IDENTIFICAÇÃO DE NOVOS TRANSCRITOS NO GENOMA HUMANO UTILIZANDO DADOS DE ALINHAMENTO ENTRE SEQÜÊNCIAS EXPRESSAS E A SEQÜÊNCIA GENÔMICA

FABIANA BETTONI

Dissertação apresentada à Fundação Antônio Prudente para a obtenção do título de Mestre em Ciências

Área de concentração: Oncologia

Orientador:
Dr. Anamaria Aranha Camargo
Co-Orientador:
Dr. Andrew John George Simpson

São Paulo
2003

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A. C. Camargo

Bettoni, Fabiana

Identificação de novos transcritos no genoma humano utilizando dados de alinhamento entre seqüências expressas e a seqüência genômica / Fabiana Bettoni – São Paulo, 2003.

154p.

Dissertação(mestrado)-Fundação Antônio Prudente.

Curso de Pós-Graduação em Ciências-Área de concentração: Oncologia.

Orientador: Anamaria Aranha Camargo

Descritores: 1. GENOMA HUMANO. 2.CROMOSSOMO 21. 3. GENES/identificação. 4. SEQÜÊNCIAS-ETIQUETAS EXPRESSAS. 5. BIOLOGIA COMPUTACIONAL.

"O que mais vale na vida é procurar dar um pouco de felicidade à vida dos outros."

B. Powell

DEDICATÓRIA

Aos meus pais e ao meu irmão, que juntos me deram força para nunca desanimar e desistir de meus sonhos. Agradeço por tudo que fizeram e renunciaram para que eu pudesse chegar até aqui e dizer que esta vitória é dedicada, com todo meu amor, a vocês.

Às minhas “nonas”, que mesmo com todas as suas dificuldades nunca deixaram de orar por mim. Apesar da distância sei que vocês sempre estiveram e estarão ao meu lado.

À minha orientadora Dra. Anamaria A. Camargo, por ter confiado em meu potencial e por representar para mim muito mais do que uma excelente profissional. Muito obrigada por ser uma amiga e mãe maravilhosa. Sem o seu apoio, a sua confiança e, principalmente, os seus ensinamentos, com certeza eu não teria chegado até aqui.

AGRADECIMENTOS

Ao Dr. Andrew J. G. Simpson pela oportunidade de ser sua aluna e poder contar com suas extraordinárias idéias e sugestões.

Ao Raphael B. Parmigiani, por ter me ensinado grande parte de tudo o que sei hoje. Muito obrigada pela sua sincera amizade, paciência, atenção e por ter ocupado várias horas de seu tempo com a correção desta tese.

À Lilian C. Pires, pela amizade sincera de todos esses anos de convivência. Obrigada pelo incentivo constante e por sempre estar do meu lado nos momentos mais difíceis. Em especial, agradeço ao meu primeiro e único "sobrinho" Vinícius pela leitura e "correção" desta tese.

À Mariana Brait, por estar sempre presente com sua amizade e atenção tornando inesquecíveis estes anos de convívio. Agradeço, ainda, ao seu pai e a suas duas mães por me receberem com tanto carinho. Sem o apoio de todos vocês com certeza tudo teria sido mais difícil.

À Márcia Dellamano, pela descoberta de uma grande amizade e pelo convívio diário no trabalho e em casa. Obrigada pelo apoio e compreensão, especialmente, durante a redação deste trabalho.

À Dirce Carraro, pelos importantes ensinamentos e conselhos além das conversas nos momentos de descontração e das salvadoras caronas de final de semana.

À todas as pessoas do Laboratório de Biologia Molecular e Genômica, Raphael, Lilian, Elisângela, Dirce, Anna Christina, Elisa, Jane, Élen, Mariana, Ana Paula, Jasna, Murilo, Fernando, Ricardo, Daniela, Newton e Fabrício, pela convivência e companheirismo. Obrigada por todo apoio e ajuda oferecidos no desenvolvimento deste trabalho.

Aos demais colegas do Instituto Ludwig pela convivência e amizade.

Às amigas Milena, Paula, Juliana e Denise, por compreenderem os tempos sem notícia, telefonemas e e-mails. Obrigada por torcerem por mim e por estarem ao meu lado em momentos marcantes como este.

A todos os amigos presentes nesta importante fase da minha vida. Obrigada pelo apoio e pelos momentos de descontração que com certeza contribuíram para que eu chegasse até aqui.

A todos os participantes do projeto "Transcript Finishing Initiative", pelos conhecimentos adquiridos e novas amizades conquistadas.

Ao Prof. Dr. Sandro de Souza e todo o seu Laboratório de Biologia Computacional pela ajuda com o Banco de Dados do Transcriptoma. Em especial, ao Jorge pela disponibilidade e boa vontade com o projeto TFI, ao Élisson pela importante ajuda com a análise dos bancos de dados de SAGE e à Natanja pelo auxílio com a análise das formas alternativas de "splicing".

Ao Prof. Dr. Milton Faria Jr e todo o seu laboratório de Bioinformática pela disponibilidade, atenção e paciência nestes dois anos do projeto TFI.

Ao Prof. Dr. André Vettore, Profa. Dra. Otávia Cabalero, Adriana Bulgarelli e Márcia Dellamano pelo importante auxílio com as análises no Real Time.

A todos os funcionários do Instituto Ludwig pelo suporte técnico e administrativo além da convivência diária. Em especial à Ana Cláudia, pela disponibilidade e auxílio nos problemas do dia-a-dia, e à Eliane, pela atenção e, principalmente, por ter me apresentado à Dra. Anamaria A. Camargo.

À Márcia Hiratoni e Ana Maria Kuninari, por me receberem com tanta disponibilidade e atenção e, também, pelo trabalho desenvolvido junto à secretaria de pós-graduação.

À Suely Francisco e todos os funcionários da biblioteca da Fundação Antonio Prudente pela disponibilidade e excelente ajuda na organização deste material bibliográfico.

Ao Prof. Dr. Ricardo R. Brentani, por administrar o Instituto Ludwig de Pesquisa sobre o Câncer e o Hospital do Câncer AC Camargo, permitindo que este trabalho fosse desenvolvido em um centro de excelência na área de pesquisa sobre o câncer.

Ao Prof. Dr. Luiz Fernando Lima Reis pela coordenação deste curso de pós-graduação.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela bolsa de estudo concedida e à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) pelo financiamento do projeto “Transcript Finishing Initiative” em colaboração com o Instituto Ludwig.

À todas as professoras da Escola Especializada "Schwester Heine" por terem me proporcionado momentos inesquecíveis junto às crianças deste hospital. Em especial agradeço a todos estes pequenos pacientes que com seus abraços e sorrisos sinceros me permitiram descobrir o verdadeiro valor deste trabalho.

RESUMO

Bettoni F. **Identificação de novos transcritos no genoma humano utilizando dados de alinhamento entre seqüências expressas e a seqüência genômica.** São Paulo; 2003. [Dissertação de Mestrado-Fundação Antonio Prudente]

O principal objetivo do Projeto Genoma Humano é a identificação de todos os genes humanos. No entanto, devido à complexidade do genoma, o processo de identificação gênica a partir da seqüência genômica não é direto e requer a utilização de estratégias complementares. Nesta dissertação, descrevemos a utilização de alinhamentos entre seqüências expressas e a seqüência genômica para a identificação de novos transcritos humanos. Inicialmente, utilizando a seqüência genômica como base para o mapeamento e clusterização das ESTs, foram realizadas RT-PCRs para o fechamento dos "gaps" existentes entre "clusters" de ESTs que derivam de um mesmo gene. Este processo foi chamado "Transcript Finishing" e gerou um total de 59.975pb de seqüências transcritas definindo a estrutura de 211 transcritos. As seqüências validadas apresentaram em média 6,1 ESTs em cada "cluster" de origem derivadas, em média, de 3,45 tecidos diferentes. Aproximadamente 21% destas seqüências representaram transcritos novos sendo que a maioria desses não foi corretamente predita por programas computacionais. A estratégia do TFI demonstrou ser particularmente eficiente tanto na caracterização de transcritos pouco abundantes e expressos em um número restrito de tecidos como também na determinação da estrutura gênica e de formas alternativas de "splicing". Os dados de alinhamento entre seqüências expressas e a seqüência genômica também foram utilizados na

identificação de novos transcritos localizados no cromossomo 21, uma vez que a identificação de todos os genes desse cromossomo é necessária para um melhor entendimento da patologia molecular da Síndrome de Down e de outras doenças mapeadas nesse cromossomo. Através de análises *in silico* seguidas de validação experimental, foi possível identificar 19 novos transcritos e 4 unidades transcricionais, que apresentavam "splicing" mas não continham uma ORF aparente. A estratégia adotada nesse trabalho é bastante direta se comparada a estratégias tradicionais e se mostrou extremamente eficiente, elevando em 10% o número total de genes no HC21. Baseando-se no perfil de expressão dos novos transcritos nós pudemos identificar dois possíveis candidatos (C21orf99 e C21orf100) a antígenos tumorais. Análise de expressão diferencial entre tecido normal e tumoral também foi realizada para todos os novos transcritos com base no banco de dados de SAGE seguida de validação experimental por PCR em Tempo Real. Em análise preliminar, nós pudemos identificar um transcrito (MCM3APAS) com maior expressão em tecido tumoral de cérebro.

SUMMARY

Bettoni F. **Identificação de novos transcritos no genoma humano utilizando dados de alinhamento entre seqüências expressas e a seqüência genômica** [Identification of new human transcripts through the alingment between transcript sequences and the human genome sequence]. São Paulo; 2003. [Dissertação de Mestrado-Fundação Antonio Prudente]

The main objective of the Human Genome Project is the identification of the complete set of human genes. However, due to the highly complex structure of the genome the process of gene identification based on the genomic sequence is not straightforward and requires the application of complementary strategies. Here we describe the utilization of the alignment between expressed sequences and the genomic sequence to identify novel human transcripts. Initially, using the genomic sequence as a scaffold for EST mapping and clustering we have performed RT-PCR to bridge gaps between EST clusters that are likely to be derived from the same gene. This process was called Transcript Finishing generating a total of 59,975bp transcribed sequences and defining the structure of 211 transcripts. Validated sequences had an average of 6.1 ESTs in each cluster derived, on average, from 3.45 distinct tissues. About 21% of these sequences still represent unreported human transcripts, most of which had not been correctly predicted by computer programs. The TF strategy appears to be particularly useful both for characterization of low abundance transcripts expressed in a restricted set of tissues and for delineation of gene boundaries and alternatively spliced isoforms. The alignment between transcript sequences and the genome

sequence was also used to identify new transcripts on human chromosome 21 (HC21) since the identification of all HC21 genes is a necessary step in understanding the molecular pathogenesis of trisomy 21 (Down syndrome) and other diseases mapped on this chromosome. Through *in silico* analysis followed by experimental validation we were able to identified 19 new transcripts and 4 transcriptional units that are spliced but contain no obvious ORF. The strategy adopted in this work is straightforward in comparison to other traditional strategies and has proven extremely efficient, increasing the HC21 gene count by 10%. Based on the expression pattern of these new transcripts we were able to identify two possible candidates (C21orf99 and C21orf100) to novel tumor antigens. Differential expression in normal and tumor tissues were analyzed for all new transcripts based on SAGE database followed by Real-Time PCR validation. On a preliminary analysis, we were able to identify one transcript (MCM3APAS) over expressed on tumor brain tissue.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

µg – Micrograma

µl – Microlitro

µM – Micromolar

APP – "Amyloid-b Precursor Protein"

ATCC – "American Type Culture Collection"

BAC – Cromossomo artificial de bactéria (do ingles "Bacterial Artificial Chromosome")

BLAST – "Basic Local Alignment Search Tool"

BLAT – "BLAST-Like Alignment Tool"

cDNA – DNA complementar

CGAP – "Cancer Genome Anatomy Project"

CT – "Cycle Treshold"

CYYR1 – "Cysteine and tyrosine-rich protein 1"

C21orf – "Chromosome 21 Open Reading Frame"

dNTP – Deoxinucleotídeo

DEPC – Di-etil pirocarbonato

DMSO – Dimetilsulfóxido

DNA – Ácido desoxirribonucléico

DNAse – Desoxirribonuclease

DO – Densidade Óptica

DSCR8 – "Down Syndrome Critical Region 8"

EDTA – Ácido etilenodiaminotetracetico disódio

ESTs – Etiqueta de seqüências expressas (do inglês "Expressed Sequence Tags")

FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo

GAPDH – Gliceraldeído 3-fosfato desidrogenase

hMLH1 - "human Mut-L Homologue 1"

HCGP – Projeto Genoma Humano do Câncer

HTGS – "High Throughput Genome Sequences"

IHGSC – "International Human Genome Sequencing Consortium"

I.M.A.G.E. - "Integrated Molecular Analysis of Genomes and their Expression"

Kb – kilobases

LINEs - "Long Interspersed Nuclear Elements"

mRNA – RNA mensageiro

Mb – megabases

MCM3AP – "Minichromosome maintenance deficient 3 acetylating protein"

MCM3APAS – "MCM3-associated protein antisense"

MGC – "Mammalian Gene Collection"

ml – Mililitros

MRPS6 - "Mitochondrial ribosomal protein S6"

NCBI – "National Center for Biotechnology Information"

NIA – Instituto Nacional do Envelhecimento (do inglês "National Institute of Aging")

ng – Nanograma

nm – Nanômetro

ONSA – "Organization for Nucleotide Sequencing and Analysis"

ORESTES – "Open Reading Frame ESTs"

ORF – Fase aberta de leitura (do inglês "Open Reading Frame")

pb – pares de bases

PCR – Reação em cadeia da polimerase (do inglês "Polimerase Chain Reaction")

RACE - "Rapid Amplification of cDNA Ends"

RNA – Ácido ribonucléico

RNAse – Ribonuclease

rpm – rotações por minuto

RT-PCR – "Reverse Transcriptase – Polimerase Chain Reaction"

RZPD - "Resource Center German Human Genome Project"

SAGE – "Serial Analysis of Gene Expression"

SDS – Dodecilsulfato de sódio

SINEs - "Short Interspersed Nuclear Elements"

SNPs – Polimorfismos de nucleotídeos únicos (do inglês "Single Nucleotide Polymorphisms")

TFI – "Transcript Finishing Initiative"

TM – "Melting Temperature"

UCSC – "University of California Santa Cruz"

UNAERP – Universidade de Ribeirão Preto

UV – Ultravioleta

ÍNDICE

ANEXOS

Anexo 1 "Transcript Finishing Initiative: Closing Gaps in the Human Transcriptome." Artigo científico submetido à revista Genome Research.

Anexo 2 Quadro referente aos dados de anotação de todas as seqüências consenso geradas pelo projeto TFI.

Anexo 3 Reymond A, Camargo AA, Deutsch S, et al Nineteen Additional Unpredicted Transcripts from Human Chromosome 21. **Genomics** 2002; 79: 824-32.

INTRODUÇÃO

1 INTRODUÇÃO

1.1 O PROJETO GENOMA HUMANO

O termo genoma é designado para representar o conjunto de genes e seqüências regulatórias de um dado organismo (BRODER e VENTER 2000). Sendo assim, os principais objetivos de um Projeto Genoma compreendem a determinação da seqüência completa do DNA de um organismo e a identificação de todos os seus genes.

Independentemente do organismo a ser estudado e da metodologia a ser aplicada, um Projeto Genoma apresenta duas etapas principais: (1) seqüenciamento e montagem computacional das seqüências e (2) identificação dos genes e anotação gênica. De uma maneira geral, pode-se subdividir a primeira etapa em três importantes fases: determinação da seqüência de pequenos fragmentos de DNA, montagem destas seqüências com o auxílio de programas de computador gerando longas seqüências genômicas conhecidas como “contigs” e obtenção da seqüência final através da análise de qualidade da seqüência dos “contigs” e subseqüente produção das seqüências de regiões genômicas descontínuas que separam os “contigs” (fechamento dos “gaps”).

Uma vez conhecida a seqüência genômica nos deparamos com um conjunto infinito de letras A, T, C e G que, na realidade, só terão algum significado quando conseguirmos identificar nesta seqüência o que realmente representa um gene. Desta forma, parte-se para a segunda etapa de um projeto genoma que consiste na identificação de seus genes e suas respectivas funções sendo a primeira realizada por programas computacionais de predição gênica e a segunda através de buscas por

similaridade dos genes identificados com seqüências de nucleotídeos e proteínas depositadas em bancos de dados públicos.

O genoma humano haplóide, especificamente, é dividido em 23 moléculas de DNA, os cromossomos, sendo a mais curta com 46Mb e a maior com 250Mb. Estes 23 cromossomos consistem em 22 autossomos e um cromossomo sexual (X ou Y) e, no total, representam aproximadamente 3 bilhões de nucleotídeos (BROWN 1999).

Ao contrário dos genomas bacterianos e de eucariotos inferiores, o genoma humano possui uma organização complexa. Em organismos procariontes os genomas são pequenos, raramente ultrapassando o tamanho de 5Mb, possuem pequenas regiões intergênicas e genes organizados em operons (grupos de genes localizados adjacentemente no genoma e representando unidades transcricionais) que não apresentam introns. Já os genomas eucarióticos são significativamente maiores, podendo variar de menos de 10 Mb a 100.000Mb, no geral não contêm operons, apresentam longas regiões intergênicas e, também, introns, além de possuírem um elevado número de seqüências altamente repetitivas (BROWN 1999). Estima-se que as seqüências transcritas correspondam somente a 3% do genoma humano sendo o restante representado em sua maior parte por seqüências repetitivas que incluem, principalmente, LINEs ("Long Interspersed Nuclear Elements"), SINEs ("Short Interspersed Nuclear Elements") e DNA satélites.

Devido a sua complexidade, o seqüenciamento e a identificação de genes no genoma humano torna-se uma tarefa difícil e, conseqüentemente, há mais de uma década o número exato de genes que compõe o genoma humano representa uma incógnita para pesquisadores de todo o mundo. Durante o início da década de 90, a estimativa mais aceita quanto ao número de genes era de, aproximadamente, 50 mil.

No entanto, este número foi gradualmente crescendo com o desenvolvimento de técnicas de seqüenciamento de DNA cada vez mais avançadas. Dentre os inúmeros trabalhos publicados recentemente que visam estimar o número de genes existentes no genoma humano, podemos ressaltar a análise realizada por EWING e GREEN (2000) que comparou um grupo de ESTs com os genes já identificados no cromossomo 22 e seqüências de mRNA obtidas do GenBank estimando um valor aproximado de 35.000 genes. Esse valor foi confirmado em um estudo realizado por ROEST et al. (2000), através da análise de seqüências conservadas entre o peixe *Tetraodon nigroviridis* e o genoma humano, a qual foi capaz de estimar a existência de 28.000 a 34.000 genes humanos. Baseando-se na análise do banco de dados do "TIGR Gene Index" (QUACKENBUSH et al. 2000), LIANG et al. (2000) estimaram a existência de um valor aproximado de 120.000 genes no genoma humano, um número quase três vezes superior ao estimado nos trabalhos anteriores. Neste contexto, vale ressaltar que mesmo após a publicação da seqüência completa do genoma humano o número de genes presentes em nosso genoma continua indeterminado.

O seqüenciamento do genoma humano foi primeiramente conduzido no meio acadêmico pelo Consórcio Internacional de Seqüenciamento do Genoma Humano (IHGSC) a partir do início da década de 90 até 1998, quando um grupo do setor privado, representado pela empresa Celera Genomics, iniciou um projeto em paralelo utilizando uma estratégia de seqüenciamento alternativa. A proposta inicial do Consórcio Público era determinar a seqüência completa do genoma humano até 2005, entretanto, com a fundação da Celera este tempo foi reduzido e as primeiras versões do rascunho do genoma humano foram publicadas, respectivamente, nas revistas Nature (LANDER et al. 2001) e Science (VENTER et al. 2001) em fevereiro de 2001.

1.1.1 Estratégias de Seqüenciamento do Genoma Humano

Como já citado anteriormente, para o seqüenciamento do genoma humano foram aplicadas duas metodologias distintas. Enquanto o Consórcio Público utilizou a técnica de "Shotgun" Hierárquico (LANDER et al. 2001), a Celera Genomics adotou para o mesmo processo a técnica de "Shotgun" simples (VENTER et al. 2001), amplamente utilizada até então no seqüenciamento de genomas bacterianos menos complexos.

Na metodologia de "Shotgun" simples, quebra-se o genoma diretamente em pequenos segmentos de DNA randômicos (2-4Kb). Estes, por sua vez, após o seqüenciamento em larga escala, são montados computacionalmente com base na sobreposição entre as seqüências para a obtenção da seqüência completa do genoma em questão (VENTER et al. 2001) (Figura 1). Já na técnica de "Shotgun" Hierárquico, primeiramente o genoma é fragmentado e clonado em sistemas que permitem a propagação de fragmentos de alto peso molecular que, em seguida, são ordenados gerando, desta maneira, um mapa físico. Este é utilizado, então, para a escolha de um conjunto mínimo de clones a serem seqüenciados através da sub-clonagem e construção de mini bibliotecas de "Shotgun". Desta forma, a seqüência genômica é obtida com base na sobreposição entre as seqüências de cada clone e a localização cromossômica dos mesmos (LANDER et al. 2001) (Figura 2).

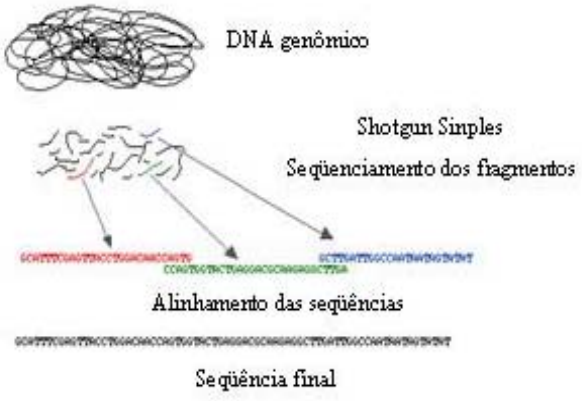


Figura 1 - Estratégia de "Shotgun" simples adotada pela empresa Celera Genomics. Representação esquemática da metodologia adotada pela empresa Celera Genomics para o seqüenciamento do Genoma Humano.

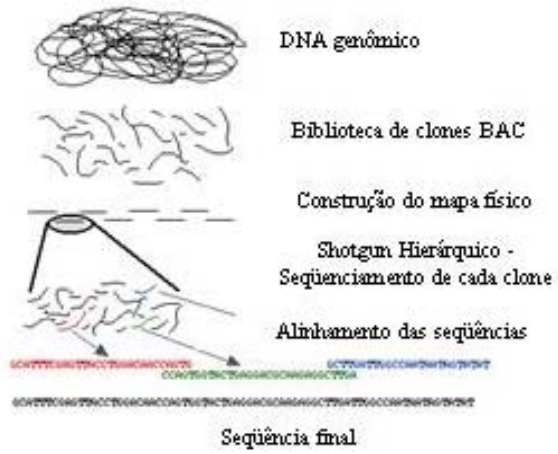


Figura 2 - Estratégia de "Shotgun" Hierárquico adotada pelo Consórcio Público. Representação esquemática da metodologia adotada pelo Consórcio Público para o seqüenciamento do Genoma Humano.

A metodologia de "Shotgun" hierárquico permite a produção de uma seqüência genômica de alta qualidade em relação ao "Shotgun" simples. Isso por que a montagem das seqüências é local e ancorada ao genoma (LANDER et al. 2001).

Uma grande vantagem desta técnica está na facilidade em resolver problemas como a presença de repetições e seqüências de baixa qualidade, uma vez que estes estão confinados a um clone individual e não a toda a seqüência genômica como seria na técnica de "Shotgun" simples (VENTER et al. 2001). No entanto, vale ressaltar que as seqüências geradas cobriram aproximadamente 94% do genoma humano com os dois rascunhos apresentando-se semelhantes quanto à qualidade das mesmas (AACH et al. 2001).

Apesar de terem utilizado estratégias de seqüenciamento diferentes, o IHGSC e a Celera utilizaram abordagens muito semelhantes para a identificação de genes e anotação. A partir da seqüência genômica, ambos basearam-se, respectivamente, em programas computacionais de predição gênica e buscas por similaridade com seqüências de nucleotídeos e proteínas depositadas em bancos de dados públicos.

1.1.2 Identificação de genes a partir da seqüência genômica e anotação

A pouco mais de quinze anos pesquisadores de todo o mundo vêm desenvolvendo programas computacionais de predição gênica para a identificação de genes a partir da seqüência genômica com o objetivo de automatizar e facilitar este processo. Assim, os primeiros programas de predição gênica foram construídos com o intuito de identificar apenas um gene em uma seqüência, raramente predizendo elementos promotores (ROGIC et al. 2001). Desta forma, este processo tornou-se muito mais eficiente em genomas procarióticos uma vez que estes apresentam uma alta densidade gênica e são constituídos por genes contínuos facilitando a análise.

Devido à complexidade do genoma humano foi necessário o desenvolvimento de novos programas com a capacidade de identificar estruturas

gênicas mais complexas. Estes foram, desta forma, desenhados para reconhecer padrões na estrutura gênica tais como a presença de uma fase aberta de leitura (ORF) e seqüências sinais relacionadas com a transcrição e o processamento do RNA (promotores, códons de início e término da transcrição, sítios aceptor e doador de "splicing" e poliadenilação).

Entretanto, apesar do desenvolvimento de programas computacionais mais eficientes, as seqüências sinais são bastante degeneradas e inespecíficas e o seu reconhecimento na seqüência genômica se torna bastante difícil. Segundo os dados obtidos por um estudo que comparou programas de predição desenvolvidos em diferentes épocas (ROGIC et al. 2001), os programas GenMark.hmm, Genie, GenScan, HMMgene, FGENES, Morgan e MZEF, por exemplo, apresentaram um aumento de 17% na acuidade e 19% na sensibilidade e especificidade em relação a programas mais antigos como GeneID, SORFIND, GeneParser2 e 3, GRAIL2, GenLang, FGENEH e Xpound (BURSET e GUIGO 1996).

Contudo, apesar de bastante direta, a identificação de genes *in silico* apresenta uma grande margem de erro e estima-se que 90% das estruturas gênicas preditas por programas de computador esteja incorreta. É importante ressaltar que durante a primeira anotação do cromossomo 22, por exemplo, apesar de 94% dos transcritos humanos validados e localizados neste cromossomo terem sido identificados pelo programa GenScan, apenas 20% dos mesmos tiveram sua estrutura corretamente predita e 16% dos exons experimentalmente validados não foram detectados *in silico* (DUNHAM et al. 1999).

Frente a todas estas dificuldades, a identificação de genes a partir da seqüência genômica humana é feita, atualmente, em duas etapas: predição gênica e confirmação da estrutura predita através da presença de similaridade com seqüências

humanas de cDNA completas e ESTs ("Expressed Sequence Tags"), assim como similaridade com genes e proteínas de outros organismos também disponíveis em bancos de dados públicos.

No processo de identificação de genes desenvolvido pelo Consórcio Público foram identificados, aproximadamente, 30 mil transcritos (LANDER et al. 2001) enquanto a Celera Genomics foi capaz de identificar perto de 40 mil genes (VENTER et al. 2001). No entanto, um estudo comparando os genes identificados pelo Consórcio Público e pela Celera Genomics comprovou que os critérios utilizados por ambos os grupos para a anotação do genoma humano são inadequados e incompletos (HOGENESCH et al. 2001). Em uma primeira comparação entre genes com estrutura já bem caracterizada e possuindo seqüência completa de mRNA catalogada no banco de dados RefSeq, aproximadamente 84% desses transcritos foram identificados por ambos os grupos. No entanto, em uma segunda análise, o mesmo estudo revelou que a maioria dos novos genes identificados pelos dois grupos era exclusiva de cada grupo. Desta forma, pode-se concluir que os respectivos catálogos de genes são individualmente incompletos e que o genoma humano deve ter um número maior de genes do que inicialmente determinado pelos dois grupos.

Em decorrência de todas as dificuldades apresentadas, diversos trabalhos têm sido publicados com o intuito de tornar o catálogo de genes humanos cada vez mais completo. WIEMANN et al. (2001), por exemplo, através do alinhamento de seqüências de cDNA com as seqüências já publicadas dos cromossomos 21 e 22, foram capazes de identificar genes não anotados na primeira versão da seqüência completa destes cromossomos e ainda genes que haviam sido preditos de maneira incorreta. Outros estudos realizados especificamente com o cromossomo 21, seja através de comparação com o cromossomo 16 correspondente de camundongo

(PLETCHER et al. 2001) ou da combinação de análises computacionais e validações experimentais (GARDINER et al. 2002; REYMOND et al. 2001) também contribuíram para a identificação de novos genes localizados neste cromossomo. Além disso, a exemplo do ocorrido para o cromossomo 21, dois grupos (HIROSAWA et al. 2001; COLLINS et al. 2003), utilizando dados de ESTs, análises comparativas com seqüências de outros organismos e validação experimental, foram capazes de identificar novos genes não anotados no cromossomo 22.

A baixa eficiência na estratégia de anotação utilizada pelos grandes grupos de seqüenciamento pode ser explicada por três fatores: a complexidade da estrutura dos genes humanos, a baixa eficiência dos programas de predição gênica e a existência de um número insuficiente de seqüências de cDNA e proteínas disponíveis em bancos de dados públicos que podem ser utilizadas como evidência experimental no processo de identificação de genes. Em outras palavras, está ficando cada vez mais evidente a necessidade de gerar seqüências de cDNA completas e ESTs ou até mesmo de seqüenciar, paralelamente, genomas de organismos evolutivamente relacionados com a espécie humana, tornando as análises *in silico* mais acuradas.

No entanto, apesar de promissora, a estratégia de seqüenciamento de genomas de organismos evolutivamente próximos à espécie humana ainda não é capaz de determinar com eficiência a estrutura exata de cada gene (limites exon-intron) muito menos as formas alternativas relacionadas a cada transcrito. Desta forma, acredita-se realmente que a determinação da estrutura gênica, assim como das formas alternativas dos transcritos, só será possível através do seqüenciamento de moléculas de cDNA.

1.2 SEQÜENCIAMENTO DE MOLÉCULAS DE cDNA EM LARGA ESCALA

Como mencionado anteriormente, os genes representam apenas 3% de todo o genoma, o que dificulta a sua identificação a partir da seqüência genômica. Neste contexto, as moléculas de cDNA são um material de extrema utilidade pois, por serem sintetizadas a partir do RNA mensageiro (mRNA), representam a porção transcrita do genoma, ou seja, correspondem à fração do genoma que carrega a informação dos genes.

O cDNA obtido através da transcrição reversa de moléculas de mRNA pode ser inserido em vetores de expressão e replicado em células hospedeiras dando origem às chamadas bibliotecas de cDNA. As moléculas de cDNA podem ser, então, seqüenciadas de duas formas: seqüenciamento completo do inserto de clones de cDNA e seqüenciamento parcial das extremidades desses clones gerando as ESTs.

A produção de seqüências completas de cDNA e ESTs vem sendo realizada por diversos grupos simultaneamente ao seqüenciamento do genoma humano. Atualmente, existem, aproximadamente, 27.887 seqüências completas de cDNA (não redundantes) e mais de 5 milhões de ESTs humanas disponíveis em bancos de dados públicos. No entanto, apesar de partirem do mesmo material (clones representados em bibliotecas de cDNA), as duas metodologias geram dados relativamente distintos e as limitações encontradas na identificação de genes também são distintas como discutido a seguir.

1.2.1 Seqüenciamento completo de clones de cDNA

O seqüenciamento completo de clones de cDNA previamente selecionados por conterem transcritos inteiros pode ser considerado uma estratégia direta e confiável para a determinação de transcritos humanos. A seleção dos clones representativos que serão completamente seqüenciados é feita após o seqüenciamento parcial das extremidades 5' e 3' e análise computacional de um grande número de clones por biblioteca. Os insertos dos clones selecionados são, então, subclonados em fragmentos menores compatíveis com o seqüenciamento. Posteriormente, as seqüências referentes a estes fragmentos são montadas com o auxílio de programas de computador para a obtenção da seqüência completa do inserto (STRAUSBERG et al. 1999).

Vários projetos que visam o seqüenciamento completo de clones de cDNA estão em andamento como, por exemplo, o "Mammalian Gene Collection" (MGC) (STRAUSBERG et al. 1999). No entanto, a construção de bibliotecas enriquecidas para mRNAs completos e normalizadas em relação à abundância dos transcritos é um processo trabalhoso acarretando na diminuição da escala de produção e limitando o número de genes que podem ser caracterizados por essa metodologia.

1.2.2 Seqüenciamento parcial de clones de cDNA: EST

A produção de ESTs (seqüências parciais de uma das extremidades de um clone de cDNA) teve suas origens em 1980, mas somente no início da década de noventa elas começaram a ser produzidas em larga escala (ADAMS et al. 1995; KORENBERG et al. 1995; HILLIER et al. 1996). Até julho de 2003 (dbEST release 071103) mais de 5.372.149 ESTs humanas obtidas a partir de diversos órgãos e tecidos foram depositadas no GenBank, derivadas, principalmente, do "Merck Gene

Index Project" (WILLIAMSON 1999; ECKMAN et al. 1998), do "Cancer Genome Anatomy Project" (CGAP) (STRAUSBERG et al. 2000) e do Projeto Genoma Humano do Câncer (HCGP) (DIAS NETO et al. 2000; CAMARGO et al. 2001).

As ESTs foram inicialmente exploradas na identificação de novos genes (BROWN et al. 2003; ADAMS et al. 1992, 1993). No entanto, atualmente, estão sendo intensamente utilizadas na construção de perfis de transcritos tecido-específicos (KATSANIS et al. 2002; MÉGY et al. 2002; HUMINIECKI e BICKNELL 2000), na construção de mapas físicos (HUDSON et al. 1994), na comparação de genomas de diferentes organismos (TUGENDREICH et al. 1994; LEE et al. 2002), no estudo qualitativo da expressão de mRNA através de variantes de "splicing" (XU et al. 2002; XIE et al. 2002; WANG et al. 2003), poliadenilação alternativa (BEAUDOING e GAUTHERET 2001; ISELI et al. 2002) e SNPs ("Single Nucleotide Polymorphisms") (CLIFFORD et al. 2000; IRIZARRY et al. 2000; HU et al. 2002) e, ainda, na predição e identificação de transcritos na seqüência genômica (BAILEY et al. 1998; JIANG e JACOB 1998).

No entanto, sabe-se que os bancos de dados de ESTs contêm inúmeros artefatos em conseqüência da presença de seqüências de baixa qualidade e de vários tipos de contaminação (SOREK e SAFER 2003). Estas incluem seqüências de vetores, seqüências de outros organismos, como bactéria e vírus, e o mais importante: seqüências intrônicas ou provenientes de regiões intergênicas derivadas de uma contaminação de DNA durante a extração do mRNA utilizado na geração das bibliotecas de cDNA.

Além disso, devido a limitações técnicas da reação de seqüenciamento, as ESTs são geradas a partir das extremidades 5′ ou 3′ das moléculas de cDNA e, desta forma, as regiões centrais dos genes estão sub-representadas em bancos de dados de

ESTs, dificultando a reconstrução do transcrito completo. Em conseqüência deste fato, apenas 14% dos genes humanos já caracterizados estão representados por ESTs em toda a sua extensão (DE SOUZA et al. 2000).

Dentro deste contexto, a FAPESP e o Instituto Ludwig lançaram o Projeto Genoma Humano do Câncer (HCGP) com o objetivo de gerar 1 milhão de ESTs utilizando uma nova metodologia denominada ORESTES (DIAS NETO et al. 2000) ("Open Reading Frame ESTs"). Esta estratégia é capaz de gerar seqüências derivadas das porções centrais dos transcritos e, simultaneamente, normalizar a população de mRNA de forma a gerar seqüências derivadas de transcritos com baixo nível de expressão. A base desta abordagem é a amplificação por PCR ("Polimerase Chain Reaction") de seqüências expressas utilizando iniciadores aleatórios sob condições de baixa estringência. Devido a essas características, as ORESTES complementam as ESTs geradas através de metodologias convencionais, que, em grande parte, são derivadas das extremidades 5' ou 3' dos transcritos e apresentam uma grande tendência para a representação de transcritos mais abundantemente expressos (DIAS NETO et al. 2000; CAMARGO et al. 2001).

As ORESTES têm sido utilizadas por diversos grupos que integram o HCGP na caracterização de perfis tecido-específicos, formas alternativas de "splicing", identificação de SNPs e, dentro do contexto do nosso grupo, na identificação de novos transcritos humanos em conjunto com as demais ESTs. Com o intuito de comprovar a utilidade das seqüências geradas pelo HCGP na identificação de novos genes, um estudo comparou a seqüência do cromossomo 22 com um banco de dados composto por 250.000 seqüências geradas na fase inicial do projeto (DE SOUZA et al. 2000). A análise confirmou a existência de um grande número de genes identificados, além de identificar 219 seqüências de transcritos não anotados no cromossomo 22. Das 219

seqüências, 48 foram exclusivamente geradas pelo HCGP. Estes resultados demonstram que a produção de ESTs em larga escala é capaz de contribuir significativamente na identificação de todos os genes humanos.

1.3 IDENTIFICAÇÃO DE GENES HUMANOS COM BASE NO ALINHAMENTO ENTRE A SEQÜÊNCIA GENÔMICA E SEQÜÊNCIAS EXPRESSAS

Uma estratégia alternativa utilizada no processo de identificação de genes e que não faz uso de programas computacionais de predição gênica é o alinhamento direto entre seqüências expressas e a seqüência genômica. Técnicas que mapeiam a posição de seqüências expressas em fragmentos de DNA podem ser utilizadas na localização de exons e introns já que as mesmas derivam de moléculas de RNA mensageiro, representantes da porção transcrita do genoma. Desta forma, a partir do alinhamento de uma seqüência expressa em determinada região genômica pode-se inferir a existência de um gene correspondente a esta EST na mesma região.

Entretanto, deve-se lembrar que este tipo de análise não fornece informações precisas quanto à definição do início e término da região transcrita de um gene uma vez que se baseia, na maioria dos casos, nos dados de seqüências parciais. Além disso, esta metodologia pode gerar inúmeros artefatos de alinhamento que devem ser cuidadosamente analisados. O fato das seqüências expressas apresentarem, no geral, baixa qualidade impossibilita a utilização de um "cutoff" de similaridade de 100% no alinhamento entre a seqüência genômica e as seqüências expressas. Assim sendo, uma similaridade de pelo menos 95% é geralmente adotada e este fato pode levar à presença de múltiplos alinhamentos em regiões gênicas diferentes para a mesma

seqüência expressa. Os alinhamentos múltiplos se devem à existência de famílias gênicas e pseudogenes de forma que seqüências correspondentes a diferentes transcritos que fazem parte de uma mesma família gênica podem acabar alinhando nas mesmas regiões genômicas e seqüências de um determinado gene podem alinhar em regiões genômicas correspondentes a seus pseudogenes.

Outro problema encontrado na estratégia de alinhamento entre seqüências expressas e a seqüência genômica são as contaminações com DNA genômico presentes nos bancos de dados de seqüências expressas e que quando alinhadas no genoma levam à falsa inferência da existência de um gene em determinada região genômica. A solução seria utilizar a presença de "splicing" nas seqüências expressas como critério na seleção de candidatos a novos transcritos. A presença de "splicing" em uma determinada seqüência expressa pode ser indiretamente inferida através da existência de interrupções no alinhamento com a seqüência genômica as quais estão relacionadas à presença de introns na seqüência genômica que são removidos das seqüências expressas após a ocorrência do "splicing". Entretanto, devido à presença de seqüências de baixa qualidade, a existência de uma interrupção no alinhamento da seqüência expressa em relação à seqüência genômica pode ser erroneamente interpretado como relacionado à presença de um intron e à ocorrência de "splicing". Neste caso, um critério mais rigoroso para a seleção de candidatos a novos transcritos seria a análise da presença dos sítio conservados de "splicing" (GT/AG).

Apesar de todas essas limitações, o alinhamento entre seqüências expressas e a seqüência genômica é capaz de fornecer dados confiáveis para a identificação de genes no genoma humano. ZHUO et al. (2001), por exemplo, através do alinhamento de seqüências consensos representativas de cada "cluster" do UniGene com a seqüência do genoma humano, mapearam um total de 59.500 "clusters" do UniGene

produzindo um rascunho do mapa físico humano com anotações para grande parte dos transcritos humanos. Posteriormente, o mesmo grupo produziu um índice de genes humanos baseado na integração dos dados públicos referentes às seqüências transcritas, proteínas e ao mapeamento destas no genoma humano juntamente com dados de predições gênicas sendo identificado um total de 75.983 genes dos quais somente 16.673 correspondiam a genes conhecidos (WRIGHT et al. 2001).

Desta forma, a partir do interesse de utilizar esta metodologia de alinhamento entre seqüências expressas e a seqüência genômica humana na identificação de novos genes humanos e, também, de explorar os dados gerados pelo HCGP, o Laboratório de Biologia Computacional do Instituto Ludwig criou um banco de dados representando o transcriptoma humano.

Assim, o banco de dados do Transcriptoma Humano contém informações sobre as regiões transcritas do genoma humano. Tais regiões foram definidas *in silico* através do mapeamento das seqüências de cDNA na seqüência genômica disponível nos bancos de dados públicos e, também, da produção e mapeamento de "tags" de 50 nucleotídeos derivadas da extremidade 3' dos diferentes transcritos humanos. A produção das "tags" é uma maneira prática e direta para definir a localização espacial dos transcritos no genoma, demarcando com precisão o final de um transcrito.

Os dados armazenados neste banco de dados foram utilizados para criar "clusters" de seqüências transcritas com base nas coordenadas dos exons definidas na seqüência genômica, de forma que dois transcritos foram considerados parte de um mesmo "cluster" (e, portanto, derivadas de um mesmo transcrito) quando compartilhavam as coordenadas de pelo menos um exon. Essa estratégia de agrupamento evita parcialmente os artefatos gerados através do alinhamento direto de seqüências transcritas, geralmente associados à baixa qualidade das seqüências, à

presença de extensas famílias gênicas no genoma e a formas alternativas de "splicing".

Desta forma, com base neste banco de dados é possível delimitar todas as regiões transcritas para um determinado clone genômico, assim como extrair várias informações, tais como: a identidade das ESTs alinhadas, a origem tecidual e o método utilizado na geração de cada uma das ESTs, além das coordenadas de alinhamento das mesmas. As coordenadas referentes aos exons e introns definidas através dos alinhamentos foram indexadas tanto em relação à seqüência genômica, quanto ao transcrito. Todas as informações sobre o mapeamento foram armazenadas em um banco de dados relacional MySQL podendo ser visualizadas através de uma Interface Gráfica.

No entanto, devido à presença de artefatos já descritos anteriormente gerados durante o processo de alinhamento, após a seleção de seqüências expressas mapeadas em uma determinada região genômica é necessário um processo de inspeção manual. Neste é realizada uma análise quanto ao alinhamento contínuo da seqüência expressa contra a seqüência genômica e também uma confirmação do melhor alinhamento desta seqüência na região determinada. Os candidatos selecionados seguem para a validação experimental e conseqüente confirmação dos mesmos como novos genes.

Desta maneira, frente ao que foi discutido até o momento surgiu o nosso interesse em aplicar esta metodologia de alinhamento entre seqüências expressas e a seqüência genômica tanto em larga escala, realizando uma análise de todo o genoma humano, quanto restrita a um cromossomo, tentando encontrar possíveis relações entre novos genes e doenças associadas e mapeadas no mesmo. A identificação de genes no genoma humano levou ao desenvolvimento de um projeto chamado

"Transcript Finishing Initiative" (TFI) que envolveu mais de 30 laboratórios de todo o estado de São Paulo. Já com relação à análise em pequena escala, em colaboração com o grupo de Bioinformática do Instituto Ludwig da Suíça e o grupo de pesquisa do Dr. Stylianos Antonarakis da Universidade de Genebra, surgiu a idéia de aplicar esta metodologia utilizando como base a seqüência do cromossomo 21.

Desta maneira, esta dissertação foi dividida em duas partes com o objetivo de facilitar a leitura e a compreensão de seu conteúdo. A primeira parte consiste na descrição dos métodos aplicados e dos resultados obtidos com o projeto TFI, enquanto em uma segunda parte foi descrito todo o processo de identificação de novos transcritos no cromossomo 21.

*PARTE I**

IDENTIFICAÇÃO DE NOVOS TRANSCRITOS HUMANOS COM BASE NO ALINHAMENTO ENTRE A SEQÜÊNCIA GENÔMICA E SEQÜÊNCIAS EXPRESSAS.

