, (2006). Contrariamente, Gimelbrant, Hutchinson *et al.*, (2007) classificaram MAPK13 como expressão bialélica. Gimelbrant, Hutchinson *et al.*, (2007) também classificaram NT5DC3, SQSTM1 e MRPS27 como expressão bialélica enquanto nossos dados experimentais demonstram um claro padrão de expressão diferencial destes genes. Além disso, todos os 11 genes apresentaram um padrão aleatório de expressão entre os diversos indivíduos, e estudos recentes acreditam que esse padrão aleatório de expressão diferencial seja um mecanismo importante para estabelecer a diversidade entre os indivíduos (Gimelbrant, Hutchinson *et al.*, 2007; Bjornsson, Albert *et al.*, 2008).

5. Conclusões.

5.1. O Impacto dos SNPs nos dados experimentais de SAGE e MPSS.

Neste trabalho, nós mostramos que a presença de SNPs dentro de seqüências de cDNAs humanas são responsáveis pela geração de 2020 etiquetas alternativas (1100 etiquetas alternativas de SAGE e 920 etiquetas alternativas de MPSS) e que 8.6% de todos os genes humanos conhecidos apresentam pelo menos uma etiqueta alternativa associada à presença de SNPs. Globalmente, tais resultados provam a viabilidade de se empregar etiquetas alternativas para corrigir as estimativas dos níveis de expressão gênica derivados a partir das técnicas de SAGE e MPSS.

5.2. A integração do banco de etiquetas alternativas com o SAGE *Genie*.

Para tornar nossa análise acessível para toda a comunidade científica, integramos o banco de dados de etiquetas alternativas de SAGE ao SAGE *Genie* (<http://cgap.nci.nih.gov/SAGE>) (Boon, Osorio *et al.*, 2002), onde os dados podem ser baixados diretamente ou visualizados no *website* (Figura 4.1.4).

5.3. Etiquetas alternativas na busca de genes com expressão monoalélica.

Após explorar o conjunto de etiquetas alternativas, descobrimos que podemos utilizá-las para identificar a expressão gênica alelo-específica. Analisando todas as bibliotecas públicas de SAGE, nosso estudo categorizou 812 genes, 440 com padrão de expressão monoalélica e 372 exibindo perda de expressão monoalélica em tumor, o que representa 43,15% do conjunto inicial de 1882 genes com etiquetas alternativas, resultados que são consistentes com outros estudos (Lo, Wang *et al.*, 2003; Ge, Gurd *et al.*, 2005; Pant, Tao *et al.*, 2006; Gimelbrant, Hutchinson *et al.*, 2007)

5.4. Banco de genes com expressão alélica diferencial.

Para sabermos quantos dos nossos genes candidatos já haviam sido descritos como exemplos de expressão alélica diferencial, compilamos uma base de dados com resultados de três experimentos de larga escala sobre expressão alélica e com os dados provenientes de um banco de *imprinting* genômico (Morison, Paton *et al.*, 2001). Estes dados estão disponíveis no endereço *web* <http://www.compbio.ludwig.org.br/~jorge/monoallelic/>. Do nosso conjunto de 812 genes candidatos, 78 genes já haviam sido descritos na literatura como apresentando expressão alélica diferencial (ver figura 4.2.5.1) e 5 genes como *imprinted*, o que corrobora a eficiência da nossa metodologia.

5.5. Validação experimental da expressão alélica diferencial.

Do repertório de genes candidatos à expressão alélica diferencial, 29 foram selecionados para validação experimental. A análise desses experimentos foi baseada na razão normalizada das diferenças de expressão entre os alelos (ver material e métodos, seção 3.12). Dentre os 17 genes provenientes de indivíduos heterozigotos, três demonstraram expressão bialélica e 14 apresentaram vários níveis de diferenças na expressão de seus alelos. Dentre estes últimos, os genes ATP5F1, PHC1 e INOC1 apresentaram um padrão de expressão monoalélica em 95% das amostras.

5.6. Observações finais

Acreditamos que nossa estratégia para utilizar etiquetas de SAGE e MPSS na identificação de genes submetidos à expressão alélica diferencial é promissora, principalmente quando vemos a proporção de genes que confirmam a expressão diferencial na nossa validação experimental (14 de 17 genes). Finalmente, ressaltamos que a indisponibilidade de dados sobre a genotipagem dos SNPs nas bibliotecas de SAGE é o maior

limitante para a utilização da nossa metodologia. Porém, o uso crescente e a contínua redução dos custos das técnicas de seqüenciamento em larga escala (454-Roche, SOLEXA-Illumina e SOLID-ABI) sugerem que, num futuro próximo, será economicamente viável a genotipagem dos SNPs nas mesmas amostras usadas para construção das bibliotecas de SAGE. Esses dados, combinados com nossa metodologia, poderão ser de grande valia para estudos dirigidos da análise da expressão alélica diferencial.

6. Referências

- Bix, M. e R. M. Locksley. Independent and epigenetic regulation of the interleukin-4 alleles in CD4⁺ T cells. Science, v.281, n.5381, Aug 28, p.1352-4. 1998.
- Bjornsson, H. T., T. J. Albert, *et al.* SNP-specific array-based allele-specific expression analysis. Genome Res, v.18, n.5, May, p.771-9. 2008.
- 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.
- Bray, N. J., P. R. Buckland, *et al.* Cis-acting variation in the expression of a high proportion of genes in human brain. Hum Genet, v.113, n.2, Jul, p.149-53. 2003.
- 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.
- Cattanach, B. M. e M. Kirk. Differential activity of maternally and paternally derived chromosome regions in mice. Nature, v.315, n.6019, Jun 6-12, p.496-8. 1985.
- Chess, A. Expansion of the allelic exclusion principle? Science, v.279, n.5359, Mar 27, p.2067-8. 1998.
- Chess, A., I. Simon, *et al.* Allelic inactivation regulates olfactory receptor gene expression. Cell, v.78, n.5, Sep 9, p.823-34. 1994.
- Cowles, C. R., J. N. Hirschhorn, *et al.* Detection of regulatory variation in mouse genes. Nat Genet, v.32, n.3, Nov, p.432-7. 2002.
- Ewing, B. e P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, v.8, n.3, Mar, p.186-94. 1998.
- Ewing, B., L. Hillier, *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, v.8, n.3, Mar, p.175-85. 1998.
- 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.
- 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., D. O. Vidal, *et al.* Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biol, v.8, n.3, p.R40. 2007.
- Ge, B., S. Gurd, *et al.* Survey of allelic expression using EST mining. Genome Res, v.15, n.11, Nov, p.1584-91. 2005.
- Gimelbrant, A., J. N. Hutchinson, *et al.* Widespread monoallelic expression on human autosomes. Science, v.318, n.5853, Nov 16, p.1136-40. 2007.