O PROJETO "TRANSCRIPT FINISHING INITIATIVE"

*** Um artigo científico contendo os resultados obtidos no desenvolvimento deste projeto foi submetido à revista Genome Research em setembro deste ano. (Anexo 1)**

INTRODUÇÃO

2.1 INTRODUÇÃO

Reconhecendo a importância das seqüências expressas na identificação de genes humanos e visando utilizar os dados gerados no HCGP, a FAPESP e o Instituto Ludwig lançaram o Projeto "Transcript Finishing Initiative" (TFI). O TFI representa uma extensão do HCGP e teve como objetivo determinar a estrutura e gerar as seqüências de novos genes humanos utilizando como base o banco de dados do Transcriptoma, ou seja, a seqüência do genoma humano e todas as seqüências expressas (mRNA, ESTs e ORESTES) disponíveis em bancos de dados públicos.

O projeto teve uma duração de aproximadamente dois anos e foi dividido em duas frentes interdependentes de trabalho: o desenvolvimento de ferramentas de bioinformática e a validação experimental, envolvendo um total de 36 grupos de pesquisa do estado de São Paulo. O TFI contou com a participação de 5 grupos de bioinformática responsáveis pelo desenvolvimento de ferramentas computacionais para a seleção dos transcritos a serem validados e avaliação e anotação dos novos transcritos, além de 2 laboratórios centrais de Coordenação (Instituto Ludwig de pesquisa sobre o Câncer e Instituto de Química da Universidade de São Paulo) e 29 laboratórios de pesquisa de todo o estado (Laboratórios de Validação) para a realização da parte experimental. A lista dos 36 laboratórios participantes do projeto pode ser encontrada no endereço http://200.18.51.201/transcript/. Assim como os demais projetos realizados dentro da rede ONSA ("Organization for Nucleotide Sequencing and Analysis"), a formação e o treinamento dos grupos foram priorizados. Todos os laboratórios de validação receberam orientação para a utilização das ferramentas de bioinformática, assim como para a amplificação e seqüenciamento dos fragmentos de validação.

Ao contrário dos projetos de produção de seqüências de cDNA completas, o TFI gerou a seqüência de fragmentos parciais de cDNA unindo dois "clusters" de seqüências expressas. Desta forma, todo o trabalho relacionado com a preparação de bibliotecas de cDNA de alta qualidade foi evitado. Esses fragmentos parciais (TFs) foram gerados através de RT-PCR ("Reverse Transcriptase PCR"), utilizando cDNAs provenientes de diversos tecidos como molde, de forma a validar e complementar a estrutura de genes humanos parcialmente representados por ESTs (Figura 3). O produto final do projeto foi um catálogo de seqüências virtuais, validadas experimentalmente, de novos transcritos humanos. Trata-se de uma abordagem inédita com profundo embasamento computacional e complementar às estratégias já disponíveis.

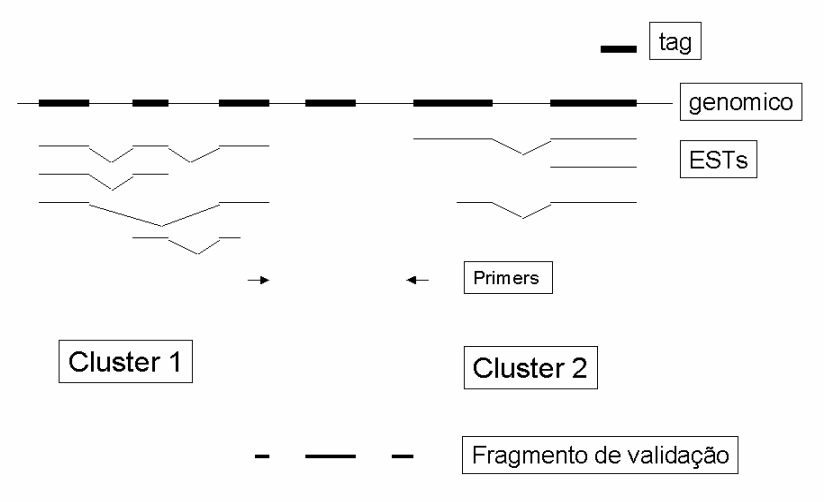


Figura 3 - Esquema representando a estratégia geral de "Transcript Finishing". Através do alinhamento entre seqüências expressas (ESTs) e a seqüência genômica humana (Genômico) é possível identificar exons (linha preta grossa) e introns (linha preta fina) assim como agrupar as ESTs com base nas coordenadas dos alinhamentos ("cluster" 1 e 2). A estratégia do TFI visa unir dois "clusters" de ESTs através de validação experimental por RT-PCR ("primers") gerando um fragmento validação e, assim, confirmando que eles fazem parte de um mesmo transcrito.

OBJETIVOS

2.2 OBJETIVOS

- Determinar a estrutura e gerar as seqüências de novos transcritos humanos baseando-se nos dados de alinhamento entre a seqüência genômica e seqüências expressas disponíveis em bancos de dados públicos.

- Avaliar a eficiência da estratégia proposta pelo projeto TFI através de buscas por similaridade com seqüências humanas depositadas em bancos de dados públicos além da existência total ou parcial de uma predição gênica referente a cada uma das seqüências.

- Verificar a ocorrência de formas alternativas de "splicing" para todos os novos transcritos através de abordagens *in silico* seguida de confirmação experimental.

MATERIAIS E MÉTODOS

2.3 MATERIAIS E MÉTODOS

2.3.1 Desenvolvimento das Ferramentas de Bioinformática

2.3.1.1 Construção do Banco de dados do Projeto TFI

A seleção de transcritos humanos parcialmente representados por ESTs para validação experimental foi feita a partir do Banco de dados do Transcriptoma Humano desenvolvido pelo Laboratório de Biologia Computacional do Instituto Ludwig (coordenação Dr. Sandro José de Souza) e que contém informações relacionadas ao alinhamento entre seqüências expressas e a seqüência genômica. Para a construção do banco, os dados correspondentes às seqüências transcritas foram obtidos a partir de diversas fontes: (1) seção de ESTs humanas do banco de dados EMBL versão 66; (2) seqüências de mRNA humano presentes no EMBL versão 66; (3) seqüências de ORESTES provenientes do HCGP e (4) seqüências de mRNA humano presentes no banco de dados RefSeq, disponibilizado pelo NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). Foram utilizadas 2.684.329 seqüências transcritas e seqüências genômicas com tamanho superior a 10Kb disponíveis nas seções HUM e HTG do banco de dados do Projeto Genoma Humano. O mapeamento das seqüências transcritas em relação à seqüência genômica foi feito com base na similaridade entre as mesmas utilizando a ferramenta BLASTN (NCBI version 2.0.12), também disponibilizada pelo NCBI.

As seqüências transcritas foram filtradas quanto à presença de contaminação com seqüências de vetores, bactérias e fungos e os elementos repetitivos foram mascarados usando o "software" PFP (Paracel, Pasadena, CA). Para cada par correspondente de seqüências transcritas e genômica foi feito um alinhamento local

utilizando o programa Sim4 (FLOREA et al. 1998), com parâmetros W=15, R=0, A=4 e P=1 a fim de que os limites entre introns e exons, assim como a direção do transcrito em relação à seqüência genômica, fossem definidos. O dado obtido a partir do Sim4 foi filtrado para eliminar todos os alinhamentos que não continham pelo menos uma região correspondente com no mínimo 95% de identidade acima de 30 nucleotídeos.

Os dados dos alinhamentos entre as seqüências expressas (ESTs e cDNAs) e a seqüência genômica foram armazenados em um banco de dados relacional MySQL (Figura 4) e utilizados para criar "clusters" de seqüências transcritas baseados nas posições de cada seqüência dentro dos clones genômicos individuais. Os membros de cada "cluster" foram determinados através das coordenadas dos exons em relação à seqüência genômica. Se as coordenadas de pelo menos um exon fossem comuns a duas seqüências transcritas, as mesmas eram consideradas parte de um mesmo "cluster". Utilizando esta estratégia, foram definidos 244.148 "clusters", dos quais 14.598 incluíam seqüências completas de mRNA já validadas experimentalmente e 229.550 "clusters" correspondiam somente a seqüências parciais. Vale ressaltar ainda, que no grupo de 14.598 "clusters", 13.149 (90%) tinham pelo menos uma EST correspondente e os 1.449 "clusters" restantes eram compostos somente por seqüências completas de mRNA.

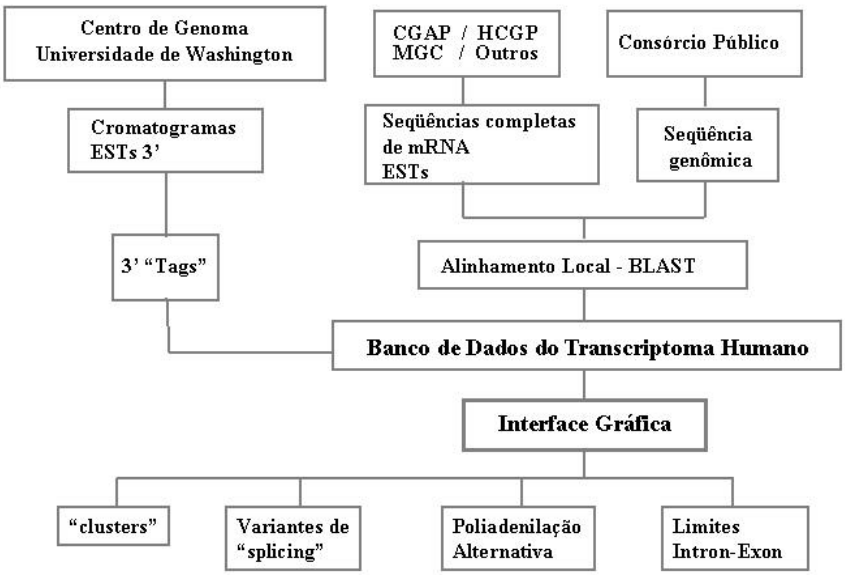


Figura 4 - Esquema representando a construção do Banco de Dados do Transcriptoma. Os dados dos alinhamentos entre as seqüências transcritas (ESTs e cDNAs) e a seqüência genômica, além do alinhamento e mapeamento das 3' "tags", foram armazenados em um banco de dados relacional MySQL podendo ser visualizados através de uma Interface Gráfica.

Além disso, também foram mapeadas no genoma as 3' "tags" que correspondem à extremidade 3' dos transcritos. Para a extração das "tags", seqüências correspondentes aos últimos 50 nucleotídeos imediatamente adjacentes à cauda poliA, foram utilizados os dados dos cromatogramas de ESTs derivadas da porção 3' dos transcritos evitando, desta forma, problemas decorrentes da qualidade de seqüência e a constante remoção da cauda de poli-A durante a submissão das seqüências de ESTs aos bancos de dados. Após a exclusão das "tags" contendo elementos repetitivos, as demais foram, então, mapeadas no genoma através de alinhamento e agrupadas em "clusters" de acordo com o seu mapeamento.

2.3.1.2 Seleção dos "clusters" de ESTs para validação experimental

A seleção de "clusters" para validação experimental foi realizada segundo critérios que visavam maximizar a eficiência de validação e o sucesso do projeto. Desta forma, foram selecionados: (1) pares de "clusters" de ESTs que não continham uma seqüência de mRNA completa, uma vez que o projeto priorizava a caracterização de novos transcritos humanos; (2) pares de "clusters" que continham ESTs com "splicing", uma vez que a presença de "splicing" é a evidência mais forte para a confirmação da origem da EST como seqüência expressa e (3) pares de "clusters" de ESTs que estivessem a uma distância máxima de 10Kb um do outro, uma vez que quanto maior a distância entre os "clusters" menor a probabilidade de eles pertencerem a um mesmo transcrito e de haver sucesso nas amplificações.

Uma vez estabelecidos os critérios de seleção, programas na linguagem Perl foram desenvolvidos pelo Laboratório de Biologia Computacional do Instituto Ludwig (coordenação Dr. Sandro José de Souza) para a seleção de pares de "clusters" a partir do Banco de Dados do Transcriptoma, descrito anteriormente. Cada par de "cluster" selecionado definiu um TF para o qual foram desenhados iniciadores posteriormente distribuídos aos grupos. Antes da distribuição para os grupos, a coordenação se encarregou da inspeção manual e individual de cada um destes pares de "clusters" selecionados automaticamente. Além da posição espacial dos "clusters" em relação à seqüência genômica e da presença de "tags" nos mesmos, dados sobre o padrão de expressão também foram observados, sendo selecionados apenas pares de "clusters" cuja expressão era compatível com os cDNAs disponibilizados pelo projeto.

2.3.1.3 Desenho automático dos iniciadores

Após a seleção manual dos "clusters", o desenho dos iniciadores para a realização das reações de RT-PCR foi feito de maneira automatizada. Uma série adicional de programas foi desenhada para extrair do banco de dados as informações e coordenadas para o desenho automático dos iniciadores. O grupo de bioinformática da Escola Paulista de Medicina (coordenação Dr. Paulo Paiva) se encarregou do processo de automatização que teve como base o programa Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi). Para a escolha dos iniciadores foram utilizados os seguintes parâmetros: tamanho do iniciador, máximo de 21 bases e mínimo de 17 bases; temperatura de anelamento, máxima de 65°C e mínima de 55°C; "GC clamp"=1. Os iniciadores desenhados pelo programa Primer3 foram excluídos quando apresentavam alinhamento inespecífico em regiões genômicas não relacionadas. Uma página foi disponibilizada (http://compbio.epm.br/tfi) para a submissão dos dados relativos aos "clusters" selecionados e para a obtenção das seqüências dos iniciadores desenhados automaticamente (Figura 5).

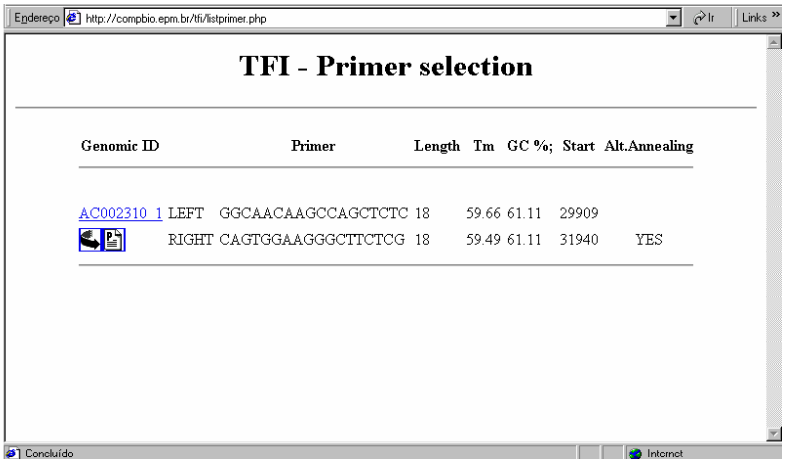


Figura 5 - "Homepage" desenvolvida para a seleção automática dos iniciadores para RT-PCR. Com base no clone genômico correspondente a cada par de "cluster" selecionado é possível obter a seqüência dos iniciadores para a realização das reações de RT-PCR.

2.3.1.4 Interface Gráfica do Banco de Dados do Projeto TFI

Uma Interface Gráfica com acesso via Internet foi desenvolvida na linguagem TCL/TK pelo grupo de Bioinformática da Unisa (Coordenação Dr. Paulo Oliveira) visando uma melhor interpretação dos resultados dos alinhamentos entre as seqüências genômicas e seqüências expressas. A interface permite a visualização dos "clusters" de ESTs, da presença de "splicing" e de formas variantes de "splicing" além das 3' "tags". Também foram disponibilizados recursos que permitem verificar a procedência de uma determinada seqüência de interesse, quer em termos do projeto responsável pela sua produção, quer em termos do tecido de origem, assim como fazer o "download" de sua seqüência a partir do GenBank (Figura 6).

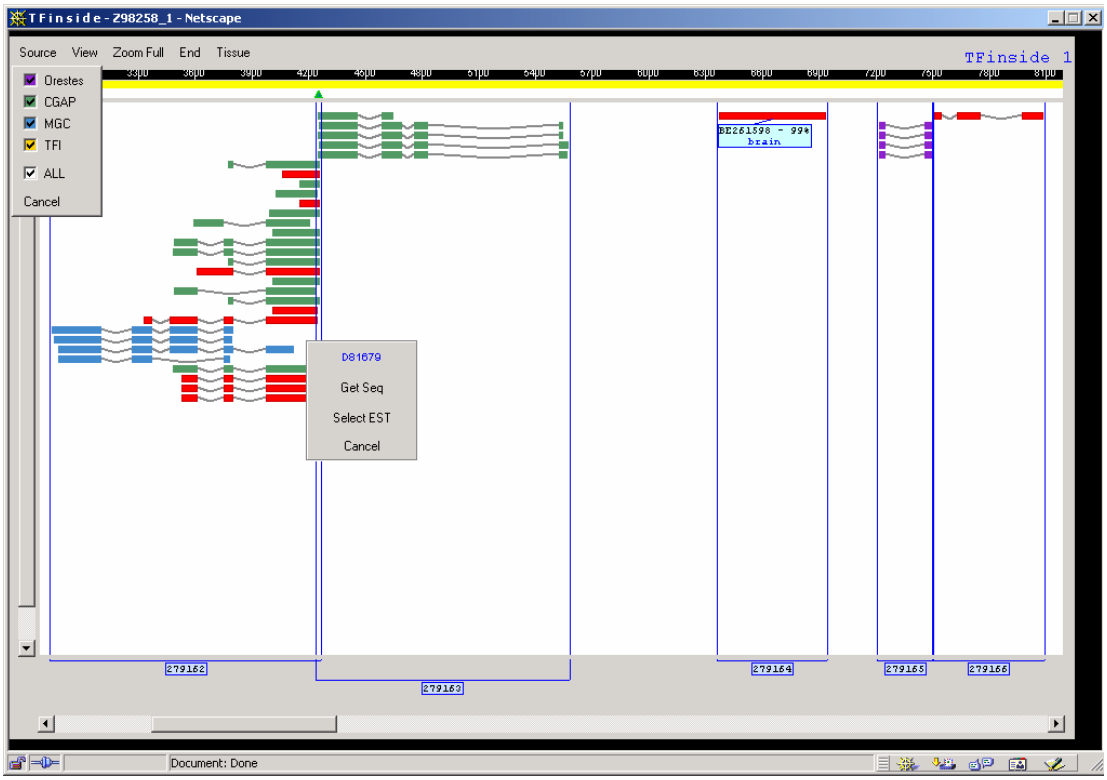


Figura 6 - Imagem da Interface Gráfica desenvolvida para o Projeto TFI. Os exons simbolizam as linhas grossas e os introns as linhas finas. A seqüência genômica está representada em amarelo e as seqüências expressas são representadas em diferentes cores de acordo com o projeto de origem (MGC, ORESTES ou CGAP). Os triângulos verdes representam a presença das 3' "tags" e delimitam as extremidades 3' dos transcritos. Através de "links" é possível, ainda, obter o número de acesso de cada seqüência no GenBank, assim como informações sobre o tecido de origem e a similaridade dos alinhamentos.

2.3.1.5 "Homepage" do projeto e submissão de seqüências de validação

A "homepage" do projeto foi desenvolvida pelo grupo da UNAERP (coordenação Dr. Milton Faria) e pode ser acessada no endereço http://200.18.51.201/transcript/. A página fornece informações gerais sobre o projeto, ferramentas e "softwares" que são necessários para a visualização da interface e submissão de seqüências. Também são disponibilizados dados individuais sobre o desempenho de cada grupo e informações sobre os transcritos validados (Figura 7).

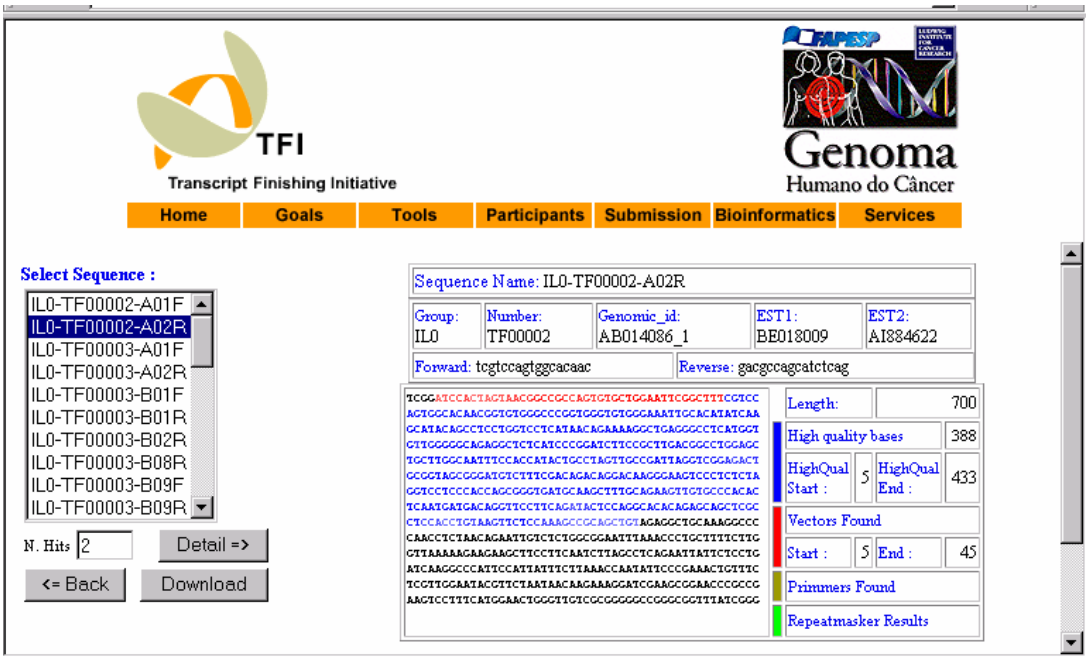


Figura 7 - "Homepage" desenvolvida para avaliação das seqüências de validação submetidas ao "pipeline" do Projeto TFI. Página do Projeto TFI onde podem ser encontrados todos os dados referentes aos TFs distribuídos: clone genômico onde está localizado o TF, número de acesso no GenBank de cada EST do par de "clusters" selecionado e seqüências dos iniciadores utilizados na validação experimental. Além disso, podem ser visualizadas as análises de qualidade das seqüências de validação depositadas por cada grupo e também a avaliação das mesmas com relação à presença de seqüências repetitivas e vetores.

2.3.1.6 Análise das seqüências de validação e atualização do Banco de Dados do Projeto

As seqüências geradas pelos laboratórios de validação foram submetidas via Internet ao Laboratório de Bioinformática da UNAERP (coordenação Dr. Milton Faria), responsável pelo armazenamento das mesmas. Em seguida, essas seqüências foram enviadas ao Laboratório de Bioinformática do Hemocentro de Ribeirão Preto (coordenação Dr. Wilson da Silva Jr.), responsável pela avaliação da qualidade das seqüências produzidas, assim como pela eliminação de eventuais seqüências contaminantes (bactérias, fungos, vetores) e elementos repetitivos. Uma série de programas que constituem o "pipeline" de submissão foi desenvolvida pelo grupo de bioinformática para a análise automática desses dados. A qualidade das seqüências foi

avaliada através do programa Phred utilizando um TrimCutOff de 0,06171 (EWING et al. 1998). Todas as seqüências com menos de 100pb foram automaticamente excluídas. Seqüências repetitivas foram identificadas e mascaradas com o auxílio do programa REPEAT-MASKER utilizando os parâmetros padrões. Seqüências contaminantes foram identificadas por meio de buscas, através do programa BLASTN, em bancos de dados correspondentes a seqüência completa do DNA mitocondrial humano e seqüências genômicas derivadas de bactérias e fungos. Os "cut-offs" para a qualidade das seqüências e análises do BLASTN ("e-value" de 10^{-5} para buscas contra DNA mitocondrial e "e-value" de 10^{-30} para buscas contra seqüências de bactérias e fungos) foram cuidadosamente estabelecidos e extensivamente revisados pela equipe de bioinformática em colaboração com a coordenação do Instituto Ludwig (Figura 8).

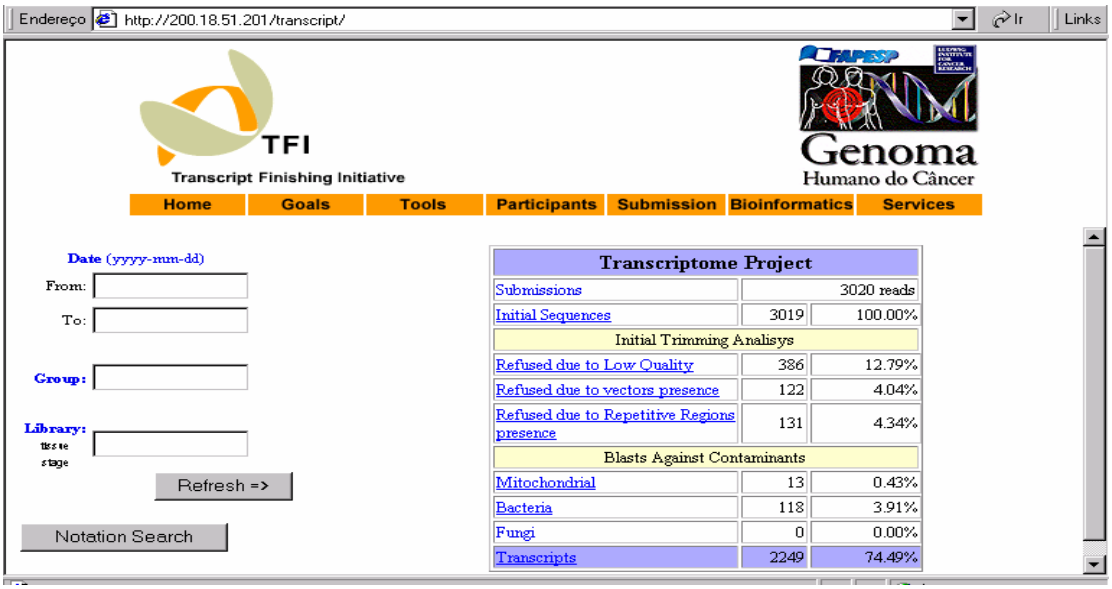


Figura 8 - "Homepage" do "pipeline" de submissão das seqüências de validação. Nesta página podem ser visualizados o número total de seqüências submetidas, o número de seqüências excluídas devido à baixa qualidade, presença de seqüência de vetor, elementos repetitivos e contaminantes além de quantas seqüências foram aprovadas e carregadas no banco de dados do projeto.

2.3.2 Validação Experimental

2.3.2.1 Linhagens Celulares

Como fonte de material biológico foram utilizadas linhagens celulares tumorais provenientes da ATCC ("American Type Culture Collection") e cultivadas em meio de cultura apropriado segundo recomendações da mesma (http:// www.atcc.org). A opção pelo uso de linhagens celulares foi tomada frente à necessidade de obtenção de grandes quantidades de RNA de alta qualidade e a fim de evitar qualquer complicação posterior com questões éticas. Foram utilizadas 20 linhagens celulares derivadas de diferentes tecidos (Tabela 1).

Tabela 1 - Linhagens celulares humanas utilizadas para a validação experimental dos pares de "clusters" selecionados. Linhagens celulares, com seus respectivos tecidos de origem, utilizadas na síntese dos cDNAs e, conseqüentemente, nas reações de RT-PCR para validação experimental dos pares de "clusters" selecionados.

1	Hep G2	fígado, hepatocarcinoma
2	ZR-75-1	mama, carcinoma ductal
3	TT	carcinoma; tireóide, medula
4	XP (transformada SV40)	Xeroderma Pigmentosum, pele
5	Hs 578T	mama, carcinoma ductal
6	U937	linfoma histiocítico, histiócitos
7	IM-9	linfoblastos B, sangue periférico
8	Hs 732.PL	placenta normal
9	Hs 1.TES	Testículo normal
10	SW480	cólon; adenocarcinoma coloretal
11	FADu	carcinoma de células escamosas; faringe
12	Saos 2	Osso; osteosarcoma
13	NCI-H1155	pulmão, carcinoma; sítio metastático: linfonodo
14	SCABER	Bexiga; carcinoma de células escamosas
15	SKmel 25	pele, melanoma maligno
16	HeLa	cervix; adenocarcinoma
17	T98G	cérebro; glioblastoma multiforme
18	A172	cérebro; glioblastoma
19	DU145	carcinoma próstata
20	rim	cultura primária de tumor de rim

2.3.2.2 Extração de RNA, avaliação da qualidade e controle da contaminação com DNA genômico

A extração de RNA das linhagens celulares acima descritas foi feita através da técnica de Sedimentação em cloreto de césio (CHIRGWIN et al. 1979) já bem estabelecida no laboratório da Dra. Mari Sogayar IQ-USP. A metodologia foi escolhida por apresentar melhor rendimento em termos de quantidade e qualidade do RNA, assim como menor grau de contaminação por DNA genômico, quando comparada com outras técnicas como, por exemplo, a extração utilizando Trizol®.

As células foram homogeneizadas em 9ml de solução de lise (4M Isotiocianato de guanidina / 25mM Citrato de Sódio – pH 7 / 0,1M β-mercaptoetanol). Em seguida, o lisado foi aplicado sobre 4ml de um colchão de cloreto de césio (5,7M CsCl / 25mM NaAc) e centrifugado a 29.000rpm/ 20°C por no mínimo 17 horas. Após a centrifugação, o RNA formou um precipitado no fundo do tubo que foi solubilizado em água biodestilada estéril (água DEPC – Di-etil pirocarbonato).

A quantificação do RNA obtido foi feita através da leitura de uma alíquota da amostra em espectrofotômetro com comprimento de onda equivalente a 260nm, considerando-se que 1 OD_{260nm} equivale a 40µg/ml de RNA. A relação entre as leituras realizadas a 260 e 280nm foi utilizada como parâmetro na estimativa do grau de contaminação do RNA por proteínas.

Para serem utilizados na síntese de cDNA, a integridade dos RNAs foi visualizada aplicando-se 1µg de RNA total em gel de agarose 1,0%. Antes de ser aplicado, o RNA foi desnaturado a 65°C por cinco minutos, sendo mantido em condição desnaturante em tampão de amostra contendo uréia (7M uréia e 30% glicerol). Todos os procedimentos foram feitos com materiais próprios para uso exclusivo de RNA. A coloração foi feita com brometo de etídio e o gel visualizado em luz ultravioleta (UV). Desta maneira, foram considerados íntegros, os RNAs que apresentaram as bandas correspondentes aos RNAs ribossômicos 28S e 18S bem evidentes, tendo a primeira o dobro da intensidade da segunda. (Figura 9).

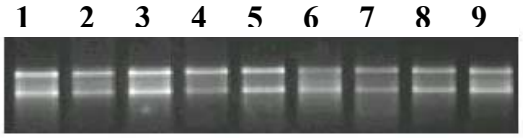


Figura 9 – Controle de qualidade aplicado após extração do RNA total. Análise do RNA total, extraído através da técnica de Sedimentação em cloreto de césio, através de um gel de agarose desnaturante no qual podem ser visualizadas as bandas correspondentes aos RNAs ribossômicos 28S e 18S. Neste gel foram analisadas as seguintes amostras de RNA: 1. SW480, 2. IM-9, 3. FADu, 4. Saos 2, 5. SCABER, 6. HeLa, 7. T98G, 8. A172, 9. DU145.

Para se ter a garantia de que o RNA extraído não estava contaminado com DNA genômico, foi aplicado o teste de hMLH1 ("human mut-L homologue 1") que consiste na amplificação por PCR de uma alíquota do RNA extraído (SILVA et al. 2003). Esta reação teve iniciadores desenhados nos introns 12 e 13 do gene hMLH1 (Forward - 5'TGG TGT CTC TAG TTC TGG3' e Reverse - 5'CAT TGT TGT AGT AGC TCT GC3') e utilizou como "template" 200ng de RNA total. Para um volume final igual a 25µl, a reação foi realizada em presença de tampão 1x Taq DNA polimerase, 1,6mM MgCl2, 0,2mM dNTPs, 0,5µM de cada iniciador e 1 unidade de Taq DNA polimerase (GIBCO/BRL). A reação teve 35 ciclos com uma temperatura de anelamento de 55°C e um tempo de extensão de 6 minutos. Caso o RNA estivesse contaminado com DNA seria amplificado um fragmento de aproximadamente 250pb visualizado em gel de poliacrilamida 8% (Figura 10). Nas amostras em que foi detectada a contaminação aplicou-se o tratamento com DNAse (Promega).

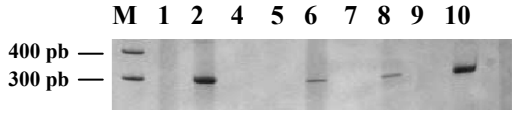


Figura 10 – Análise da presença de contaminação com DNA genômico no RNA total extraído. Visualização, em gel de poliacrilamida 8%, da amplificação do RNA total com iniciadores desenhados nos introns 12 e 13 do gene MLH1. Foram analisadas as seguintes amostras: 1. Saos-2, 2. ZR-75-1, 3. HeLa, 4. SW480, 5. Hep G2, 6. U937, 7. IM-9, 8. FADu e 9. SCABER. As amostras de RNA das linhagens 2. ZR-75-1, 6. U937 e 8. FADu apresentaram-se contaminadas com DNA genômico. A canaleta 10 corresponde ao controle positivo com DNA genômico e o peso molecular (M) aplicado foi o 100bp "ladder".

2.3.2.3 Tratamento do RNA total com DNAseI

O RNA total, quando necessário, foi tratado com DNAse livre de RNAse (Promega), utilizando 10 unidades da enzima para tratar 50µg de RNA por 60 minutos a 37°C, a fim de eliminar uma possível contaminação deste material com DNA genômico. A reação foi interrompida pela adição de uma solução de neutralização (0,05M EDTA, 1,5M NaOAC e 1% SDS) e o RNA foi extraído com fenol/clorofórmio e precipitado a –20°C por 18 horas pela adição de 250mM NaCl e etanol 100% com um volume igual a duas vezes o volume inicial. Após centrifugação, o RNA foi solubilizado em 30,0µl de água DEPC. A quantificação foi feita através da leitura de uma alíquota da amostra no espectrofotômetro a 260 e 280nm conforme descrito anteriormente. As amostras tratadas foram novamente submetidas ao teste de hMLH1 para confirmar a eliminação da contaminação.

2.3.2.4 Síntese de cDNA

Para a síntese de cDNA foram utilizados 2µg de RNA total por reação, com iniciadores oligo(dT)12-18, utilizando o kit "Superscript^{TM} Preamplification System for First-Strand Synthesis", segundo indicações do fornecedor. A síntese foi feita em larga escala para distribuição entre todos os laboratórios de validação.

Vale ressaltar que modificações no protocolo de síntese foram introduzidas no desenrolar do projeto para aumentar a eficiência de validação. Basicamente, a síntese de cDNA passou a ser feita a partir de RNA poli A^+ isolado de 200µg de RNA total utilizando-se o "PolyAttract mRNA isolation kit" (Promega) com a utilização de uma combinação de iniciadores randômicos, oligo(dT) e "SuperScipt II" (Invitrogen) seguindo instruções do fornecedor. As alterações foram introduzidas a fim de favorecer a amplificação de transcritos de baixa abundância e alto peso molecular.

Antes da distribuição, a qualidade do cDNA foi avaliada através da amplificação dos genes controles GAPDH (Gliceraldeído-3-fosfato-desidrogenase) (Forward - 5'CTG CAC CAC CAA CTG CTT A3' e Reverse - 5'CAT GAC GGC AGG TCA GGT C3') e Notch2 (Forward - 5'ACT GTG GCC AAC CAG TTC TC3' e Reverse - 5' CTC TCA CAG GTG CTC CCT TC3').

Para a amplificação do gene GAPDH foi realizada uma reação com um volume final igual a 10µl, em presença de tampão 1x Taq DNA polimerase, 1,0mM MgCl2, 0,1mM dNTPs, 0,3µM de cada iniciador e 0,5 unidade de Taq DNA polimerase (GIBCO/BRL), utilizando como "template" 1µl do cDNA sintetizado. A reação teve 35 ciclos com uma temperatura de anelamento de 60°C e um tempo de extensão final de 6 minutos. A amplificação de um fragmento de 305pb visualizado em gel de poliacrilamida 8% (Figura 11A), confirmou a eficiência da síntese de cDNA uma vez que o gene GAPDH é altamente expresso em quase todos os tecidos.

O gene Notch 2, entretanto, é um gene de 11Kb e de baixo nível de expressão de forma que a sua utilização como controle de síntese de cDNA garante a eficiência da síntese em termos da representatividade de transcritos longos e raros uma vez que os iniciadores utilizados em sua amplificação foram construídos na região 5' do gene. A reação para amplificação deste gene foi realizada com um volume final igual a 25µl, em presença de tampão 1x Taq DNA polimerase, 1,5mM MgCl2, 0,1mM dNTPs, 0,4µM de cada iniciador e 1 unidade de Taq DNA polimerase (GIBCO/BRL), utilizando como "template" 1µl do cDNA sintetizado. A reação teve uma desnaturação inicial de 94°C por 4 minutos e 55°C por 2 minutos, 40 ciclos com uma temperatura de anelamento de 55°C e um tempo de extensão de 1 minuto e um tempo de extensão final de 10 minutos. A amplificação de um fragmento de 300pb visualizado em gel de agarose 1% (Figura 11B), confirmou, desta forma, a eficiência da síntese de cDNA.

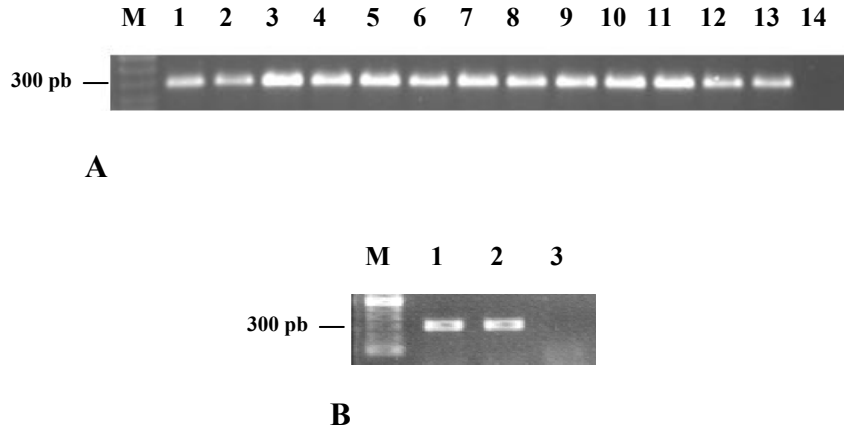


Figura 11 - Avaliação da qualidade da síntese do cDNA. (A) Visualização da amplificação do gene controle GAPDH em gel de poliacrilamida 8% nas seguintes amostras de cDNA: 1. Saos-2, 2. ZR-75-1, 3. HeLa, 4. SW480, 5. Hep G2, 6. U937, 7. IM-9, 8. FADu, 9. SCABER, 10. Skmel 25, 11. T98G, 12. A172. A canaleta 13 corresponde ao controle positivo com DNA genômico e a 14 ao controle negativo. O peso molecular (M) aplicado foi o 100bp "ladder". (B) Visualização da amplificação do gene controle Notch2 em gel agarose 1% nas amostras de linfoblastos B (1) e cultura primária de tumor de rim (2). A canaleta 3 corresponde ao controle negativo da reação e o peso molecular (M) aplicado foi o 100bp "ladder".

2.3.3 Caracterização de Novos Transcritos Humanos

Os pares de "clusters" de ESTs (TFs) foram selecionados automaticamente a partir do banco de dados e inspecionados manualmente pela coordenação através da interface gráfica. Iniciadores específicos para os "clusters" selecionados foram desenhados automaticamente e tiveram sua síntese solicitada pelos grupos da coordenação.

Além disso, um painel contendo cDNAs de diferentes linhagens tumorais e iniciadores específicos para a amplificação de diferentes TFs foram repassados periodicamente aos laboratórios de validação. Um protocolo padrão para a amplificação, assim como possíveis modificações nas condições de RT-PCR, foi elaborado pela coordenação e enviado aos grupos. A reação padrão consistiu em um

volume final de 25µl, em presença de 1µl de cDNA, tampão 1x Taq DNA polimerase, 1,5mM $MgCl_2$, 0,2mM dNTPs, 0,4µM de cada iniciador, e 1 unidade de Taq DNA polimerase (GIBCO/BRL). A reação teve uma desnaturação inicial de 94°C por 4 minutos, 35 ciclos com uma temperatura de anelamento de 55°C e um tempo de extensão final de 10 minutos.

Com base nestas informações os grupos procederam com a amplificação, clonagem, seqüenciamento e submissão das seqüências ao banco de dados. Modificações no protocolo inicial ficaram a critério dos grupos e incluíram a temperatura de anelamento, utilização de novos iniciadores, concentração de $MgCl_2$ e adição de "enhancers" como a betaína e o DMSO. Além disso, reações de PCR "nested" também foram aplicadas nos casos em que os TFs apresentavam um baixo nível de expressão.

Os produtos amplificados foram diretamente seqüenciados com os mesmos iniciadores utilizados para a reação de RT-PCR ou clonados e, em seguida, seqüenciados. As reações de seqüenciamento foram realizadas utilizando-se o "DYEnamic^{TM} ET terminator Cycle Sequencing Kit" (Amersham Pharmacia) e o seqüenciador ABI377 Prism (Perkin Elmer) seguindo as instruções do fornecedor.

As seqüências de validação submetidas pelos grupos que passaram no "pipeline" de qualidade foram incorporadas ao banco de dados e à interface gráfica do projeto depois de comprovada a especificidade das mesmas. Esta especificidade foi avaliada automaticamente através do BLASTN ("e-value" de 10^{-30}) contra a seqüência genômica a partir do qual o TF foi escolhido. Para as seqüências específicas (com BLASTN positivo) os dados referentes aos alinhamentos foram carregados no banco de dados e representados na interface gráfica como seqüências de validação.

Estas seqüências de validação, em conjunto com as seqüências das ESTs dos "clusters" selecionados, foram utilizadas para a produção de seqüências consensos dos transcritos validados. O sistema empregado para a montagem dos consensos foi baseado nos programas PhredPhrap e BLAST. O programa PhredPhrap foi utilizado para montar as seqüências de validação e o consenso gerado foi, então, alinhado com as seqüências de ambas as ESTs utilizando o BLAST. As coordenadas do BLAST foram, em seguida, utilizadas para a montagem do consenso final, incluindo as seqüências de ambas as ESTs (Figura 12). Uma interface gráfica foi desenvolvida para monitorar o processo de montagem e permitir o acesso às seqüências consensos (http://200.18.51.201/transcript/). Os consensos obtidos foram utilizados para a anotação e caracterização dos novos transcritos.

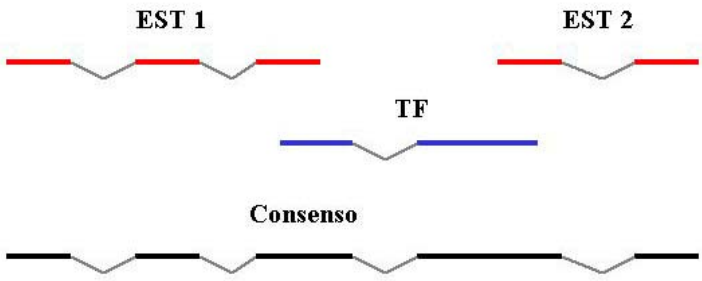


Figura 12 – Representação esquemática da montagem da seqüência consenso. Construção da seqüência consenso (linha preta) com base nas coordenadas das ESTs (linhas vermelhas) e da seqüência do TF (linha azul).

Como o objetivo geral do projeto era a caracterização de novos transcritos humanos, a anotação foi direcionada para as seqüências produzidas durante o projeto que não possuíam similares humanas depositadas em bancos de dados públicos na época da anotação. Em um primeiro momento, as seqüências consenso de cada transcrito foram mapeadas na montagem do genoma humano disponibilizada pela

Universidade de Santa Cruz – Califórnia (UCSC- montagem de Abril de 2003) (http://www.genome.ucsc.edu) (KENT et al. 2002). O mapeamento foi feito através do alinhamento contra a seqüência do genoma humano utilizando o programa BLAT. O resultado do alinhamento pode ser visualizado através da interface gráfica da UCSC que permite uma comparação visual entre o alinhamento obtido e os alinhamentos de genes humanos já conhecidos ("Known Genes", RefSeq e "Human mRNAs") e coordenadas de exons preditos por programas de computador a partir da seqüência genômica humana (GenScan, FZgenessh++, GeneID). Assim sendo, através de uma rápida inspeção manual é possível verificar a existência de transcritos humanos já caracterizados cujas coordenadas de alinhamento se sobrepõem total ou parcialmente à seqüência consenso gerada em nosso projeto. Além disso, também é possível verificar a existência total ou parcial de predição gênica para a mesma seqüência consenso.

Um transcrito validado foi considerado um gene novo se as coordenadas do alinhamento com a seqüência genômica não apresentassem sobreposição com as coordenadas de qualquer outra seqüência referente a um transcrito humano já caracterizado. Já com relação às predições, é importante ressaltar que foi considerada a predição individual dos exons ao invés da predição da seqüência completa do transcrito. Assim, um exon validado foi considerado predito se as coordenadas de alinhamento fossem definidas por pelo menos um dos três programas de predição citados anteriormente, não sendo necessária a sobreposição exata das bordas do exon. Além disso, um transcrito validado foi considerado não predito se todos os exons não fossem preditos por nenhum dos programas de predição.

Em seguida, as seqüências consensos geradas em nosso projeto e que não apresentaram sobreposição com transcritos humanos já depositados em banco de

dados públicos foram, então, anotadas através de buscas por similaridade ao nível de aminoácidos utilizando o programa BLASTX, disponibilizado pelo NCBI, e através da identificação de domínios proteicos utilizando o programa "Motif Scan in a Protein Sequence" do site http://hits.isb-sib.ch/cgi-bin/PFSCAN.

2.3.4 Caracterização de Formas Alternativas de "Splicing"

Para a validação de um único TF os grupos muitas vezes utilizaram cDNAs derivados de diferentes tecidos permitindo dessa forma, ao longo do projeto, a caracterização de formas alternativas de "splicing" em paralelo à determinação da estrutura do novo transcrito. O grau de variabilidade referente ao uso alternativo dos exons ("splicing" alternativo) foi avaliado em todas as seqüências consenso obtidas. As seqüências individuais geradas para cada TF, juntamente com as ESTs correspondentes, foram submetidas a um alinhamento múltiplo em relação à montagem do genoma humano disponibilizada pela UCSC através do programa BLAT. Foi investigada a presença de formas alternativas de "splicing" entre diferentes seqüências de uma mesma TF e das mesmas em relação as ESTs. Além disso, foram verificadas formas alternativas em relação aos transcritos humanos, quando estes já estavam descritos.

Com o intuito de eliminar artefatos de alinhamento causados por erros de seqüenciamento e problemas na montagem da seqüência genômica foram consideradas formas alternativas de "splicing" somente exons definidos pela presença dos sítios aceptores e doadores de "splicing" conservados (GT/AG). Para as seqüências que apresentaram sítios conservados de "splicing" foram desenhados iniciadores (Tabela 2) específicos para cada uma das isoformas (Figura 13) utilizando o programa Primer3 e os parâmetros padrões já descritos anteriormente. A

presença das diferentes isoformas foi analisada através de reações de RT-PCR "touchdown" utilizando-se o painel de cDNAs (descrito na tabela 1 do item 2.3.2.1) composto por 20 linhagens celulares derivadas de diferentes tecidos. Além disso, a amplificação do gene GAPDH foi utilizada como controle da integridade e quantidade de RNA utilizado na síntese dos cDNAs.

Tabela 2 - Seqüências dos iniciadores desenhados para validação experimental das formas alternativas de "splicing". Os iniciadores foram construídos para cada uma das isoformas dos TFs que apresentaram formas alternativas de "splicing" entre suas seqüências de validação.

TF	Iniciador	Seqüência	Temperatura de anelamento
TF00118	P1	GGCGCTCCAAGGCACTCTTA	72,3°C
TF00118	P2	CCCACTCCCCGAACTCTACAGC	76,9°C
TF00200	P1	TCAGATGCCACCAACATTGAGG	71,3°C
TF00200	P2	CCTCTCTGAGAACCCGATCAGC	75,0°C
TF00200	P3	CCTGCTGTTCATTCTGCTTTG	68,9°C
TF00200	P4	GGCAATGTCCCTGCCCTTTATT	71,3°C
TF00274	P1	GCTGGATGCAACCACAACAGC	72,8°C
TF00274	P2	TGTCCAGGTACTCCTCCACCACA	75,3°C
TF00274	P3	GCCGGTTTAGCGGGAACTTCAT	73,2°C
TF00351	P1	TGGTGAGTGTCCCCAAGTACCA	73,2°C
TF00351	P2	GCACCAGGTCCATGATGAACAG	73,2°C
TF01004	P1	GAGTTCAGAAACTGCCGCCGTA	73,2°C
TF01004	P2	GCAAGGTTGGCCTTCTGTTCAC	73,2°C
TF01004	P3	GTTGGCCATGGCACACATCAT	70,8°C
TF01058	P1	TGGTGAATGGTTCCTGCTTTCA	69,5°C
TF01058	P2	CCGAAGCAGGTGGATTTCTTGA	71,3°C
TF01058	P3	GGGGGTTGCTTGATCCTAGATG	73,2°C

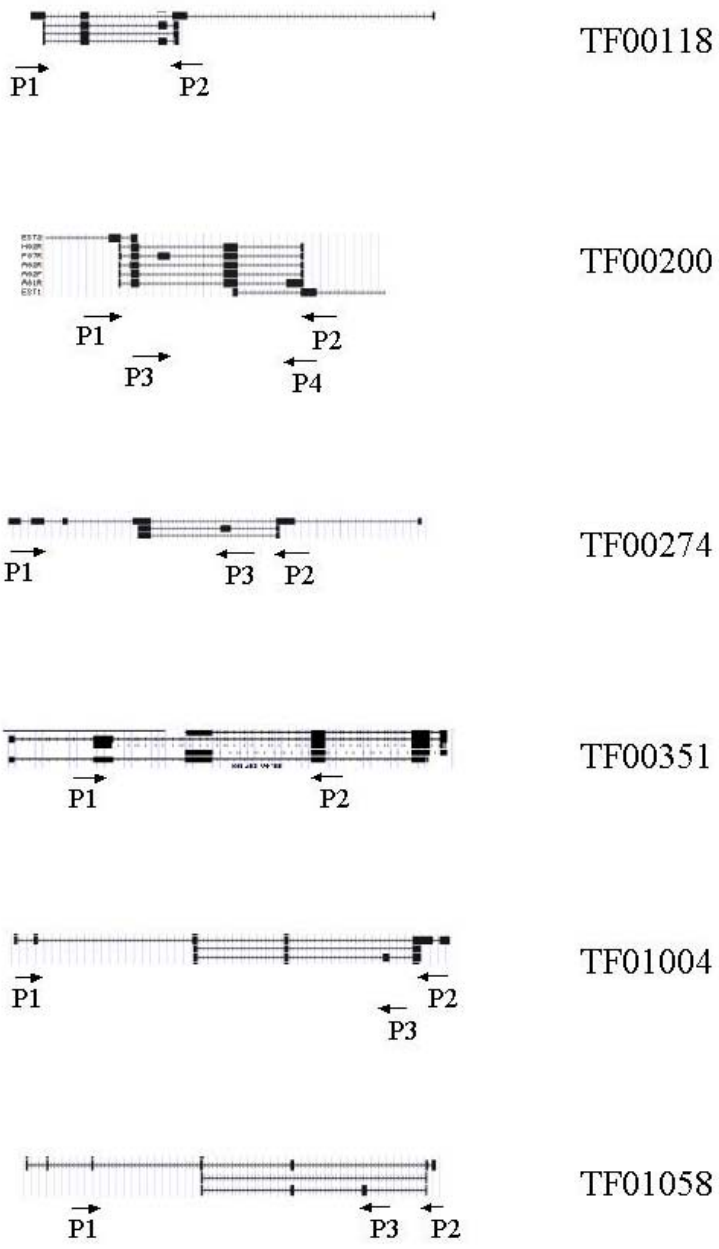


Figura 13 - Representação esquemática do alinhamento entre a seqüência genômica e as seqüências de validação correspondentes a cada um dos TFs que apresentaram formas alternativas de "splicing". Alinhamento das seqüências individuais geradas para cada TF contra a montagem do genoma humano disponibilizada pela UCSC utilizando o programa BLAT. Os quadrados pretos representam os exons e as linhas cinzas os introns. As setas indicam as regiões onde foram construídos os iniciadores (P) específicos para as isoformas presentes em cada TF.

As reações foram realizadas em um volume final de 25µl, em presença de 1µl de cDNA, tampão 1x Taq DNA polimerase, 1,5mM $MgCl_2$, 0,2mM dNTPs, 0,4µM de cada iniciador, e 1 unidade de Taq DNA polimerase (GIBCO/BRL). A reação teve uma desnaturação inicial de 94°C por 3 minutos, uma temperatura de anelamento inicial de 72°C sendo reduzida em 1°C por ciclo até alcançar a temperatura de 62°C quando iniciou os 25 ciclos com esta temperatura de anelamento. Além disso, a reação foi realizada com um tempo de extensão final de 10 minutos e os produtos amplificados foram visualizados em gel de agarose 1,5%.

A decisão de ser utilizada a RT-PCR "touchdown" deve-se ao fato de que todas as combinações possíveis entre os iniciadores foram realizadas em uma mesma placa. Assim, como os iniciadores apresentavam diferentes temperaturas de anelamento, este tipo de reação favoreceu o anelamento dos mesmos uma vez que a cada ciclo a temperatura era reduzida.

RESULTADOS E DISCUSSÃO

2.4 RESULTADOS E DISCUSSÃO

2.4.1 Construção do banco de dados do projeto TFI e seleção dos "clusters" para validação experimental

A seleção de transcritos humanos parcialmente representados por ESTs para validação experimental foi feita a partir do Banco de dados do Transcriptoma Humano desenvolvido pelo Laboratório de Biologia Computacional do Instituto Ludwig e contém informações relacionadas ao alinhamento entre seqüências expressas e a seqüência genômica. O mapeamento das seqüências transcritas na seqüência genômica foi feito com o programa BLASTN sendo, para tanto, utilizada a montagem do genoma humano fornecida pelo NCBI como fonte da informação genômica e seqüências transcritas obtidas das divisões dbEST e NR do GenBank. Os dados destes alinhamentos foram armazenados em um banco de dados relacional MySQL e utilizados para criar "clusters" de seqüências transcritas baseados nas posições de cada seqüência dentro dos clones genômicos individuais.

Os membros de cada "cluster" foram determinados através das coordenadas dos exons em relação à seqüência genômica. Se as coordenadas de pelo menos um exon fossem comuns a dois transcritos, os mesmos eram considerados parte de um mesmo "cluster". Este processo de clusterização contribui para a produção de "clusters" de alta qualidade agrupando corretamente seqüências transcritas com uma sobreposição mínima, distinguindo membros de famílias gênicas e evitando a formação de quimeras.

Além disso, a comparação com a seqüência genômica permitiu a distinção entre as seqüências transcritas e as derivadas de uma contaminação com DNA ou

moléculas de mRNA imaturo, uma vez que estas formam clusters de seqüências alinhadas continuamente, ou seja, sem “splicing”. É importante lembrar que o alinhamento entre as seqüências expressas e a seqüência genômica geralmente apresenta interrupções devido a presença de introns na seqüência genômica os quais são removidos das seqüências expressas após a ocorrência de “splicing”.

Também foram mapeadas no genoma as 3’ “tags”, derivadas das extremidades 3’ dos diferentes transcritos, servindo de âncoras para delimitar com precisão o final dos transcritos (ISELI et al. 2002). Após o mapeamento no genoma as mesmas foram agrupadas em “clusters” sendo identificadas, no total, 100.261 “tags”.

Em uma visão geral do banco, foram identificados 244.148 “clusters” dos quais 14.598 continham pelo menos uma seqüência completa de cDNA e 229.550 “clusters” eram compostos somente por seqüências parciais. Do grupo de 14.598 “clusters”, vale ressaltar que 13.149 “clusters” (90%) apresentavam pelo menos uma EST correspondente e 1.449 (10%) eram compostos somente por seqüências completas de cDNA. Estes dados demonstram, desta forma, que apesar de existirem mais de cinco milhões de seqüências expressas disponíveis em bancos de dados públicos as mesmas ainda não cobrem totalmente o transcriptoma humano. É notável que os “clusters” compostos exclusivamente por ESTs apresentam um número reduzido de seqüências (média de 5,9 seqüências) derivadas de um número reduzido de tecidos diferentes (média de 3,0 tecidos) quando comparados com os “clusters” que contêm seqüências completas de cDNA os quais apresentam uma média de 65,5 seqüências derivadas de oito diferentes tecidos. Assim, baseando-se nestas observações, podemos acreditar que os transcritos humanos ainda não identificados apresentam um nível de expressão muito baixo e são expressos em um número restrito

de tecidos. Desta forma, a caracterização desses transcritos só será possível com a utilização de metodologias diretas como a proposta por este projeto.

Uma vez desenvolvidas as ferramentas de bioinformática descritas anteriormente partiu-se para o processo de seleção de "clusters" no Banco de dados do Transcriptoma para validação experimental dos mesmos. Cada par de "cluster" selecionado definiu um TF distribuído aos grupos de validação. A seleção dos "clusters" foi feita segundo critérios que visavam maximizar a eficiência de validação e o sucesso do projeto. Desta maneira, foram selecionados:

- pares de "clusters" de ESTs que não continham uma seqüência de mRNA completa, uma vez que o projeto priorizava a caracterização de novos transcritos humanos;
- pares de "clusters" que continham ESTs com "splicing", uma vez que a presença de "splicing" é a evidência mais forte que confirma a origem da EST como seqüência expressa;
- pares de "clusters" de ESTs que estivessem a uma distância máxima de 10Kb um do outro, uma vez que quanto maior a distância entre os "clusters", menor a probabilidade de eles pertencerem a um mesmo transcrito e menor a probabilidade de sucesso nas amplificações.

Com o estabelecimento dos critérios de seleção, programas na linguagem Perl foram desenvolvidos para a seleção dos pares de "clusters". Inicialmente foram selecionados 2.373 pares de "clusters" (aproximadamente 2% do número total de "clusters" compostos somente por seqüências parciais) e a coordenação se encarregou da inspeção manual e individual de cada um dos pares selecionados antes da distribuição para os grupos.

A inspeção manual permitiu uma avaliação da similaridade do alinhamento das seqüências transcritas com a seqüência genômica e da extensão do mesmo, assim como a posição dos pares de "clusters" selecionados em relação às 3' "tags". Desta forma, foram selecionados 489 pares de "clusters" para validação experimental. Cada "cluster" estava a uma distância média de 2.879pb de seu par, sendo que o número médio de ESTs em cada "cluster" foi de 5,92. Já com relação à distribuição de tecidos, cada "cluster" selecionado tinha em média ESTs derivadas de 3 tecidos diferentes e a grande maioria dos pares de "clusters" (67.3%) não apresentava ESTs derivadas de um mesmo tecido demonstrando, assim, que, possivelmente, a grande maioria dos transcritos humanos ainda não identificada apresenta um perfil de expressão restrito e um nível de expressão muito baixo.

2.4.2 Estratégia de Validação Experimental

Uma visão geral das estratégias computacionais e de validação experimental foi resumida na Figura 14 abaixo. Um total de cinco grupos de bioinformática e 31 laboratórios de validação interligados via Internet participaram, respectivamente, da fase computacional e experimental do projeto. Em seqüência à seleção dos "clusters" e inspeção manual dos mesmos, foram, automaticamente, construídos os iniciadores para a validação dos TFs por RT-PCR. A seqüência genômica foi escolhida como base no desenho dos iniciadores uma vez que, de modo geral, a mesma apresenta um nível de qualidade melhor quando comparado com as seqüências das ESTs.

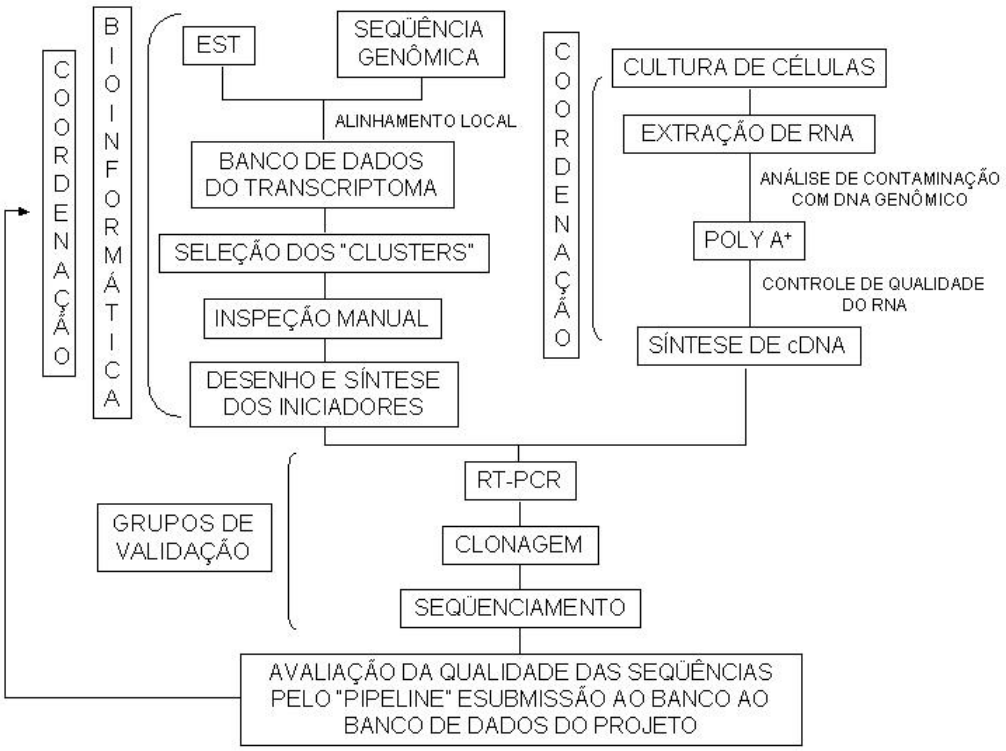


Figura 14 – Representação esquemática das estratégias computacionais e de validação experimental desenvolvidas no Projeto TFI. Esquema indicando todos os passos das estratégias computacionais e de validação experimental realizados durante o desenvolvimento deste projeto pelos grupos de bioinformática, validação e coordenação.

A síntese do cDNA foi um passo muito importante neste projeto já que tanto a qualidade quanto a representatividade dos diferentes tecidos influenciariam diretamente na eficiência de validação. Assim, vários controles foram aplicados para assegurar que o material distribuído aos grupos de validação era de alta qualidade e estava completamente livre de contaminação com DNA genômico. Vale citar que modificações no protocolo de síntese do cDNA foram adotadas durante o desenvolvimento do projeto e incluíram a utilização de RNA poliA+ e a combinação de oligo dT e iniciadores randômicos. Em dados gerais foram preparadas 22 amostras de cDNA representando 18 diferentes tecidos.

Um total de 3012 seqüências foi gerado ao longo do projeto. Desse total, 765 seqüências (25.4%) foram descartadas por apresentarem baixa qualidade ou por

corresponderem a seqüências repetitivas ou contaminantes. As demais seqüências (74.6%) foram analisadas quanto à sua especificidade através de BLASTN contra o clone genômico a partir do qual o TF foi selecionado. Uma vez comprovada a especificidade, os dados dos alinhamentos foram incluídos no banco de dados original, identificados como uma seqüência de validação e incorporados na interface gráfica do projeto (Figura 15).

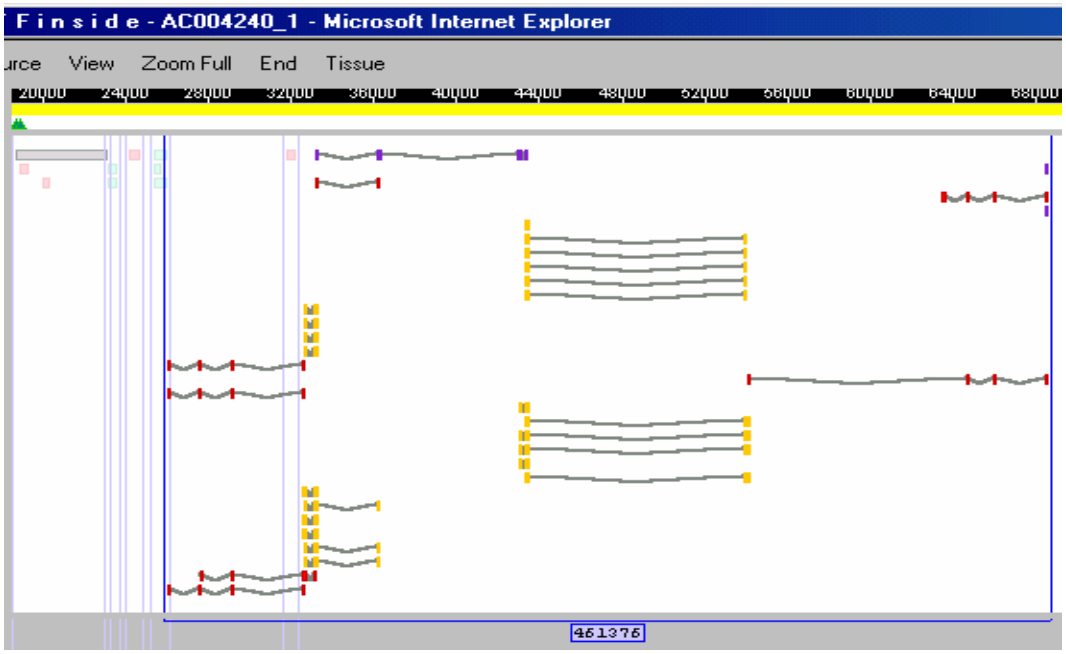


Figura 15 - Visualização da Interface gráfica do Projeto TFI após a submissão de seqüências de validação. As seqüências expressas e a seqüência genômica estão representadas como descrito na Figura 6 enquanto as seqüências de validação estão representadas em amarelo.

Dos 489 TFs distribuídos, 211 (43,1%) foram validados, com uma média de 7,5 TFs validados por grupo. Por se tratar de uma abordagem inédita, fica difícil avaliar a eficiência de amplificação obtida. É importante ressaltar que a eficiência de validação variou bastante entre os grupos (15.3 a 91.6%) devido, principalmente, a diferenças na experiência prévia dos mesmos com as técnicas de RT-PCR, clonagem e seqüenciamento.

Para tentar minimizar essas diferenças, a coordenação implantou um sistema de tutoria, no qual cada um dos membros dos laboratórios de coordenação passou a acompanhar semanalmente o progresso e as dificuldades de cada grupo de validação através de contato por e-mail ou telefone. Os tutores re-avaliaram manualmente cada TF, verificaram o desenho dos iniciadores e forneceram orientações para a realização da RT-PCR e seqüenciamento, sugerindo que um melhor treinamento dos grupos poderia elevar a eficiência obtida até o momento.

O impacto da implantação da tutoria foi, no geral, muito positivo como pode ser verificado na Tabela 3. Para avaliação deste impacto, os grupos foram classificados de acordo com a sua eficiência de validação e a distribuição dos grupos nessas classes foi comparada antes e depois da implantação da tutoria. Antes da implantação da tutoria, 16% dos grupos apresentavam eficiência de validação entre 0 e 10% enquanto somente 9,7% apresentava uma eficiência de validação entre 40 e 50% e 13% dos grupos possuíam eficiência acima de 50%. Após a implantação da tutoria, nenhum grupo continuou com uma porcentagem de eficiência de validação entre 0 e 10%. Já com relação à porcentagem de grupos com eficiência de validação entre 40 e 50%, esta subiu para 29% enquanto o número de grupos com eficiência maior que 50% subiu para 32,25%. Com a implantação da tutoria a eficiência de validação global do projeto subiu de 28.8% para 43.1%.

Tabela 3 - Avaliação do impacto da implantação da tutoria na eficiência de validação dos grupos. Os grupos foram classificados de acordo com a sua eficiência de validação e a distribuição dos grupos nessas classes pode ser comparada antes e depois da implantação da tutoria.

Eficiência de Validação	N° de grupos antes da tutoria (%)	N° de grupos após a tutoria (%)
0-10%	5 (16%)	0
10-20%	3 (9,7%)	3 (9,7%)
20-30%	10 (32,25%)	3 (9,7%)
30-40%	6 (19,35%)	6 (19,35%)
40-50%	3 (9,7%)	9 (29%)
>50%	4 (13%)	10 (32,25%)

No total foram gerados 59.975 pares de bases não redundantes de seqüências transcritas e caracterizados 432 novos exons correspondentes a 211 transcritos humanos parcialmente representados por ESTs. A média em pares de bases novos por TF foi de 281,6 e a mediana de 207 enquanto a média de exons novos foi de 2,03 por TF distribuído e a mediana de 2,0. As seqüências correspondentes aos TFs validados foram submetidas ao dbEST e os números de acesso foram de CF272536 a CF272733.

Para tentar identificar outros fatores, além da experiência dos grupos de validação, que estariam influindo na eficiência de validação dos TFs selecionados, foi realizada uma análise, através do teste de Mann-Whitney, entre um grupo de TFs validados (composto por 174 TFs) e um grupo de TFs não validados (composto por 208 TFs). Os seguintes parâmetros foram avaliados entre os dois grupos:

- Distância entre os dois "clusters" selecionados para validação. Esse critério foi escolhido pois quanto maior a distância entre os "clusters", menor a probabilidade deles pertencerem a um mesmo transcrito e maior a dificuldade encontrada na validação;

- Número de ESTs em cada “cluster”; pois sabemos que quanto menor o número de ESTs menor o nível de expressão gênica e, portanto, maior a dificuldade na validação;
- Número de tecidos diferentes em cada “cluster”; já que quanto menor o número de tecidos diferentes em cada “cluster”, mais restrito é o padrão de expressão do transcrito e mais difícil é a validação experimental;
- Presença ou ausência de ESTs derivadas de um mesmo tecido para cada um dos “clusters”. A presença de ESTs de um mesmo tecido nos dois “clusters” aumentaria as chances dos dois “clusters” pertencerem a um mesmo transcrito e, portanto, mais provável seria a validação.

Como pode ser verificado na Tabela 4, os TFs validados apresentaram em média uma menor distância entre os “clusters” selecionados, um maior número de ESTs em cada “cluster”, e um maior número de ESTs derivadas de tecidos diferentes. Todas estas diferenças foram estatisticamente significativas confirmando a tendência de validação para TFs com maior nível de expressão em um número também maior de tecidos. No entanto, com relação à presença de ESTs derivadas do mesmo tecido em ambos os “clusters”, a mesma não afetou o processo de validação.

Tabela 4 - Análise comparativa entre TFs validados e não validados. A análise estatística foi realizada, através do teste de Mann-Whitney, entre um grupo de TFs validados, composto por 174 TFs, e um grupo de TFs não validados, composto por 208 TFs. A significância estatística foi inferida através do valor de p.

Análise	TFs validados (desvio padrão)	TFs não validados (desvio padrão)	Valor de p
Distância média entre os "clusters"	2,609 (3,202)	3,105 (2,942)	0,008
Média do n° de ESTs em cada "cluster"	6,10 (8,91)	5,77 (13,23)	0,010
Média do n° de diferentes tecidos em cada "cluster"	3,45 (4,27)	2,85 (4,54)	0,002
N° de TFs com tecido comum em ambos os "clusters"	63	62	0,223

Estas informações podem ser utilizadas para melhorar os critérios de seleção de "clusters" para validação. Poderíamos, por exemplo, passar a selecionar "clusters" com uma distância máxima de 3.0kb, compostos por mais de 6 ESTs derivadas de pelo menos 4 tecidos distintos. No entanto, é importante ressaltar que esses critérios reduziriam muito o número de "clusters" selecionados e que esses dados refletem características peculiares dos transcritos humanos que ainda não foram completamente caracterizados: padrão de expressão baixo e restrito.

2.4.3 Anotação dos Novos Transcritos Humanos

As seqüências de validação, em conjunto com as seqüências de ESTs dos "clusters" selecionados, foram utilizadas para a produção de seqüências consensos dos transcritos validados. O sistema utilizado para a montagem dos consensos foi baseado nos programas PhredPhrap e BLAST. O programa PhredPhrap foi utilizado para montar as seqüências de validação e o consenso gerado foi alinhado com as

seqüências de ambas as ESTs utilizando BLAST. As coordenadas do BLAST foram, em seguida, utilizadas para a montagem do consenso final, incluindo as seqüências de ambas as ESTs. Um total de 186 consensos foi obtido a partir dos 211 TFs validados. Para 25 TFs não foi possível a obtenção de uma seqüência consenso devido a presença de seqüências repetitivas e formas alternativas de "splicing", assim como seqüências incompletas dos fragmentos de validação. Os consensos obtidos apresentaram uma média de 1.240pb e foram disponibilizados na Internet através do endereço http://200.18.51.201/transcript/ (Figura 16)

Figura 16 - "Homepage" construída para visualização das montagens das seqüências dos TFs e seus respectivos consensos. Nesta página é possível obter a seqüência referente a cada uma das ESTs utilizadas para seleção dos TFs assim como a montagem das seqüências de validação dos TFs e da seqüência consenso. Nesta figura podem ser visualizados a seqüência do TF00002 e os dados dos alinhamentos entre esta e as seqüências das ESTs 1 e 2 correspondentes a este TF.

Como o objetivo geral do projeto é a caracterização de novos transcritos humanos, a anotação foi direcionada para as seqüências produzidas durante o projeto

que não possuíam similares humanas depositadas em bancos de dados públicos na época da anotação. Em um primeiro momento, a seqüência consenso de cada transcrito foi mapeada na montagem do genoma humano disponibilizada pela Universidade de Santa Cruz – Califórnia (UCSC- montagem de Abril de 2003) (KENT et al. 2002). O mapeamento foi feito através do alinhamento contra a seqüência do genoma humano utilizando o programa BLAT. O resultado deste alinhamento pode ser visualizado através da interface gráfica da UCSC (Figura 17). A interface, por sua vez, permite uma comparação visual entre o alinhamento obtido e os alinhamentos de genes humanos já conhecidos e coordenadas de exons preditos por programas de computador a partir da seqüência genômica humana (GenScan, FZgenessh++, GeneID).

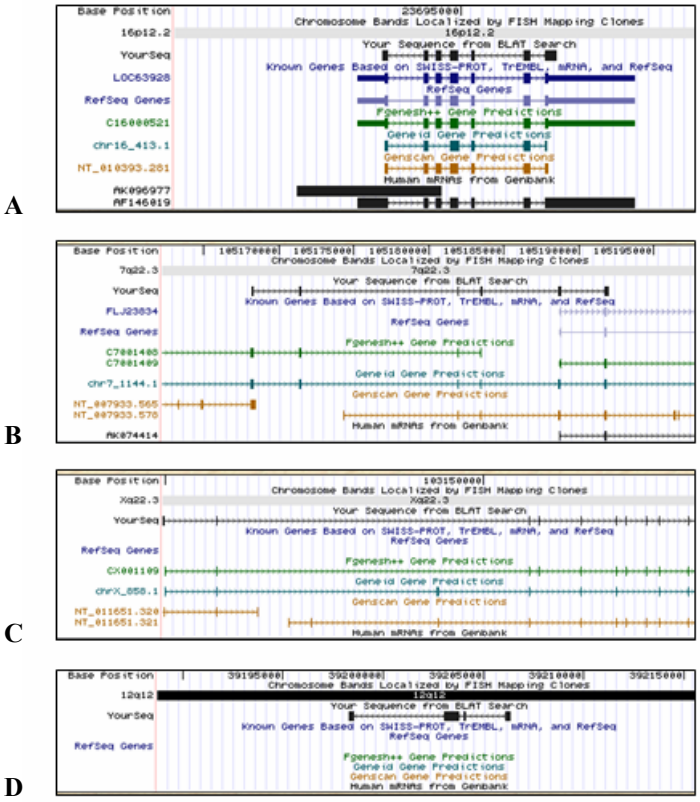


Figura 17 - Alinhamento das seqüências consensos contra a seqüência genômica. O alinhamento das seqüências consensos foi realizado contra a montagem do genoma humano disponibilizada pela UCSC utilizando o programa BLAT. As seqüências consenso estão representadas em preto ("Your sequence"), predições gênicas estão representadas em verde ("Fgenesh++") e em areia ("GenScan"), seqüências de transcritos humanas já depositadas em bancos de dados públicos em azul ("Known genes") e em preto (human mRNAs). (A) Seqüência consenso do TF00023 com sobreposição total com transcrito humano já conhecido. (B) Seqüência consenso do TF01102 com sobreposição parcial com transcrito humano já conhecido (C) Seqüência consenso do TF01013 sem sobreposição com transcrito humano já conhecido mas com sobreposição com predição gênica (D) Seqüência consenso do TF00125 sem sobreposição com transcrito humano já conhecido e sem sobreposição com predição gênica.

Uma fração significativa dos consensos (68.8%) apresentou uma sobreposição completa com as coordenadas de genes conhecidos ou seqüências completas de mRNA humano (Figura 17a). Após a análise manual de alguns casos, verificamos que a maior parte dessas seqüências havia sido depositada recentemente durante o desenvolvimento do nosso projeto, sendo a grande parte delas derivadas de

projetos de seqüenciamento de clones de cDNA completos (MGC e Rinken). Além disso, uma fração bem reduzida dos nossos consensos (10.2%) apresentava sobreposição parcial com seqüências de mRNA já descritas representando, assim, extensões (maior parte da extremidade 5') das seqüências publicadas (Figura 17b).

Dos 186 consensos analisados, 39 (21%) representavam seqüências completamente novas sem nenhuma sobreposição com seqüências de mRNA humanas disponíveis em bancos de dados públicos até julho de 2003. Desses consensos, 12 (6,5 %) e 15 (8%) foram, respectivamente, total e parcialmente preditos por programas de computador (Figura 17c). Além disso, doze consensos obtidos (6,5%) não foram preditos por nenhum programa de predição gênica e a existência desses transcritos só pôde ser confirmada através de dados gerados pelo nosso projeto (Figura 17d). Esses dados demonstram que os programas de predição não são 100% eficientes e comprovam a necessidade do seqüenciamento direto de moléculas de cDNA para a caracterização de novos transcritos humanos. A tabela completa com os dados de anotação de todos os consensos segue como anexo 2 sendo apresentado na tabela 5 um resumo dos dados discutidos anteriormente.

Tabela 5 - Classificação das TFs analisadas e anotadas quanto à descrição prévia e existência de predição.

Seqüência consenso	N° absoluto	Porcentagem
Gene conhecido	128	68.8%
Extensão de gene conhecido	19	10.2%
Transcrito novo com predição total	12	6.5%
Transcrito novo com predição parcial	15	8.0%
Transcrito novo sem predição	12	6.5%
Total	186	100%

Em seguida, as seqüências consensos geradas em nosso projeto e que não apresentaram sobreposição com transcritos humanos já depositados em banco de dados públicos foram anotadas através de buscas por similaridade em nível de aminoácidos utilizando o programa BLASTX e através da identificação de domínios proteicos utilizando o programa "Motif Scan in a Protein Sequence" do site http://hits.isb-sib.ch/cgi-bin/PFSCAN. Das 39 seqüências consensos que não apresentaram sobreposição com transcritos humanos conhecidos, 27 (69,2%) apresentaram uma fase aberta de leitura de pelo menos 100 aminoácidos e 8 (20,5%) apresentaram um domínio protéico conservado (anexo 1). Vale ressaltar que alguns dos domínios encontrados são típicos de famílias de proteínas envolvidas em processos importantes como transcrição, transdução de sinais, ciclo celular, entre outros.

Como exemplo dessa abordagem podemos tomar o TF00318 cuja seqüência consenso montada apresenta 1296pb e alinha com alta similaridade com um clone genômico localizado no cromossomo 14 região 14q32.13. A tradução dessa seqüência de nucleotídeos gera uma seqüência de 418 aminoácidos que quando analisada para a presença de domínios protéicos conhecidos apresenta uma região de alta similaridade com uma família de proteínas chamadas Serpinas que são inibidoras de serina protease (Figura 18). Vale citar que estas proteínas estão envolvidas em uma variedade de processos biológicos incluindo a coagulação sangüínea, ativação do sistema complemento, angiogênese, inflamação e supressão de tumores (VAN GENT et al. 2003).

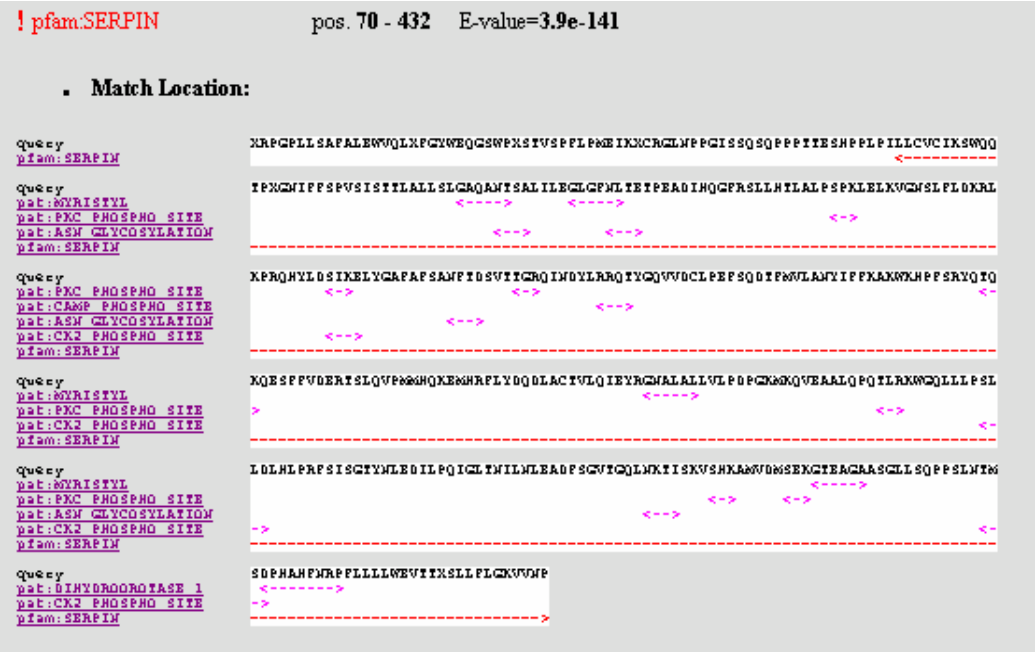


Figura 18 - Visualização do alinhamento das Serpinas com a seqüência de aminoácidos correspondente ao TF00318. Resultado da anotação obtida para o TF00318 a partir da utilização do programa "Motif Scan in a Protein Sequence" do site http://hits.isb-sib.ch/cgi-bin/PFSCAN. Na figura podemos observar representada por uma linha vermelha tracejada a similaridade do TF com as Serpinas ao longo da seqüência de aminoácidos.

2.4.4 Caracterização de Formas Alternativas de "Splicing"

Vários estudos têm sugerido que pelo menos 30 a 35% dos genes humanos sofrem o processo de "splicing" alternativo (BRETT et al. 2000; MODREK et al. 2001). Entretanto, pode-se dizer que este valor está provavelmente subestimado uma vez que inúmeros tipos celulares ainda não foram completamente explorados através do seqüenciamento de moléculas de cDNA. Assim, para a validação de um único TF os grupos muitas vezes utilizaram cDNAs derivados de diferentes tecidos permitindo dessa forma, ao longo do projeto, a caracterização de formas alternativas de "splicing" em paralelo à determinação da estrutura do novo transcrito.

O grau de variabilidade referente ao uso alternativo dos exons ("splicing" alternativo) foi avaliado nas 186 TFs com consenso disponibilizado. As seqüências

geradas para cada TF, juntamente com as ESTs correspondentes, foram submetidas a alinhamento múltiplo em relação à montagem do genoma humano disponibilizada pela UCSC através do programa BLAT (KENT et al. 2002) (Figura 19).

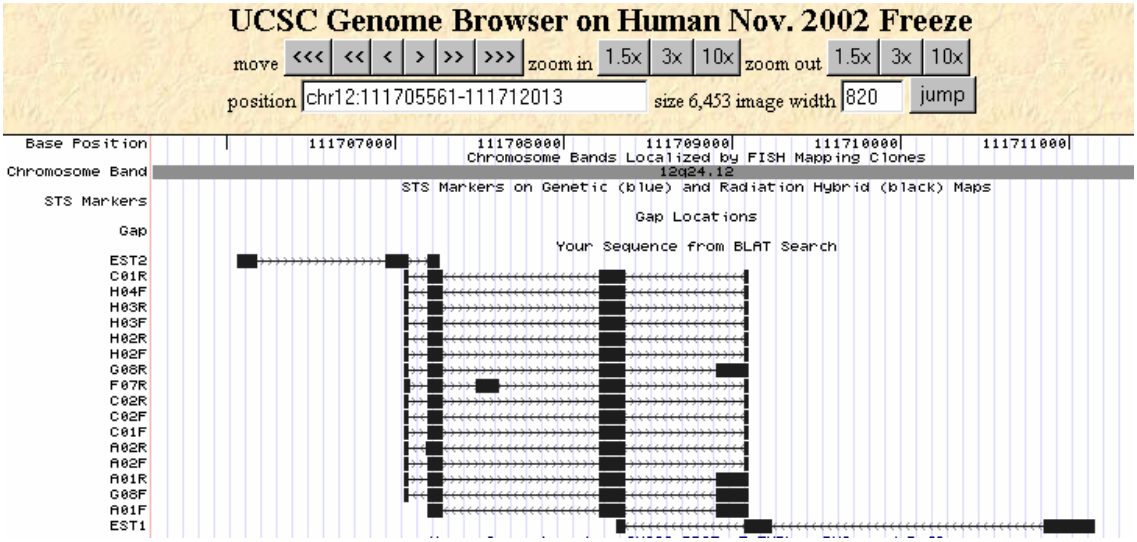


Figura 19 - Alinhamento das seqüências de validação e ESTs do TF00200 contra o genoma humano, utilizando o programa BLAT. Com base na visualização desta interface é possível verificar a presença de um exon a mais na seqüência F07R e a utilização de um sítio aceptor alternativo nas seqüências G08R, A01R, G08F, A01F.

A análise *in silico* identificou evidências de "splicing" alternativo em 22 TFs (12%), sendo que em 13 TFs foi detectada a retenção de introns e em nove o uso alternativo de exons (Tabela 6). É importante ressaltar que os sítios aceptores e doadores conservados (GT/AG) estavam presentes em todos os TFs que apresentaram o uso alternativo de exons.

Tabela 6 - Avaliação das formas alternativas de "splicing" encontradas em 22 TFs. Relação dos TFs que apresentaram formas alternativas de "splicing" com o respectivo número de isoformas encontradas e validadas. Nd, "not done"; * não houve amplificação do fragmento.

Consensos Validados	Tipos de "splicing" alternativo	Presença do sítio aceptor e doador de "splicing"	N° de isoformas encontradas	N° de isoformas validadas
TF00118	Inserção de exon	Sim	2	1
TF00200	Inserção de exon	Sim	4	4
TF00274	Inserção de exon	Sim	2	2
TF00351	Inserção de exon	Sim	2	2
TF01004	Inserção de exon	Sim	2	1
TF01058	Inserção de exon	Sim	3	0*
TF00155	Inserção de exon	Sim	2	nd
TF00238	Inserção de exon	Sim	2	nd
TF00308	Inserção de exon	Sim	2	nd
TF00003	Retenção de intron	nd	nd	nd
TF00019	Retenção de intron	nd	nd	nd
TF00035	Retenção de intron	nd	nd	nd
TF00052	Retenção de intron	nd	nd	nd
TF00099	Retenção de intron	nd	nd	nd
TF00112	Retenção de intron	nd	nd	nd
TF00125	Retenção de intron	nd	nd	nd
TF00131	Retenção de intron	nd	nd	nd
TF00209	Retenção de intron	nd	nd	nd
TF00285	Retenção de intron	nd	nd	nd
TF00371	Retenção de intron	nd	nd	nd
TF00148	Perda de exon	nd	nd	nd
TF01061	Perda de exon	nd	nd	nd

Para a validação experimental foram selecionados seis TFs (TF00118, TF00200, TF00274, TF00351, TF01004 e TF01058) com o uso alternativo de exons, representando um total de 14 isoformas. A RT-PCR "touchdown" confirmou dez (83%) das 12 isoformas investigadas sendo que algumas destas apresentaram um

perfil de expressão muito restrito detectado somente em um tecido ou em um número reduzido dos mesmos. O TF01058 especificamente não foi amplificado e, desta forma, não foi possível confirmamos a existência de suas duas isoformas.