- Hardy, G. H. Mendelian Proportions in a Mixed Population. Science, v.28, n.706, Jul 10, p.49-50. 1908.
- Hollander, G. A., S. Zuklys, *et al.* Monoallelic expression of the interleukin-2 locus. Science, v.279, n.5359, Mar 27, p.2118-21. 1998.
- Jongeneel, C. V., M. Delorenzi, *et al.* An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res, v.15, n.7, Jul, p.1007-14. 2005.
- Khatib, H. Is it genomic imprinting or preferential expression? Bioessays, v.29, n.10, Oct, p.1022-8. 2007.
- Knight, J. C. Allele-specific gene expression uncovered. Trends Genet, v.20, n.3, Mar, p.113-6. 2004.
- Leighton, P. A., J. R. Saam, *et al.* Genomic imprinting in mice: its function and mechanism. Biol Reprod, v.54, n.2, Feb, p.273-8. 1996.
- Li, E., C. Beard, *et al.* Role for DNA methylation in genomic imprinting. Nature, v.366, n.6453, Nov 25, p.362-5. 1993.
- Lo, H. S., Z. Wang, *et al.* Allelic variation in gene expression is common in the human genome. Genome Res, v.13, n.8, Aug, p.1855-62. 2003.
- Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). Nature, v.190, Apr 22, p.372-3. 1961.
- Mcgrath, J. e D. Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell, v.37, n.1, May, p.179-83. 1984.
- Monk, D., P. Arnaud, *et al.* Limited evolutionary conservation of imprinting in the human placenta. Proc Natl Acad Sci U S A, v.103, n.17, Apr 25, p.6623-8. 2006.
- Morison, I. M., C. J. Paton, *et al.* The imprinted gene and parent-of-origin effect database. Nucleic Acids Res, v.29, n.1, Jan 1, p.275-6. 2001.
- Morison, I. M., J. P. Ramsay, *et al.* A census of mammalian imprinting. Trends Genet, v.21, n.8, Aug, p.457-65. 2005.
- Oakey, R. J. e C. V. Beechey. Imprinted genes: identification by chromosome rearrangements and post-genomic strategies. Trends Genet, v.18, n.7, Jul, p.359-66. 2002.
- Pant, P. V., H. Tao, *et al.* Analysis of allelic differential expression in human white blood cells. Genome Res, v.16, n.3, Mar, p.331-9. 2006.
- Pernis, B., G. Chiappino, *et al.* Cellular localization of immunoglobulins with different allotypic specificities in rabbit lymphoid tissues. J Exp Med, v.122, n.5, Nov 1, p.853-76. 1965.

- Prawitt, D., T. Enklaar, *et al.* Identification and characterization of MTR1, a novel gene with homology to melastatin (MLSN1) and the trp gene family located in the BWS-WT2 critical region on chromosome 11p15.5 and showing allele-specific expression. Hum Mol Genet, v.9, n.2, Jan 22, p.203-16. 2000.
- Rajewsky, K. Clonal selection and learning in the antibody system. Nature, v.381, n.6585, Jun 27, p.751-8. 1996.
- Reik, W. e J. Walter. Genomic imprinting: parental influence on the genome. Nat Rev Genet, v.2, n.1, Jan, p.21-32. 2001.
- Sakabe, N. J., J. E. De Souza, *et al.* ORESTES are enriched in rare exon usage variants affecting the encoded proteins. C R Biol, v.326, n.10-11, Oct-Nov, p.979-85. 2003.
- 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.
- Surani, M. A., S. C. Barton, *et al.* Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature, v.308, n.5959, Apr 5-11, p.548-50. 1984.
- Velculescu, V. E., L. Zhang, *et al.* Serial analysis of gene expression. Science, v.270, n.5235, Oct 20, p.484-7. 1995.
- Wang, Z., H. Fan, *et al.* Comparative sequence analysis of imprinted genes between human and mouse to reveal imprinting signatures. Genomics, v.83, n.3, Mar, p.395-401. 2004.
- Weinberg, W. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg v.64, January 13, p.368–382. 1908.
- Wrzeska, M. e B. Rejduch. Genomic imprinting in mammals. J Appl Genet, v.45, n.4, p.427-33. 2004.
- Yan, H., W. Yuan, *et al.* Allelic variation in human gene expression. Science, v.297, n.5584, Aug 16, p.1143. 2002.
- Yang, H. H., Y. Hu, *et al.* Computation method to identify differential allelic gene expression and novel imprinted genes. Bioinformatics, v.19, n.8, May 22, p.952-5. 2003.

Lista de Anexos.

1. Súmula Curricular.

2. Artigo incluído nesta tese:

- a. A.P.M. Silva, **J.E.S. De Souza**, P.A.F. Galante, S. J. De Souza, A.A. Camargo (2004). **The impact of SNPs on the generation of SAGE and MPSS alternative tags. Nucleic Acid Research.** v.32, n.20, p.6104-10.

Anexo 1 – SÚMULA CURRICULAR

DADOS PESSOAIS

Nome Jorge Estefano Santana de Souza

Nasci.: São Paulo-SP, 14/02/1977

EDUCAÇÃO

2002 – 2008 Doutorado em Bioinformática.
Universidade de São Paulo, USP, São Paulo, Brasil.
Orientador: Sandro Jose de Souza.
Laboratório de Bioinformática do Instituto Ludwig de Pesquisa sobre o Câncer.

Co-Orientador: Junior Barrera.
Instituto de Matemática e Estatística da Universidade de São Paulo.

1998 – 2002 Graduação em Ciência da Computação.
Universidade de Santo Amaro, UNISA, São Paulo, Brasil.

BOLSAS RECEBIDAS

1	Doutorado Direto
Outubro/2002 - Setembro/2006
CAPES.
2	Monitoria PAE
Outubro/2004 - Janeiro/2005.
Programa de Aperfeiçoamento de Ensino
3	Bolsa técnica
Novembro/2000 - setembro/2001
Instituto Butantã.
4	Bolsa técnica
Janeiro/2000 - Dezembro/2000.
UNISA – Universidade de Santo Amaro.

PUBLICAÇÕES (Artigos Completos e Resumos em Congressos)

1 Artigos completos publicados em periódicos

08: P.A.F. Galante, D.O. Vidal, **J.E.S. De Souza**, A.A. Camargo, S. J. De Souza (2007). **Sense-antisense pairs in mammals: functional and evolutionary considerations.** *Genome Biology*. V8, I.3, p.r40.1-r40.14.

07: A.P.M. Silva, **J.E.S. De Souza**, P.A.F. Galante, S. J. De Souza, A.A. Camargo (2004). **The impact of SNPs on the generation of SAGE and MPSS alternative tags.** *Nucleic Acid Research*. v.32, n.20, p.6104-10.

06: Sogayar MC, Camargo AA, Bettoni F, Carraro DM, Pires LC, Parmigiani RB, Ferreira EN, de Sá Moreira E, do Rosário D de O Latorre M, Simpson AJ, Cruz LO, Degaki TL, Festa F, Massirer KB, Sogayar MC, Filho FC, Camargo LP, Cunha MA, De Souza SJ, Faria M Jr, Giuliatti S, Kopp L, de Oliveira PS, Paiva PB, Pereira AA, Pinheiro DG, Puga RD, **S de Souza JE**, Albuquerque DM, Andrade LE, Baia GS, Briones MR, Cavaleiro-Luna AM, Cerutti JM, Costa FF, Costanzi-Strauss E, Espreafico EM, Ferrasi AC, Ferro ES, Fortes MA, Furchi JR, Giannella-Neto D, Goldman GH, Goldman MH, Gruber A, Guimarães GS, Hackel C, Henrique-Silva F, Kimura ET, Leoni SG, Macedo C, Malnic B, Manzini B CV, Marie SK, Martinez-Rossi NM, Menossi M, Miracca EC, Nagai MA, Nobrega FG, Nobrega MP, Oba-Shinjo SM, Oliveira MK, Orabona GM, Otsuka AY, Paço-Larson ML, Paixão BM, Pandolfi JR, Pardini MI, Passos Bueno MR, Passos GA, Pesquero JB, Pessoa JG, Rahal P, Rainho CA, Reis CP, Ricca TI, Rodrigues V, Rogatto SR, Romano CM, Romeiro JG, Rossi A, Sá RG, Sales MM, Sant'Anna SC, Santarosa PL, Segato F, Silva WA Jr, Silva ID, Silva NP, Soares-Costa A, Sonati MF, Strauss BE, Tajara EH, Valentini SR, Villanova FE, Ward LS, Zanette

DL; Ludwig-FAPESP Transcript Finishing Initiative (2004). **A transcript finishing initiative for closing gaps in the human transcriptome.** *Genome Res.* 14(7):1413-23.