Um exemplo da validação experimental descrita está ilustrado na Figura 20 para o TF0200. A análise por BLAT das seqüências geradas para a validação do TF0200 revelou a presença de duas formas alternativas para esse transcrito. Uma das formas apresentava um exon a mais de 138pb entre os exons XVII e XVIII e a outra apresentava um sítio aceptor alternativo no exon XIX que acrescentava a esse exon uma extensão de 21pb. Iniciadores específicos (P3 e P4) para cada uma das isoformas foram desenhados e utilizados em reações de RT-PCR juntamente com iniciadores (P1 e P2) desenhados para a amplificação do transcrito protótipo (já descrito em banco de dados públicos) (figura 20A). Através dessas reações, foram validadas a existência das 4 formas de "splicing" possíveis para essas combinações de exons. Na Figura 20B podemos verificar a amplificação de um fragmento de tamanho esperado para a forma protótipo (388pb), assim como a amplificação de fragmentos específicos para a forma com o exon a mais (370pb) e para a forma com o exon extendido (314pb). Também é possível observar que mais de uma isoforma está presente nos diferentes tecidos analisados e que tecidos diferentes possuem isoformas em comum.

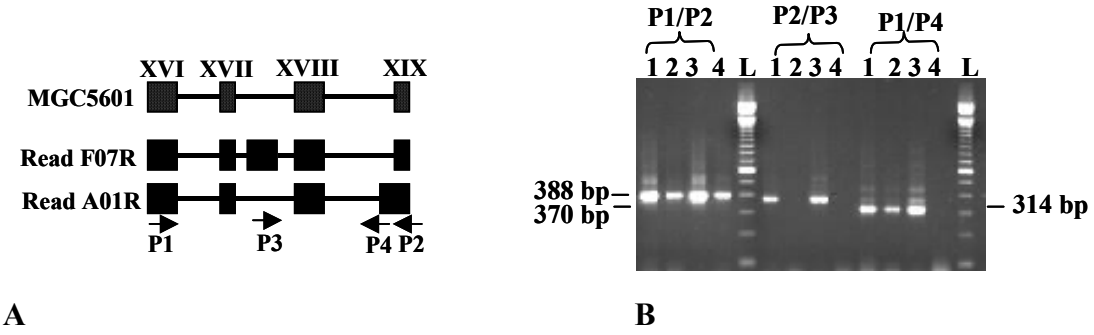


Figura 20 - Validação experimental de formas alternativas de "splicing" para o TF0200. (A) Representação esquemática do alinhamento das seqüências do TF0200 e suas respectivas ESTs com a seqüência genômica humana através do programa BLAT. Duas formas alternativas de "splicing" podem ser observadas. (B) reações de RT-PCR com iniciadores específicos para a isoforma protótipo (P1 e P2) e para as isoformas alternativas (P3 e P4). Os produtos de PCR foram analisados em gel de agarose 1,5%. Quatro tecidos distintos foram utilizados (1: glioblastoma multiforme; 2: glioblastoma; 3: carcinoma de próstata; 4: rim; L: 100bp "Ladder").

CONSIDERAÇÕES FINAIS

2.5 CONSIDERAÇÕES FINAIS

Atualmente, pesquisadores de todo o mundo têm direcionado seus estudos na determinação completa do catálogo de genes humanos. É certo que esta informação terá um profundo impacto em diversas áreas da Biologia como a Evolução Humana, Genômica Estrutural e Medicina, no entanto, devido à estrutura complexa de nossos genes, a identificação de regiões transcritas no genoma humano torna-se extremamente difícil.

Estimativas baseadas em predições gênicas relacionadas aos cromossomos 21 e 22, já completamente seqüenciados, assim como em relação aos rascunhos do genoma humano, demonstram que nosso genoma contém menos de 35.000 genes (HATTORI et al. 2000; DUNHAM et al. 1999; LANDER et al. 2001; VENTER et al. 2001). Além disso, este número é mantido por uma análise preliminar de genes já conhecidos que apresentam uma cobertura de ESTs (EWING e GREEN 2000) e também por comparações entre diferentes genomas (ROEST et al. 2000). Vale ressaltar, que a maioria destes 35.000 genes já está representada por uma seqüência completa de cDNA em bancos de dados de seqüências expressas. No banco de dados do UniGene (http://www.ncbi.nlm.nih/Unigene/), por exemplo, existem 28.412 "clusters" representados por pelo menos uma seqüência completa de cDNA.

Entretanto, outras análises baseadas na clusterização (LIANG et al. 2000) e mapeamento das ESTs no genoma (WRIGHT et al. 2001) sugerem que o catálogo de genes humanos é definitivamente muito maior, variando de 60.000 a 100.000. Estes dados são mantidos pelo fato dos bancos de dados de seqüências expressas também serem compostos em grande parte por "clusters" contendo somente ESTs. No caso do

UniGene, por exemplo, existem 79.628 “clusters” compostos somente por ESTs sendo uma fração significativa (37,3%) correspondente à “clusters” com somente uma seqüência.

Contudo, devido à presença de inúmeros artefatos relacionados à baixa qualidade das seqüências e de vários tipos de contaminação nos bancos de dados de ESTs, não é possível determinar com precisão o número real de genes humanos representados por seqüências expressas. Além disso, o número reduzido de seqüências nos “clusters” compostos somente por ESTs sugere que a maioria dos transcritos ainda não identificados apresenta um perfil de expressão muito baixo e restrito. Desta forma, está cada vez mais clara a necessidade de se utilizar abordagens múltiplas e complementares para a identificação de todos os genes humanos.

Neste projeto demonstramos a utilidade da estratégia do TFI na caracterização de novos transcritos humanos e de formas alternativas de “splicing”. Trata-se de uma abordagem inédita com profundo embasamento computacional e complementar às estratégias já disponíveis. A estratégia permite a caracterização de transcritos com baixo nível de expressão e com padrões de introns e exons que não são reconhecidos por programas de predição gênica sem a necessidade de gerar clones de moléculas de cDNA completas. Com o crescente número de seqüências expressas depositadas em bancos de dados públicos e devido à sua fácil implementação, a estratégia do TFI certamente representará uma estratégia eficiente para completarmos o tão esperado catálogo de genes humanos.

*PARTE II**

UTILIZAÇÃO DOS DADOS DE SEQÜÊNCIAS EXPRESSAS NA IDENTIFICAÇÃO DE NOVOS TRANSCRITOS NO CROMOSSOMO 21

*** Os resultados referentes à identificação, caracterização e determinação do perfil de expre dos novos transcritos foram publicados na revista Genomics de junho de 2002. (Anexo 3)**

INTRODUÇÃO

3.1 INTRODUÇÃO

O cromossomo 21 é o menor cromossomo autossomo humano representando, aproximadamente, 1,5% do genoma (HATTORI et al. 2000). O primeiro marco no estudo deste cromossomo ocorreu em 1959 quando pesquisadores franceses descobriram que indivíduos com três cópias do cromossomo 21 desenvolviam uma síndrome descrita em 1866 pelo médico britânico John L. H. Down e que, atualmente, é conhecida como Síndrome de Down (LEJEUNE et al. 1959). Desde então, mais de vinte doenças já foram mapeadas neste cromossomo. Dentre elas podemos citar a desordem bipolar (STRAUB et al. 1994), certas deficiências imunológicas (ANTONARAKIS 2001) e várias desordens monogênicas, como uma das formas do Mal de Alzheimer (APP) (HATTORI et al. 2000) e a epilepsia mioclônica progressiva (PENNACCHIO et al. 1996).

Várias anormalidades cromossômicas relacionadas a neoplasias têm sido observadas no cromossomo 21. Perda de heterozigosidade, por exemplo, foi observada em regiões específicas do cromossomo 21 em inúmeros tumores sólidos, tais como o câncer de mama, pâncreas, cabeça e pescoço, estômago, pulmão, entre outros, sugerindo a existência de pelo menos um gene supressor de tumor ainda não identificado no mesmo. As mais recentes descobertas sugerem deleções nas regiões 21q11-q21, relacionada com câncer de pulmão (KOHNO et al. 1998), e 21q22.1, relacionada com adenocarcinoma de estômago (SAKATA et al. 1997).

Além disso, estudos epidemiológicos revelaram que indivíduos com Síndrome de Down demonstram ter uma certa "proteção" contra o desenvolvimento de tumores sólidos, visto a baixa incidência dos mesmos. Em contraposição, estes

pacientes apresentam um risco vinte vezes maior de desenvolver leucemia durante a infância (HASLE et al. 2000). Vale ressaltar que a trissomia do 21 é mais freqüente aneuploidia cromossômica encontrada nos casos de leucemia (NIZETIC 2001).

Em maio de 2000, o cromossomo 21 tornou-se o segundo cromossomo humano a ter sua seqüência completamente determinada e publicada (HATTORI et al. 2000), sendo o primeiro o cromossomo 22 (DUNHAM et al. 1999). O seqüenciamento foi realizado através da seleção e ordenação de um conjunto mínimo de clones BACs ("Bacterial Artificial Chromosome") que cobria toda a extensão do cromossomo e que foi posteriormente seqüenciado por "Shotgun" hierárquico ou deleções seriais utilizando o sistema de transposons. A seqüência disponibilizada está dividida em 4 "contigs" e totaliza 33.546.361 nucleotídeos (HATTORI et al. 2000).

A identificação de genes a partir da seqüência genômica foi feita através da utilização de programas de predição gênica e confirmação dessas predições através da presença de similaridade com seqüência de nucleotídeos e/ou proteínas disponíveis em banco de dados públicos. Os parâmetros aplicados incluíam a definição de exons quando preditos por pelo menos dois dos programas de predição gênica GRAIL, GenScan e MZEF; utilização de ESTs para a validação das predições somente quando apresentassem "splicing" entre dois ou mais exons e uma similaridade maior que 95% na região de "overlap"; e similaridade com proteínas conhecidas ou domínios funcionais quando apresentassem identidade maior que 25% em seqüências com mais de 50 aminoácidos (HATTORI et al. 2000).

Foram anotados um total de 225 genes compreendendo 127 genes com seqüência completamente conhecida e 98 genes com estrutura parcialmente predita (HATTORI et al. 2000), sendo este número relativamente pequeno se comparado com

os 545 genes encontrados no cromossomo 22 (DUNHAM et al. 1999). Foram, ainda, identificados 59 pseudogenes no cromossomo 21 (HATTORI et al. 2000).

Os genes identificados foram classificados em 4 categorias: 1) genes com alta similaridade com genes humanos já completamente caracterizados (127 genes), 2) genes com similaridade com seqüências parciais de cDNAs ou ORFs de outros organismos (13 genes), 3) genes cuja seqüência predita apresentava similaridade com domínios protéicos conhecidos (17 genes), 4) genes com similaridade com ESTs ou somente identificados por programas de predição gênica (68 genes) (HATTORI et al. 2000).

Entretanto, devido às limitações existentes no processo de identificação de genes a partir da seqüência genômica, é muito provável que existam, ainda, genes não caracterizados no cromossomo 21. Acredita-se que a caracterização desses novos genes será crucial para um melhor entendimento das inúmeras doenças genéticas associadas a este cromossomo.

Desta forma, pesquisadores de todo o mundo têm intensificado seus estudos na revisão e atualização do catálogo de genes do cromossomo 21 através da descoberta de genes não encontrados anteriormente seguida de confirmação experimental. Isto está sendo possível devido, principalmente, à produção de um elevado número de ESTs e à disponibilização da seqüência genômica de outros organismos.

Uma recente pesquisa analisando a anotação do cromossomo 21, através do estudo de 34 genes localizados neste cromossomo, demonstrou que a anotação completa e acurada necessita da inspeção visual da anotação genômica para interpretação dos dados seguida de análise experimental para correção dos resultados (GARDINER et al. 2002). O estudo foi capaz de corrigir e completar a seqüência

genômica de 16 genes, identificar seis transcritos que codificam ORFs pequenas ou ambíguas e três casos nos quais o processo de "splicing" alternativo produz duas seqüências protéicas estruturalmente não relacionadas. Além disso, foram identificados seis genes codificando proteínas com motivos funcionais, dois genes com baixa similaridade com seus ortólogos protéicos de camundongo e quatro com conservação significante em *Drosophila melanogaster*. Estes resultados, conseqüentemente, representam implicações na geração do mapa completo de transcritos do cromossomo 21 e, também, para todo o genoma humano.

Além disso, uma análise comparativa da seqüência do cromossomo 21 com seu maior correspondente em camundongos, o cromossomo 16 (30Mb), por exemplo, demonstrou a existência de 3 novos genes não descritos na anotação inicial do cromossomo 21 (PLETCHER et al. 2001). Essa análise permitiu, ainda, a confirmação da existência de 15 genes preditos somente por programas de computador assim como a existência de predições incorretas.

Da mesma forma, com o objetivo de atualizar a anotação do cromossomo 21 e avaliar a qualidade das predições gênicas, REYMOND et al. (2001) demonstraram que a maioria das "open reading frames" definidas originalmente com base na sobreposição de ESTs com "splicing" foram corretamente preditas. No entanto, a maioria das predições definidas somente *in silico* não corresponderam a genes verdadeiros, confirmando, desta maneira, os resultados obtidos anteriormente por PLETCHER et al. (2001) com parte do cromossomo. Além disso, o grupo identificou três novos genes que não haviam sido preditos por nenhum programa computacional.

Com base nestes estudos surgiu, então, o interesse de identificar novos genes no cromossomo 21 através do alinhamento entre seqüências expressas e a seqüência genômica. Em colaboração com o grupo de Bioinformática do Instituto Ludwig da

Suíça e o grupo de pesquisa do Dr. Stylianos Antonarakis da Universidade de Genebra, conhecido pelos inúmeros trabalhos publicados com o cromossomo 21, realizou-se um estudo aprofundado neste cromossomo com o objetivo de encontrar novos transcritos que pudessem estar relacionados com os diferentes fenótipos identificados na Síndrome de Down e as inúmeras doenças já mapeadas neste cromossomo.

A partir dos mesmos dados de alinhamento entre a seqüência genômica e seqüências expressas foram construídos dois bancos de dados independentes: o banco de dados do Transcriptoma (descrito no item 2.3.1.1 desta dissertação) e o banco de dados 21Ace, desenvolvido pelo grupo de Bioinformática do Instituto Ludwig da Suíça. Este banco foi organizado com base nos dados gerados a partir de uma re-análise da seqüência do cromossomo 21 incorporando novas ESTs e da criação de um algoritmo para extração e mapeamento das 3' "tags" destas seqüências. Um gráfico foi criado permitindo, desta forma, a identificação de ORFs, das 3' "tags", da presença de ilhas de CpG e das predições do GenScan.

A seleção de candidatos a novos transcritos foi independente uma vez que cada grupo utilizou critérios diferentes com base no seu banco de dados. Quanto à validação e caracterização dos novos transcritos, o grupo da Suíça foi responsável pela obtenção da seqüência completa de cada um dos transcritos enquanto nós determinamos o perfil de expressão tecidual dos mesmos.

OBJETIVOS

3.2 OBJETIVOS

- Identificar novos transcritos localizados no cromossomo 21 utilizando as ferramentas disponibilizadas pelo projeto TFI.

- Avaliar o perfil de expressão dos novos transcritos em diferentes tecidos.

- Verificar a ocorrência de expressão diferencial dos transcritos identificados em tecido normal e tumoral através de abordagens *in silico* e validação experimental.

MATERIAIS E MÉTODOS

3.3 MATERIAIS E MÉTODOS

3.3.1 Identificação de Novos Transcritos no Cromossomo 21

3.3.1.1 Busca no banco de dados do Transcriptoma

A identificação de novos transcritos localizados no cromossomo 21 foi realizada através de buscas, utilizando a linguagem MySQL, no banco de dados do Projeto Transcriptoma Humano descrito em detalhes no item 2.3.1.1 desta dissertação. Foram selecionados "clusters" de ESTs que alinhavam com clones genômicos correspondentes ao cromossomo 21, que quando alinhados com a seqüência genômica apresentavam interrupções no alinhamento indicando a presença de "splicing" e, também, que não apresentassem similaridade com seqüências de genes já anotados ou preditos gerando, desta maneira, uma lista preliminar de "clusters" de ESTs candidatas.

3.3.1.2 Inspeção manual dos candidatos selecionados e seleção de novos transcritos localizados no cromossomo 21

A seqüência de uma EST representativa de cada "cluster" foi comparada com um banco de dados de seqüências (com e sem evidência experimental), relacionadas na publicação original sobre o seqüenciamento completo e a anotação do cromossomo 21, através da ferramenta BLASTN. As ESTs que não apresentaram similaridade com as seqüências de genes anotados no cromossomo 21 foram reanalisadas individualmente.

Os alinhamentos entre as seqüências destas ESTs e a seqüência genômica foram repetidos manualmente para identificar possíveis artefatos gerados durante o processo de alinhamento. Novamente, utilizou-se a ferramenta BLASTN e os bancos de dados NR e HTGS ("HighThroughput Genomic Sequence") do "GenBank". Através desses alinhamentos foram confirmadas a presença de "splicing" e a ausência de melhor alinhamento com seqüências de outras regiões do genoma.

3.3.2 Produção das Seqüências Completas dos Novos Transcritos Identificados no Cromossomo 21

A seqüência completa dos novos transcritos identificados no cromossomo 21 foi produzida através de duas estratégias: seqüenciamento completo de insertos de clones de cDNA correspondentes às ESTs selecionadas e extensão da seqüência obtida através da técnica de RACE.

3.3.2.1 Seqüenciamento completo dos clones de cDNA disponibilizados através do I.M.A.G.E. ("Integrated Molecular Analysis of Genomes and their Expression")

Com base na seqüência de cada EST candidata foram escolhidos os clones de cDNA correspondentes através de buscas por BLASTN no banco de dados de clones de cDNA disponibilizados pelo consórcio I.M.A.G.E. (http://www.rzpd.de/dist/html/clones). Posteriormente, os clones foram solicitados ao RZPD ("Resource Center German Human Genome Project") pelo grupo do Dr. Stylianos Antonarakis da Universidade de Genebra e, em seguida, foram completamente seqüenciados.

De 300 a 500 nanogramas dos plasmídeos contendo os insertos de interesse foram seqüenciados utilizando o kit "DYEnamic^{TM} ET terminator Cycle Sequencing Kit" (Amersham Pharmacia) disponível para o seqüenciamento automático no seqüenciador ABI377 Prism (Perkin Elmer) segundo instruções do fornecedor. Insertos pequenos foram seqüenciados a partir de suas extremidades utilizando-se os iniciadores T3 (5'ATT AAC CCT CAC TAA AGG GA3') e T7 (5'TAA TAC GAC TCA CTA TAG GG3') enquanto para insertos com um tamanho maior foi utilizada a estratégia de "primer walking". A Tabela 7 mostra a relação dos candidatos com seus respectivos clones do I.M.A.G.E..

Tabela 7 - Clones do I.M.A.G.E. correspondentes a cada candidato selecionado.
* Clone de cDNA de camundongo obtido através do Instituto Nacional do Envelhecimento (NIA) (http://lgsun.grc.nia.nih.gov/cDNA/15k.html).

Candidato	Clone I.M.A.G.E.
C21orf65	781289
C21orf81	2304590 / 4798734 / 5289033
C21orf82	429071 / 2097381
C21orf83	814590 / 1853490 / 1467262 / 3477187
C21orf84	2723484 / 2723574
C21orf85	2781245
C21orf86	1756203
C21orf87	705099 / 2461608
C21orf88	526903 / 2097151
C21orf89	2934152
C21orf90	430058
C21orf93	1756203 / 2172104
C21orf94	2464987 / 1170079
C21orf95	4250623 / 4106483 / 162308 / 2338862
C21orf99	1461135 / 2909444
C21orf100	3289153
C21orf101	5285646 / 1744284 / 4764462
C21orf102	H3100H09*
D21S2088E	307666 / 773409
D21S2089E	1621839 / 2910016 / 1755433
D21S2090E	1881464
D21S2091E	1468536

3.3.2.2 RACE ("Rapid Amplification of cDNA Ends")

A seqüência completa da cada transcrito também foi obtida através da técnica de RACE, que permite a amplificação das extremidades 5' e 3' de seqüências transcritas, tendo sido aplicada aos candidatos: C21orf81, C21orf83, C21orf84 e C21orf87. Para tanto, foi utilizado o "Marathon™ Amplification cDNA kit" (CLONTECH® n°K1802-1). A síntese da fita dupla de cDNA a partir de RNA poli A^+ e a ligação de adaptadores, importantes na obtenção dos fragmentos 5' e 3' dos transcritos, às moléculas de cDNAs sintetizadas foram realizadas segundo instruções do fornecedor.

Primeiramente, foi feita uma PCR com iniciadores específicos para os adaptadores e para a porção interna do transcrito de interesse em presença de 5µl de cDNA, 0,2mM de dNTPs, 0,2µM de cada iniciador e 1 unidade de Advantage Taq DNA Polimerase (CLONTECH®), em tampão apropriado. Os produtos de PCR foram analisados em gel de poliacrilamida 8% e, então, clonados utilizando o TOPO TA Cloning Kit (Invitrogen) segundo especificações do fornecedor. A transformação bacteriana foi realizada utilizando-se 5µl da reação de ligação e 100µl de bactérias competentes preparadas a partir de células JM109 e segundo o protocolo descrito por INOUE et al. (1990). Em seguida, os plasmídeos obtidos na clonagem contendo os insertos de interesse foram seqüenciados como descrito no item anterior.

3.3.3 Avaliação do Padrão de Expressão de Cada Candidato

O padrão de expressão tecidual dos candidatos foi determinado através de RT-PCR e "Nested-PCR". Os iniciadores específicos para cada candidato foram

desenhados e testados em cDNAs de diferentes tecidos produzidos a partir de RNA total obtido comercialmente da CLONTECH®.

3.3.3.1 Painel de RNAs da CLONTECH®

Foram utilizados RNAs provenientes de tecidos humanos normais comercializados pela CLONTECH® e provenientes dos mais variados tecidos: próstata, mama, intestino delgado, cérebro, cérebro fetal, pulmão, testículo, coração, rim, fígado, fígado fetal, traquéia, cólon, medula óssea, baço, timo, músculo esquelético, útero, placenta, glândula adrenal, glândula salivar e medula espinhal. Estes foram enviados em Etanol 70% e 0,08M Acetato de Sódio, e para a utilização, os mesmos foram precipitados e ressuspendidos em água DEPC. Antes de serem utilizados para a síntese de cDNA, a qualidade destes RNAs também foi avaliada em gel preparativo de agarose como descrito no item 2.3.2.2 desta dissertação.

3.3.3.2 Síntese de cDNA

A síntese da primeira fita de cDNA foi realizada como descrito no item 2.3.2.4 desta dissertação. No entanto, neste caso, para avaliar a eficiência de síntese do cDNA, foi aplicado somente o teste GAPDH (também detalhado no mesmo item).

3.3.3.3 RT-PCR e "Nested-PCR"

O cDNA sintetizado foi utilizado em reações de PCR com iniciadores específicos para cada candidato selecionado (Tabela 8). Foram realizadas RT-PCRs seguidas de "nested-PCRs" com iniciadores internos aos utilizados na primeira

amplificação. Os iniciadores foram construídos com base nas seqüências obtidas a partir do alinhamento das ESTs com a seqüência genômica. Os mesmos foram desenhados em exons diferentes tornando possível, desta forma, a distinção da amplificação do DNA genômico com a amplificação do cDNA.

Tabela 8 - Seqüências dos iniciadores utilizados para avaliação do perfil de expressão dos candidatos selecionados. Para cada candidato estão especificados os iniciadores utilizados na primeira reação e na "nested" com suas respectivas temperaturas de anelamento e tamanho dos fragmentos amplificados. Nd = "not done"

Candidato	Iniciadores RT-PCR	Iniciadores "Nested-PCR"	Temperatura de anelamento	Tamanho do amplicon
C21orf65	F1 aacatgggtggcaaaaagag R1 aggtgcttaacaaatgccatc	FN gagctgccatttagaacatgc RN aagttgggtccaggctcc	60°C	1ª reação – 153pb "Nested" – 73pb
C21orf81	F1 gcaggtgcaaaaaggaaaac R1 cctgctgtgaaaggagcag	FN tttctgtcagagtgcatgttc RN tgtgaaaggagcagtacaaag	60°C	1ª reação – 202pb "Nested" – 147pb
C21orf82	F1 gcagattcttggcagacctc R1 gggttcccctcagaaagc	FN tcgatgttgctggctttg RN agcagacaagagtcgggg	60°C	1ª reação – 282pb "Nested" – 138pb
C21orf83	F1 aacacgggtcacagcacc R1 tggacccatctcagtggag	FN aatccaatgggaaagccc RN gtgccatccactggtgtg	65°C	1ª reação – 366pb "Nested" – 212pb
C21orf84	F1 attgccttagcagcgcc R1 acaccccggatcgagaag	FN gccgctgttggtttcatc RN cgaggcttgcgtgacag	60°C	1ª reação – 283pb "Nested" – 191pb
C21orf85	Nd	nd	-	-
C21orf86	F1 gtcatggatggggctcac R1 aggactgctgtgggtcttg	FN acgctggcagaggtgaag RN aacgtgtaggacccgctg	60°C	1ª reação – 307pb "Nested" – 195pb
C21orf87	F1 gcttcgaggacaggaaagc R1 tcgtccacctctccaacc	FN aagtgaggattggctcgc RN tgactccctagtttccggc	RT-PCR ► 60°C Nested ► 65°C	1ª reação – 422pb "Nested" – 340pb
C21orf88	F1 gctccagcattccctgtg R1 acgattcccatttcacgg	FN tggtgcaccctgaagtcc RN tgggtcacacagccacag	55°C	1ª reação – 156pb "Nested" – 50pb
C21orf89	F1 ccttgatttctcctcctttgg R1 aatgatgtccggctgtgc	FN tgggcttcctgactctttg RN tcattagcacgtgcagctc	60°C	1ª reação – 292pb "Nested" – 196pb
C21orf90	F1 tttgtcaaacatccgggg R1 gcaggacttgaggctttatcc	FN atggcatgggtgacacg RN aggcagtcacttccaatcatc	60°C	1ª reação – 184pb "Nested" – 115pb

C21orf93	F1 tcgagccatgtcttggtg R1 tgtggctgggtgcctatc	FN acctcctttccatgtgcg RN agcccctggaccatgac	60°C	1ª reação – 379pb "Nested" – 270pb
C21orf94	F1 gcaaatattgccctaaaatg R1 agaccaaatgtatgtgttgcc	FN tgttgaagattgcagtgatgg RN ggagaggggacatgacag	55°C	1ª reação – 198pb "Nested" – 91pb
C21orf95	F1 ccctgaatgtgcttaggtgg R1 ggtcattgctgggattgc	FN gcagggatattccacctgac RN catcaacaccgtctcctcc	60°C	1ª reação – 494pb "Nested" – 295pb
C21orf99	F1 aagactgaatgagtggcag R1 ctgattcaaattacttcttacag	FN ggaacatctaagatgatca RN atattcggagagtgagc	60°C	1ª reação – 900pb "Nested" – 800pb
C21orf100	F1 caacgtgacattgtttggag R1 tgtccatctgaatgcccac	FN tggtgtctgtgaaaagggg RN tggttcattcagtagctccac	60°C	1ª reação – 579pb "Nested" – 447pb
C21orf101	F1 aggccctgatggacagag R1 aaggctaaaatctggcgaatc	FN ggaaaacctgggtgaacg RN ttcctgggtcagagggtg	60°C	1ª reação – 322pb "Nested" – 186pb
C21orf102	F1 ctcagggagctggatctgtc R1 cgatggggtccttggag	FN agaggatccccaaggacg RN aaccagccgaacatggtg	60°C	1ª reação – 492pb "Nested" – 276pb
MCM3APAS	F1 ggtgcatgtgaagcaatgg R1 tcagccccttgttggatg	Nd	60°C	1ª reação – 157pb
D21S2088E	F1 ttgtgctcacttgcagagc R1 tgtgttgttggatttctgc	FN ccaatccgtgccatattttc RN gcaaatcacatggccaaac	55°C	1ª reação – 313pb "Nested" – 183pb
D21S2089E	F1 tgatcacaatcagcattgg R1 tggatggcacaagtcacc	FN ttccccaaggcaagtgag RN cgccctgactggtaaagc	55°C	1ª reação – 385pb "Nested" – 208pb
D21S2090E	F1 agctttgcgatgagacgg R1 aatggctgtttcctttccg	FN accatccagctaagcccc RN ttccagcccagctacacag	55°C	1ª reação – 206pb "Nested" – 85pb
D21S2091E	F1 ctggtttgttttcccatgc R1 agtatccctgcaaccccc	FN tttgaggatcaactttgcc RN tccgggaagaaagagaagc	60°C	1ª reação – 308pb "Nested" – 121pb

Na primeira reação foi utilizado 1µl de cDNA para um volume final igual a 25µl enquanto na "nested-PCR" utilizou-se como molde 1µl da primeira reação. As PCRs foram realizadas em presença de tampão 1x Taq DNA polimerase, 1,6mM $MgCl_2$, 0,2mM dNTPs, 0,4µM de cada iniciador e 1 unidade de Taq DNA polimerase (GIBCO/BRL).

As reações de PCR tiveram 35 ciclos enquanto as "nested-PCRs" foram realizadas em 30 ciclos. A temperatura de anelamento foi determinada pelo programa

OLIGOTECH e o tempo de extensão final de ambas as reações foi de 6 minutos. Na maioria dos casos, a temperatura utilizada para a RT-PCR e a "nested-PCR" foi a mesma, com exceção apenas do candidato C21orf87. Em seguida, os produtos amplificados foram visualizados em gel de poliacrilamida 8%.

3.3.4 Anotação Funcional dos Novos Transcritos

A anotação funcional dos novos transcritos foi realizada pelo grupo da Suíça através de busca de similaridade com seqüências já caracterizadas. Ferramentas computacionais disponíveis via Internet foram utilizadas para a análise preliminar das seqüências completas dos novos transcritos. Busca de similaridade com outras seqüências de nucleotídeos e proteínas disponíveis em bancos de dados públicos foram realizadas através dos programas BLASTN e BLASTX, respectivamente. Já com relação à presença de domínios protéicos, a busca foi realizada nos bancos de dados InterPro, PFAM e ProSite.

3.3.5 Avaliação do Padrão de Expressão Diferencial entre Tecido Normal e Tumoral dos Novos Transcritos

A avaliação baseou-se, primeiramente, na análise *in silico* através da metodologia de SAGE ("Serial Analysis of Gene Expression") seguida por comprovação experimental através da técnica de PCR semiquantitativa e PCR em Tempo Real (Real-Time PCR). Para a realização destas técnicas foram utilizados, além dos tecidos normais de cérebro e próstata provenientes da CLONTECH®, as linhagens celulares de adenocarcinoma de próstata (PC3), carcinoma de próstata (DU145), glioblastoma (A172) e glioblastoma multiforme (T98G), provenientes da

ATCC. As metodologias utilizadas para extração do RNA e síntese do cDNA foram descritas, respectivamente, nos itens 2.3.2.2 e 2.3.2.4 da parte I desta dissertação.

3.3.5.1 SAGE Genie

A técnica de SAGE permite uma análise simultânea e quantitativa de um grande número de transcritos sendo que os dados gerados através desta metodologia são mantidos pelo CGAP ("Cancer Genome Anatomy Project) e disponibilizados via Internet pelo NCBI (http://www.ncbi.nlm.nih.gov/SAGE) (LIANG 2002). Para facilitar as buscas no banco de SAGE, o CGAP criou o SAGE Genie (http://cgap.nci.nih.gov/SAGE) que compreende um conjunto de ferramentas envolvidas no processamento dos dados gerados pela técnica de SAGE (BOON et al. 2002).

A técnica de SAGE identifica uma etiqueta ("tag") em cada um dos transcritos que é correspondente a uma seqüência de 10 nucleotídeos, provenientes da extremidade 3' do transcrito adjacente ao último sítio de restrição da enzima NlaIII (CATG) (VELCULESCU et al. 1995). Os dados produzidos através desta técnica apresentam-se como uma lista de etiquetas. Desta forma, o nível de expressão de cada transcrito em um determinado tecido está diretamente ligado ao número de etiquetas específicas desses transcritos identificadas em uma determinada biblioteca. Essas informações, como citado anteriormente, são, então, depositadas em bancos de dados para análise e comparações digitais via "web" (BOON et al. 2002).

A análise da expressão diferencial dos transcritos identificados neste trabalho foi feita através da identificação da "tag" virtual de cada candidato. A busca no banco de SAGE foi realizada através do SAGE Genie sendo excluídas da análise as bibliotecas derivadas somente de linhagens celulares. Vale ressaltar, neste

momento, que o banco de SAGE compreende um número crescente de mais de 5,2 milhões de SAGE "tags" determinadas a partir de 114 tipos celulares (BOON et al. 2002).

Esta análise foi realizada com colaboração do Laboratório de Biologia Computacional do Instituto Ludwig, um dos grupos responsáveis pelo desenvolvimento do SAGE Genie. A partir dos resultados obtidos, foram selecionados candidatos que apresentaram expressão diferencial ("cutoff" = 4) entre tecido normal e tumoral para serem analisados através da técnica de PCR semiquantitativa.

3.3.5.2 PCR semiquantitativa

Para esta reação foram utilizados os mesmos iniciadores selecionados para a RT-PCR já descrita anteriormente além de iniciadores específicos para GAPDH utilizados como controle de amplificação na reação. Para um volume final igual a 20µl foram utilizados 2µl da reação de cDNA. A PCR foi realizada em presença de tampão 1x Taq DNA polimerase, 1,5mM $MgCl_2$, 0,2mM dNTPs, 0,4µM de cada iniciador e 1 unidade de Taq DNA polimerase (GIBCO/BRL).

As reações foram feitas em triplicata sendo cada amostra retirada e analisada em ciclos sucessivos da PCR de forma que a amplificação pudesse ser progressivamente acompanhada. No caso das amostras analisadas com o gene constitutivo GAPDH, as alíquotas foram retiradas nos ciclos 22, 25 e 28. Já nas reações com iniciadores específicos para os candidatos, as amostras foram retiradas nos ciclos 28, 30 e 35. Deste modo, é possível visualizar um aumento gradual na expressão do transcrito analisado a cada ciclo em que é retirada uma alíquota.

3.3.5.3 "Real-Time PCR"

A confirmação dos resultados obtidos na PCR semiquantitativa para os genes que apresentaram diferenças de expressão entre tecido normal e tumoral foi realizada através de uma PCR em Tempo Real ("Real-Time PCR"). Nesta reação, há um monitoramento da fluorescência emitida a cada ciclo de amplificação permitindo, desta forma, uma quantificação precisa do produto amplificado ciclo a ciclo.

A metodologia utilizada na análise de detecção da fluorescência emitida foi o SYBR® Green I (MORRISON et al. 1998). Neste sistema o SYBR® Green se liga à moléculas de fita dupla emitindo uma fluorescência de maneira que a cada ciclo da reação novas moléculas são formadas e, conseqüentemente, o nível de fluorescência emitida aumenta. Assim, quanto menor o ciclo em que a fluorescência atinge o limite determinado pelo aparelho maior é a expressão do gene analisado (MORRISON et al. 1998).

Desta forma, valores quantitativos foram obtidos no ponto durante a ciclagem no qual a amplificação do produto foi detectada ao invés da quantidade do produto acumulado depois de um número fixo de ciclos. O parâmetro CP ("Crossing Point") ou CT ("Cycle Treshold") foi definido como uma fração do número de ciclos na qual a fluorescência gerada ultrapassou um padrão fixo sendo os dados de quantificação analisados pelo "software" de análise do aparelho (MORRISON et al. 1998).

A amplificação dos genes analisados foi quantificada e relacionada com os valores obtidos com cada gene constitutivo utilizando-se a equação $2^{-\Delta\Delta CT}$ para ser adquirido um valor normalizado de cada amostra quanto à variabilidade na quantidade e à integridade do RNA (LIVAK e SCHMITTGEN 2001). Vale destacar que o valor inferido à ΔCT correspondeu à diferença entre a média dos CTs do gene de interesse e

a média dos CTs obtidos para cada um dos normalizadores. Já o cálculo da fórmula ΔΔCT envolveu a subtração entre o valor de ΔCT para cada amostra de tecido tumoral e o valor de ΔCT calculado para o tecido normal correspondente.

Os genes constitutivos utilizados para corrigir variações nas quantidades iniciais de RNA foram a β-actina (Forward 5'CAC TGT GTT GGC GTA CAG GT 3' e Reverse 5'TCA TCA CCA TTG GCA ATG AG 3'), a ciclofilina (Forward 5' TGA GAC AGC AGA TAG AGC CAA GC 3' e Reverse 5' TCC CTG CCA ATT TGA CAT CTT C 3') e o GAPDH (Forward - 5'CTG CAC CAC CAA CTG CTT A 3' e Reverse - 5'CAT GAC GGC AGG TCA GGT C 3').

As primeiras reações foram realizadas no aparelho LigthCycler™ (Roche Diagnostics, Mannheim, Germany) em um volume total de 20μl de reação por capilar. Cada reação foi realizada em duplicata na presença de tampão 10x, 3mM $MgCl_2$, 0,2mM de dNTPs (Promega), 0,4μM de cada iniciador, 5% de DMSO, 0,1μl de SYBR® Green I (Sigma) diluído 1:100, 0,75U de Taq Platinum DNA Polimerase (GIBCO/BRL) e 2μl de cDNA preparado como descrito anteriormente.

As condições de amplificação consistiram em dois minutos de denaturação inicial a 95°C, seguida de 45 ciclos de denaturação a 94°C por 15 segundos, anelamento dos iniciadores durante 20 segundos (Tabela F) e um tempo de extensão de 30 segundos a 72°C. A detecção do produto amplificado fluorescente foi realizada na última etapa de cada ciclo a uma temperatura que varia entre 3 e 4°C abaixo da TM ("Melting Temperature") do produto. Para confirmar a especificidade da amplificação, os produtos amplificados foram submetidos à análise da curva de "melting" ao final de cada reação através do resfriamento das amostras a 65°C seguido por um aumento até 95°C a 0,1°C por segundo com a contínua aquisição da fluorescência. Os cDNAs utilizados nas reações foram provenientes dos RNAs de

cérebro normal da CLONTECH®, das linhagens celulares A172 e T98G, cinco amostras de tecido cerebral cedidas pelo laboratório da Dra. Mari Sogayar do Instituto de Química da USP, sendo duas referentes ao tecido normal, uma à astrocitoma de grau II e duas à meningioma, além de cinco amostras de glioblastomas cedidas pelo Dr. Greg Riggins da Universidade de Duke.

Devido a problemas na reprodutibilidade dos resultados obtidos nestes experimentos e também no processo de aquisição de fluorescência do aparelho, a análise foi novamente realizada no aparelho ABI Prism® 7000 Sequence Detection System. As reações foram realizadas em um volume de 25µl por amostra sendo que cada reação foi feita em duplicata seguindo o mesmo protocolo aplicado no aparelho LigthCycler™. As condições de amplificação consistiram em dois minutos de denaturação inicial a 95°C, seguida de 40 ciclos de denaturação a 94°C por 15 segundos, anelamento dos iniciadores durante 30 segundos (Tabela 9) e um tempo de extensão de 30 segundos a 72°C. Para confirmar a especificidade da amplificação, os produtos amplificados foram submetidos à mesma análise da curva de "melting" descrita anteriormente.

Tabela 9 - Temperaturas de anelamento dos iniciadores referentes aos genes MCM3APAS, MCM3AP, β-actina, ciclofilina e GAPDH.

Gene	Temperatura de anelamento
MCM3APAS	60°C
MCM3AP	65°C
β-actina	62°C
Ciclofilina	62°C
GAPDH	62°C

RESULTADOS E DISCUSSÃO

3.4 RESULTADOS E DISCUSSÃO

3.4.1 Identificação de Novos Transcritos no Cromossomo 21

A identificação de novos transcritos localizados no cromossomo 21 baseou-se em dois bancos de dados: o banco de dados do Transcriptoma já descrito detalhadamente no item 2.3.1.1 desta dissertação e o banco de dados 21Ace desenvolvido pelo grupo de Bioinformática do Instituto Ludwig da Suíça. A seleção automática de candidatos a novos transcritos foi independente uma vez que cada grupo utilizou critérios diferentes com base no seu banco de dados.

Após a seleção dos candidatos, cada grupo revisou manualmente sua lista de possíveis novos transcritos e, em seguida, estes dados foram agrupados para facilitar o processo de validação experimental. A produção da seqüência completa de todos os transcritos, assim como a anotação dos mesmos, foi realizada pelo grupo da Suíça enquanto o padrão de expressão tecidual dos novos transcritos foi determinado em nosso laboratório.

3.4.1.1 Seleção automática de candidatos no banco de dados do Transcriptoma

Primeiramente, foram realizadas buscas no banco de dados do Projeto Transcriptoma Humano utilizando a linguagem MySQL. Os critérios aplicados para a seleção das ESTs candidatas consistiram na identificação de ESTs que alinhavam com clones genômicos correspondentes ao cromossomo 21, que apresentassem "splicing" e que não correspondessem à seqüência completa de genes anotados previamente neste cromossomo, gerando, assim, uma lista preliminar de candidatos.

Dentre os 9.732 “clusters” correspondentes ao cromossomo 21 e presentes no banco de dados do TFI, 8.627 representavam “clusters” compostos somente de ESTs. Já dentre estes, apenas 758 apresentavam ESTs com “splicing”, cerca de 7,8% do número total de clusters.

O fato de serem selecionadas apenas ESTs com “splicing” elevou a confiabilidade com relação aos resultados gerados já que, desta maneira, evitamos a possível seleção de seqüências contaminantes de DNA genômico. Com certeza, dentre as ESTs excluídas por não apresentarem “splicing” podem ser encontradas algumas possíveis candidatas a verdadeiros transcritos, entretanto, tal critério foi necessário para evitarmos a presença de falso positivos e facilitar a etapa de validação experimental.

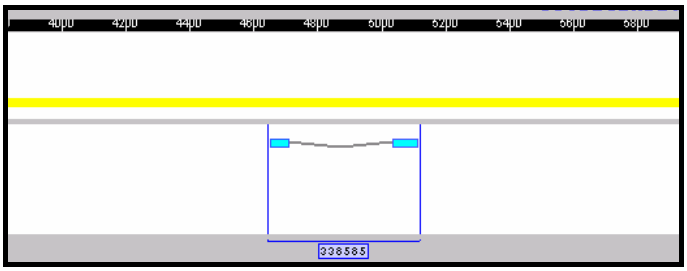
3.4.1.2 Inspeção manual dos candidatos selecionados

Com base na lista preliminar, as ESTs candidatas foram revistas manualmente para evitar possíveis artefatos gerados durante o processo de alinhamento, eliminando, desta forma, a existência de um melhor alinhamento em outra região do genoma, assim como, a presença de seqüências com falso “splicing”.

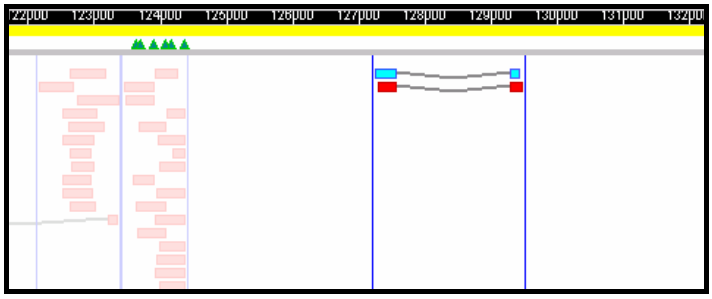
Após esta análise visual o número de ESTs candidatas foi reduzido a 71. Estas foram identificadas como SPO e, então, enviadas ao Grupo de Bioinformática do Instituto Ludwig da Suíça para comparação, através do BLASTN, com as seqüências dos genes já anotados ou preditos no cromossomo 21 e as dos candidatos selecionados pelo referido grupo através do banco de dados 21Ace, identificados como GVA. Dentre os 71 candidatos SPO, cinco correspondiam a candidatos GVA e 41 a genes anotados ou preditos anteriormente. O número reduzido de candidatos em comum entre os dois grupos deve-se ao fato do grupo da Suíça ter utilizado alguns

parâmetros diferentes dos aplicados para os SPOs. No caso dos GVAs foi levado em conta a presença de uma fase aberta de leitura e de pelo menos um dos seguintes critérios: presença de ilha de CpG, de uma 3' "tag" ou de predições do GenScan.