05: Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA Jr, Dias Neto E, Grivet M, Gruber A, Guimaraes PE, Hide W, Iseli C, Jongeneel CV, Kelso J, Nagai MA, Ojopi EP, Osorio EC, Reis EM, Riggins GJ, Simpson AJ, de Souza S, Stevenson BJ, Strausberg RL, Tajara EH, Verjovski-Almeida S, Acencio ML, Bengtson MH, Bettoni 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, Sa Ferreira K, Furnari FB, Faria M Jr, Galante PA, Guimaraes GS, Holanda AJ, Kimura ET, Leerkes MR, Lu X, Maciel RM, Martins EA, Massirer KB, Melo AS, Mestriner CA, Miracca EC, Miranda LL, Nobrega FG, Oliveira PS, Paquola AC, Pandolfi JR, Campos Pardini MI, Passeti F, Quackenbush J, Schnabel B, Sogayar MC, **Souza JE**, Valentini SR, Zaiats AC, Amaral EJ, Arnaldi LA, de Araújo AG, de Bessa SA, Bicknell DC, Ribeiro de Camaro ME, Carraro DM, Carrer H, Carvalho AF, Colin C, Costa F, Curcio C, Guerreiro da Silva ID, Pereira da Silva N, Dellamano M, El-Dorri H, Espreafico EM, Scattone Ferreira AJ, Ayres Ferreira C, Fortes MA, Gama AH, Giannella-Neto D, Giannella ML, Giorgi RR, Goldman GH, Goldman MH, Hackel C, Ho PL, Kimura EM, Kowalski LP, Krieger JE, Leite LC, Lopes A, Luna AM, Mackay A, Mari SK, Marques AA, Martins WK, Montagnini A, Mourão Neto M, Nascimento AL, Neville AM, Nobrega MP, O'Hare MJ, Otsuka AY, Ruas de Melo AI, Paco-Larson ML, Guimarães Pereira G, Pereira da Silva N, Pesquero JB, Pessoa JG, Rahal P, Rainho CA, Rodrigues V, Rogatto SR, Romano CM, Romeiro JG, Rossi BM, Rusticci M, Guerra de Sá R, Sant' Anna SC, Sarmazo ML, Silva TC, Soares FA, Sonati Mde F, de Freitas Sousa J, Queiroz D, Valente V, Vettore AL, Villanova FE, Zago MA, Zalcberg H; Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium; Human

Cancer Genome Project Sequencing Consortium (2003). **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*; 100(23):13418-23.

04: N.J. Sakabe, **J.E.S. De Souza**, P.F.A. Galante, P.S.L. de Oliveira, F. Passetti, H. Brentani, E.C. Osório, A.C. Zaiats, M.R. Leerkes, J.P. Kitajima, R.R. Brentani, R.L. Strausberg, A.J.G. Simpson & S.J. de Souza (2003). **ORESTES are enriched in rare exon usage variants affecting the encoded proteins.** *C.R. Biol*; 326(10-11):979-85.

03: A.P. Silva, A.C. Salim, A. Bulgarelli, **J.E.S. De Souza**, E.C. Osorio, O.L. Caballero, C. Iseli, B.J. Stevenson, C.V. Jongeneel, S.J. de Souza, A.J. Simpson, A.A. Camargo (2003). **Identification of 9 novel transcripts and two RGSL genes within the hereditary prostate cancer region (HPC1) at 1q25.** *Gene*. 310:49-57.

02: E.C. Osorio, **J.E.S. de Souza**, A.C. Zaiats, P.S. de Oliveira, S.J. de Souza (2003). **pp-Blast: a "pseudo-parallel" Blast.** *Braz J Med Biol Res*. 463-4.

01: M. Sakharkar, F. Passetti, **J.E.S. de Souza**, M. Long, S. J. de Souza (2002). **ExInt: an Exon Intron Database.** *Nucleic Acid Research*. v.30, n.1, p.191 - 194.

10: R. Ramalho, J.E.S. de Souza, D. Meyer, S.J. de Souza. The frequency spectrum of polymorphisms of ESEs in constitutive and alternative exons of the human genome – Annual Meeting of the Society for Molecular Biology and Evolution “SMBE”. June 5th-8th, 2008 Barcelona, Spain. Resumo: P-86.

09: J.E.S. de Souza, D.O. Vidal, P.A.F. Galante, S.J. de Souza, A.A Camargo. Identification of Mono-allelic Gene Expression and Differential Allelic Gene Expression in the Human Genome – 3st International Conference of the AB3C “X-meeting”, November 1th –3th, 2007, São Paulo, SP, Brazil. Resumo: p.181.

08: S. Ezquina, J.E.S. de Souza, S. J. Souza. The effect of polymorphisms and somatic mutations in the repertoire of polyadenylation variants – 3st International Conference of the AB3C “X-meeting”, November 1th –3th, 2007, São Paulo, SP, Brazil. Resumo: p.194.

07: D.O. Vidal, P.A.F. Galante, L.C. Pires, J.E.S. de Souza, A.A. Camargo, S.J. de Souza. High-throughput identification and gene expression analysis of novel antisense transcripts using MPSS – 2st International Conference of the AB3C “X-meeting”, August 6th –10th, 2006, Fortaleza, CE, Brazil. Resumo: LB-21.

06: J.E.S. de Souza, P.A.F. Galante , L.C. Pires, S.J. de Souza, A.A. Camargo. Identification of novel imprinted genes in human. – 2st International Conference of the AB3C “X-meeting”, August 6th –10th, 2006, Fortaleza, CE, Brazil. Resumo: I-17.

05: J.E.S. de Souza, O.M. Chaim, P.A.F. Galante, A.P.M. Silva, S.J. de Souza, A.A. Camargo. Identification of novel imprinted genes using allele-specific SAGE and MPSS

tags – 1st International Conference of the AB3C “X-meeting”, October 4th –7th 2005, Caxambu, MG, Brazil. Resumo Pag.77.

04: A.P.M. Silva, J.E. de Souza, P.A.F. Galante, S.J. de Souza, A.A. Camargo. The impact of SNPs on the interpretation of SAGE and MPSS experimental data – 2st International Conference on Bioinformatics and Computational Biology, October 25th –28th 2004, Angra dos Reis, RJ, Brazil. Resumo P.53.

03: J.E. de Souza, N.J. Sakabe, S.J. de Souza. Study of the correlation of ESEs disrupted by SNPs and alternative splicing – 2st International Conference on Bioinformatics and Computational Biology, October 25th –28th 2004, Angra dos Reis, RJ, Brazil. Resumo P.59.