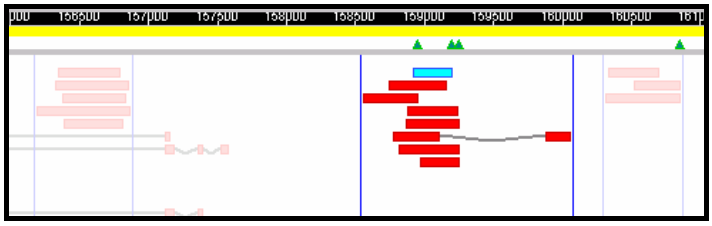
Os 25 candidatos SPO restantes foram, em seguida, reanalisados individualmente através do alinhamento entre suas seqüências e a seqüência genômica para confirmação da presença de "splicing" e do melhor alinhamento no cromossomo 21. Nesta fase foram excluídos três candidatos por alinharem em outros cromossomos, dois por apresentarem seqüências repetitivas e um por não apresentar "splicing". Oito candidatos foram eliminados posteriormente na fase experimental por não apresentarem amplificações específicas em reações de RT-PCR e/ou um clone I.M.A.G.E. correspondente. Sendo assim, foram selecionados 11 candidatos a verdadeiros transcritos. A Figura 21 abaixo mostra a visualização na Interface Gráfica dos "clusters" correspondentes a cada um dos 11 candidatos selecionados.



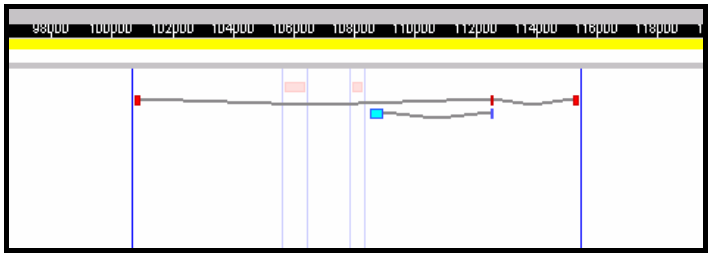
A



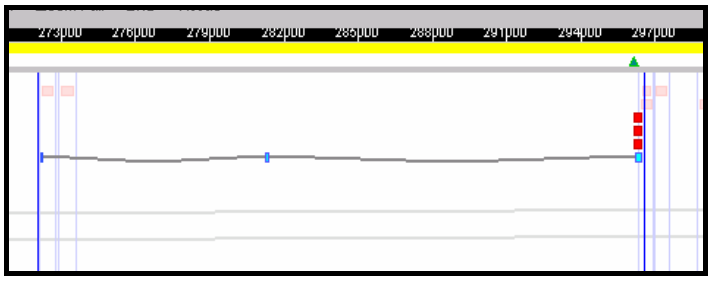
B



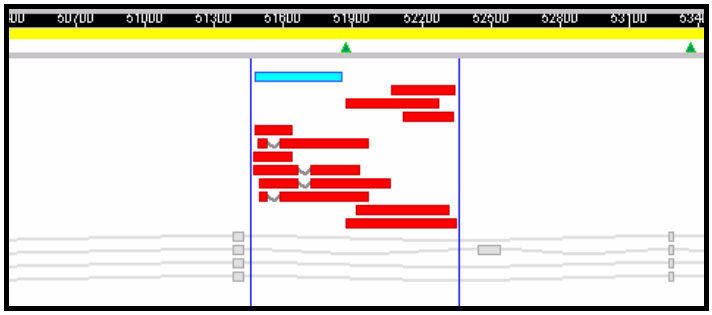
C



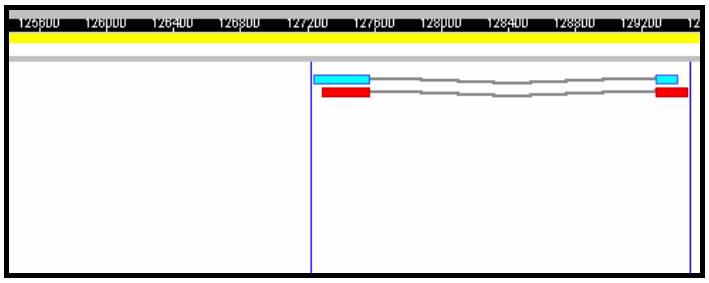
D



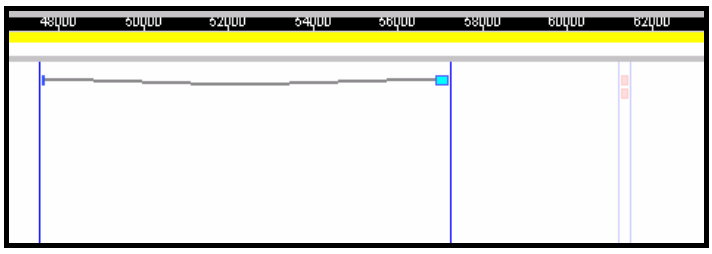
E



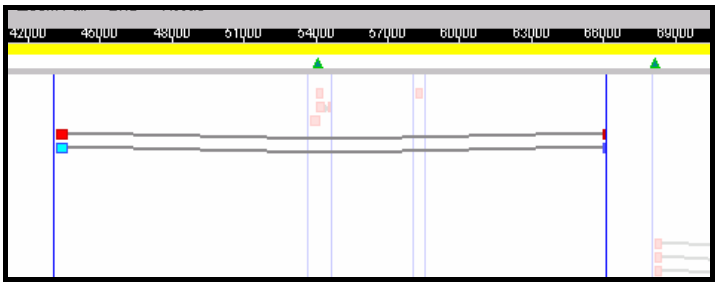
F



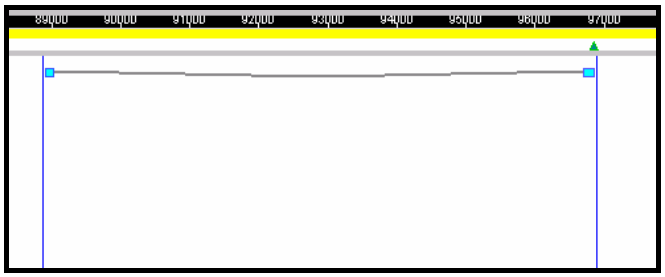
G



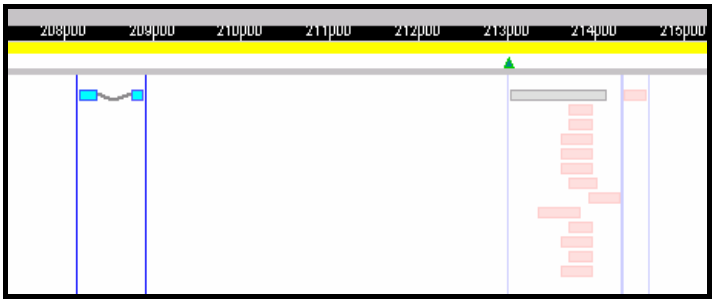
H



I



J



L

Figura 21 - Visualização através da Interface Gráfica dos "clusters" correspondentes a cada um dos 11 candidatos SPO selecionados. (A) C21orf65, (B) C21orf86, (C) C21orf87, (D) C21orf88, (E) C21orf89, (F) C21orf90, (G) C21orf93, (H) C21orf94, (I) D21S2089E, (J) D21S2090E, (L) D21S2091E.

É importante ressaltar, neste momento, com base na análise do "cluster" correspondente a cada candidato, que a maioria apresentava um número muito reduzido de ESTs com uma média de 3,2 ESTs por "cluster". Desta forma, este valor demonstrou que tais transcritos provavelmente apresentavam um baixo nível de expressão e talvez por este motivo ainda não haviam sido identificados anteriormente.

3.4.2 Validação Experimental dos Candidatos Identificados

3.4.2.1 Confirmação dos candidatos identificados

Para uma análise mais precisa e para a produção da seqüência completa dos 11 candidatos SPO e 11 candidatos GVA selecionados, foram identificados os clones I.M.A.G.E. correspondentes a cada uma das ESTs. Nesta fase da análise todas as ESTs foram classificadas com um nível de qualidade 1. Após a obtenção dos clones, estes foram completamente seqüenciados e, com base nesta seqüência, seguiram-se, então, novas análises.

Inicialmente, as seqüências completas dos clones de cDNA foram analisadas quanto ao alinhamento no cromossomo 21. As seqüências que apresentaram um melhor alinhamento neste cromossomo foram mantidas em nossa lista e classificadas com um nível de qualidade 2. Estas foram, então, reanalisadas quanto à presença de uma ORF e também de "splicing", uma vez que a seqüência de baixa qualidade de uma EST pode criar uma interrupção no alinhamento contra a seqüência genômica gerando um falso "splicing". Desta forma, com base na seqüência obtida a partir do clone I.M.A.G.E. é possível confirmar a existência de um "splicing" real. As ESTs que apresentaram "splicing" mas não tinham ORF foram classificadas como unidades transcricionais (D21S2088E à D21S2091E), enquanto as que apresentaram ORF

foram identificadas como C21orfs. Estas foram, então, classificadas com um nível de qualidade 3.

Após confirmação dessas ESTs como transcritos "bona fide" através da análise de expressão por RT-PCR, as mesmas foram classificadas com um nível de qualidade 4. Apenas o candidato C21orf87 foi excluído desta última classificação, uma vez que não foi identificada a expressão do mesmo em pelo menos um dos 22 tecidos analisados. A figura 22 resume a estratégia utilizada para confirmação dos candidatos selecionados.

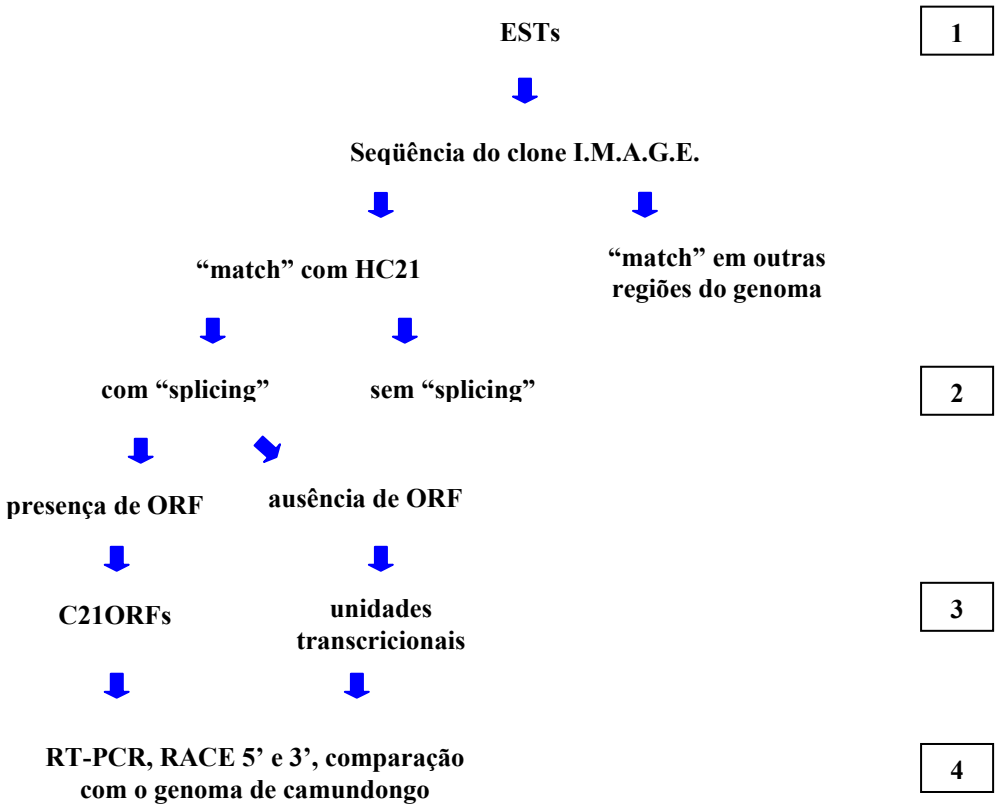


Figura 22 - Representação esquemática da estratégia utilizada na confirmação dos candidatos selecionados. Os números à direita representam o nível de qualidade dado aos transcritos em cada fase da análise.

3.4.2.2 Produção da seqüência completa dos novos transcritos e anotação

Além da obtenção das seqüências dos clones I.M.A.G.E., algumas das ESTs que representaram transcritos "bona fide" tiveram suas seqüências completas obtidas através da técnica de RACE. A seqüência do transcrito C21orf81, por exemplo, foi obtida através do produto de RACE 5' derivado de cDNAs de pulmão, cérebro e coração. Já o transcrito C21orf83 teve sua seqüência obtida dos produtos de RACE derivados somente de testículo, enquanto para o candidato C21orf84 foi utilizado somente cDNA de placenta. No caso do C21orf87, a obtenção de seqüências, por RACE 5', derivadas de cDNA de pulmão e placenta foi de grande importância, uma vez que comprovou a existência do mesmo e permitiu que este transcrito fosse elevado ao nível de qualidade 4.

Após a obtenção das seqüências completas de todos os transcritos identificados partiu-se para uma estratégia comparativa com a seqüência genômica de camundongo com o objetivo de confirmar a identidade de nossos transcritos e, até mesmo, de identificar novos genes no cromossomo 21. Primeiramente, através da ferramenta BLASTN, foi possível identificar regiões homólogas aos genes C21orf65, C21orf83, C21orf95 e C21orf101 na seqüência genômica de camundongo. Com base nas seqüências destes transcritos e das ESTs de camundongo encontradas nas regiões identificadas foi possível construir as seqüências de cDNA dos ortólogos em camundongo dos genes C21orf83, C21orf95 e C21orf101. Os números de acesso no GenBank para as seqüências dos ortólogos correspondentes são, respectivamente, AY063457, AY061854 e AY061856.

Com base no fato de que a ferramenta utilizada acima, BLASTN, pode não reconhecer pequenas homologias "escondidas" em longas seqüências de DNA, nosso estudo comparativo seguiu, então, para uma segunda parte. Nesta nova análise foi

realizado um alinhamento entre a seqüência genômica humana e a seqüência genômica de camundongo. Seqüências genômicas humanas delimitadas por dois genes conhecidos, representando "âncoras", e contendo um dos novos transcritos identificados foram comparadas com seqüências de camundongo delimitadas por ortólogos destas "âncoras" através do programa DOTTER (SONNHAMMER e DURBIN 1996). Este representa um programa dot-plot através do qual podem ser comparadas seqüências de DNA ou proteínas, assim como seqüências de DNA contra seqüências de proteínas. Esta metodologia permitiu a identificação de seqüências similares aos transcritos C21orf65, C21orf82, C21orf83, C21orf85, C21orf95 e C21orf101 em regiões sintênicas de camundongo demonstrando, desta forma, que estes genes apresentam-se conservados entre roedores e primatas. Além disso, foi possível identificar um "contig" de ESTs de camundongo que mapeava na região sintênica do cromossomo 10 de camundongo com o cromossomo 21 humano. A seqüência consenso deste contig apresentou uma fase aberta de leitura (número de acesso no "GenBank" AY061858) conservada entre estas espécies. Esta descoberta nos permitiu a reconstrução da seqüência de cDNA do correspondente gene humano, agora identificado como C21orf102, aumentando o número de transcritos identificados para 23.

Os resultados obtidos com base nas análises descritas acima nos faz avaliar até que ponto a seqüência genômica de um novo organismo pode contribuir na identificação de genes humanos. É certo que esta contribuição existe mas, na verdade, de uma forma muito reduzida do realmente esperado por inúmeros pesquisadores. Vale lembrar que dentre os 23 transcritos identificados neste trabalho, apenas seis (26%) apresentaram-se conservados em camundongo. Além disso, um estudo comparativo entre o cromossomo 21 humano e suas regiões sintênicas em

camundongo identificou uma grande quantidade de “blocos” conservados mas com função totalmente desconhecida (DERMITZAKIS et al. 2002). De acordo com os resultados obtidos, há um elevado número de regiões conservadas entre tais organismos, entretanto, não correspondendo a genes conhecidos e na sua maioria não representando regiões transcritas.

Desta maneira, foram, então, encontrados 23 novos transcritos no cromossomo 21, sendo que 19 apresentam uma fase aberta de leitura e 4 foram identificados por unidades transcricionais. Dentre os 19 transcritos identificados, todos foram nomeados C21ORFs, com exceção do MCM3APAS (“MCM3-associated protein antisense”) por compartilhar dois exons com o gene MCM3AP (“minichromosome maintenance deficient 3 acetylating protein”) e estar localizado na fita oposta do mesmo. O número de acesso de cada transcrito está presente na Tabela 10 assim como os dados referentes ao número de exons, tamanho da ORF e eventual presença de domínio protéico.

Tabela 10 -Descrição dos 23 transcritos identificados no cromossomo 21 com o correspondente número de acesso no GenBank. Dados referentes ao número de exons, tamanho da ORF e eventual presença de domínio protéico de cada transcrito também estão descritos na tabela.

GENE	Nº de acesso	Nº exons	Nº aa	DOMÍNIOS
C21orf65	AF426256	4	91	-
C21orf81	AF426257	6	89	-
C21orf82	AF426258	2	64	-
C21orf83	AF426259	8	353	"zinc finger" C_2H_2
C21orf84	AF426261	4	77	-
C21orf85	AF426262	2	96	-
C21orf86	AF426264	2	165	-
C21orf87	AF426265	1	145	-
C21orf88	AF426266	3	145	-
C21orf89	AF426268	3	33	-
C21orf90	AF426269	3	65	-
C21orf93	AF427488	3	139	-
C21orf94	AF427489	2	62	-
C21orf95	AF401639	3	154	Prolina
C21orf99	AF427490	4	68	-
C21orf100	AY063459	2	55	-
C21orf101	AY061855	3	125	Família ribosomal S6
C21orf102	AY061857	1	257	Leucina
MCM3APAS	AF426263	4	123	-
D21S2088E	AY063451	5	-	-
D21S2089E	AY063452	3	-	-
D21S2090E	AY063454	2	-	-
D21S2091E	AY063455	3	-	-

Em características gerais, estes novos genes apresentam ORFs de tamanho pequeno (média de 122 aminoácidos) e número baixo de exons (média de 3,3 exons), além de um nível de expressão baixo (descrito detalhadamente no próximo item). Estes dados confirmam que os algoritmos desenvolvidos para a predição gênica apresentam um "bias" contra genes com ORFs pequenas. Além disso, as predições necessitam de validação experimental que acabam também apresentando um "bias" agora para genes mais abundantes. Assim, existe uma tendência na identificação de genes que apresentam ORFs longas e com um alto nível de expressão. É importante ressaltar que os genes já anotados no cromossomo 21 (HATTORI et al. 2000), por exemplo, apresentam uma média de 575 aminoácidos e com relação às seqüências dos bancos de dados RefSeq, SwissProt e TrEMBL esta média é reduzida para 469

aminoácidos (LANDER et al. 2001), um valor ainda muito mais alto que a média de 122 aa encontrada entre nossos transcritos. Além disso, com relação ao número de exons, enquanto os 23 transcritos apresentaram uma média de 3,3 exons, as seqüências dos bancos de dados citados anteriormente apresentam uma média de 8,8 exons. Desta maneira, frente a estes dados é possível justificar a razão pela qual os programas de predição gênica aplicados anteriormente ao cromossomo 21 não foram capazes de identificar os transcritos descritos neste trabalho.

Vale ressaltar neste momento que a confiabilidade de nossa metodologia foi comprovada durante a análise final destes transcritos através da identificação paralela de três de nossos genes por outros pesquisadores. O candidato C21orf65 foi identificado como DSCR8 ("Down Syndrome Critical Region 8") (TOYODA et al. 2002), o C21orf95 recebeu o nome de CYYR1 ("Cysteine and tyrosine-rich protein 1") (VITALE et al. 2002) e o C21orf101 foi chamado de MRPS6 ("Mitochondrial ribosomal protein S6") (SUZUKI et al. 2001).

A maioria dos C21ORFs não demonstrou similaridade significativa com qualquer proteína já caracterizada com exceção de quatro transcritos. A análise foi feita com base na similaridade com produtos protéicos e domínios funcionais disponíveis nos bancos de dados públicos Pfam, InterPro e Prosite. O produto protéico do C21orf83 contém três domínios "zinc fingers" do tipo C_2H_2 (PF00096), a proteína do C21orf101 representa um membro da família ribosomal S6 (IPR000529), o transcrito C21orf102 contém trechos ricos em leucina (IPR000372 e IPR001611) e a proteína do C21orf95 contém um domínio rico em prolina (PS50099).

3.4.2.3 Determinação do padrão de expressão dos novos transcritos

O perfil de expressão tecidual para cada um dos transcritos identificados foi analisado em 22 tecidos normais: testículo, mama, pulmão, próstata, intestino delgado, placenta, cólon, cérebro, cérebro fetal, fígado, fígado fetal, timo, glândula salivar, coração, útero, tecido ósseo, medula espinhal, rim, baço, músculo esquelético, traquéia e glândula adrenal. Para cada candidato foram construídos dois pares de iniciadores, F1/R1 e FN/RN para as reações de RT-PCR e "Nested-PCR" respectivamente, já citados anteriormente. Os mesmos foram construídos em exons diferentes e próximos à extremidade 3' do transcrito de acordo com o alinhamento com a seqüência genômica para confirmar a amplificação da molécula de cDNA (Figura 23).

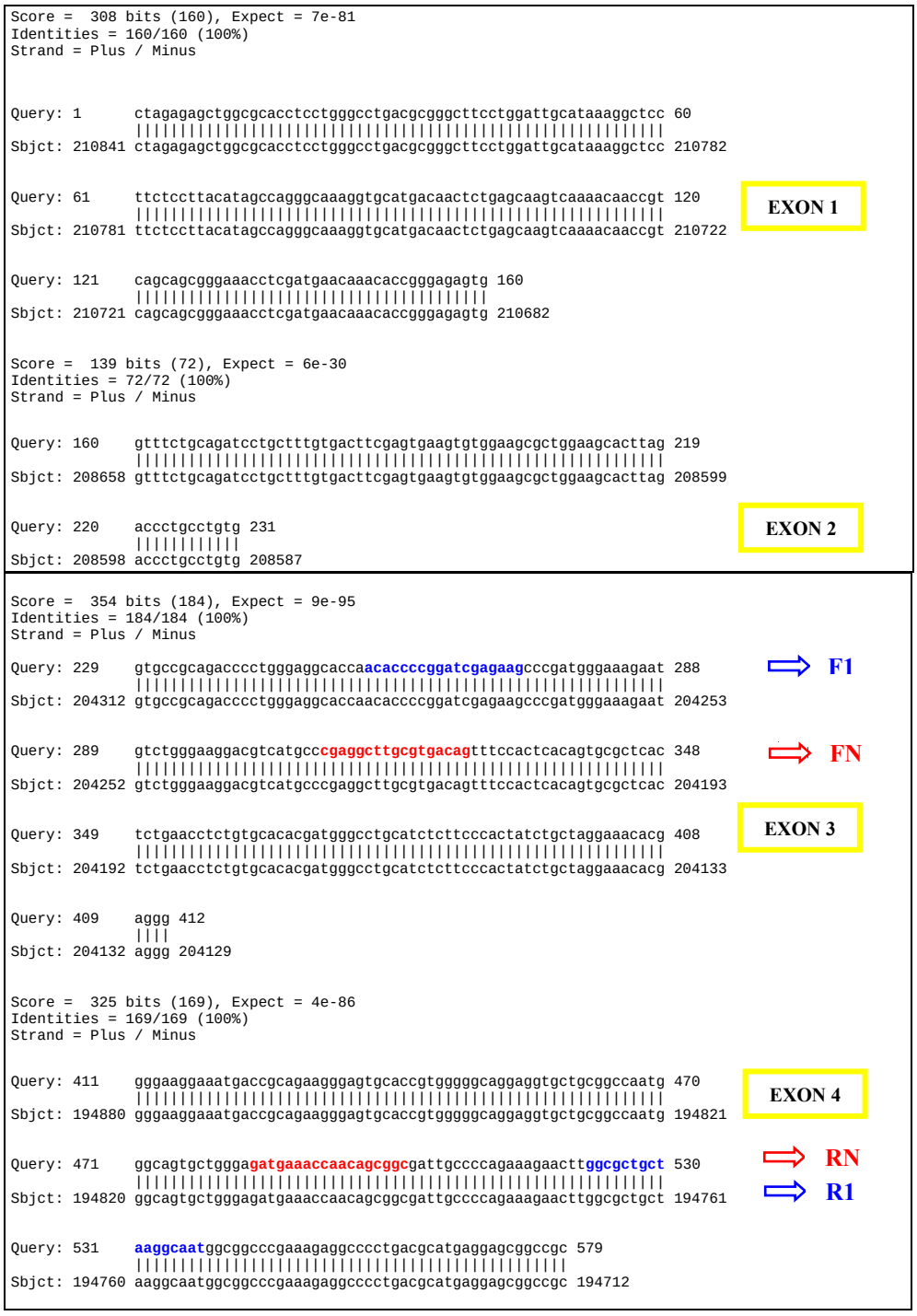


Figura 23 - Alinhamento da seqüência do transcrito C21orf84 com a seqüência genômica. O alinhamento foi feito através da ferramenta BLAST 2 sequences sendo que em azul estão identificados os iniciadores “Forward”1 e “Reverse”1 utilizados na RT-PCR e em vermelho os iniciadores “Forward”N e “Reverse”N utilizados na “Nested-PCR”.

De acordo com a posição dos iniciadores calculou-se o tamanho dos fragmentos a serem amplificados em cada reação de forma que a visualização dos mesmos em gel de agarose 1% confirmou a expressão no tecido correspondente (Figura 24).

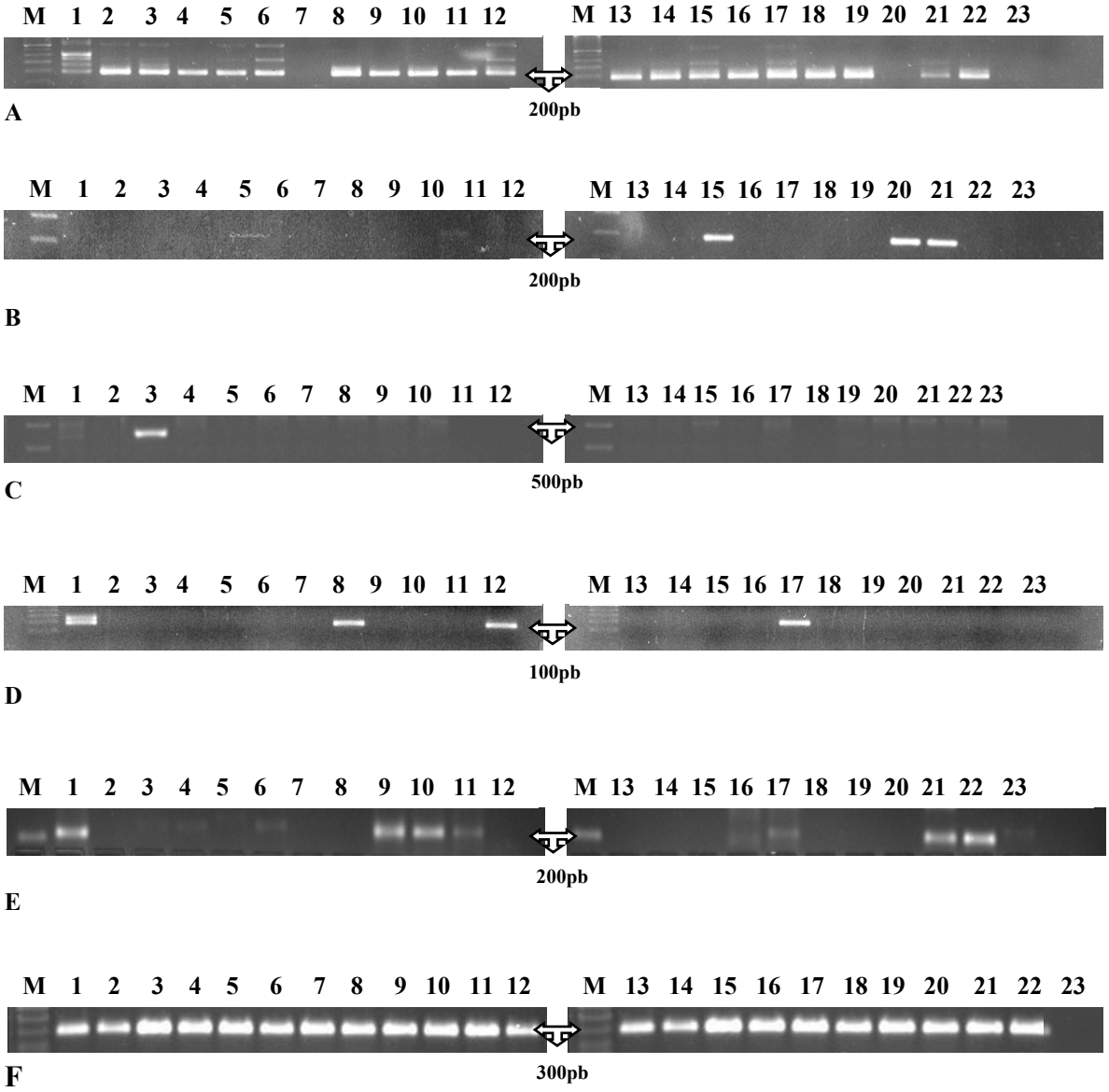


Figura 24 – Visualização do padrão de expressão dos novos transcritos localizados no cromossomo 21. Visualização em gel de agarose 1% do perfil de expressão dos transcritos C21orf84 (A), C21orf86 (B), C21orf100 (C), D21S2088E (D) e D21S2089E (E) nos seguintes tecidos: 1.testículo, 2.pulmão, 3.próstata, 4.intestino delgado, 5.mama, 6.cérebro, 7.coração, 8.útero, 9.medula óssea, 10.placenta, 11.cólon, 12.cérebro fetal, 13.fígado, 14.fígado fetal, 15.timo, 16.glândula salivar, 17.medula espinhal, 18.rim, 19.baço, 20.músculo esquelético, 21.traquéia, 22.glândula adrenal. O peso molecular(M) aplicado foi o 100bp "ladder" e a canaleta 23

corresponde ao controle com DNA genômico. (F) Visualização em gel de agarose 1% da amplificação do gene GAPDH nas amostras de cDNA dos tecidos citados anteriormente.

Vale ressaltar, que a determinação do padrão de expressão tecidual da maioria dos candidatos foi possível apenas com as reações "nested", demonstrando, assim, a baixa expressão dos mesmos. Este fato pode justificar, como discutido anteriormente, a razão pela qual os transcritos identificados neste trabalho não foram encontrados durante o projeto de seqüenciamento do cromossomo 21.

Tomando-se por base apenas o perfil de expressão de todos os transcritos analisados (Tabela 11) podemos destacar, em especial, os transcritos C21orf99 e C21orf100. O primeiro começou a ser intensamente estudado a partir de uma análise de sua seqüência de nucleotídeos onde foi possível detectar uma alta similaridade com o antígeno tumoral de diferenciação de mama NYBR.1. Frente ao perfil de expressão e com base nos dados obtidos no decorrer do projeto de doutorado do aluno Raphael B. Parmigiani de nosso laboratório, que identificou e caracterizou este transcrito, foi possível concluir que o transcrito C21orf99 corresponde a um antígeno tumoral da classe "cancer-testis". Antígenos desta classe apresentam como principal característica um perfil de expressão em tecidos normais restrito a testículo e em tecidos tumorais abrangendo um grande número de tipos histológicos sendo considerados, atualmente, candidatos ideais para o desenvolvimento de vacinas e imunização passiva (SCANLAN et al. 2002).

Tabela 11 - Perfil de expressão tecidual de todos os transcritos analisados.

Gene	Testis	Lung	Prostate	Small Intestin	Breast	Brain	Heart	Uterus	Bone Marrow	Placenta	Colon	Fetal Brain	Liver	Fetal Liver	Thymus	Salivary Glan	Spinal Cord	Kidney	Spleen	Muscle	Trachea	Adrenal Glan
C21orf65	+	+	+	-	-	+	-	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+
C21orf81	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+
C21orf82	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf83	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf84	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
C21orf86	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-
C21orf87	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C21orf88	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf89	+	+	-	+	-	+	-	-	-	+	-	+	-	-	-	-	+	-	+	+	+	+
C21orf90	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	+
C21orf93	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf94	+	-	+	-	-	+	-	+	+	+	+	+	-	+	+	+	+	-	-	+	+	+
C21orf95	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf99	+	-	+	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-
C21orf100	-	-	+	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C21orf101	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C21orf102	+	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	+	-	+	+
D21S2088E	+	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-
D21S2089E	+	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	+	-	-	-	+	+
D21S2090E	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-
D21S2091E	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

3.4.2.4 Caracterização preliminar do transcrito C21orf100 como um possível antígeno de diferenciação

O perfil de expressão restrito à próstata do transcrito C21orf100 nos fez acreditar que o mesmo pudesse representar um membro da classe dos antígenos de diferenciação. Os antígenos de diferenciação são antígenos que apresentam expressão em tipos celulares específicos ou em estágios específicos de diferenciação de um determinado tecido (RETTIG e OLD 1989). Além disso, a expressão destes antígenos em células normais geralmente é preservada nas células tumorais, o que os torna importantes marcadores no diagnóstico imunopatológico diferencial do câncer.

Estes antígenos podem, ainda, ser alvo para a imunoterapia específica. Isto porque se considerarmos que apenas o tecido normal do qual o tumor se originou será

afetado, já seria uma grande vantagem sobre a quimioterapia, completamente inespecífica. Podemos citar como exemplo o anticorpo anti-CD20, que reconhece um antígeno de diferenciação de linfócito, o primeiro anticorpo monoclonal aprovado pelo FDA ("Food and Drugs Administration") para imunoterapia de linfoma (GRILLO-LOPEZ et al. 1999).

Desta maneira, partiu-se para uma análise do perfil de expressão deste transcrito em um painel de amostras de tumores do Banco de Tumores do Hospital do Câncer. Para tanto foram utilizadas primeiramente 25 amostras de tumores de próstata e, em seguida, oito amostras de cada um dos respectivos tumores: tireóide, estômago, útero, melanoma e mama. As metodologias utilizadas para extração do RNA, síntese do cDNA e RT-PCR já foram descritas anteriormente e seguiram os mesmo padrões.

A análise de expressão em tumores de próstata apresentou positividade em 80% das amostras (Figura 25). Já com relação as 40 amostras correspondentes aos cinco tipos de tumores citados anteriormente, o transcrito apresentou-se expresso, ainda que em um nível muito reduzido, apenas em uma amostra de tumor de estômago e uma de mama.

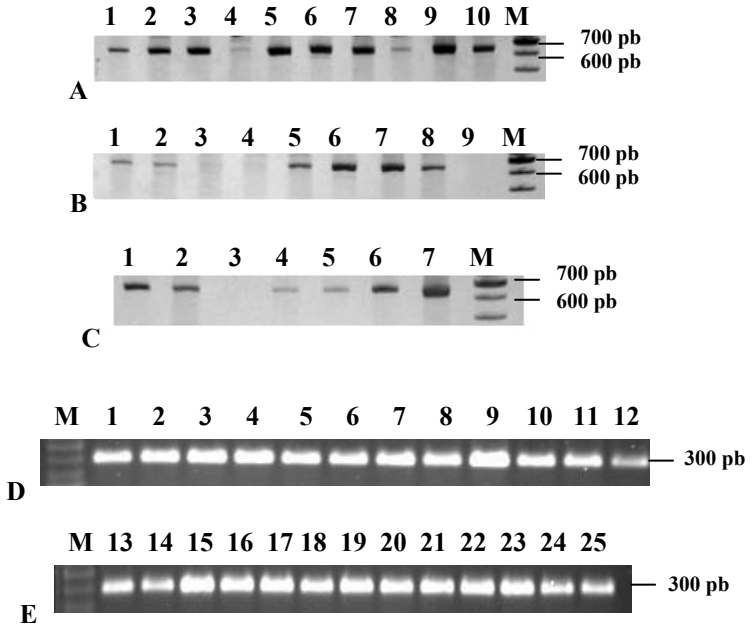


Figura 25 - Análise da expressão do transcrito C21orf100 em tumores de próstata. (A) Visualização em gel de poliacrilamida 8% da expressão do transcrito C21orf 100 em 10 amostras de tumores de próstata, (B) Visualização em gel de poliacrilamida 8% da expressão do transcrito C21orf 100 em 9 amostras de tumores de próstata, (C) Visualização em gel de poliacrilamida 8% da expressão do transcrito C21orf 100 em 6 amostras de tumores de próstata sendo a canaleta 7 referente à expressão do transcrito C21orf100 em tecido normal de próstata. (D) e (E) Visualização em gel de agarose 1% da amplificação do gene GAPDH nas amostras de cDNA dos tecidos citados anteriormente. O peso molecular (M) aplicado foi o 100bp "ladder".

Frente a estes resultados podemos supor que o transcrito possa realmente representar um novo antígeno de diferenciação. Entretanto, análise em um número maior de amostras, assim como dados de expressão da proteína codificada por este gene além da localização celular da mesma, são necessários para que este fato seja realmente confirmado (BERA et al. 2002; IAVARONE et al. 2002; OLSSON et al. 2001, 2003).

3.4.3 Análise da Expressão Diferencial dos Novos Transcritos em Tumores

3.4.3.1 Análise in silico da expressão diferencial dos transcritos identificados

Para tentarmos estabelecer uma possível correlação entre o câncer e os transcritos identificados neste trabalho partimos para uma avaliação da expressão diferencial entre tecido normal e tumoral destes transcritos com base em uma análise *in silico* seguida por validação experimental. A avaliação de cada transcrito quanto à sua expressão diferencial em tecido normal e tumoral baseou-se, inicialmente, em uma análise no banco de dados de SAGE disponibilizado pelo CGAP. Os dados contidos neste banco são processados através de várias ferramentas disponibilizadas pelo "SAGE Genie" como o "SAGE Anatomic Viewer" que permite visualizar a comparação dos níveis de expressão de qualquer transcrito entre tecido normal e tumoral (Figura 26) (BOON et al. 2002; LIANG 2002).

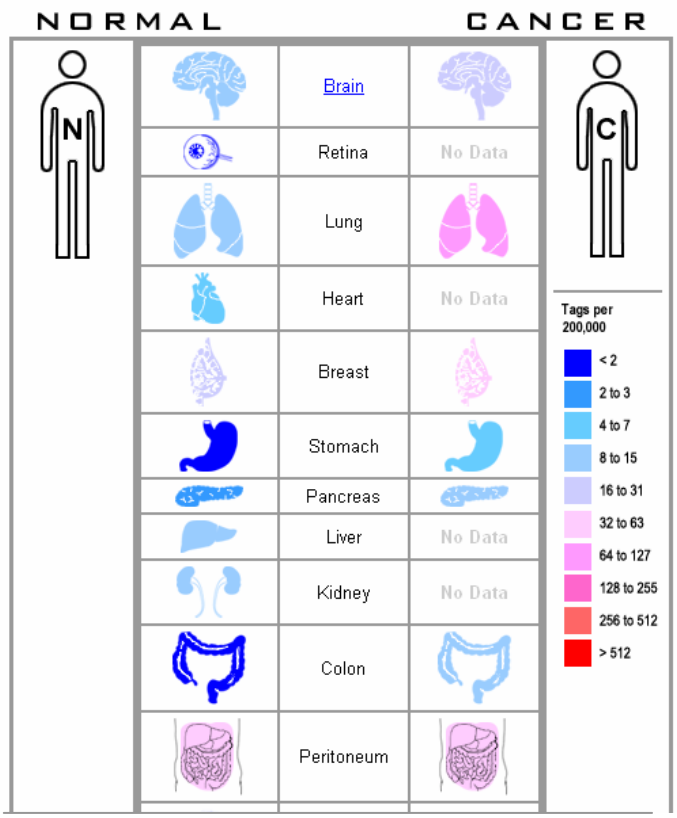


Figura 26 – "Homepage" da ferramenta "SAGE Anatomic Viewer" disponibilizada pelo "SAGE Genie". Visualização dos resultados obtidos utilizando a ferramenta "SAGE Anatomic Viewer". Para cada órgão analisado é possível detectar diferenças nos níveis de expressão de um determinado transcrito entre tecido normal e tumoral a partir da avaliação do número de "tags" correspondentes a cada tecido.

Apesar de toda a tecnologia e facilidade envolvida na utilização desta metodologia para a obtenção de dados de expressão entre tecidos normais e tumorais podemos encontrar algumas limitações neste tipo de análise. Embora não existam dúvidas quanto à utilidade de SAGE na análise global da expressão gênica, alguns desafios ainda precisam ser vencidos. Isto inclui seqüências que não apresentam o sítio da enzima NlaIII (menos de 1% das seqüências analisadas) e, principalmente, a existência de um número reduzido de bibliotecas presentes neste banco. Ainda que não seja pequeno, este representa um número muito reduzido de tecidos o que acaba

dificultando a análise de transcritos com um padrão de expressão restrito. Além disso, o fato da "tag" ser representada por uma seqüência de bases de apenas dez nucleotídeos leva a uma redução na especificidade da análise e, conseqüentemente, gera a presença de "tags" conflitantes, ou seja, uma mesma "tag" representando transcritos diferentes. Nestes casos, acaba não sendo possível avaliar a expressão diferencial de um determinado transcrito uma vez que os dados obtidos não são específicos.

Desta forma, com base na seqüência de cada um dos 19 transcritos que apresentavam ORF (C21orfs) foram identificadas as "tags" correspondentes aos dez nucleotídeos adjacentes ao sítio da enzima NlaIII em direção à extremidade 3' dos genes (Tabela 12). Para cinco transcritos (C21orf87, C21orf89, C21orf93, C21orf99 e C21orf102) não foi possível determinar a seqüência da "tag" correspondente uma vez que os mesmos não apresentavam o sítio da enzima NlaIII. Para os 14 candidatos restantes foi possível identificar a "tag" real correspondente não sendo encontrado nenhum caso onde foi possível identificar apenas a "tag" virtual, isto é, "tag" extraída *in silico* ainda não validada experimentalmente.

Tabela 12 - Seqüências das "Tags" referentes aos 19 transcritos identificados

Transcrito	SAGE "tag"
C21orf65	CCGTGACAAT
C21orf81	CACTCTGACA
C21orf82	ATGCTGAAAA
C21orf83	GCTTCCCCAC
C21orf84	AGGAGCGGCC
C21orf85	AACCTCACCG
C21orf86	ACCCCTGCTC
C21orf87	"tag" não identificada
C21orf88	CTGTTCTTGT
C21orf89	"tag" não identificada
C21orf90	GGTGACACGC
C21orf93	"tag" não identificada
C21orf94	TCCCCCTCTC
C21orf95	TATTTATAAA
C21orf99	"tag" não identificada
C21orf100	CAGCAGCCAT
C21orf101	TGGAAATAAA
C21orf102	"tag" não identificada
MCM3APAS	AACTTGACTT

Das 14 seqüências com "tags" apenas duas eram conflitantes (C21orf82 e C21orf83) não fornecendo, assim, dados claros sobre a expressão dos referentes transcritos (Tabela 13). No entanto, os dois candidatos não foram excluídos e seguiram para a validação experimental com o intuito de avaliar melhor cada um destes casos.

Posteriormente, através do SAGE Genie foi feito uma busca no banco de dados de SAGE com base nas "tags" já determinadas a fim de encontrar informações com relação à expressão diferencial das mesmas. A diferença na expressão entre tecido normal e tumoral foi considerada significativa acima de um "cutoff" de 4 por se tratarem de transcritos com baixo nível de expressão (Tabela 13). Com base nesta análise, foram identificados, então, três candidatos com expressão diferencial: C21orf83, C21orf100 e MCM3APAS. Como o transcrito C21orf100 já foi anteriormente analisado no item IV.2.4 e representa um possível novo antígeno de diferenciação, seguiram com a validação experimental somente os candidatos C21orf83 e MCM3APAS.

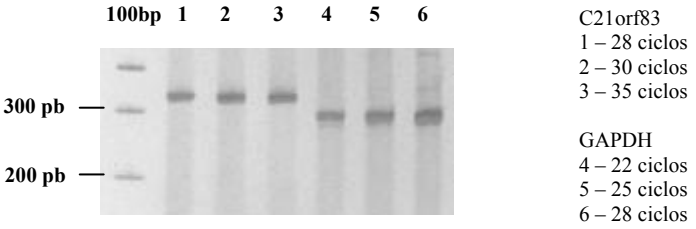
Tabela 13 - Resultados obtidos com base no banco de dados de SAGE para os 19 transcritos. Na coluna referente ao "cutoff" 4, os tecidos relacionados apresentaram diferença na expressão entre normal e tumoral igual ou superior a 4.