02: F. Prosdocimi, G.C. Cerqueira, L.P. Camargo, F. Camargo, R.G.M. Ferreira, A.C.M. Junqueira, Á.V.F. Flatschart, A.F. Silva, A.N. dos Reis, A.C.F. dos Santos, A.N. Júnior, C.I. Wust, E. Binneck, J.L. Kessedjian, J.H. Petretski, R.P. Lima, R.M. Pereira, S. Jardim, V.S. Sampaio, J.E.S. de Souza, A.T. Vasconcelos, H. Brentani, A.A. Camargo. Candidate genes for the late onset Alzheimer disease in human chromosome 10 – 1st International Conference on Bioinformatics and Computational Biology, May 14th –16th 2003, Ribeirão Preto, SP, Brazil. Resumo 10.39

01: J.E. de Souza, N.J. Sakabe, J. Barrera, S.J. de Souza. A large-scale study of SNPs in regulatory elements of alternative splicing and their possible association to human diseases – 1st International Conference on Bioinformatics and Computational Biology, May 14th –16th 2003, Ribeirão Preto, SP, Brazil. Resumo 10.114

Anexo 2 – Artigo: A.P.M. Silva, **J.E.S. De Souza**, P.A.F. Galante, S. J. De Souza, A.A. Camargo (2004). **The impact of SNPs on the generation of SAGE and MPSS alternative tags. Nucleic Acid Research.** v.32, n.20, p.6104-10.

Nucleic Acids Research

The impact of SNPs on the interpretation of SAGE and MPSS experimental data

Ana Paula M. Silva, Jorge E. S. De Souza, Pedro A. F. Galante, Gregory J. Riggins, Sandro J. De Souza and Anamaria A. Camargo

Nucleic Acids Res. 32:6104-6110, 2004.

doi:10.1093/nar/gkh937

The full text of this article, along with updated information and services is available online at <http://nar.oxfordjournals.org/cgi/content/full/32/20/6104>

References

This article cites 16 references, 9 of which can be accessed free at <http://nar.oxfordjournals.org/cgi/content/full/32/20/6104#BIBL>

Cited by

This article has been cited by 7 articles at 6 October 2008 . View these citations at <http://nar.oxfordjournals.org/cgi/content/full/32/20/6104#otherarticles>

Reprints

Reprints of this article can be ordered at http://www.oxfordjournals.org/corporate_services/reprints.html

Email and RSS alerting

Sign up for email alerts, and subscribe to this journal's RSS feeds at <http://nar.oxfordjournals.org>

PowerPoint® image downloads

Images from this journal can be downloaded with one click as a PowerPoint slide.

Journal information

Additional information about Nucleic Acids Research, including how to subscribe can be found at <http://nar.oxfordjournals.org>

Published on behalf of

Oxford University Press
<http://www.oxfordjournals.org>

The impact of SNPs on the interpretation of SAGE and MPSS experimental data

Ana Paula M. Silva, Jorge E. S. De Souza^{1,2}, Pedro A. F. Galante^{1,3}, Gregory J. Riggins⁴, Sandro J. De Souza¹ and Anamaria A. Camargo*

Laboratory of Molecular Biology and Genomics and ¹Laboratory of Computational Biology, Ludwig Institute for Cancer Research, 01509-010, São Paulo, SP, Brazil, ²Interunit in Bioinformatics and ³Department of Biochemistry, University of São Paulo, 05508-900, São Paulo, SP, Brazil and ⁴John Hopkins University School of Medicine, 21224, Baltimore, MD, USA

Received August 5, 2004; Revised September 24, 2004; Accepted October 25, 2004

ABSTRACT

Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS) are powerful techniques for gene expression analysis. A crucial step in analyzing SAGE and MPSS data is the assignment of experimentally obtained tags to a known transcript. However, tag to transcript assignment is not a straightforward process since alternative tags for a given transcript can also be experimentally obtained. Here, we have evaluated the impact of Single Nucleotide Polymorphisms (SNPs) on the generation of alternative SAGE and MPSS tags. This was achieved through the construction of a reference database of SNP-associated alternative tags, which has been integrated with SAGE Genie. A total of 2020 SNP-associated alternative tags were catalogued in our reference database and at least one SNP-associated alternative tag was observed for ~8.6% of all known human genes. A significant fraction (61.9%) of these alternative tags matched a list of experimentally obtained tags, validating their existence. In addition, the origin of four out of five SNP-associated alternative MPSS tags was experimentally confirmed through the use of the GLGI-MPSS protocol (Generation of Long cDNA fragments for Gene Identification). The availability of our SNP-associated alternative tag database will certainly improve the interpretation of SAGE and MPSS experiments.

INTRODUCTION

The determination of gene expression profiles under normal and pathological conditions is one of the major challenges of the post-genomic era (1,2). A key point for this achievement is the development of techniques that are able to detect all transcripts expressed in a cell population in an unbiased manner, and to precisely determine significant differences in the

expression level of all transcripts, including those expressed at very low levels (2).

SAGE (3) and MPSS (4) are powerful techniques developed for a genome-wide analysis of gene expression. Both methods are capable of uniformly analyzing gene expression irrespective of mRNA abundance and without a priori knowledge of the transcript sequence. In the SAGE technique, a short sequence tag with 10 nt adjacent to the 3'-most NlaIII restriction site is extracted from each expressed sequence (3). The extracted tags are then concatenated for high-throughput sequencing analysis and tag counts are used to measure the relative abundance of their corresponding transcripts. Usually >50 000 tags are generated within a single SAGE experiment.

Similar to SAGE, MPSS also relies on the production of short tags adjacent to the 3'-most DpnII restriction site in transcripts (4). However, due to the combination of *in vitro* cloning of cDNA molecules on the surface of microbeads (5) with non-gel-based high-throughput signature sequencing, a single MPSS experiment can generate over 10⁷ tags, providing a 10-fold coverage of the transcripts expressed in a human cell (4).

SAGE and MPSS data interpretation relies on efficient computational tools for the extraction and counting of tag sequences from raw sequence files, as well as for establishing comparisons of tag abundances between different libraries (3,4). Another important step in analyzing SAGE and MPSS data is the assignment of experimentally obtained tags to a known human transcript (6). This is achieved through the construction of a tag–transcript reference database (7,8). These databases are usually constructed by scanning publicly available mRNA sequences for the presence of the 3'-most restriction sites for the enzymes used for SAGE (NlaIII) and MPSS (DpnII) library construction. A virtual tag sequence downstream to the restriction site is then extracted from each mRNA sequence and stored, together with sequence annotation, in the tag–transcript reference database. Matching experimentally obtained tags to the tag–transcript reference database reveals the identity of the corresponding transcript (7,8).