Gene	TAG	TAG confl.	Cutoff 4	Nº tags por 200.000
C21orf65	OK	NO	-	
C21orf81	OK	NO	-	
C21orf82	OK	OK	-	
C21orf83	OK	OK	Cérebro tumoral	Normal - 4 tags Tumoral – 7 tags
C21orf84	OK	NO	-	
C21orf85	OK	NO	-	
C21orf86	OK	NO	-	
C21orf87	NO	-	-	
C21orf88	OK	NO	-	
C21orf89	NO	-	-	
C21orf90	OK	NO	-	
C21orf93	NO	-	-	
C21orf94	OK	NO	-	
C21orf95	OK	NO	-	
C21orf99	NO	-	-	
C21orf100	OK	NO	Próstata normal	Normal - 6 tags Tumoral - < 2 tags
C21orf101	OK	NO	-	
C21orf102	NO	-	-	
MCM3APAS	OK	NO	Próstata normal	Normal - 6 tags Tumoral - 3 tags

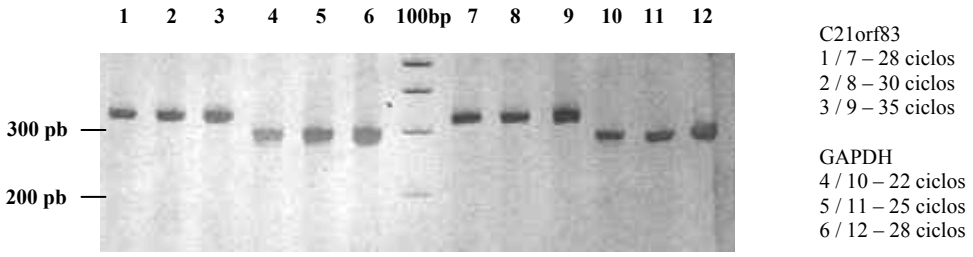
3.4.3.2 Validação experimental dos dados obtidos na análise *in silico* com relação à expressão diferencial dos transcritos identificados

Os candidatos C21orf83 e MCM3APAS selecionados pela análise de SAGE foram, primeiramente, avaliados através de uma PCR semiquantitativa. A expressão de cada transcrito foi avaliada em tecido normal de próstata e cérebro em comparação com linhagens celulares tumorais dos mesmos tecidos (DU145 e PC3, A172 e T98G, respectivamente). Além disso, utilizou-se o gene constitutivo, GAPDH, como controle de amplificação na análise dos resultados obtidos.

O candidato C21orf83 não demonstrou expressão diferencial nos tecidos analisados (Figura 27) enquanto o transcrito MCM3APAS mostrou-se mais expresso nas linhagens tumorais de cérebro (Figura 28).

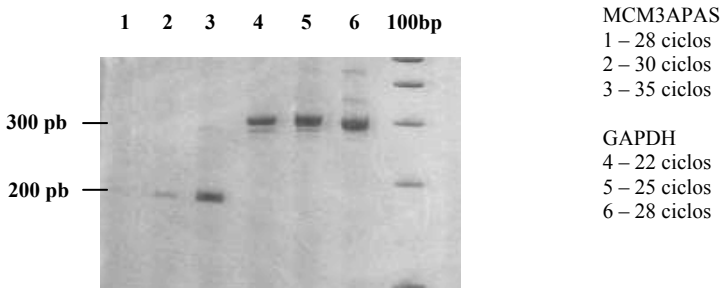


A

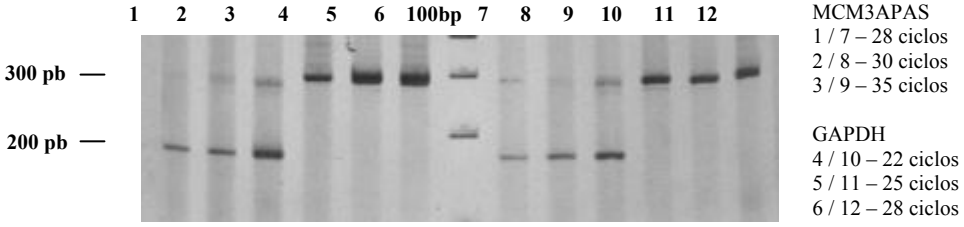


B

Figura 27 - Análise de expressão por PCR semiquantitativa do transcrito C21orf83. (A) Visualização em gel de acrilamida 8% da expressão do transcrito C21orf83 em tecido normal de cérebro. (B) Visualização em gel de acrilamida 8% da expressão do transcrito C21orf83 nas linhagens tumorais de cérebro A172 (1 a 6) e T98G (7 a 12).



A



B

Figura 28 - Análise de expressão por PCR semiquantitativa do transcrito MCM3APAS. (A) Visualização em gel de acrilamida 8% da expressão do transcrito MCM3APAS em tecido normal de cérebro. (B) Visualização em gel de acrilamida 8% da expressão do transcrito MCM3APAS nas linhagens tumorais de cérebro A172 (1 a 6) e T98G (7 a 12).

Com base nos resultados obtidos nas reações acima foi possível observar que não houve uma correlação entre estes e os dados referentes à análise de SAGE uma vez que para o transcrito C21orf83 os dados *in silico* não foram confirmados enquanto para o candidato MCM3APAS, que apresentou expressão diferencial em próstata segundo os dados de SAGE, a validação experimental demonstrou uma diferença nos níveis de expressão entre cérebro normal e tumoral. Como estes transcritos apresentam um baixo nível de expressão é possível que a cobertura dos dados de SAGE tenha sido insuficiente para gerar dados mais confiáveis e que o número de amostras analisadas seja limitado para a confirmação destes dados.

Em seguida, com o objetivo de obter uma quantificação precisa da expressão do transcrito MCM3APAS, a mesma foi analisada através de uma PCR em tempo real.

3.4.3.3 Quantificação da expressão do transcrito MCM3APAS em "Real-Time" PCR

Para medir os níveis de expressão do transcrito MCM3APAS através da PCR em Tempo Real foram utilizados, em um primeiro momento, o tecido normal de cérebro proveniente da CLONTECH® (correspondente a um "pool" de 4 amostras), as linhagens tumorais A172 e T98G e amostras de tecido normal de cérebro, astrocitoma de grau II e meningioma cedidas pelo laboratório da Dra. Mari Sogayar do Instituto de Química da Universidade de São Paulo.

Esta análise foi realizada no aparelho LigthCycler™ (Roche Diagnostics, Mannheim, Germany) sendo apresentados na Tabela 14 os valores de CT obtidos para cada amostra analisada. O que podemos observar é a ausência de reprodutibilidade entre os CTs de cada duplicata correspondente à análise do transcrito MCM3APAS onde encontramos amostras que apresentaram uma diferença de até oito ciclos entre as reações. Entretanto, com relação aos CTs correspondentes à análise da β-actina, os dados foram completamente reprodutíveis apresentando uma diferença mínima entre as duplicatas. Por ser um transcrito muito pouco expresso podemos demonstrar que o aparelho não apresentou uma sensibilidade suficiente para determinar o CT das amostras analisadas uma vez que ainda em muitos casos o valor obtido foi > 41.0 ou indeterminado.

Tabela 14 - Valores de CTs obtidos para os genes MCM3APAS e β-actina nas reações de "real time" realizadas no aparelho LigthCycler™. CT1 e CT2 correspondem às duplicatas de cada análise. Indet = valor não determinado pelo aparelho. Nd = "not done" (a reação não foi realizada).

Amostra	MCM3APAS CT1	MCM3APAS CT2	β-actina CT1	β-actina CT2
Cérebro normal CLONTECH®	Indet.	Indet.	14.83	14.84
Linhagem celular A172	> 41.0	35.84	11.96	12.35
Linhagem celular T98G	> 41.0	34.10	10.15	10.52
Tecido normal de cérebro – I	> 41.0	> 41.0	nd	nd
Tecido normal de cérebro – II	> 41.0	> 41.0	nd	nd
Astrocitoma de grau II	36.83	37.11	nd	nd
Meningioma – I	> 41.0	> 41.0	nd	nd
Meningioma - II	> 41.0	32.65	nd	nd

Frente a estas dificuldades a análise foi realizada agora no aparelho ABI Prism® 7000 Sequence Detection System. No entanto, como as amostras I e II de tecido normal de cérebro, astrocitoma de grau II e meningioma I e II foram completamente utilizadas nas análises anteriores as mesmas não seguiram na avaliação tendo sido utilizadas amostras de glioblastomas cedidas pelo Dr. Greg Riggins da Universidade de Duke. A Tabela 15 apresenta os dados dos CTs correspondentes à análise do gene MCM3APAS além dos normalizadores β-actina, ciclofilina e GAPDH.

Tabela 15 - Valores de CTs obtidos para o gene MCM3APAS e os normalizadores β-actina, ciclofilina e GAPDH nas reações realizadas no aparelho ABI Prism. CT1 e CT2 correspondem às duplicatas de cada análise.

Amostra	MCM3APAS (CT1 / CT2)	β-actina (CT1 / CT2)	Ciclofilina (CT1 / CT2)	GAPDH (CT1 / CT2)
Cérebro normal CLONTECH®	36.08 / 36.13	15.46 / 15.31	24.53 / 24.51	19.94 / 19.05
Linhagem celular A172	33.67 / 33.92	14.36 / 14.87	25.71 / 25.26	17.63 / 17.28
Linhagem celular T98G	31.40 / 31.72	12.86 / 11.92	24.12 / 24.17	17.00 / 17.26
Glioblastoma – I	33.99 / 34.01	15.07 / 14.75	25.68 / 25.41	20.80 / 20.79
Glioblastoma – II	34.01 / 34.17	15.10 / 14.76	26.07 / 26.52	17.55 / 18.26
Glioblastoma – III	32.67 / 33.22	15.05 / 14.91	24.12 / 24.30	18.46 / 18.28
Glioblastoma – IV	36.07 / 36.82	15.20 / 15.16	27.28 / 27.43	18.42 / 18.31
Glioblastoma - V	35.33 / 35.34	17.83 / 17.84	25.82 / 26.11	14.79 / 14.63

Nesta nova análise podemos observar que a reprodutibilidade antes presente somente nos genes constitutivos agora também pode ser encontrada nos dados obtidos para o transcrito MCM3APAS demonstrando, assim, uma maior sensibilidade do aparelho ABI Prism com relação ao LigthCycler™. Desta forma, os valores obtidos foram considerados confiáveis partindo-se, então, para a quantificação real da expressão do referido gene.

A amplificação do gene MCM3APAS foi quantificada e relacionada com os valores obtidos com cada gene constitutivo utilizando-se a equação $2^{-\Delta\Delta CT}$ para ser adquirido um valor normalizado de cada amostra quanto à variabilidade na quantidade e à integridade do RNA (LIVAK e SCHMITTGEN 2001). Vale destacar que o valor inferido à ΔCT correspondeu à diferença entre a média dos CTs do gene de interesse e a média dos CTs obtidos para cada um dos constitutivos. Já o cálculo da fórmula ΔΔCT envolveu a subtração entre o valor de ΔCT para cada amostra de tecido

tumoral e o valor de ΔCT calculado para o tecido normal correspondente. A Tabela 16 apresenta os valores normalizados do transcrito MCM3APAS obtidos com relação a cada gene constitutivo.

Tabela 16 - Valores normalizados do transcrito MCM3APAS com relação aos genes constitutivos β-actina, cilcofilina e GAPDH. Valores com sinal positivo indicam um nível de expressão elevado em número de vezes da correspondente amostra com relação ao tecido normal. Valores com sinal negativo indicam um nível de expressão reduzido em número de vezes da correspondente amostra com relação ao tecido normal.

Amostras	β-actina	Ciclofilina	GAPDH
Cérebro normal CLONTECH®	1	1	1
Linhagem celular A172	+ 2.91	+ 9.65	+ 1.21
Linhagem celular T98G	+ 2.93	+ 17.88	+ 4.53
Glioblastoma – I	+ 3.10	+ 8.69	+ 10.56
Glioblastoma – II	+ 2.95	+ 13.74	+ 1.34
Glioblastoma – III	+ 6.77	+ 7.21	+ 4.11
Glioblastoma – IV	- 1.44	+ 5.62	- 2.77
Glioblastoma – V	+ 9.32	+ 4.63	- 16.67

Frente aos resultados obtidos e apresentados na tabela anterior torna-se difícil concluir qual seria o melhor gene constitutivo a ser utilizado em estudos de expressão diferencial em cérebro. Diante dos dados discrepantes encontrados não foi possível inferir um valor referente ao nível de expressão do gene MCM3APAS, entretanto, o que podemos concluir é que o transcrito MCM3APAS se apresenta mais expresso em amostras de tecido tumoral de cérebro do que no normal correspondente uma vez que, independentemente do gene constitutivo utilizado, em cinco das sete amostras analisadas foi possível observar um nível de expressão mais elevado. Ainda

assim, a análise de um número maior de amostras é necessária para que estes dados sejam confirmados.

3.4.3.4 Análise do perfil de expressão do transcrito MCM3AP

O transcrito MCM3APAS ("MCM3-associated protein antisense") foi assim nomeado por estar localizado na fita oposta do gene MCM3AP ("minichromosome maintenance deficient 3 acetylating protein") com o qual compartilha dois exons (Figura 29). Sabe-se que este transcrito é uma acetiltransferase que atua inibindo o início da replicação do DNA agindo através da interação com proteínas MCM ("minichromosome maintenance"), componentes essenciais do complexo de pré-replicação (TAKEI et al. 2001, 2002). Desta forma, com base nestes dados e frente a relatos da literatura que demonstram a existência da regulação da expressão de um gene por seu correspondente "antisense" afetando a transcrição, o processamento e a tradução do mesmo (YELIN et al. 2003), foi questionada a possibilidade de regulação da expressão entre os genes MCM3AP e MCM3APAS levando-nos, assim, a avaliar o perfil de expressão do gene MCM3AP para fins de comparação.

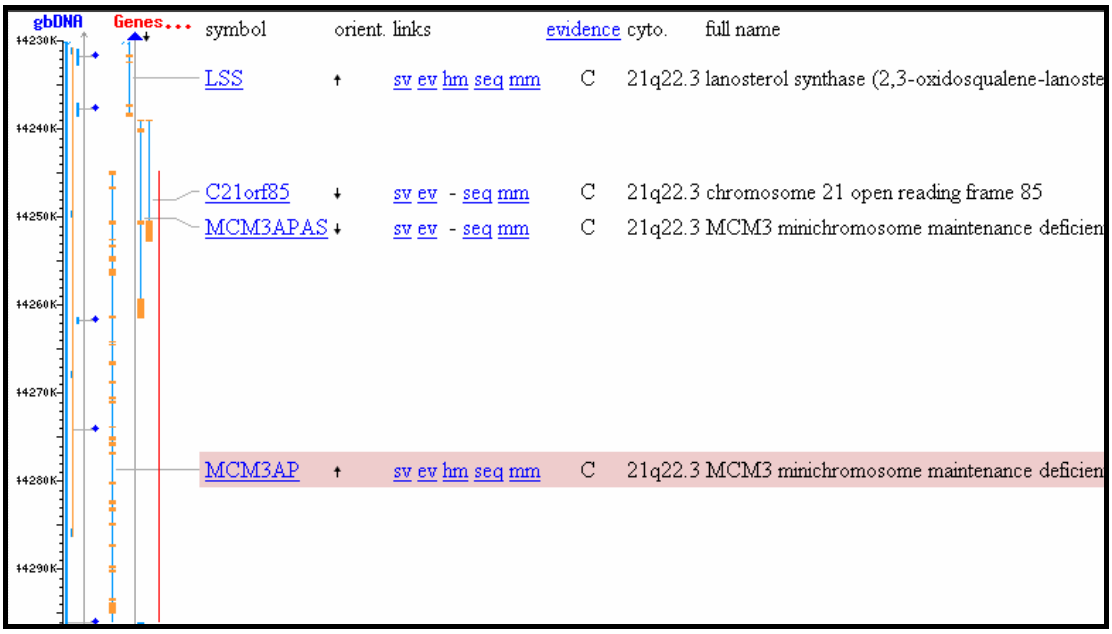


Figura 29 - Representação esquemática da localização cromossômica dos transcritos MCM3AP e MCM3APAS utilizando a ferramenta "Map viewer" disponibilizada pelo NCBI. Localização dos transcritos MCM3AP e MCM3APAS no cromossomo 21 indicando a localização em fitas opostas e os exons em comum.

Em um primeiro momento, a expressão do gene MCM3AP foi avaliada nos tecidos: testículo, pulmão, próstata, mama, cérebro, cólon, fígado, rim, baço e placenta. Os dois transcritos apresentaram um padrão de expressão similar sendo que o MCM3AP mostrou-se expresso em todos os tecidos analisados enquanto o transcrito MCM3APAS não apresentou expressão apenas em pulmão e fígado (Tabela 17).

Tabela 17 - Perfil de expressão dos transcritos MCM3AP e MCM3APAS. O sinal positivo representa a expressão no referido tecido enquanto o sinal negativo representa a ausência de expressão no mesmo.

Tecido	MCM3AP	MCM3APAS
Testículo	+	+
Pulmão	+	-
Próstata	+	+
Mama	+	+
Cérebro	+	+
Cólon	+	+
Fígado	+	-
Rim	+	+
Baço	+	+
Placenta	+	+

O gene MCM3AP também teve sua expressão quantificada pela PCR em tempo real utilizando-se o aparelho ABI Prism® 7000 Sequence Detection System e as mesmas amostras de linhagens celulares de cérebro e glioblastomas. A Tabela 18 apresenta os dados dos CTs correspondentes à análise do gene MCM3AP além dos constitutivos β-actina, ciclofilina e GAPDH.

Tabela 18 - Valores de CTs obtidos para o gene MCM3AP e os constitutivos β-actina, ciclofilina e GAPDH nas reações realizadas no aparelho ABI Prism. CT1 e CT2 correspondem às duplicatas de cada análise.

Amostra	MCM3AP (CT1 / CT2)	β-actina (CT1 / CT2)	Ciclofilina (CT1 / CT2)	GAPDH (CT1 / CT2)
Cérebro normal CLONTECH®	24.80 / 24.29	15.46 / 15.31	24.53 / 24.51	19.94 / 19.05
Linhagem celular A172	24.04 / 23.50	14.36 / 14.87	25.71 / 25.26	17.63 / 17.28
Linhagem celular T98G	24.00 / 24.04	12.86 / 11.92	24.12 / 24.17	17.00 / 17.26
Glioblastoma – I	24.98 / 25.09	15.07 / 14.75	25.68 / 25.41	20.80 / 20.79
Glioblastoma – II	26.07 / 26.18	15.10 / 14.76	26.07 / 26.52	17.55 / 18.26
Glioblastoma – III	25.37 / 25.18	15.05 / 14.91	24.12 / 24.30	18.46 / 18.28
Glioblastoma – IV	26.07 / 25.37	15.20 / 15.16	27.28 / 27.43	18.42 / 18.31
Glioblastoma - V	23.17 / 22.44	17.83 / 17.84	25.82 / 26.11	14.79 / 14.63

Novamente podemos observar a reprodutibilidade dos dados obtidos tanto com relação ao transcrito MCM3AP como para os genes constitutivos. Assim, a amplificação do gene MCM3AP foi quantificada e relacionada com os valores obtidos com cada gene constitutivo utilizando-se a mesma equação $2^{-\Delta\Delta CT}$ para ser adquirido um valor normalizado de cada amostra quanto à variabilidade na quantidade e à integridade do RNA. A Tabela 19 apresenta os valores normalizados do transcrito MCM3AP obtidos com relação a cada gene constitutivo.

Tabela 19 - Valores normalizados do transcrito MCM3AP com relação aos genes de referência β-actina, ciclofilina e GAPDH. Valores com sinal positivo indicam um nível de expressão elevado em número de vezes da correspondente amostra com relação ao tecido normal. Valores com sinal negativo indicam um nível de expressão reduzido em número de vezes da correspondente amostra com relação ao tecido normal.

Amostras	β-actina	Ciclofilina	GAPDH
Cérebro normal CLONTECH®	1	1	1
Linhagem celular A172	+ 1.00	+ 3.32	- 2.43
Linhagem celular T98G	- 5.55	+ 1.10	- 3.57
Glioblastoma – I	- 2.32	+ 1.20	+ 1.45
Glioblastoma – II	- 4.16	+ 1.14	- 9.09
Glioblastoma – III	- 2.17	- 2.04	- 3.57
Glioblastoma – IV	- 2.63	+ 3.14	- 5.0
Glioblastoma - V	+ 18.25	+ 9.06	- 8.33

Neste caso, assim como nos dados obtidos com o transcrito MCM3APAS, há uma grande variação entre os níveis de expressão com relação a cada um dos constitutivos sendo impossível inferir um valor referente ao nível de expressão do gene MCM3AP. O que podemos concluir, tomando-se por base a β-actina, é que o transcrito MCM3APAS se apresenta mais expresso em amostras de tecido tumoral de cérebro do que no normal correspondente enquanto o MCM3AP está menos expresso em tumores de cérebro. Se levássemos em conta apenas o gene ciclofilina seria impossível obter um resultado conclusivo pois enquanto o gene MCM3APAS apresenta-se mais expresso em tumores de cérebro, o MCM3AP na maioria das amostras analisadas não apresenta uma diferença significativa de expressão (> 2.0). Já no caso do GAPDH, cada amostra apresentou um comportamento diferente entre os transcritos dificultando ainda mais nossa análise. Desta maneira, com tantas variações

de valores tal análise torna-se por fim subjetiva e limita nossa interpretação dos dados obtidos.

CONSIDERAÇÕES FINAIS

3.5 CONSIDERAÇÕES FINAIS

Embora o cromossomo 21 já esteja relativamente bem anotado, o alinhamento entre a seqüência genômica e seqüências expressas demonstrou grande eficiência na identificação de novos transcritos elevando em 10% o número total de genes neste cromossomo. Isto demonstra que a identificação completa do conjunto de genes humanos não dependerá somente de métodos computacionais mas sim destes ligados à verificação experimental. Frente aos resultados apresentados neste trabalho podemos ainda concluir que o número de genes no genoma humano pode ser pelo menos 10% maior do que se espera.

Os 23 genes identificados apresentaram características peculiares como ORFs pequenas (ou mesmo ausência de ORF), baixo número de exons e padrão de expressão baixo e restrito o que dificultaria a identificação dos mesmos por programas de predição gênica. Além disso, a análise de expressão destes transcritos revelou dois genes, C21orf99 e C21orf100, que apresentaram um perfil de expressão restrito em tecidos normais levando-nos a caracterizá-los como possíveis antígenos tumorais.

Frente ao perfil de expressão e com base nos dados obtidos no decorrer do projeto de doutorado do aluno Raphael B. Parmigiani de nosso laboratório, que identificou e caracterizou o transcrito C21orf99, foi possível concluir que este transcrito corresponde a um antígeno tumoral da classe "cancer-testis", sendo importante ressaltar que antígenos desta classe são considerados, atualmente, candidatos ideais para o desenvolvimento de vacinas e imunização passiva (SCANLAN et al. 2002). Já com relação ao C21orf100, por apresentar um perfil de expressão restrito a próstata normal e com base nas análises de expressão

subseqüentes em amostras de tumores de próstata, tireóide, estômago, útero, melanoma e mama acreditamos que o mesmo possa representar um possível membro da classe de antígenos de diferenciação (RETTIG e OLD 1989), importantes marcadores no diagnóstico imunopatológico diferencial do câncer.

Através da análise de expressão diferencial dos novos transcritos entre tecido normal e tumoral foi possível identificar um transcrito, MCM3APAS, com expressão diferencial entre cérebro normal e tumoral. Este gene, diferentemente dos outros identificados no cromossomo 21, foi assim nomeado por estar localizado na fita oposta do gene MCM3AP com o qual compartilha dois exons. Sabe-se que o transcrito MCM3AP é uma acetiltransferase que atua inibindo o início da replicação do DNA agindo através da interação com proteínas MCM ("minichromosome maintenance"), componentes essenciais do complexo de pré-replicação (TAKEI et al. 2001, 2002), levando-nos, desta forma, a supor que um gene pudesse estar regulando a ação do outro. Entretanto, com base na análise de expressão de ambos não foi possível obter dados conclusivos quanto a este fato.

REFERÊNCIAS BIBLIOGRÁFICAS

4 REFERÊNCIAS BIBLIOGRÁFICAS

Aach J, Bulyk ML, Church GM, Comander J, Derti A, Shendure J. Computational comparison of two draft sequences of the human genome. **Nature** 2001; 409:856-9.

Adams MD, Dubnick M, Kerlavage AR, et al. Sequence identification of 2,375 human brain genes. **Nature** 1992; 355:632-4.

Adams MD, Kerlavage AR, Fields C, Venter JC. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. **Nat Genet** 1993; 4:256-67.

Adams MD, Kerlavage AR, Fleischmann RD, et al. Initial assessment of the human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. **Nature** 1995; 377:3-174. Supplement.

Antonarakis SE. Chromosome 21: from sequence to applications. **Curr Opin Genet Dev** 2001; 11:241-6.

Bailey LC, Searls Jr DB, Overton GC. Analysis of EST-driven gene annotation in human genomic sequence. **Genome Res** 1998; 8:362-76.

Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. **Genome Res** 2001; 11:1520-6.

Bera TK, Maitra R, Iavarone C, et al. PATE, a gene expressed in prostate cancer, normal prostate, and testis, identified by a functional genomic approach. **Proc Natl Acad Sci U.S.A** 2002; 99:3058-63.

Boon K, Osorio EC, Greenhut SF, et al. An anatomy of normal and malignant gene expression. **Proc Natl Acad Sci U.S.A** 2002; 99:11287-92.

Brett D, Hanke J, Lehmann G, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. **FEBS Lett** 2000; 474:83-6.

Broder S, Venter JC. Whole genomes: the foundation of new biology and medicine. **Curr Opin Biotechnol** 2000; 11:581-5.

Brown TA. **Genomes**. Oxford: BIOS Scientific; 1999. p.1-12: What is a genome?.

Brown S, Chang JL, Sadee W, Babbitt PC. A semiautomated approach to gene discovery through expressed sequence tags data mining: discovery of new human transporter genes. **AAPS PharmSci** 2003; 5:1-18.

Burset M, Guigo R. Evaluation of gene structure prediction programs. **Genomics** 1996; 34:353-67.

Camargo AA, Samaia HP, Dias Neto E, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci U.S.A** 2001; 98:12103-8.

Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. **Biochemistry** 1979; 18:5294-9.

Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH. Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. **Genome Res** 2000; 10:1259-65.

Collins JE, Goward ME, Cole CG, et al. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. **Genome Res** 2003; 13:27-36.

Dermitzakis ET, Reymond A, Lyle R, et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. **Nature** 2002; 420:578-82.

De Souza SJ, Camargo AA, Briones MR, et al. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. **Proc Natl Acad Sci U.S.A** 2000; 97:12690-3.

Dias Neto E, Garcia CR, Verjovski-Almeida S, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci U.S.A** 2000; 97:3491-6.

Dunham I, Shimizu N, Roe BA, et al. The DNA sequence of human chromosome 22. **Nature** 1999; 402:489-95.

Eckman BA, Aaronson JS, Borkowski JA, et al. The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. **Bioinformatics** 1998; 14:2-13.

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Res** 1998; 8:175-85.

Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. **Nat Genet** 2000; 25:232-4.

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. **Genome Res** 1998; 8:967-74.

Gardiner K, Slavov D, Bechtel L, Davisson M. Annotation of human chromosome 21 for relevance to Down syndrome: gene structure and expression analysis. **Genomics** 2002; 79:833-43.

Grillo-Lopez AJ, White CA, Varns C, et al. Overview of the clinical development of rituximab: first monoclonal antibody approved for the treatment of lymphoma. **Semin Oncol** 1999; 26:66-73.

Hasle H, Clemmensen IH, Mikkelsen M. Risks of leukaemia and solid tumours in individuals with Down`s syndrome. **Lancet** 2000; 355: 165-9.

Hattori M, Fujiyama A, Taylor TD, et al. The DNA sequence of human chromosome 21. **Nature** 2000; 405: 311-9.

Hillier LD, Lennon G, Becker M, et al. Generation and analysis of 280,000 human expressed sequence tags. **Genome Res** 1996; 6:807-28.

Hirosawa M, Nagase T, Murahashi Y, Kikuno R, Ohara O. Identification of novel transcribed sequences on human chromosome 22 by expressed sequence tag mapping. **DNA Res** 2001; 8:1-9.

Hogenesch JB, Ching KA, Batalov S, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. **Cell** 2001; 106:413-5.

Hu G, Modrek B, Riise Stensland HM, et al. Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. **Pharmacogenomics J** 2002; 2:236-42.

Hudson TJ, Colbert AM, Reeve MP, et al. Isolation and regional mapping of 110 chromosome 22 STSs. **Genomics** 1994; 24:588-92.

Huminiecki L, Bicknell R. *In silico* cloning of novel endothelial-specific genes. **Genome Res** 2000; 10:1796-806.

Iavarone C, Wolfgang C, Kumar V, et al. PAGE4 is a cytoplasmic protein that is expressed in normal prostate and in prostate cancers. **Mol Cancer Ther** 2002; 1:329-35.

Inoue H, Nojima H, Okayama H. High efficiency transformation of Escherichia coli with plasmids. **Gene** 1990; 96:23-8.

Irizarry K, Kustanovich V, Li C, et al. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. **Nat Genet** 2000; 26:233-6.

Iseli C, Stevenson BJ, De Souza SJ, et al. Long-range heterogeneity at the 3' ends of human mRNAs. **Genome Res** 2002; 12:1068-74.

Jiang J, Jacob HJ. EbEST: an automated tool using expressed sequence tags to delineate gene structure. **Genome Res** 1998; 8:268-75.

Katsanis N, Worley KC, Gonzalez G, Ansley SJ, Lupski JR. A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. **Proc Natl Acad Sci U.S.A** 2002; 99:14326-31.

Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. **Genome Res** 2002; 12:996-1006.

Kohno T, Kawanishi M, Matsuda S, et al. Homozygous deletion and frequent allelic loss of the 21q11.1-q21.1 region including the ANA gene in human lung carcinoma. **Genes Chromosomes Cancer** 1998; 21:236-43.

Korenberg JR, Chen XN, Adams MD, Venter JC. Toward a cDNA map of the human genome. **Genomics** 1995; 29:364-70.

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

Lee Y, Sultana R, Pertea G, et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). **Genome Res** 2002; 12:493-502.

Lejeune J, Gautier M, Turpin R. Etude des chromosomes somatique des neufs enfants mongoliens. **CR Acad Sci Paris** 1959; 248:1721-2.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. **Nat Genet** 2000; 25:239-40.

Liang P. SAGE Genie: a suite with panoramic view of gene expression. **Proc Natl Acad Sci U.S.A** 2002; 99:11547-8.

Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\,\Delta CT}$ method. **Methods** 2001; 25:402-8.

Mégy K, Audic S, Claverie JM. Heart-specific genes revealed by expressed sequence tags (EST) sampling. **Genome Biol** 2002; 3:1-11.

Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. **Nucleic Acids Res** 2001; 29:2850-9.

Morrison TB, Weis JJ, Wittwer CT. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. **Biotechniques** 1998; 24:954-9.

Nizetic D. Functional genomics of the Down syndrome. **Croat Med J** 2001; 42:421-7.

Olsson P, Bera TK, Essand M, et al. GDEP, a new gene differentially expressed in normal prostate and prostate cancer. **Prostate** 2001; 48:231-41.

Olsson P, Motegi A, bera TK, Lee B, Pastan I. PRAC2: a new gene expressed in human prostate and prostate cancer. **Prostate** 2003; 56:123-30.

Pennacchio LA, Lehesjoki AE, Stone NE, et al. Mutations in the gene encoding cystatin B in progressive myoclonus epilepsy (EPM1). **Science** 1996; 271:1731-4.

Pletcher MT, Wiltshire T, Cabin DE, Villanueva M, Reeves RH. Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. **Genomics** 2001; 74:45-54.

Quackenbush J, Liang F, Holt I, Pertea G, Upton J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. **Nucleic Acids Res** 2000; 28:141-5.

Rettig WJ, Old LJ. Immunogenetics of human cell surface differentiation. **Annu Rev Immunol** 1989; 7:481-511.

Reymond A, Friedli M, Henrichsen CN, et al. From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. **Genomics** 2001; 78:46-54.

Roest CH, Jaillon O, Bernot A, et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. **Nat Genet** 2000; 25:235-8.

Rogic S, Mackworth AK, Ouellette FB. Evaluation of gene-finding programs on mammalian sequences. **Genome Res** 2001; 11:817-32.

Sakata K, Tamura G, Nishizuka S, et al. Commonly deleted regions on the long arm of chromosome 21 in differentiated adenocarcinoma of the stomach. **Genes Chromosomes Cancer** 1997; 18:318-21.

Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. **Immunol Rev** 2002; 188:22-32.

Silva AP, Salim AC, Bulgarelli A, et al. Identification of 9 novel transcripts and two RGSL genes within the hereditary prostate cancer region (HPC1) at 1q25. **Gene** 2003; 310:49-57.

Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic treshold control suited for genomic DNA and protein sequence analysis. **Gene** 1996; 167:GC1-10.

Sorek R, Safer HM. A novel algorithm for computational identification of contaminated EST libraries. **Nucleic Acids Res** 2003; 31:1067-74.

Straub RE, Lehner T, Luo Y, et al. A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. **Nat Genet** 1994; 8:291-6.

Strausberg RL, Feingold EA, Klausner RD, Collins FS. The mammalian gene collection. **Science** 1999; 286:455-7.

Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. The cancer genome anatomy project: building an annotated gene index. **Trends Genet** 2000; 16:103-6.

Suzuki T, Terasaki M, Takemoto-Hori C, et al. Proteomic analysis of the mammalian mitochondrial ribosome: identification of protein components in the 28S small subunit. **J Biol Chem** 2001; 276:33181-95.

Takei Y, Swietlik M, Tanoue A, Tsujimoto G, Kouzarides T, Laskey R. MCM3AP, a novel acetyltransferase that acetylates replication protein MCM3. **EMBO Rep** 2001; 2:119-23.

Takei Y, Assenberg M, Tsujimoto G, Laskey R. The MCM3 acetylase MCM3AP inhibits initiation, but not elongation, of DNA replication via interaction with MCM3. **J Biol Chem** 2002; 277:43121-5.

Toyoda A, Noguchi H, Taylor TD, et al. Comparative genomic sequence analysis of the human chromosome 21 Down syndrome critical region. **Genome Res** 2002; 12:1323-32.

Tugendreich S, Bassett DE Jr, McKusick VA, Boguski MS, Hieter P Genes conserved in yeast and humans. **Hum Mol Genet** 1994; 3:1509-17.

Van Gent D, Sharp P, Morgan K, Kalsheker N. Serpins: structure, function and molecular evolution. **Int J Biochem Cell Biol** 2003; 35:1536-47.

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

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

Vitale L, Casadei R, Canaider S, et al. Cysteine and tyrosine-rich 1 (CYYR1), a novel unpredicted gene on human chromosome 21 (21q21.2), encodes a cysteine and tyrosine-rich protein and defines a new family of highly conserved vertebrate-specific genes. **Gene** 2002; 290:141-51.

Wang Z, Lo HS, Yang H, et al. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. **Cancer Res** 2003; 63:655-7.

Wiemann S, Weil B, Wellenreuther R, et al. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. **Genome Res** 2001; 11:422-35.

Williamson AR. The Merck Gene Index project. **Drug Discov Today** 1999; 4:115-122.

Wright FA, Lemon WJ, Zhao WD, et al. A draft annotation and overview of the human genome. **Genome Biol** 2001; 2 (7): research0025.1–research0025.18.

Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A, Mintz L. Computational analysis of alternative splicing using EST tissue information. **Genomics** 2002; 80:326-30.

Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. **Nucleic Acids Res** 2002; 30:3754-66.

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

Zhuo D, Zhao WD, Wright FA, et al. Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. **Genome Res** 2001; 11:904-18.

ANEXOS

ANEXO 1

A Transcript Finishing Initiative for Closing Gaps in the Human Transcriptome.

The Ludwig – FAPESP Transcript Finishing Initiative *.

* A complete list of authors appears at the end of this manuscript

Running Title: Transcribed regions in the human genome.

Key words: human genome, transcriptome, validation, alternative splicing

Corresponding Authors:

Anamaria A. Camargo

Ludwig Institute for Cancer Reserach

Rua Prof. Antonio Prudente 109, 4th floor

01509-010 São Paulo SP Brazil

Phone 55 11 3388-3248

Fax 55 11 3207-7001

anamaria@compbio.ludwig.org.br

Mari Cleide Sogayar

Chemistry Institute

University of São Paulo

Av. Prof. Lineu Prestes 748

05508-900 São Paulo SP Brasil

Phone 55 11 3091 3820

Fax 55 11 3091 3820

mcsoga@iq.usp.br

ABSTRACT

We report the results of a Transcript Finishing Initiative, undertaken for the purpose of identifying and characterizing novel human transcripts, in which RT-PCR was used to bridge gaps between paired EST clusters, mapped against the genomic sequence. Each pair of EST clusters selected for experimental validation was designated a Transcript Finishing Unit (TFU). A total of 489 TFUs were selected for validation, and an overall efficiency of 43.1% was achieved. We generated a total of 59,975 bp of transcribed sequences organized into 432 exons, contributing to the definition of the structure of 211 human transcripts. The structure of several transcripts reported here was confirmed during the course of this project, through the generation of their corresponding full-length cDNA sequences. Nevertheless, about 21% of the sequences we generated represent still unreported human transcripts, most of which had not been correctly predicted by computer programs. The TFI strategy provides a significant contribution to the definition of the complete catalog of human genes and transcripts, since it appears to be particularly useful for identification of low abundance transcripts expressed in a restricted set of tissues as well as for the delineation of gene boundaries and alternatively spliced isoforms.

Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. CF272536 to CF272733.

INTRODUCTION

A primary objective of the Human Genome Project has been the identification of the complete set of human genes and their derived transcripts. A major step towards this goal was achieved at the beginning of 2001 with the publication of two independent draft versions of the human genome sequence and the identification of more than 30,000 genes (Lander et al., 2001; Venter et al., 2001). However, it became apparent that extracting exonic sequences directly from the human genome is not straightforward and that a variety of complementary strategies are required for gene identification and characterization.

In this context, microarrays (Penn et al., 2000; Dennis, 2001; Schoemaker et al., 2001; Kapranov et al., 2002) and sequence comparisons with other organisms at an appropriate evolutionary distance (Batzoglou et al., 2000; Roest et al., 2000) constitute powerful preliminary approaches to identifying transcribed regions within the genome sequence. Nevertheless, transcript sequencing is necessary both for the final proof of the existence of an expressed gene and for the precise identification of intron/exon boundaries and alternatively spliced forms (Camargo et al., 2002).

A full-length cDNA sequence, ideally including a transcription initiation site and a polyadenylation site, is the gold standard for transcript definition. Considerable progress has been made in the generation of representative full-length cDNA sequences, especially following the development of sophisticated protocols for obtaining full-length transcript molecules and to correct for transcript expression bias (Bonaldo et al., 1996, Carninci et al., 2000). Currently, several major projects are systematically generating full-length cDNA sequences (Strausberg et al., 1999; Wiemann et al., 2001; Kikuno et al., 2002; Nakajima et al., 2002; Strausberg et al., 2002) and their contribution to complete the human gene catalog has been remarkable.

Expressed Sequence Tags (ESTs) are another major source of transcript sequence. ESTs are either single-pass, partial sequences derived either from the 5′ and 3′ extremities of cDNA clones (Adams et al., 1991) or are specifically directed towards the central coding regions of transcripts, in the case of Open Reading Frame ESTs (ORESTES) (Dias et al., 2000; Camargo et al., 2001). Initially, ESTs were exploited as a source for gene discovery (Adams et al., 1992; Adams et al., 1993), but have also been widely used to build tissue-specific transcript profiles (Bortoluzzi et al., 2000a Bortoluzzi et al., 2000b; Bortoluzzi et al., 2000c; Huminiecki and Bicknell, 2000; Phillips et al., 2000; Yu et al., 2001; Katsanis et al., 2002; Megy et al., 2003), to construct gene-based physical maps (Hudson et al., 1994), to compare genomes of different organisms (Tugendreich et al., 1994; Lee et al., 2002), to accurately identify transcripts in genomic sequences (Bailey, Jr. et al. 1998; Jiang and Jacob, 1998; Kan et al., 2001) and to study aspects of mRNA structure, such as splicing variants (Modrek et al., 2001; Hide et al., 2001, Clark and Thanaraj, 2002; Kan et al., 2002; Xie et al., 2002; Xu et al., 2002 Wang et al., 2003), alternative polyadenylation (Gautheret et al., 1998; Beaudoing and Gautheret, 2001; Iseli et al., 2002) and single nucleotide polymorphisms (Picoult-Newberg et al., 1999; Garg et al., 1999; Clifford et al., 2000; Hu et al., 2002; Irizarry et al., 2000;).

To date, more than 5,200,000 human ESTs have been generated from different organs and tissues, deriving mainly from the Merck Gene Index Project (Williamson et al. 1999), the Cancer Genome Anatomy Project (CGAP) (Strausberg et al., 2000), and the Human Cancer Genome Project Ludwig/FAPESP (HCGP) (Dias et al., 2000; Camargo et al., 2001). Nevertheless, it is widely recognized that EST databases are subjected to artifacts related to the partial, low quality, nature of the sequences and the presence of various kinds of contamination (Sorek and Safer, 2003).Typical

contaminants include vector sequences, sequences from other organisms, such as bacteria and viruses and, most importantly, sequences from intronic or intergenic DNA derived from DNA contamination and/or the presence of significant amounts of heterogeneous nuclear RNA in mRNA preparations used to generate the cDNA libraries. Additionally, because of the large differences in abundance between RNA species, the coverage of individual transcripts by ESTs is highly variable. Despite that, it is believed that the vast majority of transcripts have been sampled at least once by either a full-length cDNA or EST sequence (Ewing and Green, 2000; Liang et al., 2000).

Although the amount of transcript data currently available is not sufficient to identify all human genes, the judicious use of this dataset, in conjunction with the draft sequences of the human genome, has been highly informative in the characterization of new human genes. For example, we have recently built a Transcriptome Database, based on exhaustive pair-wise comparison between transcript sequences and the publicly available human genome sequence and have successfully used it to identify novel genes located on chromosome 21 (Reymond et al., 2002) and at the Hereditary Prostate Cancer Locus (HPC-1) at 1q25 (Silva et al., 2003).

Here we describe the utilization of the Transcriptome Database to guide the generation of novel human transcript sequences on a genome-wide basis. Using the genomic sequence as a scaffold for EST mapping and clustering, we have used RT-PCR to bridge gaps between EST clusters that we judged as likely to be derived from the same genes. The resulting novel sequence confirms that the ESTs from different clusters are, in fact, derived from a common transcript and defines the intervening region between them. Since this process is very similar to the finishing phase of

genome projects, we called it Transcript Finishing. This powerful, albeit laborious, approach allows the characterization of novel human transcripts and splicing isoforms, which appear to be generally expressed at a low abundance level and/or in a restricted set of tissues and avoids the necessity of a full-length cDNA clone in order to confirm the structure of a gene.

RESULTS

Generation of the Transcriptome Database and EST Cluster Selection for Experimental Validation

We have utilized the publicly available human genome and transcript sequences to identify and experimentally validate additional transcribed regions in the human genome. The two datasets were integrated into the Transcriptome Database by using the BLASTN program to map all transcript sequences onto the assembled version of the human genome available from the NCBI. We have also mapped to the genome, using the raw data generated by EST sequencing projects, a set of trusted 3' tags that provide unique identifiers for transcript 3' ends (Iseli et al., 2002). The tags were used for positional orientation and as a start point for transcript reconstruction. To facilitate visualization of the alignments and the access to information such as project and tissue source of the sequences, alignment scores, and the position of 3'tags, a graphical interface was also developed (Figure1).

We identified 244,148 human transcript clusters, of which 14,598 contained at least one full-length cDNA sequence, and 229,550 clusters which were composed exclusively of partial transcript sequences. Of the set of 14,598 clusters containing full-length sequences, 13,149 (90%) had at least one corresponding EST and the remaining 1,449 (10%) were composed only of full-length cDNA sequences. These

data demonstrate that, despite the fact that over 5 million EST sequences are available, they do not fully cover the human transcriptome and that the generation of additional transcribed sequences is still required.