A reliable tag to transcript assignment is, thus, crucial for the correct interpretation of SAGE and MPSS data. However,

*To whom correspondence should be addressed at Rua Prof. Antonio Prudente 109, 4th floor, 01509-010 São Paulo, SP, Brazil. Tel: +55 11 3388 3248; Fax: +55 11 3207 7001; Email: anamaria@compbio.ludwig.org.br

tag to transcript assignment is not a straightforward process; since many SAGE and MPSS tags can ambiguously match multiple known transcripts (usually due to the presence of repetitive elements in the 3' UTR of human transcripts) and a significant portion of these tags can have no match to the tag-transcript reference database (7,8). In addition, alternative tags other than the 3'-most predicted virtual tag can be experimentally obtained in SAGE and MPSS experiments. There are both artifactual and biological reasons why alternative tags are generated in SAGE and MPSS experiments. Artifactual alternative tags can be generated during SAGE and MPSS library construction if, e.g. cDNA synthesis is primed from internal poly A stretches within the mRNA sequence, or if digestion with the corresponding restriction enzyme is incomplete, producing an alternative tag which is not adjacent to the 3'-most restriction site (7). On the other hand, genuine alternative tags can be generated in cases of alternative polyadenylation and alternative splicing near the 3' end of the transcript [(7), Galante, P.A.F., Guimarães, G.S., Kirschbaum-Slager, N., Riggins, G.J., Cerruti, J. and De Souza, S.J., manuscript submitted].

In theory, alternative tags can also be associated with the presence of SNPs within the tag sequence or within the restriction enzyme sites used for SAGE and MPSS library construction. SNPs are the most common genetic variation present in the human genome, occurring once every 100–300 bases (9–12). In this work, we have evaluated the impact of SNPs on the generation of genuine alternative SAGE and MPSS tags. For the purpose of this analysis, we have considered single base substitutions and small insertion/deletion polymorphisms as SNPs and have named the genuine alternative tags as SNP-associated alternative tags.

The identification of SNP-associated alternative tags was achieved through the construction of a reference database in which the analysis of mRNA sequences from UniGene was combined with information available from the NCBI SNP database. Our results highlight the importance of considering the occurrence of SNPs in tag to transcript assignments, since at least one SNP-associated alternative tag was observed for 8.6% of all known human genes. Our reference database contains 2020 SNP-associated alternative tags and can be accessed through SAGE Genie (<http://cgap.nci.nih.gov/SAGE/>).

MATERIALS AND METHODS

A reference database for SNP-associated alternative tags

A total of 130 148 mRNA sequences catalogued at UniGene (Build #163) and 5 789 183 SNPs from the NCBI-SNP database (Build #118) were mapped onto the publicly available human genome sequence (Build #134). The mapping of mRNA sequences to the human genome was carried as previously described (13,14), and the mapping of SNPs was achieved through the alignment of sequences flanking the SNPs according to the NCBI criteria for SNP mapping (<http://www.ncbi.nlm.nih.gov/SNP>). We have considered single base substitutions and small insertion/deletion polymorphisms as SNPs, and restricted our analysis to SNPs mapped only once to the human genome sequence. A MySQL database was loaded with mapping information of all mRNAs and SNPs that shared an overlap in genomic coordinates. To accurately represent the 3' end of a transcript, only

mRNA sequences containing a poly-A tail were selected from the initial set of 130 148 sequences. A total of 54 645 mRNA sequences (corresponding to 20 300 human genes according to UniGene) was scanned for the presence of NlaIII (for the SAGE analysis) and DpnII (for the MPSS analysis) restriction sites, and virtual tags downstream to the 3'-most site were extracted and considered as the original tags. The analysis was conducted with the dataset containing the coordinates of SNPs, restriction site and original tag position for each mRNAs sequence.

The identification of SNP-associated alternative tags was divided into three major categories as illustrated in Figure 1. First, we identified mRNA sequences in which the presence of an SNP generated a new restriction site downstream to the original tag, producing in this way a 3' SNP-associated alternative tag. Second, we identified mRNA sequences in which the presence of an SNP disrupted the 3'-most restriction site associated with the original tag and, as a consequence, the restriction site immediately upstream to the 3'-most site was used for the generation of the SNP-associated alternative tag. Finally, we identified mRNA sequences in which the SNP did not affect the restriction sites, but occurred within the adjacent tag sequence, producing an SNP-associated alternative tag with a single base substitution as compared to the original tag. SNP-associated alternative tags catalogued in the reference database were compared to a list of experimentally obtained SAGE and MPSS tags.

Experimental SAGE and MPSS databases

SAGE and MPSS tags that have been reliably obtained from human mRNA samples were used as experimental evidence to validate the SNP-associated alternative tags. The criteria for the selection of these tags were previously described (7,15). Experimental SAGE data was obtained from SAGE Genie (<http://cgap.nci.nih.gov/SAGE>) and comprised 586 144 unique tags generated from 260 SAGE libraries, which

Figure 1. The impact of SNPs on tag to gene assignments. For the analysis, SNPs were divided into three major categories: (A) SNPs that generate a new restriction enzyme site downstream to the original tag; (B) SNPs that disrupted the 3'-most restriction site associated with the original tag; (C) SNPs that did not affect the restriction sites, but occurred within the adjacent tag sequence. Restriction sites are represented by gray boxes, original tags by hatched boxes and SNP-associated alternative tags by open boxes. The location of the SNPs within mRNA sequences is indicated by arrows.

were derived from 25 different tissues. MPSS experimental data was extracted from the Ludwig Institute for Cancer Research and the National Cancer Institute MPSS database, and comprise 84 555 unique tags generated from six MPSS libraries derived from two different tissues (colon and breast).

Experimental validation of SNP-associated alternative MPSS tags

The specificity of five SNP-associated alternative MPSS tags (Table 4), derived from the HB4a breast cell line was experimentally confirmed by GLGI-MPSS (16). This technique allows the conversion of MPSS tags into their corresponding 3' cDNA fragments. A sense primer including 17 bases of the MPSS tag sequence and an antisense primer (ACTATCTAGAGCGGCCGCTT) present in the 3' end of all cDNA molecules and incorporated from reverse transcription primers were used for GLGI-MPSS amplification. The reaction mixture was prepared in a final volume of 30 μ l, including 1 $\times$ Taq Platinum DNA polymerase buffer (Invitrogen), 2.0 mM MgCl₂, 83 μ M dNTPs, 2.3 ng/ μ l antisense primer, 2.3 ng/ μ l sense primer, 1.5 U of Taq Platinum DNA polymerase (Invitrogen) and 0.5–0.8 μ l of the same cDNA source used for MPSS library construction. PCR conditions used for amplification were 94°C for 2 min, 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 min after the last cycle. The amplified products were ethanol precipitated and cloned into the pGEM[®]-T Easy vector (Promega). Eight 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) and an ABI3100 sequencer (Applied Biosystems). Sequences were searched against GenBank (nr and dbEST databases) using BLASTN (<http://www.ncbi.nlm.nih.gov/BLAST/>) to confirm the identity of the fragments.

SNP typing

All SNPs associated with the MPSS alternative tags selected for GLGI-MPSS analysis were typed. Four SNPs (rs1053941, rs2362587, rs6961 and rs7110) were typed by genomic DNA amplification followed by restriction digestion with DpnII, and the remaining SNP (rs2422) was typed by direct DNA sequencing since the restriction analysis was not possible due to the presence of several DpnII sites within the amplified sequence. For both genotyping strategies, genomic DNA (~100 ng) from the HB4a cell line was amplified by PCR using primers flanking the SNPs (rs1053941 FW 5'-GAT GGT TCT TGT CCT ATA TC-3', rs1053941 REV 5'-CAG CCT AAG ACC CCA CT-3', rs2362587 FW 5'-AGC ACA GGC CTG GTT AC-3', rs2362587 REV 5'-TGT ATG GCT CCA TGG TCC-3', rs2422 FW 5'-GAG CTT GGA AGA TGG CG-3', rs2422 REV 5'-CAT TCC TCT TTC AAA CAG CC-3', rs6961 FW 5'-TGA ATG TCA TGC TGG TGC-3', rs6961 REV 5'-AGA GTG CAG AAG CGT ATG-3', rs7110 FW 5'-GCA ACC CTA GCA ATA CCA-3', rs7110 REV 5'-TAG CAG TGA CCT AAG TCC-3'). The amplification mixture was prepared in a final volume of 25 μ l, containing 1 $\times$ Taq Platinum DNA polymerase buffer (Invitrogen), 1.4 mM MgCl₂, 0.1 mM dNTPs, 20 μ M of each primer and 1 U of Taq DNA polymerase (Invitrogen). PCR conditions used for amplification were 94°C for 4 min, followed by

40 cycles at 94°C for 40 s, 57°C for 40 s and 72°C for 1 min. Reactions were kept at 72°C for 6 min after the last cycle. The amplified products were then either digested with DpnII or used for direct sequencing as described above. DpnII digestion was carried in a final volume of 20 μ l, including 1 $\times$ buffer, 10 U of DpnII (New England Biolabs) and 4 μ l of each PCR product. Reactions were kept at 37°C for 2 h and were analyzed on 8% polyacrylamide gel stained with silver.

RESULTS

In theory, genuine alternative tags can be associated with the presence of SNPs within the tag sequence or within the restriction enzyme sites used for SAGE and MPSS library construction. To analyze the impact of SNPs on the generation of genuine alternative SAGE and MPSS tags, we have constructed a reference database of SNP-associated alternative tags. For the construction of this database 54 645 mRNA sequences containing a poly-A tail and corresponding to 20 300 UniGene clusters (Build #163) were initially scanned for the presence of NlaIII and DpnII restriction sites. Of the 54 645 mRNA sequences analyzed, 54 124 (99.0%) contained an NlaIII restriction site and 52 779 (96.6%) contained a DpnII site. mRNA sequences were then searched for the presence of SNPs according to the NCBI SNP Database (Build #118). Of the 54 124 mRNA sequences presenting NlaIII sites, 44 033 (81.4%) contained at least one SNP and of the 52 779 mRNA sequences with DpnII sites, 43 125 (81.7%) contained at least one SNP. The analysis for the identification of SNP-associated alternative tags was divided into three major categories as illustrated in Figure 1, and described in Materials and Methods.

Creation of a new 3'-most restriction site

First, we have analyzed mRNA sequences in which the presence of an SNP generated a new restriction enzyme site downstream to the original tag, producing in this way a 3' SNP-associated alternative tag (Figure 1A). From the 44 033 mRNA sequences containing both an NlaIII site and an SNP, we have identified 573 (1.3%) sequences in which the presence of SNPs created a new NlaIII restriction site downstream to the original tag. These 573 mRNA sequences correspond to 294 unique human genes according to UniGene. A total of 305 unique SNP-associated alternative SAGE tags were extracted from these 573 mRNA sequences (Table 1). A similar analysis was carried for the 43 125 mRNA sequences containing both a DpnII site and an SNP. In this case, we have identified 393 (0.9%) mRNA sequences corresponding to 205 UniGene clusters. The presence of an SNP within these sequences generated 217 unique SNP-associated alternative MPSS tags (Table 1). We also included in this category 56 mRNA sequences that did not have an NlaIII or a

Table 1. SNP-associated tags generated by the creation of a new restriction enzyme site downstream of the position of the original tag

	SNP-associated alternative tag	Experimentally obtained tags	Ambiguous tags
SAGE analysis	305	275 (90.2%)	38 (12.8%)
MPSS analysis	217	40 (18.4%)	7 (3.2%)
Both analysis	522	315 (60.3%)	45 (8.6%)

DpnII restriction site, but acquired one because of the presence of an SNP.

In order to validate the existence of these SNP-associated alternative tags, we have compared them to a list of experimentally obtained SAGE and MPSS tags. This list included 586 144 unique SAGE tags derived from 260 SAGE libraries and 84 555 unique MPSS tags derived from six MPSS libraries. Of the 305 SNP-associated alternative SAGE tags catalogued in our reference database, 275 (90.2%) were found in the list of experimentally obtained SAGE tags, and of the 217 SNP-associated alternative MPSS tags, 40 (18.4%) matched the list of experimentally obtained MPSS tags (Table 1).

However, the presence of an SNP-associated alternative tag within a dataset of experimentally obtained tags is not always an irrefutable evidence for existence of the alternative tag and can also occur in cases of tag sequence ambiguity, when two distinct transcripts contain by chance an identical tag sequence. In order to further validate our analysis, we have determined the percentage of the SNP-associated alternative tags that also correspond to the 3'-most original tag of a distinct human transcript. A small percentage (12.8%) of the 305 SNP-associated alternative SAGE tags, corresponded to the 3' original tag of another transcript. This percentage was even smaller (3.2%) for the SNP-associated alternative MPSS tags due to the longer size and higher specificity of the tag sequence (Table 1). The presence of a high percentage of unambiguous SNP-associated alternative tag within a list of experimentally obtained SAGE and MPSS tags can thus be used to show that the occurrence of SNPs within transcript sequences is, indeed, an important source for the generation of alternative tags.

Destruction of the original 3' restriction site

We have then analyzed mRNA sequences in which the presence of an SNP disrupted the 3'-most restriction site associated with the original tag and, as a consequence, the second 3'-most site was used for the generation of the SNP-associated alternative tag (Figure 1B). We have identified, from the 44 033 mRNA sequences containing both an NlaIII site and an SNP, 498 (1.1%) sequences in which the presence of an SNP disrupted the 3'-most restriction site associated with the original tag. These 498 mRNA sequences correspond to 236 unique human genes according to UniGene, and a total of 235 unique SNP-associated alternative SAGE tags were extracted from them (Table 2). Of these 235 tags, 218 (92.8%) matched our list of experimentally obtained SAGE tags, and only a small fraction (13.2%) corresponded to the 3'-most original tag of another transcript, thus validating them as genuine SNP-associated alternative tags (Table 2).

For the 43 125 mRNA sequences containing both a DpnII site and an SNP, we have identified 422 (1%) sequences in

which the 3'-most restriction site was disrupted by an SNP. These 422 mRNA sequences correspond to 208 UniGene clusters and the presence of an SNP within these sequences generated 196 unique SNP-associated alternative MPSS tags of which 78 (39.8%) were found within experimentally obtained MPSS tags. As expected, the frequency of tag ambiguity for these SNP-associated alternative MPSS tags was very low (4.6%) (Table 2).

Single base substitutions within original tag sequence

Finally, we have analyzed mRNA sequences in which the SNP did not affect the restriction sites, but occurred within the adjacent tag sequence, producing an SNP-associated alternative tag with a single base substitution as compared to the original tag (Figure 1C). From the 44 033 mRNA sequences containing both an NlaIII site and an SNP, we have identified 1136 (2.6%) sequences in which the presence of an SNP occurred within the adjacent tag sequence. These 1136 mRNA sequences correspond to 543 unique human genes according to UniGene, and generated 560 unique SNP-associated alternative SAGE tags (Table 3). Of these 560 tags, 512 (91.4%) were found within the experimentally obtained SAGE tag dataset and only 92 (16.4%) also corresponded to the 3'-most original tag of another transcript (Table 3).