It is noteworthy that clusters composed exclusively by ESTs have a reduced number of sequences (AVG 5.9 sequences) derived from fewer different tissues (AVG 3.0), as compared to clusters containing a full-length cDNA, which have an average of 65.5 sequences derived from 8 different tissues. Based on these observations, we conclude that the human transcripts that remain to be defined are expressed at low levels in a restricted set of tissues and that their characterization will benefit from a direct approach such as the Transcript Finishing.

Since EST databases contain a significant fraction of artifactual and contaminant sequences, we selected, for experimental validation, pairs of clusters that consist of ESTs that align non-contiguously to the genome, consistent with the presence of a splicing structure. We also restricted our validation to pairs of clusters that map at a maximum distance of 10 kb from each other, in order to increase the probability that these clusters belong to the same transcript. Using these criteria, a total of 2,373 pairs of clusters (~ 2% of the total number of clusters composed of partial sequences) were initially selected and subjected to manual inspection using our graphical interface.

Manual inspection allowed the assessment of similarity and extension of the alignments, as well as the position of the selected pair of clusters relative to the 3' tags. Following this procedure, 489 pairs of clusters were initially selected for experimental validation. Selected clusters were separated from each other by an average 2,879 bp of intervening genomic sequence and were composed by an average of 5.92 EST sequences derived from an average of 3 distinct tissues. Each pair of EST

clusters selected for experimental validation was designated as a single Transcript Finishing Unit (TFU). Information related to the 489 TFUs selected for validation can be accessed at http://200.18.51.201/viewtfi.

Experimental Validation and the Generation of New Transcribed Sequences

A general overview of the computational and experimental validation strategies is presented in Figure 2. A total of 2 Coordination Groups, 4 Bioinformatics Groups and 29 Validation Laboratories, linked through the Internet, participated in the computational and experimental phase of the project (http://200.18.51.201/transcript/Participants.html). Following cluster selection and manual inspection, primers for RT-PCR validation of each TFU were designed automatically. The genomic sequence was chosen as a template for primer design since it is generally of a higher quality than EST sequences. cDNA preparation was also a critical issue, since both the quality and the representation of different tissues directly influence the validation efficiency. Several controls were implemented to ensure that the material distributed to the validation groups were of high quality and totally free of DNA contamination. A total of 22 cDNA preparations, representing 18 distinct tissues, were utilized.

The total of 3,019 sequences, generated during the project, was subjected to an automated cleaning protocol. High-quality, sequences were aligned against the genomic sequence and the alignment coordinates and scores for validated sequences were loaded into the Transcriptome Database and displayed on the graphical interface (Figure 1). We successfully validated 211 of the 489 TFUs that were distributed, yielding an overall validation efficiency of 43.1%. A total of 59,975 bp of transcribed sequence, organized into 432 exons, were generated contributing to the definition of

the structure of 211 distinct human transcripts. Each validated TFU had a mean of 281.6 bp and a median of 207 bp of novel sequence not represented by the original EST clusters and a mean of 2.03 and a median of 2 newly defined exons. Sequences for validated TFU have been submitted to GenBank under the accession numbers CF272536 to CF272733.

In order to identify variables related to the expression pattern of the novel transcripts that influence the efficiency of validation, two sets composed of 174 validated TFUs and 208 non-validated TFUs were compared. As shown in Table 1, the validated TFUs had, on average, more ESTs in each cluster derived from a larger number of different tissues. Both of these differences were statistically significant according to Mann-Whitney tests, indicating that a higher expression level and a broader expression pattern of the selected transcripts favored validation. The presence of ESTs derived from the same tissue in both clusters did not influence the likelihood of validation according to qui-square tests.

Consensus Sequences Generation and Annotation of the Validated Human Transcripts

Consensus sequences produced by assembling the sequences derived from the validation fragment and the sequences from all ESTs in both clusters were obtained for 186 of the 211 validated TFUs. Assembly of a consensus sequence was not possible for 25 TFUs, due mainly to the presence of repetitive sequences and alternative splicing forms. Consensus sequences, with an average of 1,240 bp, can be accessed at (http://200.18.51.201/viewconsensus/).

Consensus sequences derived from the validated TFUs were aligned to the July 2003 version of human genome sequence assembly provided by the UCSC, using

the BLAT search tool (http://genome.ucsc.edu) to compare the validated consensus sequences with known genes and gene predictions (Table 2). A significant fraction (68.8%) of the validated transcripts completely overlapped with the alignment coordinates of a known gene or full length human mRNA submitted to the GenBank during the course of our project (Figure 3a) and a smaller fraction (10.2%) represented extensions (mostly 5´) to partial cDNA sequences deposited in public databases (Figure 3b). However, 21% of the validated sequences represent new human transcripts as of July 2003 for which no full-length cDNA is yet available (Figure 3c, 3d). The structure of the majority (69.2%) of these new human transcripts had not been correctly predicted by *ab initio* gene prediction programs such as Fgenesh++, Geneid and GenScan.

The consensus sequences corresponding to new human transcripts were further characterized by BLASTX analysis and protein domains were predicted using the Pfam and Prosite databases. Of the 39 consensus sequences representing new human transcripts, 27 (69.2%) contain an open reading frame of at least 100 amino acids and 8 (20.5%) contained a clearly defined protein domain including three IG-like domains and a protein kinase. Complete information on the characterization of the validated TFUs, including consensus size, annotation, chromosomal location and expression pattern based on ESTs distribution are provide as Supplemental Material – Table1.

Identification and Experimental Validation of Alternatively Spliced Isoforms

Several reports have suggested that at least 30-35% of human genes undergo alternative splicing (Brett et al., 2000; Modrek et al., 2001), nevertheless, this value is probably underestimated since many cell types have not yet been fully explored by

cDNA sequencing. The use of different cDNA sources during the experimental validation phase of the new human transcripts, allowed us to identify many new splicing variants. We explored the degree of sequence variability due to alternative splicing in the set of 186 consensus sequences that we generated and found evidence for alternative splicing in 22 (12%) cases (Table 3). Intron retention was observed in 11 TFUs and alternative exon usage was detected in 11 of the 22 TFUs with alternative splicing. Conserved GT-AG splice junctions were present in all TFUs with alternative exon usage.

We selected 6 TFUs with alternative exon usage, representing a total of 14 splicing isoforms for further experimental validation. Touchdown PCR confirmed 10 (83%) of the putative investigated isoforms. No PCR amplification was achieved for one TFU. Some splicing isoforms were expressed in a restricted pattern, being detected in one or a few of the tissues analyzed by RT-PCR (data not shown). None of these splicing isoforms had been previously identified, highlighting the potential use of the TF strategy for uncovering the genetic variability generated at the transcriptome level.

A typical example of this experimental validation is illustrated in Figure 4. In this case, we were able to identify two alternative exons, one of which presents an extra exon of 138 bp, and the other a 21 bp extension of an exon already represented by EST sequences. The possible combination of these variants results in four splicing isoforms. Figure 4 shows a 388 bp product (obtained with primers P1 and P2) corresponding to the prototype isoform, a 370 bp product (primers P2 and P3) corresponding to the isoform containing the additional exon, a 314 bp product (primers P1 and P4) corresponding to the isoform with the extended exon and a 452

bp product corresponding to the isoform containing both the additional exon and the extended exon.

DISCUSSION

Currently, intense activity is directed towards defining the complete set of genes and their derived transcripts in the human genome. This information will have a profound impact in diverse areas of Biology such as Human Evolution, Structural Genomics and Medicine. However, due to the highly dispersed and complex structure of human genes, it is extremely difficult to correctly identify transcribed regions within the genome (Camargo et al., 2002).

Estimates based on gene prediction both within individual finished chromosomes (Dunham et al., 1999; Hattori et al., 2000), as well as in the draft human genome sequences (Lander et al., 2001; Venter et al., 2001), have uniformly concluded that the human genome possesses less than 35,000 genes. This number has been supported by a preliminary analysis of EST coverage of known genes (Ewing and Green, 2000) as well as comparative genomics analysis (Roest et al., 2000). Most of these 35,000 genes are already represented by a full-length cDNA sequence in transcript databases. In Unigene (http://www.ncbi.nlm.nih/Unigene/), for example, there are currently 28,412 transcript clusters represented by at least one full length cDNA sequence.

However, other analyses based on ESTs clustering (Liang et al., 2000) and mapping to the genome (Das et al., 2001; Wright et al., 2001), suggest that human transcripts are derived from a much larger gene set, ranging from 60,000 to 100,000. These analyses are supported by the fact that transcript sequences databases are also composed by a high proportion of clusters containing exclusively EST sequences. For

instance, the Unigene database contains 79,628 clusters composed only of EST sequences and a significant fraction (37.3%) of these clusters are composed of only one sequence. The reduced number of sequences in clusters composed exclusively by ESTs suggests that most of the yet to be defined human transcripts are expressed at low level and in a restricted set of tissues. Nevertheless, since EST databases are subjected to many artifacts, it is not possible to accurately determine the number of human genes represented by these sequences. Characterization of these transcripts will, therefore, require the use of varied and complementary strategies and will certainly benefit from a direct approach with a planned experimental support such as the Transcript Finishing.

Here we have proposed and validated the use of the TF strategy for characterization of new human transcripts which are only partially represented by ESTs. Since EST databases contain a significant fraction of artifactual and contaminant sequences, we selected pairs of clusters for experimental validation that exhibited a clear splicing structure when aligned to the genome. By requiring the occurrence of splicing, the level of contamination in the EST databases is significantly reduced, although at the expense of eliminating many genuine 3' ESTs. The selection criteria used in our initial analysis are very restrictive and the adoption of less stringent criteria (including clusters without a splicing structure) will certainly be required to complete the catalog of human genes using the strategy we described. Given the 2,373 initially selected clusters, of which 489 were subjected to experimental validation, 1,884 pairs of clusters remain to be validated. If we assume an overall validation efficiency of 43%, we can estimate that the TF strategy might contribute to the definition of at least 791 additional genes in the human genome.

Several factors may have influenced our validation efficiency including experimental limitations related to primer and cDNA synthesis, the particular characteristics of human transcripts such as low expression level and the existence of a significant proportion of sense-antisense transcriptional units on opposite DNA strands of the same genomic locus (Yelin et al 2003). We found that validation efficiency was enhanced by implementation of quality controls for cDNA synthesis, the use of polyA+ derived cDNA and a combination of both oligo dT and random primers for cDNA synthesis and also the use of nested RT-PCR. We also observed that validated pairs of clusters had a higher average number of ESTs per cluster, and a higher number of different tissues represented by the clusters as compared with pairs of clusters that we were not able to validate. Validated TFUs had on average 6.1 ESTs in each cluster derived on average from 3.45 distinct tissues. Noteworthy, in 41% of the validated TFUs one of the two EST clusters was composed of a single EST and in 13% of the cases both clusters corresponded to singleton ESTs, indicating the often overlooked importance of this kind of data.

A reasonable fraction (21%) of the validated sequences represented novel human transcripts which were only partially represented by ESTs. The structure of the majority (69.2%) of these new human transcripts had not been correctly predicted by ab initio gene prediction programs and, consequently, was not annotated in the human genome. Additionally, the use of different cDNA sources in the validation process, allowed us to identify many splicing variants that were further validated by RT-PCR. As for 21% of validated sequences, none of these splicing variants had been previously identified.

We conclude that the Transcript Finishing strategy provides a convenient and unique means for delineating gene boundaries and new transcript sequences. The TF

strategy permits the characterization of new human transcripts and splicing isoforms expressed at a low level and in a restricted set of tissues and will certainly continue to contribute to the definition of the complete catalog of human genes and transcripts.

METHODS

Cell Culture

Human cell lines were obtained from the American Type Culture Collection (ATCC) and cultured as recommended (www.atcc.org). The following cell lines were used in order to generate a cDNA panel representing different tissues: A172 glioblastoma; T98G multiform glioblastoma; FaDu squamous cell carcinoma; SW480 colorectal adenocarcinoma; Skmel-25 malignant melanoma; DU145 prostate carcinoma; HeLa cervix adenocarcinoma; XP Xeroderma pigmentosum fibroblasts; ZR-75-1, MCF-7 and Hs578T breast ductal carcinoma; IM9 B transformed lymphoblasts; TT thyroid carcinoma; U937 histyocytic lymphoma; Hs1.Tes normal testis; Hs732.PL normal placenta; Hep G2 hepatocarcinoma; NCI-H1155 and H358 lung carcinoma; SCaBER urinary bladder carcinoma; SAOS 2 osteosarcoma and Tu-rim primary culture of a kidney tumor.

RNA Extraction and cDNA Synthesis

Total RNA was prepared from cultured cells seeded in four 150 mm diameter (P150) plates using the cesium chloride cushion technique (Chirgwin et al. 1979). Poly A^+ RNA was isolated from 200µg of total RNA with the PolyAttract mRNA isolation kit (Promega) and the total yield of this purification used for cDNA synthesis. For cDNA synthesis, 100-200µg of total RNA or the corresponding purified mRNA were treated

with 100 units of *DNAse* I (FPLC-pure, Amersham) and reverse transcribed using oligo(dT)12-18, random primer and *SuperScript* II (Invitrogen), following the manufacturer's instructions. The resulting cDNA was then subjected to *RNase* H treatment and distributed among the 31 validation laboratories involved in the project. The quality of the cDNA synthesis and the absence of genomic DNA contamination were evaluated for each preparation as described previously (Silva et al. 2003).

RT-PCR and Sequencing

RT-PCR were carried out in 25μL reaction mixtures containing 1μL of cDNA, 10x *Taq* DNA polymerase buffer, 200μM dNTP, 6 pmols of primers, 1.5mM $MgCl_2$ and 1 unit *Taq* DNA polymerase (GIBCO/BRL). Standard PCR conditions were: 4 min. at 94°C (initial denaturation), 40 sec. at 94°C, 40 sec. at 55°C and 1 min. at 72°C for 35 cycles and a final extension step of 10 min. at 72°C. Modifications of the standard protocol included annealing temperature, $MgCl_2$ concentration and addition of PCR enhancers such as betaine. PCR products were directly sequenced with the same primers used for RT-PCR or cloned before sequencing. Sequencing reactions were carried out using the DYEnamic™ ET terminator Cycle Sequencing Kit (Amersham Pharmacia) and separated by electrophoresis using an ABI 377 Prism Sequencer (Applied Biosystems) according to supplier's recommendations.

Transcriptome Database and Graphical Interface

BLASTN was used to identify pair-wise similarities between all known transcript sequences and the draft genome sequence deposited in release 66 (March 2001) of the EMBL database. Transcribed sequence data were extracted from several sources: (i) the human EST section of EMBL release 66; (ii) human mRNA documented in the

human section of EMBL release 66; (iii) ORESTES sequences from the LICR/FAPESP Human Cancer Genome project; (iv) human mRNAs documented in the NCBI curated RefSeq database (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). For genomic sequence, we used contigs of at least 10 kb deposited in the HUM and HTG sections. Those HTG entries that had not been fully assembled were split into individual components. Therefore, the human genome dataset used is highly redundant, but can easily be reduced to one of the available assemblies.

The transcript sequences were filtered for contaminants, and repetitive elements were masked out using the PFP software package (Paracel, Pasadena, CA). For each pair of matching transcribed and genomic sequences, local alignments were generated using sim4 (Florea et al. 1998), with parameters W=15 R=0 A=4 P=1. The output of sim4 was filtered to eliminate all alignments that did not contain at least one matching region within the genome with at least 95% identity over 30nt. The alignment coordinates and related information were uploaded into a MySQL relational database. We used the data stored in the relational database to create clusters of transcribed sequences, based on their position within individual genomic contigs. The coordinates of the putative exons on the genome sequence were used to determine membership in a cluster. If coordinates of at least one exon were common to two transcripts, then these were considered to be part of the same cluster.

The 3' tags were generated as previously described (Iseli et al. 2002). Briefly, poly(A) or poly(T) were identified from original sequence trace files and the 50 nucleotides immediately adjacent to it were recorded as a candidate tag (after obtaining the reverse complement for poly(T) tracts). Duplicate tags were eliminated as were the tags matching LINE and Alu repetitive elements, ribosomal or

mitochondrial sequences, and those containing simple repeats. Matches for the remaining tags were mapped to the genome and the 50 nucleotides found downstream of the match were also recorded. Individual tags were incorporated into the MySQL database. A graphical interface was developed in TCL/TK language in order to visualize the 3′ tags, EST alignments and related information, such as tissue origin and project source of the sequences.

By querying the Transcriptome Database, we were able to select EST clusters that do not correspond to known full-length mRNA for validation. These were at a maximum of 10 kb apart from each other and exhibited a clear splicing structure when aligned to the genome. Clusters selected for validation were visually inspected before ordering primers. All systems used in this work were developed using PERL and PHP programming languages on a Linux based server running the MySQL database management system and the Apache web server.

Cluster Selection and Primers Design

The automated primer protocol received a fixed format file containing the accession number of the genomic clones and the genomic interval where the two non-contiguous EST clusters map and where the system searched for primers. The Primer3 program (version 0.9) developed by the Whitehead Institute for Biomedical Research was used for primer design adopting the following parameters: primer size of a minimum of 17 bp, optimal 18 bp and maximum 21 bp; melting temperature of a minimum of 55^{o}C, optimal 60^{o}C and maximum 65^{o}C; and GC clamp set to 1. The output of Primer3 was processed in order to filter primers that had alternate annealing sites in the given genomic sequence. The system uses a web-based interface that

allows submission of files containing information on primer design, retrieval of primers found and the modification of default parameters for primer picking.

Sequence Analysis and Database Update

Sequences were subjected to an automated protocol to: (a) assess sequence quality, (b) trim vector sequences, (c) mask repetitive elements and (d) remove undesirable sequences such as bacterial, mitochondrial, and fungi sequences. The sequence quality was determined by Phred analysis using a trimCutOff of 0.06171 (Ewing et al. 1998; Ewing and Green 1998). Sequences with less than 100 bases were excluded. Mitochondrial, bacterial and fungi sequences were identified by BLAST searches against the GenBank entry corresponding to the human mitochondrial complete genome sequence and against a locally developed human bacterial and fungal database, respectively. Significant hits were determined by using an E value of 10^{-5} for searches against mitochondrial genome and an E value of 10^{-30} for searches against bacterial databases. Masking of repetitive elements was undertaken by using the REPEAT-MASKER (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) under default parameters. The remaining, high-quality sequences, were aligned against the original genomic clone using the BLASTN program and alignment coordinates and scores were loaded into the MySQL database on a daily basis.

Consensus Assembly

The reads corresponding to validated TFs were assembled into a contig using the PhredPhrap. The contig sequence was aligned with both EST clusters, by using the BLASTN program and alignment coordinates were used for consensus generation. A

web-based interface was developed to monitor the assembly and access the consensus sequences (http://200.18.51.201/viewconsensus/).

Characterization of Validated Transcripts

Characterization of validated transcripts was pursued using the UCSC Genome Browser (Kent et al. 2002), which is available at http://genome.ucsc.edu. This allowed determination of sequence overlap between the validated consensus sequences, known genes and gene predictions. Consensus sequences derived from the validated TFUs were aligned to the July 2003 version of the human genome sequence assembly provided by UCSC using the BLAT search tool. The annotation tracks used for comparison to already known genes were: known genes, RefSeq genes and human mRNAs from the GenBank. A validated transcript was considered a NEW gene if its alignment coordinates did not match the coordinates of any other sequence available through the known genes, RefSeq genes or human mRNA annotation tracks. For comparison to gene predictions, the following tracks were used: Fgenesh++, Geneid and GenScan predictions. The prediction of individual exons instead of the full transcript prediction was considered. A validated exon was considered as predicted if it aligned within the coordinates defined by any of the three gene prediction programs (not necessarily sharing borders) and a NEW validated transcript was considered NOT PREDICTED if all exons were not predicted by the computer programs. The consensus sequences corresponding to NEW validated transcripts were further characterized by BLASTX analysis and protein domains were determined using the Pfam and Prosite databases.

Characterization and Validation of Alternatively Splicing Forms.

The individual sequences generated during the process of validation of each TFU were aligned to the human genome assembly using the BLAT search tool, together with the final consensus sequence and representative sequences derived from both EST clusters. Alternatively, spliced isoforms were visually identified using the UCSC browser. In order to eliminate alignment artifacts caused by sequencing errors and problems in the genome assembly, we have considered as alternatively spliced forms only exons defined by conserved acceptor and donor splicing sites (GT/AG). Primers for validation of predicted alternative splicing isoforms were designed using Primer3 with default parameters. The presence of alternative isoforms was analyzed using a cDNA panel composed of 20 different normal and tumor tissues. GAPDH amplification was used as a control for integrity and quantification of the RNA used for cDNA synthesis. RT-PCR products obtained in touchdown reactions were analyzed on 1.5% agarose gels.

COMPLETE LIST OF AUTHORS

Coordination Group Ludwig Institute
Fabiana Bettoni[1], Dirce Maria Carraro[1], Lilian C. Pires[1], Raphael B. Parmigiani[1], Elisa N. Ferreira[1], Eloísa de Sá Moreira[1,30], Maria do Rosário D. de O. Latorre[2], Andrew J.G.Simpson[1], Anamaria A. Camargo[1]

Coordination Group University of São Paulo Chemistry Institute
Luciana Oliveira Cruz[3], Theri Leica Degaki[3], Fernanda Festa[3], Katlin B. Massirer[3], Mari C. Sogayar[3]

Bioinformatics groups:
Fernando Camargo Filho[4], Luiz Paulo Camargo[4], Marco A. V. Cunha[5], Sandro J. De Souza[6], Milton Faria Junior[4], Silvana Giuliatti[4], Leonardo Kopp[7], Paulo S.L. de Oliveira[7], Paulo B. Paiva[8], Anderson A. Pereira[4], Daniel G. Pinheiro[5], Renato D. Puga[4], Jorge Estefano S. de Souza[6]

Validation groups:
Dulcineia M. Albuquerque[9], Luís E. C. Andrade[10], Gilson S. Baia[11], Marcelo R.S. Briones[12], Ana M.S. Cavaleiro – Luna[13], Janete M. Cerutti[14], Fernando F. Costa[9], Eugenia Costanzi-Strauss[15], Enilza M. Espreafico[16,] Adriana C. Ferrasi[17], Emer S. Ferro[11], Maria A.H.Z. Fortes[13], Joelma R.F. Furchi[18], Daniel Giannella-Neto[13], Gustavo H. Goldman[19], Maria H.S. Goldman[20], Arthur Gruber[21], Gustavo S. Guimarães[14], Christine Hackel[22], Flavio Henrique-Silva[18], Edna T. Kimura[11], Suzana G. Leoni[9], Cláudia Macedo[23,] Bettina Malnic[24], Carina V. Manzini B.[24], Suely K. N. Marie[25], Nilce M. Martinez-Rossi[23] , Marcelo Menossi[26,27], Elisabete C. Miracca[28], Maria A. Nagai[28], Francisco G. Nobrega[29], Marina P. Nobrega[29], Sueli M. Oba-Shinjo[25], Márika K. Oliveira[16], Guilherme M. Orabona[30], Audrey Y. Otsuka[31], Maria L. Paço-Larson[16], Beatriz M.C. Paixão[5], Jose R. C. Pandolfi[32], Maria

I.M.C.Pardini[17], Maria R. Passos Bueno[30], Geraldo A. S. Passos[33], Joao B. Pesquero[34], Juliana G. Pessoa[34], Paula Rahal[35], Cláudia A. Rainho[36], Caroline P. Reis[26], Tatiana I. Ricca[12], Vanderlei Rodrigues[37], Silvia R. Rogatto[36], Camila M. Romano[21], Janaína G. Romeiro[35], Antonio Rossi[37], Renata G. Sá[37], Magaly M. Sales[17], Simone C. Sant'Anna[22], Patrícia L. Santarosa[38], Fernando Segato[23], Wilson A Silva Junior[5,23,] Ismael D.C.G. Silva[31], Neusa P.Silva[10], Andrea Soares-Costa[18,] Maria F. Sonati[39], Bryan E. Strauss[40,] Eloiza H. Tajara[35,] Sandro R. Valentini[32], Fabiola E. Villanova[31], Laura S. Ward[38], Dalila L. Zanette[5] ,

[1] Ludwig Institute for Cancer Research. Rua Prof. Antonio Prudente, 109, 4ºfloor-01509-010 SP – Brazil
[2] Department of Epidemiology – School of Public Health – University of São Paulo
[3] Instituto de Quimica, Universidade de Sao Paulo, CP 26077, Sao Paulo 05513-970, SP, Brazil.
[4] Dep. de Eng. Química e de Informática, Bioinformática Universidade de Ribeirao Preto UNAERP.
[5] Centro de Terapia Celular, Hemocentro e Departamento de Clínica Médica, Faculdade de Meidicina de Ribeirão Preto, Universidade de São Paulo.
[6] Laboratório de Biologia Computacional, Instituto Ludwig, São Paulo, SP Brazil.
[7] Laboratório de Genética e Cardiologia Molecular, Instituto do Coração (INCOR), Universidade de São Paulo, SP Brazil.
[8]Bioinformatics Laboratory, Health Informatics Department, Federal
University of São Paulo, Brazil
[9] Departamento de Clinica Medica, Hemocentro, Faculdade de Ciencias Medicas, UNICAMP
[10] Rheumatology Division, Universidade Federal de São Paulo, Rua Botucatu 740 São Paulo, SP, 04113-001
[11] Departamento de Histologia e Embriologia, Instituto de Ciencias Biomédicas, Universidade de Sao Paulo
[12] Department of Microbiology, Immunology and Parasitology, Universidade Federal de São Paulo, Rua Botucatu, 862, 3º andar. CEP 04023-062, São Paulo, S.P. Brazil.
[13] Laboratory for Cellular and Molecular Endocrinology (LIM-25/HC-FMUSP). University of Sao Paulo School of Medicine. Av. Dr. Arnaldo, 455 #4305 01246-903
[14] Laboratorio de Endocrinologia Molecular, Disciplina de Endocrinologia, Departamento de Medicina, Escola Paulista de Medicina, Universidade Federal de São Paulo.
[15] Laboratório de Transferência Gênica, ICB – USP
[16] Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos da Faculdade de Medicina de Ribeirão Preto-USP - SP
[17] Laboratório de Biologia Molecular, Hemocentro, Faculdade de Medicina, UNESP, Botucatu.
[18] Departamento de Genética e Evolução, UFSCar, Rod. Washington Luis, Km 235 São Carlos, SP CEP 13 565-905
[19] Faculdade de Ciencias Farmaceuticas de Ribeirao Preto, USP
[20] Faculdade de Filosofia, Ciencias e Letras de Ribeirao Preto, USP
[21] Depto. de Patologia, Faculdade de Medicina Veterinária e Zootecnia USP. Av. Prof. Orlando Marques de Paiva 87, São Paulo SP 05508-000
[22] Departamento de Genética Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas (UNICAMP).
[23] Departamento de Genética, Faculdade de Medicina, 14040-900, Ribeirão Preto, SP, Brasil.
[24] Departamento de Bioquimica, Instituto de Quimica, Universidade de Sao Paulo, CP 26.077, 05599-970 Sao Paulo SP, Brazil.
[25] Departamento de Neurologia, Faculdade de Medicina da Universidade de São Paulo
[26] Laboratório de Genoma Funcional, Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas -SP – Brazil
[27] Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas -SP – Brazil
[28] Departamento de Radiologia, Disciplina de Oncologia da Faculdade de Medicina da Universidade de Sao Paulo, 01246.903, Sao Paulo, SP, Brazil.
[29] Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraiba, Sao Jose dos Campos, Brazil 12244-000
[30] Departamento de Biologia, Centro de Estudos do Genoma Humano, Instituto de Biociencias, Universidade de Sao Paulo, Sao Paulo, Brazil.
[31] Molecular Gynecology Laboratory, Gynecology Department, Federal University of São Paulo, Brazil

[32] Departament of Biological Sciences School of Pharmacy - Sao Paulo State University - UNESP Araraquara, SP, 14801-902
[33] Disciplina de Genética, Faculdade de Odontologia, Universidade de São Paulo, 14040-900, Ribeirão Preto, SP, Brasil
[34] Departamento de Biofísica, Universidade Federal de São Paulo
[35] Departamento de Biologia Instituto de Biociencias, Letras e Ciencias Exatas (IBILCE) Universidade Estadual Paulista (UNESP) Sao Jose do Rio Preto, SP
[36] Departamento de Genética - Instituto de Biociências UNESP - Botucatu – SP
[37] Departamento de Bioquímica e Imunologia, Faculdade de Medicina de Ribeirão Preto da USP, Av. Bandeirantes, 3.900, 14.049-900 - Ribeirão Preto - S.P.
[38] Laboratório de Genética Molecular do Cancer - Departamento de Clínica Médica, Faculdade de Ciências Médicas da Universidade Estadual de Campinas
[39] Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas (UNICAMP)
[40] Setor de Vetores Virais, Lab. de Cardiologia Molecular, InCor, Faculdade de Medicina, USP

ACKNOWLEDGMENTS

We dedicate this work to Dr. Ricardo R. Brentani and Dr. José Fernando Perez for unconditional support and constant incentive to the Brazilian Genome Initiative. We thank Fernanda G. Barbuzano, Mário H. Bengtson, Ana P. Bogassian, Miriam S. Carmo, Christian Colin, Débora C.J. Costa, Leslie E. Ferreira, Cristiane A Ferreira, Mariana C. Frigieri, Hellen T. Fuzii, Augusto D. Luchessi, Claudia R. Madella, Adriana A. Marques, Zizi de Mendonça, Camila C.B.O. Menezes, Alessandra Splendore , Flavia I.V. Errera, Julio C. Moreira, Irenice C. Silva, Sandra R. Souza and Fabiana Granja for dedicated and expert technical assistance and/or critical discussions. We also thank Dr. Winston Hide and Dr. Helena Brentani for important comments and corrections on the manuscript and Juçara Parra for acting as the administrative coordinator of this project. The work was equally supported by the Ludwig Institute for Cancer Research and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

REFERENCES

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355**: 632-634.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., and Moreno, R.F. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651-1656.

Adams, M.D., Kerlavage, A.R., Fields, C., and Venter, J.C. 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4**: 256-267.

Bailey, L.C., Searls Jr., D.B., and Overton, G.C. 1998. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* **8**: 362-376.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**: 950-958.

Beaudoing, E. and Gautheret, D. 2001. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.* **11**: 1520-1526.

Bonaldo, M.F., Lennon, G., Soares, M.B. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6**:791-806.

Bortoluzzi, S., d'Alessi, F., and Danieli, G.A. 2000a. A computational reconstruction of the adult human heart transcriptional profile. *J. Mol. Cell Cardiol.* **32**: 1931-1938.

Bortoluzzi, S., d'Alessi, F., and Danieli, G.A. 2000b. A novel resource for the study of genes expressed in the adult human retina. I*nvest Ophthalmol. Vis. Sci.* **41**: 3305-3308.

Bortoluzzi, S., d'Alessi, F., Romualdi, C., and Danieli, G.A. 2000c. The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.* **10**: 344-349.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**: 83-86.

Camargo, A.A., de Souza, S.J., Brentani, R.R., and Simpson, A.J. 2002. Human gene discovery through experimental definition of transcribed regions of the human genome. *Curr. Opin. Chem. Biol.* **6**: 13-16.

Camargo, A.A., Samaia, H.P., Dias-Neto, E., Simao, D.F., Migotto, I.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A. et al. 2001. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. U. S. A* **98**: 12103-12108.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., Hayashizaki, Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**:1617-30.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter WJ. 1979. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry **18**: 5294-5299.

Clark, F. and Thanaraj, T.A. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* **11**: 451-464.

Clifford, R., Edmonson, M., Hu, Y., Nguyen, C., Scherpbier, T., and Buetow, K.H. 2000. Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. *Genome Res.* **10**: 1259-1265.

Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. 2001. Assessment of the total number of human transcription units. *Genomics* **77**: 71-78.

Dennis,C. 2001. Tiled arrays for gene hunting. *Nat. Rev. Genet.* **2**, 161.

Dias, N.E., Garcia, C.R., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva, W.Jr., Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., et al. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. U. S. A* **97**: 3491-3496.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402**: 489-495.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**: 186-194.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**: 232-234.

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

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**: 967-974.

Garg, K., Green, P., and Nickerson, D.A. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res.* **9**: 1087-1092.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* **8**: 524-530.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405**: 311-319.

Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C., Kelso, J.F. 2001. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res*.**11**:1848-53.

Hu, G., Modrek, B., Riise Stensland, H.M., Saarela, J., Pajukanta, P., Kustanovich, V., Peltonen, L., Nelson, S.F., and Lee, C. 2002. Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. *Pharmacogenomics. J.* **2**: 236-242.

Hudson, T.J., Colbert, A.M., Reeve, M.P., Bae, J.S., Lee, M.K., Nussbaum, R.L., Budarf, M.L., Emanuel, B.S., and Foote, S. 1994. Isolation and regional mapping of 110 chromosome 22 STSs. *Genomics* **24**: 588-592.

Huminiecki, L. and Bicknell, R. 2000. In silico cloning of novel endothelial-specific genes. *Genome Res.* **10**: 1796-1806.

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W., and Lee, C.J. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26**: 233-236.

Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. 2002. Long-range heterogeneity at the 3' ends of human mRNAs. *Genome Res.* **12**: 1068-1074.

Jiang,J. and Jacob,H.J. 1998. EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Res.* **8**: 268-275.

Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **11**: 889-900.

Kan, Z., States, D., and Gish, W. 2002. Selecting for functional alternative splices in ESTs. *Genome Res.* **1**2: 1837-1845.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor,S .P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916-919.

Katsanis, N., Worley, K.C., Gonzalez, G., Ansley, S.J., and Lupski, J.R. 2002. A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. *Proc. Natl. Acad. Sci. U. S. A* **99**: 14326-14331.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12**: 996-1006.

Kikuno, R., Nagase, T., Waki, M., and Ohara, O. 2002. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **30**: 166-168.

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

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J., et al. 2002. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12**: 493-502.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**: 239-240.

Megy, K., Audic, S., and Claverie, J.M. 2003. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. *Genome Biol.* **4**: P1.

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

Nakajima, D., Okazaki, N., Yamakawa, H., Kikuno, R., Ohara, O., and Nagase, T. 2002. Construction of expression-ready cDNA clones for KIAA genes: manual curation of 330 KIAA cDNA clones. *DNA Res.* **9**: 99-106.

Penn, S.G., Rank, D.R., Hanzel, D.K., and Barker, D.L. 2000. Mining the human genome using microarrays of open reading frames. *Nat. Genet.* **26**: 315-318.

Phillips, R.L., Ernst, R.E., Brunk, B., Ivanova, N., Mahan, M.A., Deanehan, J.K., Moore, K.A., Overton, G.C., and Lemischka, I.R. 2000. The genetic program of hematopoietic stem cells. *Science* **288**: 1635-1640.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. 1999. Mining SNPs from EST databases. *Genome Res.* **9**: 167-174.

Reymond, A., Camargo, A.A., Deutsch, S., Stevenson, B.J., Parmigiani, R.B., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., et al. 2002. Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79**: 824-832.

Roest, C.H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet.* **25**: 235-238.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409**: 922-927.

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

Sorek, R. and Safer, H.M. 2003. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31**: 1067-1074.

Strausberg, R.L., Buetow, K.H., Emmert-Buck, M.R., and Klausner, R.D. 2000. The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* **16**: 103-106.

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

Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286**: 455-457.

Tugendreich, S., Bassett Jr, D.E., McKusick, V.A., Boguski, M.S., and Hieter, P. 1994. Genes conserved in yeast and humans. *Hum. Mol. Genet.* 3: 1509-1517.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et al. 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Wang, Z., Lo, H.S., Yang, H., Gere, S., Hu, Y., Buetow, K.H., and Lee,M.P. 2003. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.* **63**: 655-657.

Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., et al. 2001. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11**: 422-435.

Williamson, A.R. 1999. The Merck Gene Index project. *Drug Discov. Today* **4**:115-122.

Wright, F.A., Lemon, W.J., Zhao, W.D., Sears, R., Zhuo, D., Wang, J.P., Yang, H.Y., Baer, T., Stredney, D., Spitzner, J. et al., 2001. A draft annotation and overview of the human genome. *Genome Biol.* **2**: eletronic citation 0025.1-0025.18.

Xie, H., Zhu, W.Y., Wasserman, A., Grebinskiy, V., Olson, A., and Mintz, L. 2002. Computational analysis of alternative splicing using EST tissue information. *Genomics* **80**: 326-330.

Xu, Q., Modrek, B., and Lee, C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*. **30**: 3754-3766.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R. et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol*. **21**: 379-386.

Yu, Y., Zhang, C., Zhou, G., Wu, S., Qu, X., Wei, H., Xing, G., Dong, C., Zhai, Y., Wan, J. et al. 2001. Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. *Genome Res.* **11**: 1392-1403.

WEBSITE REFERENCES

http://www.ncbi.nlm.nih/Unigene/, Unigene Home page
http://www.atcc.org , American Type Culture Collection Home Page
http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html, RefSeq Home page
http://ftp.genome.washington.edu/cgi-bin/RepeatMasker, Repeat Masker Program
http://genome.ucsc.edu, University of California Santa Cruz – Genome Browser

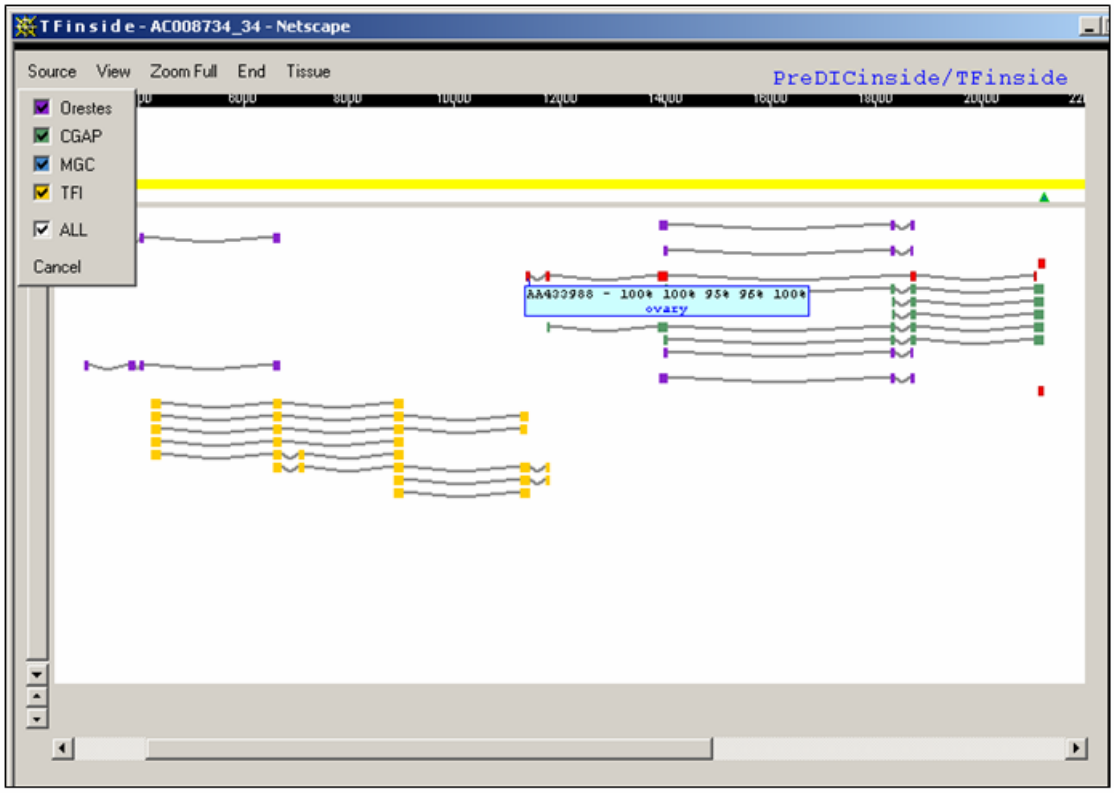


Figure 1. TFI graphical interface. The TFI graphical interface displays a region of the human genome sequence as a yellow line, with a scale in base pairs (bp). Expressed Sequence Tags (ESTs) that align with the genome sequence are shown in different colors, depending on the project of origin: ORESTES from the FAPESP/LICR Human Cancer Genome Project in purple; CGAP in green, MGC in blue and TFI in yellow, with splicing structures represented as gray lines. The interface shows an experimentally validated TFU (number 171) joining two EST clusters. The TFI interface also provides information on the tissue of origin of the transcript sequences, the percent similarity of each exon with the human genome sequence and the presence of 3´tags represented as green triangles.

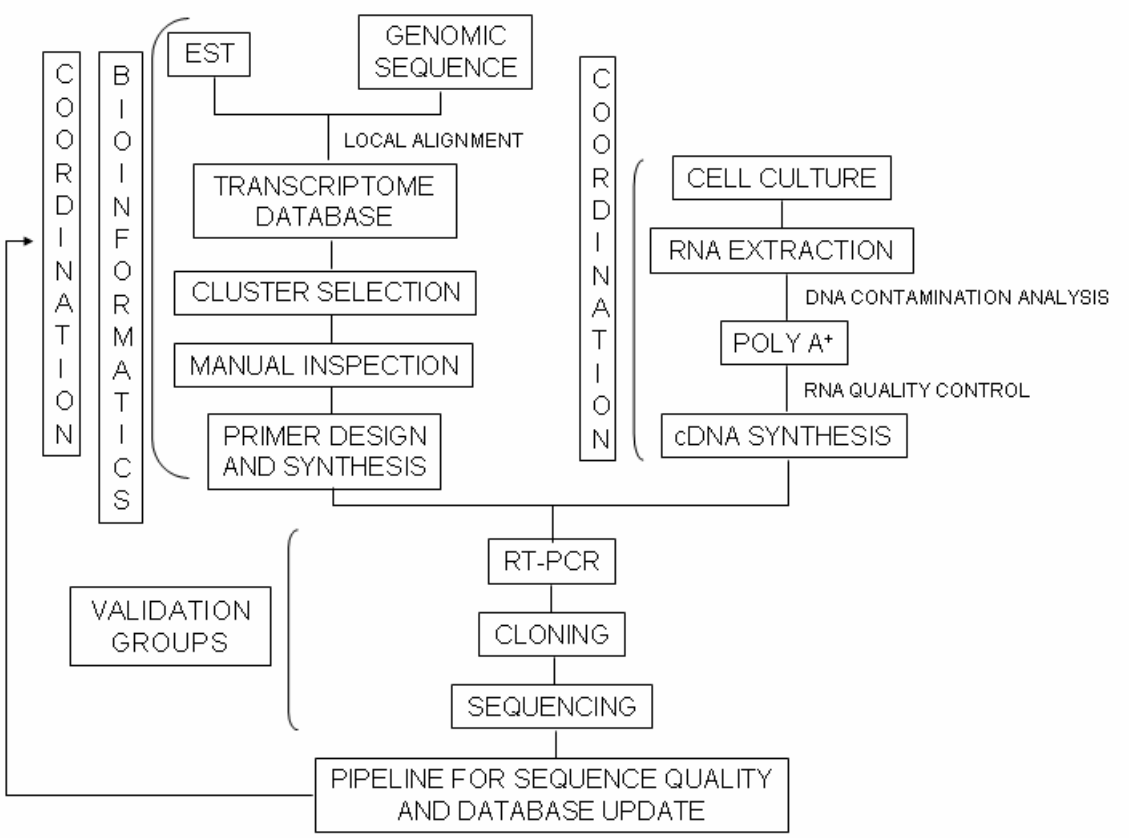


Figure2. General scheme of the TFI strategy. Schematic outline of the strategy used for computational and experimental validation of TFU sequences. Following the development of Bioinformatics tools, the generation of the Transcriptome Database and automatic cluster selection, the project tasks were divided between the Coordination and the Validation Laboratories.