Similarly, for the 43 125 mRNA sequences containing both a DpnII site and an SNP, we have identified 1009 (2.3%) mRNA sequences in which the presence of an SNP occurred within the adjacent tag sequence. These 1009 mRNA sequences correspond to 481 unique human transcripts according to UniGene, and generated 507 unique SNP-associated alternative MPSS tags of which 127 (25.0%) were experimentally obtained and 33 (6.5%) corresponded to the 3'-most tag of another transcript (Table 3).

Integration of the reference database of SNP-associated alternative tags to SAGE Genie

To make our analysis accessible to the research community, we have integrated the database of SNP-associated alternative SAGE tag to SAGE Genie (<http://cgap.nci.nih.gov/SAGE>) (7). The data can be directly downloaded as flat-files or visualized in the 'Ludwig Transcript Viewer' as exemplified in Figure 2. The information related to the existence of an SNP-associated alternative tag for a given transcript is presented in the 'Ludwig Transcript Viewer' as a separated table just below the schematic representation of the transcript sequence and its corresponding virtual SAGE tags. This table describes the impact of the SNP on the transcript sequence (e.g. creates a new 3'-most NlaIII site), and includes additional information related to the SNP-associated alternative tag, such as its sequence and position within the transcript sequence and its

Table 2. SNP-associated tags generated by the disruption of the 3'-most restriction site associated with the original tag SNP

	SNP-associated alternative tag	Experimentally obtained tags	Ambiguous tags
SAGE analysis	235	218 (92.8%)	31 (13.2%)
MPSS analysis	196	78 (39.8%)	9 (4.6%)
Both analysis	431	296 (68.7%)	40 (9.3%)

Table 3. SNP-associated tags generated by SNPs that occurred within the adjacent tag sequence

	SNP-associated alternative tag	Experimentally obtained tags	Ambiguous tags
SAGE analysis	560	512 (91.4%)	92 (16.4%)
MPSS analysis	507	127 (25.0%)	33 (6.5%)
Both analysis	1067	639 (59.9%)	125 (11.7%)

frequency in the SAGE Genie database. Information about the SNP related to the alternative tag, such as the SNP accession number, the base substitution and the position of the SNP in the transcript sequence is also provided in the table, as well as a direct link to the NCBI-SNP database. Using the 'Ludwig Transcript Viewer', the SAGE Genie user can now easily check whether an mRNA sequence presents an SNP-associated alternative tag or if a specific SAGE tag corresponds to an SNP-associated alternative tag of a known human gene.

Experimental validation of SNP-associated alternative MPSS tags

To further confirm the impact of SNPs on the generation of genuine alternative tags, we have used the GLGI-MPSS technique (16) to convert five SNP-associated alternative tags observed in the HB4a MPSS library into their corresponding 3' cDNA fragments. These extended 3' cDNA fragments were then used in similarity searches against public databases in order to confirm their specificity (Table 4).

Figure 2. Integration of the database of SNP-associated alternative SAGE tags into SAGE Genie. A representative example of the Ludwig Transcript Viewer showing the transcript encoded by the MAF1 gene (NM_032272) as a blue line and the colored boxes represent the last four virtual tags relative to the 3' end of the transcript. The expression levels for each of the four virtual tags as well as the tag position in the transcript sequence are provided in the Tag Info Summary. The existence of an SNP-associated alternative tag for the MAF1 transcript is in a specific table (as indicated by the arrow), which includes the tag sequence, tag position within the transcript, tag frequency in the SAGE Genie database, the SNP id associated with the alternative tag, the base substitution and the position of the SNP within the transcript sequence.

Table 4. Experimental validation of SNP-associated MPSS tags by GLGI-MPSS

Accession no.	SNP ID	SNP position	SNP	SNP-associated alternative tag ^a	Alternative tag count	Original tag	Original tag position	Original tag count
NM_002482	rs1053941	3039	T/G	GATCTCTGGTTTGAAG	9	GATCTTGCTCTTCAGTG	2862	6
AK023594	rs2362587	1288	G/A	GATCATAAGCAGCAATT	2	GATCTGCAACTCTTTCA	889	5
D86973	rs2422	8494	T/C	GATCCCTGTCAGATGAA	113	GATCTGACCCCTGTGAC	8487	0
NM_004168	rs6961	1956	G/A	GATCGACAAAACCTTTGA	73	GATCAGATTGTGCCCGG	1309	68
AK092889	rs7110	3205	G/A	GATCATGTTTACAGACC	5	GATCAGCACTGCAGCAA	2516	6

^aThe bold character corresponds to the polymorphic nucleotide, which created the DpnII restriction site and generated the SNP-associated alternative tag.

A sense primer corresponding to the SNP-associated alternative MPSS tag was used for GLGI-MPSS amplification as described in Materials and Methods. As can be seen in Figure 3, a predominant band was obtained for all GLGI-MPSS reactions. Bands were excised from the gel, cloned, sequenced and searched for sequence similarity against GenBank (nr and dbEST). With the exception of the SNP-associated alternative tag corresponding to the AK092889 transcript, all the others were validated and produced a 3' cDNA fragment matching the expected transcript sequence and confirming the origin of the SNP-associated alternative tag. The 3' cDNA fragment generated with the SNP-associated alternative tag corresponding to the AK092889 transcript matched an unrelated cDNA sequence (BC064564) in which the sequence corresponding to the alternative tag could not be found. This fragment should then be considered as an artifact generated by unspecific GLGI-MPSS amplification.

We then decided to genotype the HB4a cell line for the presence of the SNPs associated with these alternative tags. All of these five selected SNPs created a new restriction enzyme site downstream of the position of the original tag. Primers flanking the SNPs were designed and used to amplify HB4a genomic DNA. Amplified fragments were either digested with DpnII or used for direct sequencing. As can be seen in Figure 4, the occurrence of the four SNPs (rs1053941, rs2362587, rs6961 and rs7110) in the HB4a cell line was confirmed after restriction digestion. The observed restriction digestion pattern suggests that the HB4a cell line is heterozygous for all SNPs analyzed by restriction digestion.

Figure 3. GLGI-MPSS amplifications of five SNP-associated alternative MPSS tags. GLGI-MPSS amplifications for SNP-associated alternative MPSS tags listed in Table 4 and corresponding to the mRNA sequences NM_002482 (1); AK023594 (2); D86973 (3); NM_004168 (4); AK092889 (5) were analyzed on 1% agarose gel stained with ethidium bromide; 100 bp ladder (M) was used as molecular weight marker.

Figure 4. SNP typing by genomic DNA amplification followed by restriction enzyme digestion. The genomic region flanking the SNPs rs1053941 (SNP1), rs2362587 (SNP2), rs6961 (SNP4) and rs7110 (SNP5) was amplified using specific primers and genomic DNA from the HB4a cell line. PCR fragments were digested with DpnII, and analyzed on 8% polyacrylamide gels stained with silver; 100 bp ladder (M) was used as molecular weight marker and bands corresponding to the restriction fragments are indicated by arrows.

The presence of the remaining SNP (rs2422) was confirmed by direct sequencing (data not shown), and the HB4a turn out to be homozygous for this polymorphism. According to the SNP genotyping results in cases of heterozygosis, the occurrence of both the original and the SNP-associated tags within the HB4a MPSS library is expected. As shown on Table 4, both the original and SNP-associated tags were found in the HB4a MPSS library at approximately the same frequency for all cases of heterozygosis.

DISCUSSION

The impact of alternative polyadenylation and alternative splicing on the generation of genuine alternative tags has already been studied [(7), Galante, P.A.F., Guimarães, G.S., Kirschbaum-Slager, N., Riggins, G.J., Cerruti, J. and De Souza, S.J., manuscript submitted]. However, alternative tags can also be generated by the presence of SNPs within the tag sequence or within the restriction enzyme sites used for SAGE and MPSS library construction.

In this work, we found that the presence of SNPs within human mRNA sequences was responsible for the generation of 2020 SNP-associated alternative tags (1100 SNP-associated alternative SAGE tags and 920 SNP-associated alternative MPSS tags) and that 8.6% of all known human genes present at least one SNP-associated alternative tag. It should be noted, however, that this number is certainly underestimated because the growth of the NCBI SNP database has not yet reached a plateau (statistics available at http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi) suggesting that just a fraction of the whole repertoire of SNPs present in the human genome is so far reported. On the other hand, one should also be aware that only a fraction of the NCBI-SNP database (40.7%) is validated, what could potentially lead to the identification of artifactual SNP-associated alternative tags. However, the SNPs included in our database show the same proportion of validation (44.5%), suggesting that they are a fair representation of the whole set of SNPs. In spite of that, we expect that

the growth in the collection of validated SNPs, as well as the availability of information related to allele frequencies in specific populations will enrich and better refine the analyses presented here. We have restricted our analysis to 54 645 mRNA sequences containing a poly-A tail and, thus, considered to represent the 3' end of a transcript. However, poly-A tails are sometimes removed from transcript sequences during the database submission process, and it is likely that among the 75 503 mRNA sequences excluded from our analysis (corresponding to 8677 UniGene clusters) there are several representing genuine 3' transcript ends. If we analyze all the 130 148 mRNA sequences catalogued at UniGene, irrespective of the presence of a poly-A tail, a total of 3520 SNP-associated alternative tags can be identified (1950 SNP-associated alternative SAGE tags and 1570 SNP-associated alternative MPSS tags). This number corresponds to an increase of 74.3% in the analysis using sequences with poly-A tail.

Interestingly, we did not observe a significant decrease in the percentage of SNP-associated alternative tags that are experimentally documented. Approximately 62% of the SNP-associated alternative tags identified from mRNA sequences with a poly-A tail were experimentally obtained and this number decrease to 57.8% if we consider all mRNA sequences. Taken together, these results suggest that the majority of the mRNA sequences without a poly-A tail indeed represent a genuine 3' transcript end and that the number of SNP-associated alternative tags reported here is conservative and probably underestimated.

A significant fraction (91.4%) of the SNP-associated alternative SAGE tags were found within the CGAP SAGE database, but only 26.6% of the SNP-associated alternative MPSS tags could be experimentally obtained. This difference can be explained by the larger number of SAGE libraries used to generate our list of experimentally validated tags. If we assume that each of these SAGE and MPSS libraries were derived from a single individual, the genetic variability represented within the SAGE dataset of experimentally obtained tags (extracted from 260 different libraries) is much higher than that represented within the MPSS dataset (extracted from six different libraries), thus increasing the chance of a given polymorphism (and consequently the corresponding SNP-associated alternative tag) being represented within the experimentally obtained dataset.

To overcome the problem of the limited number of MPSS libraries available, we have further confirmed the origin of five SNP-associated alternative MPSS tags found in the HB4a MPSS library through the use of GLGI-MPSS (16). With the exception of the SNP-associated alternative tag corresponding to the AK092889 mRNA, all the others produced a GLGI-MPSS 3' cDNA fragment matching the expected known human mRNA. The HB4a cell line was also genotyped for the presence of the SNPs and shown to be heterozygous for 4 out of the 5 SNPs analyzed. As expected, both the original and SNP-associated tags were found in the HB4a MPSS library in the cases of heterozygosis. These preliminary results suggest that the existence of SNP-associated alternative tags can be used to study allele-specific gene expression. Allele-specific variations in gene expression have been classically associated with X-chromosome inactivation and genomic imprinting, and recent studies have also shown that it is relatively common among

non-imprinted autosomal genes (17,18). We are currently using our reference database of SNP-associated alternative tags to study allele-specific gene expression in a genome-wide context. To enhance the utility of the analysis presented in this work, we have integrated our database of SNP-associated alternative tags into SAGE Genie, a web site for the analysis and presentation of SAGE data (7). SNP-associated alternative tags can now be easily identified and correctly assigned to human transcripts allowing an improvement of the interpretation of SAGE experiments. Planned updates of our reference database with sequence data generated from full-length cDNA sequencing projects as well as with new releases of the NCBI SNP database will increase the accuracy of our analysis. These updates will be periodically available through SAGE Genie and will certainly improve the interpretation of SAGE and MPSS experiments.

ACKNOWLEDGEMENTS

The authors would like to thank Daniela Gerhard, Susan Greenhut and Carl Schaefer from the National Cancer Institute for help in making our data available through SAGE Genie. Funding was provided by the CEPID Program from the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP 98/14335-2). The Ludwig Institute for Cancer Research and the National Cancer Institute funded the construction of the MPSS libraries for breast and colon cell lines, respectively.

REFERENCES

1. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
2. Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998–2003. *Science*, **282**, 682–689.
3. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
4. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
5. Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.I., Luo, S., Kirchner, J.J., Eletr, S. *et al.* (2000) *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl Acad. Sci. USA*, **97**, 1665–1670.
6. Madden, S.L., Wang, C.J. and Landes, G. (2000) Serial analysis of gene expression: from gene discovery to target identification. *Drug Discov. Today*, **9**, 415–425.
7. Boon, K., Osorio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Buetow, K.H., Strausberg, R.L., De Souza, S.J., *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
8. Clark, T., Lee, S., Ridgway, S.L. and Wang, S.M. (2002) Computational analysis of gene identification with SAGE. *J. Comput. Biol.*, **9**, 513–526.
9. Wang, D.G., Fan, J.B., Siao, C.J., Bero, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.
10. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.*, **22**, 231–238.
11. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
12. Shastri, B.S. (2002) SNP alleles in human disease and evolution. *J. Hum. Genet.*, **47**, 561–566.
13. Galante, P.A., Sakabe, N.J., Kirschbaum-Slager, N. and de Souza, S.J. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **5**, 757–765.
14. Sakabe, N.J., de Souza, J.E., Galante, P.A., de Oliveira, P.S., 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 proteins. *C R Biol.*, **326**, 979–985.
15. Jongeneel, C.V., Iseli, C., Stevenson, B.J., Riggins, G.J., Lal, A., Mackay, A., Harris, R.A., O'Hare, M.J., Neville, A.M., Simpson, A.J. *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.
16. Silva, A.P.M., Chen, J., Carraro, D.M., Wang, S.M. and Camargo, A.A. (2004) Generation of longer 3' cDNA fragments from Massive Parallel Signature Sequencing Tags. *Nucleic Acids Res.*, **32**, e94.
17. Knight, J.C. (2004) Allele-specific gene expression uncovered. *Trends Genet.*, **20**, 113–116.
18. Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H. and Lee, M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.*, **13**, 1855–1862.