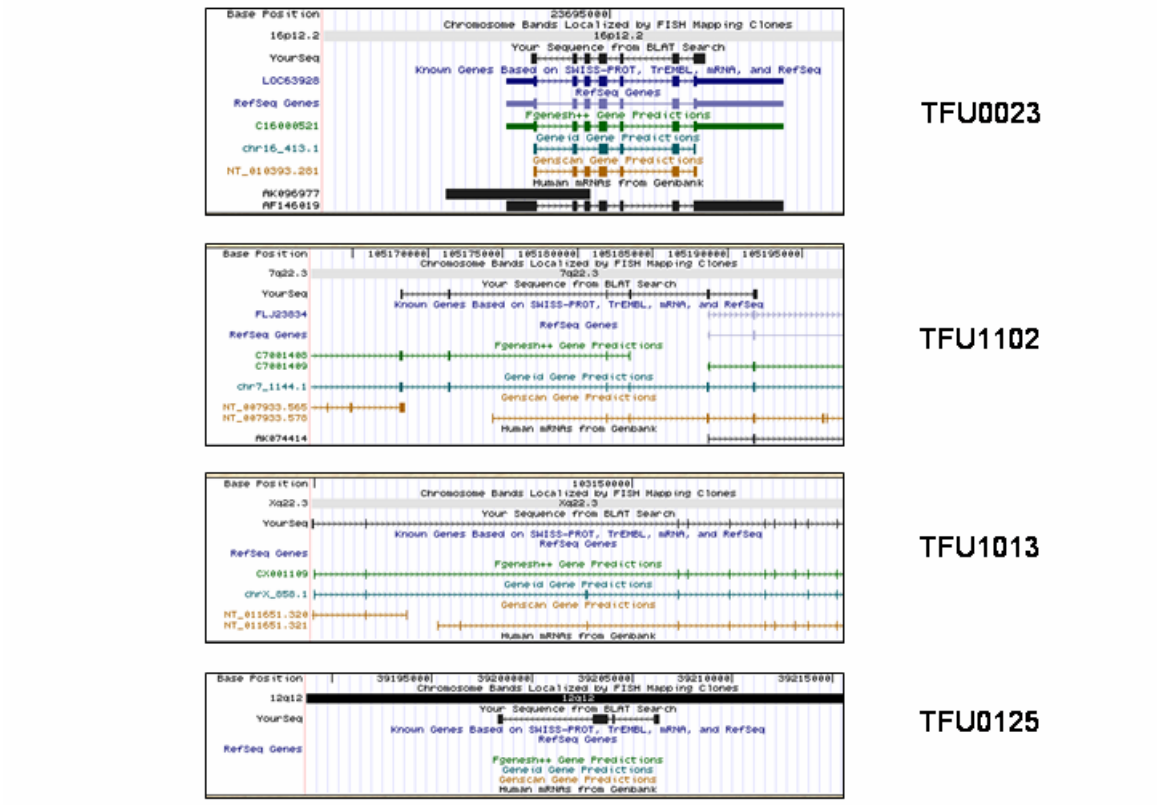


Figure 3. Characterization and annotation of validated TFUs. Alignment of four consensus sequences, derived from the validated TFUs, to the July 2003 version of the UCSC human genome sequence assembly, using the BLAT search tool. A) TFU00023 corresponds to YourSeq (black) completely overlapping with known genes based, on SWISS-PROT, TrEMEL, mRNA and RefSeq (dark blue); B) TFU01102 represents a 5' extension of a partial cDNA (FLJ23834); C) TFU01013 represents a new human transcript structure which was correctly predicted by *ab initio* gene predition transcripts, such as Fgenesh++ (green); D) TFU00125 represents a new human transcript with no predicted transcripts described by gene prediction programs.

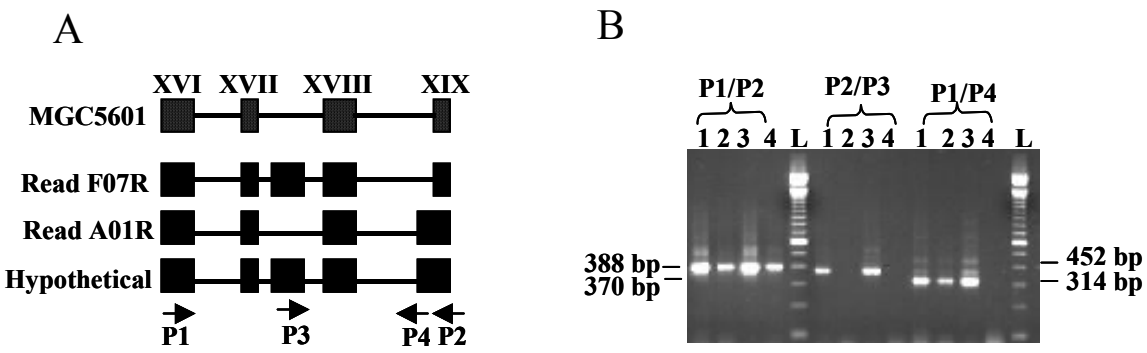


Figure 4: Experimental validation of MGC5601 gene alternative splicing isoforms. (A) Gene structure for exons XVI-XIX (boxes) of the MGC5601 gene located on chromosome 12. Introns are represented by lines. Two alternative exons are shown on TFU reads and a hypothetical combination of these two exons is also shown. Sequence F07R has an extra exon between exon XVII and XVIII. Sequence A01R has an extended exon XIX. Four primers were designed for validation tests, as indicated in the figure (P1-P4). (B) We detected all four of these alternative splicing isoforms in MGC5601. Numbers 1-4 indicate the tissues from which the cDNA was obtained (1: multiform glioblastoma; 2: glioblastoma; 3: prostate carcinoma; 4: primary kidney cell culture). The sizes of the bands obtained are indicated. L: 100 bp Ladder.

Table 1. Comparision between validated and non-validated TFUs.

	Validated TFUs (Std deviation)	**Non-Validated TFUs (Std deviation)**	**P value**
Average distance between clusters	2,609 (3,202)	3,105 (2,942)	0.008
AVG no. of ESTs in each cluster	6.10 (8.91)	5.77 (13,23)	0.010
AVG no. of distinct tissues in each cluster	3.45 (4.27)	2.85 (4.54)	0.002
Presence of a common tissue in both clusters	Yes 63 No 111	Yes 62 No 146	0.223

Table 2. Annotation of validated consensus.

Categories	**Number of Consensus Sequences**	**Percentage (%)**
Known gene	128	68.8
Extension of a known gene	19	10.2
New transcript w/ total prediction	12	6.5
New transciprt w/ partial prediction	15	8.0
New transcript w/o prediction	12	6.5
Total	186	100

Table 3. Alternative splicing forms within validated TFs.

Validated Consensus	Type of alternative splicing	Presence of conserved acceptor and donor sites	No. of alternative isoforms	No. of validated isoforms
TFU0118	Exon usage	Yes	2	1
TFU0200	Exon usage	Yes	4	4
TFU0274	Exon usage	Yes	2	2
TFU0351	Exon usage	Yes	2	2
TFU1004	Exon usage	Yes	2	1
TFU1058	Exon usage	Yes	3	0*
TFU0155	Exon usage	Yes	2	nd
TFU0238	Exon usage	Yes	2	nd
TFU0308	Exon usage	Yes	2	nd
TFU0003	Intron retention	nd	nd	nd
TFU0019	Intron retention	nd	nd	nd
TFU0035	Intron retention	nd	nd	nd
TFU0052	Intron retention	nd	nd	nd
TFU0099	Intron retention	nd	nd	nd
TFU0112	Intron retention	nd	nd	nd
TFU0125	Intron retention	nd	nd	nd
TFU0131	Intron retention	nd	nd	nd
TFU0209	Intron retention	nd	nd	nd
TFU0285	Intron retention	nd	nd	nd
TFU0371	Intron retention	nd	nd	nd
TFU0148	Exon skipping	nd	nd	nd
TFU1061	Exon skipping	nd	nd	nd

Nd, not done; * no PCR amplification.

Supplemental Material. Table1. Annotation of the consensus sequences corresponding to validated TFUs.

TFU	Consensus size in bp	Total no. of exons	Chrom Band	Status (July 2003)	Extension (exons)	Prediction	ORF/ Protein domain	Expression based on EST distibution
TF00002	1647	8	Not available	Extension (NM_021253)	3	Partial		Head and neck, heart, tonsil, germinal center b-cell, whole embryo, testis, parathyroid, pool from lung+testis+b-cell, brain ,colon, kidney, lung, placenta, eye, pool from liver+spleen
TF00003	2125	2	21q22.3	NEW	-	Not predicted	No ORF	Muscle, head_neck, pool from liver+spleen
TF00009	943	8	21q22.3	Known (AL833323	-	-		lymph, spleen
TF00011	1412	8	21q11.2	Known (NM_022136)	-	-		uterus, heart, tonsil, germinal center b-cell, pool from lung+testis+b-cell, lymph, pancreas, kidney, pooled from germ cell, pooled germinal b cell, whole blood, uterus, ear, placenta
TF00013	939	5	22q21.2	Extension (NM_153773)	1	-		testis, pool from lung+testis+b-cell, germ cell
TF00017	1018	5	17q21.31	Known (NM_133373)	-	-		stomach, head_neck, colon, senescent fibroblast, brain, germinal b cells, eye, foreskin, melanocyte
TF00018	790	10	17q21.31	NEW	-	Total	ORF=175aa	testis, pool from

								lung+testis+b-cell, lymph, tonsil, germinal center b-cell, ovary, cervix
TF00019	1053	6	17q21.32	Known (NM_017731)	-	-		prostate, colon, brain, hippocampus, kidney, pool from lung+testis+b-cell, stomach, lymph, uterus, germ cell, whole blood, subthalamic nucleus
TF00020	1368	4	7q31.1	NEW		Not predicted	No ORF	head_neck, colon, breast, adipose, white adipose tissue, tonsil, germinal center b-cell, testis, uterus, lung, brain, polled germ cell, kidney
TF00021	10115	5	17q21.2	Known (NM_152349)	-	-		pool from liver + spleen, schizophrenic brain S-11 frontal lobe, brain
TF00023	766	7	16p12.2	Known (NM_022097)	-	-		small intestine, whole embryo
TF00027	1063	9	Xq24	Known (NM_023010)	-	-		ovary, lymph, lung, heart, muscle, brain
TF00028	1125	9	Xq24	Known (NM_024528)	-	-		pool from lung+testis+b-cell, cervix, germ cell, kidney, whole embryo, uterus, brain, foreskin, melanocyte, lung, pool from liver +spleen, blood, kidney, germinal center B cell, pool from melanocyte+heart+uterus, prostate, thymus, ovary, pool from lung+testis+b-cell, testis, senescent

								fibroblast, lymph, ear, cochlea, heart
TF00032	1515	9	16p11.2	Known (NM_024048)	-	-		colon, pancreas, cervix, prostate, lung, germ cell, adipose, pool from liver+spleen, tonsil, germinal center b-cell, kidney, pool from lung+testis+b-cell, uterus, blood, lymphocyte, uterus, brain, lymph, muscle
TF00033	1364	6	11p11.2	Known (NM_032592)	-	-		stomach, pool from liver +spleen, brain, uterus
TF00035	2304	4	19p13.11	Known (AK097070)	-	-		ovary, kidney
TF00036	1653	4	16p13.3	Known (NM_032296	-	-		brain, cervix
TF00037	991	9	7q31.2	Known (NM)033427)	-	-		head_neck, colon, brain, kidney, lung
TF00038	1103	8	7q31.2	Known (NM_033427)	-	-		head_neck, colon, brain, kidney, lung
TF00039	1949	6	9p13.3	Known (NM_032634)	-	-		colon, head_neck, lung, parathyroid, uterus, brain, miningioma, pool from lung+testis+b-cell, ear, cochlea, pancreas, genitourinary tract, ovary, eye, kidney
TF00040	1141	7	19p13.3	Known (NM_032482)	-	-		breast, colon, tonsil, germinal center b-cell, brain, blood, lymphocyte, testis, lymph, senescent fibroblast
TF00041	1421	12	19p13.3	Known	-	-		breast, colon, tonsil,

				(NM_032482)				germinal center b-cell, brain, blood, lymphocyte, testis, lymph, senescent fibroblast
TF00042	1051	6	16p13.3	Known (NM_032444)_	-	-		germinal center b-cell, pool from lung+testis+b-cell, brain, blood, lymphocyte
TF00045	1174	6	7p22.1	Known (BC022378)	-	-		breast, head_neck, prostate_tumor, pool, liver+spleen, pancreas, tonsil, enriched for germinal center b-cell, lung, pool from melanocyte+heart+uterus, foreskin, melanocyte, liver
TF00047	655	7	22q12.2	Known (NM_030758)	-	-		brain, pnet
TF00050	1288	1	16q13.3	Known (NM_021195)	-	-		breast, amnion, lung, colon, germ cell, pool from lung+testis+b-cell, placenta
TF00051	1291	1	16p13.3	Known (NM_020982)	-	-		breast, amnion, lung, colon, germ cell, pool from lung+testis+b-cell, placenta
TF00053	923	8	16p13.3	Known (NM_024534)	-	-		germinal center b-cell, placenta, pancreas, muscle, total brain
TF00055	1083	8	7q34	Known (AB051505)	-	-		brain, colon, adrenal gland, kidney, uterus, pancreas, connective tissue, germ cell, whole blood, eye, retina, pool, placenta, heart,

TF00056	1368	3	7q34	Known (AB051505)	-	-		brain, colon, adrenal gland, kidney, uterus, pancreas, connective tissue, germ cell, whole blood, eye, retina, pool, placenta, heart
TF00057	938	5	21q11.2	Extension (NM_153773)	1	-		testis, pool from lung+testis+b-cell, germ cell
TF00062	1989	4	1q41	Known (AK096265)	-	-		lung, brain, epididymis, germ cell, brain, uterus, breast, pool from melanocyte+heart+uterus, kidney, heart, ovary
TF00063	1636	5	7p15.1	NEW	-	Not predicted	ORF109aa	testis, head_neck, kidney, pool from lung+testis+b-cell
TF00064	1101	5	19q13.2	Known (BC039652)	3	-		prostate, heart, uterus, lymph, t-cell, adrenal gland, breast, lung, pool from melanocyte+heart+uterus, kidney, stomach, brain, colon, cervix, muscle
TF00065	1187	2	19q13.2	Known (AK091041)	-	-		prostate, heart, uterus, lymph, t-cell, adrenal gland, breast, lung, pool from melanocyte+heart+uterus, kidney, stomach, brain, colon, cervix, muscle
TF00068	1089	8	19q13.32	Extension (BC039061)	3	-		eye, retina, brain, pituitary, pancreas, uterus, kidney, thymus, germ cell, cervix, skin,

								heart
TF00070	956	4	17p12	NEW	-	Not predicted	No ORF	Testis
TF00076	1482	2	22q13.33	Known (AK096019)	-	-		breast, brain, lung
TF00078	1173	11	22q13.33	Known (NM_020461)	-	-		kidney, uterus, whole embryo, prostate, pancreas, breast, esophagus, germ cell, brain, lung, testis, endometrium, lymph, ovary, pool from melanocyte+heart+uterus, cerebellum, synovial membrane, denis_drash, stomach, colon, head_neck
TF00079	1679	4	22q13.33	NEW	-	Partial	No ORF	brain, testis, lung, colon, heart, prostate, kidney, pool from lung+testis+b-cell, pool from melanocyte+heart+uterus
TF00080	2236	5	22q12.1	Known (AK098833)	-	-		pool from lung+testis+b-cell, brain, lung, lymph, t-cell, germ cell, pool from liver+spleen, head_neck
TF00084	2364	2	22q13.2	Known (NM_005297)	-	-		germ cell, testis, colon, brain, pancreas
TF00085	945	7	22q13.1	NEW	-	Partial	ORF=175aa	heart, pool from lung+testis+b-cell, pool from melanocyte+heart+uterus, kidney, testis
TF00086	700	4	22q13.1	Known (NM_004900)	-	-		colon, lymph, testis, pool from liver+spleen, uterus,

								prostate, lung, brain, germ cell, marrow, parathyroid, heart, pool from melanocyte+heart+uterus, tonsil, germinal center b-cell, pool from lung+testis+b-cell, bone, uterus, epithelium, ear, cochlea, head_neck
TF00097	1019	3	14q22.1	Known (NM_021818)	-	-		breast, testis, whole embryo, germ cell, pool from lung+testis+b-cell, prostate, pool from melanocyte+heart+uterus, kidney, pool from liver+spleen, cns, multiple sclerosis lesions
TF00099	1932	1	1p35.1	Known (AK097707)	-	-		testis, pool from lung+testis+b-cell, germ cell, brain, colon
TF00100	615	6	1p21.2	Known (NM_133496)	-	-		breast
TF00101	1257	10	1p36.32	Known (AK122589)	6	-		eukopheresis, eye
TF00102	1487	8	1q22	Known (BC033549)	-	-		colon, tonsil, germinal center b-cell, blood, lymphocyte, testis
TF00103	698	3	2p21	NEW	-	Not predicted	ORF=102aa	lung, placenta
TF00106	995	7	20p11.23	Known (NM_016652)	-	-		breast, head_neck, pancreas, heart
TF00107	1426	8	10q22.2	Extension (AB020720)	4	-		breast, brain, uterus, lymph, lung
TF00108	1369	4	2q37.3	Known	-	-		colon, prostate, lung, pool

				(NM_006037)				from lung+testis+b-cell, ovary, testis, amygdala, kidney, uterus
TF00110	2721	2	2p24.1	NEW	-	Not predicted	No ORF	testis, pool from lung+testis+b-cell
TF00112	1439	2	2p15	Known (AK075484)	-	-		breast, testis
TF00116	1922	4	17p12	NEW	-	Partial	ORF=128aa	pool from lung+testis+b-cell
TF00117	745	5	11p15.2	Known (BC033239)	-	-		heart, placenta, thyroid gland, prostate, colon, testis, kidney, lung, uterus, germ cell, brain, pool from liver+spleen
TF00118	649	5	11p15.2	Known (BC033239)	-	-		heart, placenta, thyroid gland, prostate, colon, testis, kidney, lung, uterus, germ cell, brain, pool from liver+spleen
TF00119	1386	2	11p15.2	Known (BC033239)	-	-		heart, placenta, thyroid gland, prostate, colon, testis, kidney, lung, uterus, germ cell, brain, pool from liver+spleen
TF00120	1018	7	Xq26.3	Known (AK096544)	-	-		lung, ovary, pool from liver+spleen, blood, brain, heart, breast
TF00122	1378	4	11p13	NEW	-	Partial	No ORF	testis, germ cell, kidney, pool from lung+testis+b-cell
TF00124	855	8	12p13.31	Known (AL832139)	-	-		head_neck, schizophrenic brain S-11 frontal lobe
TF00125	1308	4	12q12	NEW	-	Not predicted	No ORF	testis, pool from lung+testis+b-cell
TF00126	1101	5	19q13.2	Known	-	-		prostate, adrenal gland,

				(BC039652)				muscle
TF00128	1001	6	12q13.2	Known (NM_173596)	-	-		uterus, kidney, colon, pool from liver+spleen
TF00129	1831	7	12q13.2	Known (AK094340)	-	-		breast, uterus, lung, prostate, germ cell, pool from liver+spleen, multiple sclerosis lesions, brain
TF00131	906	6	14q24.1	Extension (AK090706)	1	-		placenta
TF00134	1246	10	14q23.2	Known (NM_015180)	-	-		prostate, head_neck, placenta, tonsil, germinal center b-cell, cervix
TF00140	943	5	16q24.1	Known (NM_139174)	-	-		testis, pool from lung+testis+b-cell, brain, placenta
TF00146	1345	4	14q32.31	NEW	-	Total	ORF=430aa	breast, colon
TF00147	872	6	17q21.31	Known (BC026187)	-	-		pool from lung+testis+b-cell, testis, colon, germ cell
TF00148	654	6	17q21.31	Known (NM_032387)	-	-		breast, prostate, pool from liver+spleen, kidney, colon, lung, testis, germ cell
TF00155	827	6	19q13.41	NEW	-	Partial	ORF=210aa IG-like domain profile PS50835 Evalue=0.00021	lung, prostate
TF00156	1182	8	19p13.12	Known (NM_145046)	-	-		Pool from fetal lung + testis + Bcell
TF00157	888	8	19p13.12	Known AK075541	-	-		head_neck, stomach, marrow, testis, colon, lung, genitourinary tract, brain, muscle

TF00162	1929	7	19p13.3	Known NM_079834	-	-		prostate, testis, pancreas, ovary, kidney, heart, placenta, brain, miningioma, lung, pool from lung+testis+b-cell, stomach, bone, trabecular bone cells, hip, skin, uterus, liver, leukopheresis, colon, breast, foreskin, melanocyte, cns, multiple sclerosis lesions
TF00165	948	2	19q13.43	Known BC033149	-	-		pancreas, brain
TF00168	1010	4	2p25.1	NEW	-	Not predicted	No ORF	testis, germinal center b-cell, kidney, germ cell
TF00171	1328	12	19p13.2	Known NM_024690	-	-		breast, stomach, ovary, uterus, lung
TF00175	1459	3	17p13.1	NEW	-	Not predicted	ORF=161aa	testis, pool from lung+testis+b-cell, lymph,
TF00178	1334	5	19q13.2	Known BC045649	-	-		uterus
TF00179	917	7	17q24.2	Known NM_080284	-	-		retina, pool from liver + spleen
TF00181	1117	12	17q24.3	Known NM_080282	-	-		prostate, adrenal gland, germinal center b-cell
TF00182	1018	9	17q24.3	Known NM_018672	-	-		brain
TF00187	1110	7	19q13.43	Extension NM_152478	4	-		heart, pool from melanocyte+heart+uterus, whole embryo, pool from lung+testis+b-cell, uterus, brain, ovary
TF00189	1093	6	7p21.2	Known BC034288	-	-		testis

TF00190	1355	4	7q22.1	Extension NM_133457	1	-		whole embryo, germ cell, kidney, brain
TF00192	1205	8	7q31.1	Known NM_024814	-	-		head_neck, colon, breast, adipose, white adipose tissue, tonsil, germinal center b-cell, testis, uterus, lung, brain, germ cell, kidney
TF00194	1410	10	19q13.11	Extension AL833713	4	-		germinal center b-cell, uterus
TF00195	1549	5	17p12	Known NM_153604	-	-		testis
TF00198	932	5	19q13.31	NEW	-	Partial	ORF=163aa	pool from liver+spleen
TF00200	993	6	12q24.12	Known NM_025247	-	-		breast, head_neck, brain, tonsil, germinal center b-cell, germ cell, colon, heart
TF00202	710	4	Xq22.2	Known AK091221	-	-		germ cell, germinal center b-cell, brain, breast
TF00206	1010	7	19p13.3	NEW	-	Total	ORF=328aa Leucine-rich region profile PS50319 Evalue 2.1e-14	breast, colon, tonsil, germinal center b-cell, pool from lung+testis+b-cell, thymus, lymph
TF00209	1451	6	4q25	Known NM_025144	-	-		head_neck, uterus, colon, prostate
TF00212	1173	6	Xp22.22	Known NM_152581	-	-		pool from melanocyte+heart+uterus, pool from liver+spleen, foreskin, melanocyte, brain, prostate, uterus, testis, tonsil, germinal center b-cell, pool from lung+testis+b-cell, whole embryo, placenta, colon,

								cervix, eye, retina, connective tissue, stomach, skin, ovary
TF00216	1043	6	12q24.13	Extension AB014514	2	-		prostate, head_neck, brain, uterus, pool from liver +speen
TF00217	1140	6	12q24.13	Known AB014514	-	-		prostate, head_neck, brain, uterus, pool from liver +spleen
TF00220	1284	8	16p13.3	Known NM_152459	-	-		lung, pool from melanocyte+heart+uterus, pool from lung+testis+b-cell, kidney, thymus, uterus, pancreas, esophagus, cns, multiple sclerosis lesions
TF00223	1245	5	21q22.13	NEW	-	Not predicted	No ORF	hip, colon
TF00224	735	7	21q22.3	Known NM_080860	-	-		ovary, thyroid gland, pool from lung+testis+b-cell, pool from melanocyte+heart+uterus, uterus, kidney, colon, lung, brain
TF00227	1056	5	21q22.11	Known AK096601	-	-		lymph, t-cell, prostate, pool from liver+spleen, lung, kidney, testis
TF00230	1533	12	19q13.12	Extension AK090617	2	-		testis, lymph
TF00232	1196	10	19q13.11	NEW	-	Partial	ORF=142aa	tonsil, germinal center b-cell, uterus
TF00236	1172	7	16p12.2	Known AB075850	-	-		head_neck, eye, retina, pool from liver+spleen, testis, umbilical cord, endothelium, cervix,

								tonsil, germinal center b-cell, whole embryo, colon, Ewing's sarcoma, kidney, pool from lung+testis+b-cell, placenta, brain, prostate, lung, kidney, pancreas, amygdala, germ cell, leukopheresis, muscle, leg skeletal muscle, foreskin, melanocyte, heart
TF00238	944	7	7q11.23	NEW	-	Partial	No ORF	prostate, tonsil, germinal center b-cell, lung, pool from liver+spleen, brain
TF00239	1136	12	7q11.23	Extension AL834358	1	-		colon, tonsil, germinal center b-cell, lung, pool, liver+spleen, brain
TF00241	1919	5	22q12.1	Known AK098833	-	-		pool from lung+testis+b-cell, brain, lung, lymph, t-cell, germ cell, pool from liver+spleen
TF00245	1780	4	19p13.12	Known AK091183	-	-		ovary, lung, pool from melanocyte+heart+uterus, tonsil, germinal center b-cell, brain, germ cell, pool from lung+testis+b-cell, prostate
TF00250	1411	13	19q13.12	Extension NM_173636	10	-		testis
TF00251	1521	10	7q2111	Known NM_152754	-	-		amnion, ovary, whole embryo, spleen
TF00253	1142	8	12q24.13	NEW	-	Partial	ORF=369aa	stomach, pool from liver+spleen, tonsil, germinal center b-cell, marrow, parathyroid

TF00263	954	4	2q33.2	NEW	-	Partial	ORF=203aa Serine-rich region profile PS50324 Evalue=0.21	breast, lung
TF00267	781	6	1p13.3	NEW	-	Partial	ORF=189aa IG-like domain profile PS50835 Evalue=0.0021	brain, ovary, lung, stomach
TF00274	574	6	16	Known AK094783	-	-		marrow, brain, kidney, lung, uterus, pool from liver + spleen
TF00280	1302	7	10q24.32	Extension NM_024789	3	-		germinal center b-cell, larynx, leukopheresis, head_neck, brain
TF00283	1924	12	14q23.2	Known NM_015180	-	-		breast, lung, head_neck, testis, uterus
TF00285	2264	11	19p13.2	Known BC037565	-	-		umbilical cord, endothelium, tonsil, germinal center b-cell, lymph, t-cell, colon, brain, cerebellum, whole embryo, kidney, pool from melanocyte+heart+uterus, breast, pool from liver+spleen, pineal gland, bone, placenta, germ cell, pool from lung+testis+b-cell, pancreas, uterus, germ cell, amygdala, thymus, uterus, lung, lymph, leukopheresis, eye, muscle, skin
TF00288	932	9	12q14.1	Known	-	-		breast

				AB017814				
TF00289	2342	6	17p13.1	NEW	-	Total	ORF=125aa	eye, pool from lung+testis+b-cell, testis, uterus, heart
TF00292	880	9	15q11.2	NEW	-	Total	ORF=207aa	testis, prostate, head_neck
TF00294	1291	10	1p32.3	NEW	-	Total	ORF=230aa	head_neck, marrow, brain
TF00295	1636	14	1p32.3	NEW	-	Partial	ORF=198aa Serine-rich region profile PS50324 Evalue=0.00021	head_neck, marrow, brain
TF00297	1235	12	1p32.3	NEW	-	Partial	ORF=398aa	colon, Ewing's sarcoma
TF00305	919	7	10q23.33	Known NM_022451	-	-		tonsil, germinal center b-cell, head_neck
TF00309	1366	12	16q22.1	Known AK074773	-	-		bone, kidney, thyroid, testis, germ cell, pool from lung+testis+b-cell, heart
TF00312	923	6	16p12.2	Known BC038400	-	-		lung, multiple sclerosis lesions
TF00313	1383	6	20q13.13	Known NM_080829	-	-		colon, kidney, testis
TF00314	739	4	15q15.3	Known AL832683	-	-		breast, head_neck, lymph, cervix, brain
TF00315	961	5	9q21.11	Known AL832333	-	-		amnion, adrenal gland, colon, testis
TF00318	1296	5	14q32.13	NEW	-	Total	ORF=418aa Serpin (serine protease inhibitor) IPR000215 Evalue=1.7e-158	endometrium, liver, uterus, germ cell
TF00327	1090	12	5p15.33	Known	-	-		stomach, lung, cervix, eye

				NM_024830				
TF00350	2098	5	9q12	NEW	-	Partial	ORF=130aa	prostate, colon, testis, lung, pool from lung+testis+b-cell, stomach, prostate, cns, multiple sclerosis lesions
TF00351	1059	6	3p21.31	NEW	-	Total	ORF=330aa	brain
TF00355	777	4	5q14.3	Known NM_153354	-	-		prostate, germinal center b-cell, kidney, adrenal gland, lung, pool from lung+testis+b-cell, uterus, pancreas, cns, multiple sclerosis lesions, placenta
TF00359	1008	9	12p13.31	Known AL832139	-	-		head_neck, colon
TF00363	944	4	22q11.21	Known NM_153334	-	-		Pool from lung+testis+b-cell, pool from melanocyte+heart+uterus, blood, lymphocyte, senescent fibroblast, kidney
TF00364	912	5	17q23.3	Known NM_153335	-	-		head_neck, colon, pool from lung+testis+b-cell, blood, lymphocyte, b-cell, chronic lymphotic leukemia, brain, lymph, pool from liver+spleen, endothelial
TF00371	655	4	19q13.2	NEW	-	Total	ORF=130aa	pancreas, Ewing's sarcoma, b-cell, chronic lymphotic leukemia, uterus
TF00379	457	4	17p13.1	Known AB076580	-	-		lung
TF00380	949	6	Xq23	Known	-	-		colon, ovary

				AF286598				
TF00393	811	8	9p24.3	Known AK093572	-	-		brain, head_neck, tonsil, germinal center b-cell, pool from melanocyte+heart+uterus, pool from liver+spleen
TF00394	1010	9	1q42.3	Known NM_152490	-	-		breast, uterus, ovary, testis, bone, trabecular bone cells, kidney
TF00395	1986	6	6p21.32	Known AL713634	-	-		brain, head_neck
TF00396	616	4	12q21.1	NEW	-	Total	ORF=116aa	testis, breast, stomach
TF00398	1216	5	1p36.12	Known AB007947	-	-		head_neck, stomach, colon, marrow, pool from liver+spleen, heart, tonsil, germinal center b-cell, lymph, t-cell, kidney, parathyroid, pool from lung+testis+b-cell, brain, lung, prostate, ovary, germ cell, pancreas, genitourinary tract, placenta, muscle, pectoral muscle, foreskin, melanocyte, uterus
TF00404	1124	7	4q34.2	Known BC038536	-	-		testis
TF00411	1441	3	11p13	Extension NM_024081	1	-		colon, uterus, prostate, heart, lung, placenta
TF00501	1013	8	14q23.2	Known (NM_015180)	-	-		prostate, head_neck, placenta, tonsil germinal center b-cell, cervix
TF00502	1612	2	16p12.2	Known (NM_145865)	-	-		colon
TF00513	998	5	15q21.3	NEW	-	Not	ORF=126aa	testis

						predicted		
TF00517	1411	3	16p13.3	Known (NM_032444)	-	-		brain, pool from melanocyte+heart+uterus, peripheral nervous system, dorsal root ganglia, kidney, pool from lung+testis+b-cell, germ cell
TF00518	920	8	2q33.2	Extension (AK096293)	3	-		stomach, lung, ovary
TF00519	536	3	8p21.3	Known (AK092034)	-	-		endometrium, germ cell, colon, uterus, ovary, head_neck, testis, brain, pool from lung+testis+b-cell, lung, blood, lymphocyte, kidney, prostate, pool from liver+spleen
TF00523	1058	6	16p13.3	Known (AL833717)	-	-		ovary, pool from melanocyte+heart+uterus, esophagus, germ cell, uterus, breast, lung, stomach, colon, pancreas
TF00525	974	4	19q13.43	Known (BC033149)	-	-		pancreas, brain
TF01004	943	6	Xp11.22	Known (BC036767)	-	-		germ cell, pool from lung+testis+b-cell, uterus, testis, kidney, head_neck
TF01007	1128	6	16p13.3	Known (NM_145294)	-	-		tonsil, germinal center b-cell, ovary, parathyroid, thymus, testis, pool from lung+testis+b-cell, lung, brain, colon, pool from liver+spleen, uterus, pancreas, lymph, kidney,

								heart
TF01008	1032	6	16p13.3	Known (NM_153239)	-	-		tonsil, germinal center b-cell, pool from melanocyte+heart+uterus, kidney, colon, brain, pool from lung+testis+b-cell, uterus, thymus, prostate, lymph, lung, breast
TF01009	1287	4	22q13.1	Known (BC031099)	-	-		germ cell, prostate, lung, pool from lung+testis+b-cell, colon, germ cell, kidney, head_neck, breast
TF01013	1282	11	Xq22.3	NEW	-	Total	ORF=122aa Protein Kinase Domain Evalue=2.1e-19	kidney
TF01016	686	1	6p24.1	Known (BC001646)	-	-		eye, retina, kidney, colon, prostate, lymph, heart, liver, lung
TF01018	1079	6	Xp11.3	Known (NM_032591)	-	-		tonsil, germinal center b-cell, blood, lymphocyte
TF01024	739	4	21q22.3	Known (BC038504)	-	-		breast, brain, head_neck, lymph
TF01034	1054	7	21q11.2	Known (BC036510)	-	-		testis, pool from lung+testis+b-cell, germ cell
TF01036	484	4	6p22.1	Known (NM_032507)	-	-		lung, senescent fibroblast
TF01052	1085	8	2p22.3	Known (NM_016252)	-	-		denis_drash, colon
TF01054	1229	6	2p22.3	Known (NM_016252)	-	-		head_neck, breast, brain
TF01057	1204	9	6p22.3	Extension (NM_153042)	3	-		colon, breast

TF01058	863	7	20q13.2	NEW	-	Not predicted	No ORF	testis
TF01061	1175	9	1p36.33	Known (NM_023018)	-	-		breast, prostate, head_neck, uterus, brain, ovary, colon, blood, lymphocyte, lung, leukopheresis, parathyroid, heart
TF01074	769	6	10q23.1	NEW	-	Partial	No ORF	testis
TF01081	1331	5	9q34.3	Known (AK092639)	-	-		lymph
TF01083	968	6	4q21.22	Known (AK091412)	-	-		embryo, pool from lung+testis+b-cell
TF01087	1875	2	5q14.1	Known (NM_152405?)	-	-		Placenta, uterus
TF01089	727	6	6q25.3	Known (NM_032861)	-	-		uterus, pool from lung+testis+b-cell, muscle
TF01092	1201	10	16q22.1	AL832446	-	-		tonsil, germinal center b-cell, pool from lung+testis+b-cell, blood, lymphocyte, thymus, germ cell, cervix, colon, lymph, ung, brain
TF01100	1539	7	2p13.1	Known (NM_133637)	-	-		colon, head_neck, pool from liver+ spleen
TF01102	960	6	7q22.3	Extension (NM_152750)	4	-		lung
TF01105	1090	8	17q23.2	Known (NM_022070)	-	-		head_neck, germinal center b-cell, colon, lung
TF01112	1158	7	12q23.3	NEW	-	Total	ORF=355aa Ankirin repeat region circular profile Evalue=2.1e-18	germ cell
TF01125	884	7	11q13.1	Extension	4	-		stomach, germinal center

				(AF001543)				b-cell, uterus, placenta
TF01130	1078	3	2q32.1	Known (AL832632)	-	-		uterus, lung, pool from lung+testis+b-cell, kidney, germ cell
TF01132	707	5	14q21.2	Known (BC036056)	-	-		testis
TF01140	1488	7	1q43	Known (AK095692)	-	-		testis, liver, colon
TF01145	754	4	7q11.23	Known (BC022886)	-	-		colon, head_neck, marrow, testis, placenta, breast, uterus, germinal center b-cell, pool from lung+testis+b-cell

ANEXO 2

TF	Consenso (pb)	N° total de exons	Localização Cromossômica	Status (Abril - 2003)	Extensão (n°exons)	Predição	ORF/ domínio proteico
TF00002	1647	8	Não disponível	Extensão (NM_021253)	3	Parcial	
TF00003	2125	2	21q22.3	NOVO	-	Não predito	Não contém ORF
TF00009	943	8	21q22.3	Conhecido (AL833323	-	-	
TF00011	1412	8	21q11.2	Conhecido (NM_022136)	-	-	
TF00013	939	5	22q21.2	Extensão (NM_153773)	1	-	
TF00017	1018	5	17q21.31	Conhecido (NM_133373)	-	-	
TF00018	790	10	17q21.31	NOVO	-	Total	ORF=175aa
TF00019	1053	6	17q21.32	Conhecido (NM_017731)	-	-	
TF00020	1368	4	7q31.1	NOVO		Não predito	Não contém ORF
TF00021	10115	5	17q21.2	Conhecido (NM_152349)	-	-	
TF00023	766	7	16p12.2	Conhecido (NM_022097)	-	-	
TF00027	1063	9	Xq24	Conhecido (NM_023010)	-	-	
TF00028	1125	9	Xq24	Conhecido (NM_024528)	-	-	
TF00032	1515	9	16p11.2	Conhecido (NM_024048)	-	-	
TF00033	1364	6	11p11.2	Conhecido (NM_032592)	-	-	
TF00035	2304	4	19p13.11	Conhecido (AK097070)	-	-	
TF00036	1653	4	16p13.3	Conhecido (NM_032296	-	-	
TF00037	991	9	7q31.2	Conhecido (NM)033427)	-	-	
TF00038	1103	8	7q31.2	Conhecido (NM_033427)	-	-	
TF00039	1949	6	9p13.3	Conhecido (NM_032634)	-	-	
TF00040	1141	7	19p13.3	Conhecido (NM_032482)	-	-	
TF00041	1421	12	19p13.3	Conhecido (NM_032482)	-	-	
TF00042	1051	6	16p13.3	Conhecido (NM_032444)	-	-	

TF00045	1174	6	7p22.1	Conhecido (BC022378)	-	-	
TF00047	655	7	22q12.2	Conhecido (NM_030758)	-	-	
TF00050	1288	1	16q13.3	Conhecido (NM_021195)	-	-	
TF00051	1291	1	16p13.3	Conhecido (NM_020982)	-	-	
TF00053	923	8	16p13.3	Conhecido (NM_024534)	-	-	
TF00055	1083	8	7q34	Conhecido (AB051505)	-	-	
TF00056	1368	3	7q34	Conhecido (AB051505)	-	-	
TF00057	938	5	21q11.2	Extensão (NM_153773)	1	-	
TF00062	1989	4	1q41	Conhecido (AK096265)	-	-	
TF00063	1636	5	7p15.1	NOVO	-	Não predito	ORF109aa
TF00064	1101	5	19q13.2	Conhecido (BC039652)	3	-	
TF00065	1187	2	19q13.2	Conhecido (AK091041)	-	-	
TF00068	1089	8	19q13.32	Extensão (BC039061)	3	-	
TF00070	956	4	17p12	NOVO	-	Não predito	Não contém ORF
TF00076	1482	2	22q13.33	Conhecido (AK096019)	-	-	
TF00078	1173	11	22q13.33	Conhecido (NM_020461)	-	-	
TF00079	1679	4	22q13.33	NOVO	-	Parcial	Não contém ORF
TF00080	2236	5	22q12.1	Conhecido (AK098833)	-	-	
TF00084	2364	2	22q13.2	Conhecido (NM_005297)	-	-	
TF00085	945	7	22q13.1	NOVO	-	Parcial	ORF=175aa
TF00086	700	4	22q13.1	Conhecido (NM_004900)	-	-	
TF00097	1019	3	14q22.1	Conhecido (NM_021818)	-	-	
TF00099	1932	1	1p35.1	Conhecido (AK097707)	-	-	
TF00100	615	6	1p21.2	Conhecido (NM_133496)	-	-	

TF00101	1257	10	1p36.32	Conhecido (AK122589)	6	-	
TF00102	1487	8	1q22	Conhecido (BC033549)	-	-	
TF00103	698	3	2p21	NOVO	-	Não predito	ORF=102aa
TF00106	995	7	20p11.23	Conhecido (NM_016652)	-	-	
TF00107	1426	8	10q22.2	Extensão (AB020720)	4	-	
TF00108	1369	4	2q37.3	Conhecido (NM_006037)	-	-	
TF00110	2721	2	2p24.1	NOVO	-	Não predito	Não contém ORF
TF00112	1439	2	2p15	Conhecido (AK075484)	-	-	
TF00116	1922	4	17p12	NOVO	-	Parcial	ORF=128aa
TF00117	745	5	11p15.2	Conhecido (BC033239)	-	-	
TF00118	649	5	11p15.2	Conhecido (BC033239)	-	-	
TF00119	1386	2	11p15.2	Conhecido (BC033239)	-	-	
TF00120	1018	7	Xq26.3	Conhecido (AK096544)	-	-	
TF00122	1378	4	11p13	NOVO	-	Parcial	Não contém ORF
TF00124	855	8	12p13.31	Conhecido (AL832139)	-	-	
TF00125	1308	4	12q12	NOVO	-	Não predito	Não contém ORF
TF00126	1101	5	19q13.2	Conhecido (BC039652)	-	-	
TF00128	1001	6	12q13.2	Conhecido (NM_173596)	-	-	
TF00129	1831	7	12q13.2	Conhecido (AK094340)	-	-	
TF00131	906	6	14q24.1	Extensão (AK090706)	1	-	
TF00134	1246	10	14q23.2	Conhecido (NM_015180)	-	-	
TF00140	943	5	16q24.1	Conhecido (NM_139174)	-	-	
TF00146	1345	4	14q32.31	NOVO	-	Total	ORF=430aa

TF00147	872	6	17q21.31	Conhecido (BC026187)	-	-	
TF00148	654	6	17q21.31	Conhecido (NM_032387)	-	-	
TF00155	827	6	19q13.41	NOVO	-	Parcial	ORF=210aa "IG-like domain profile" PS50835 E value=0,00021
TF00156	1182	8	19p13.12	Conhecido (NM_145046)	-	-	
TF00157	888	8	19p13.12	Conhecido AK075541	-	-	
TF00162	1929	7	19p13.3	Conhecido NM_079834	-	-	
TF00165	948	2	19q13.43	Conhecido BC033149	-	-	
TF00168	1010	4	2p25.1	NOVO	-	Não predito	Não contém ORF
TF00171	1328	12	19p13.2	Conhecido NM_024690	-	-	
TF00175	1459	3	17p13.1	NOVO	-	Não predito	ORF=161aa
TF00178	1334	5	19q13.2	Conhecido BC045649	-	-	
TF00179	917	7	17q24.2	Conhecido NM_080284	-	-	
TF00181	1117	12	17q24.3	Conhecido NM_080282	-	-	
TF00182	1018	9	17q24.3	Conhecido NM_018672	-	-	
TF00187	1110	7	19q13.43	Extensão NM_152478	4	-	
TF00189	1093	6	7p21.2	Conhecido BC034288	-	-	
TF00190	1355	4	7q22.1	Extensão NM_133457	1	-	
TF00192	1205	8	7q31.1	Conhecido NM_024814	-	-	
TF00194	1410	10	19q13.11	Extensão AL833713	4	-	
TF00195	1549	5	17p12	Conhecido NM_153604	-	-	
TF00198	932	5	19q13.31	NOVO	-	Parcial	ORF=163aa

TF00200	993	6	12q24.12	Conhecido NM_025247	-	-	
TF00202	710	4	Xq22.2	Conhecido AK091221	-	-	
TF00206	1010	7	19p13.3	NOVO	-	Total	ORF=328aa "Leucine-rich region profile" PS50319 E value=2,1e^{-14}
TF00209	1451	6	4q25	Conhecido NM_025144	-	-	
TF00212	1173	6	Xp22.22	Conhecido NM_152581	-	-	
TF00216	1043	6	12q24.13	Extensão AB014514	2	-	
TF00217	1140	6	12q24.13	Conhecido AB014514	-	-	
TF00220	1284	8	16p13.3	Conhecido NM_152459	-	-	
TF00223	1245	5	21q22.13	NOVO	-	Não predito	Não contém ORF
TF00224	735	7	21q22.3	Conhecido NM_080860	-	-	
TF00227	1056	5	21q22.11	Conhecido AK096601	-	-	
TF00230	1533	12	19q13.12	Extensão AK090617	2	-	
TF00232	1196	10	19q13.11	NOVO	-	Parcial	ORF=142aa
TF00236	1172	7	16p12.2	Conhecido AB075850	-	-	
TF00238	944	7	7q11.23	NOVO	-	Parcial	Não contém ORF
TF00239	1136	12	7q11.23	Extensão AL834358	1	-	
TF00241	1919	5	22q12.1	Conhecido AK098833	-	-	
TF00245	1780	4	19p13.12	Conhecido AK091183	-	-	
TF00250	1411	13	19q13.12	Extensão NM_173636	10	-	
TF00251	1521	10	7q2111	Conhecido NM_152754	-	-	
TF00253	1142	8	12q24.13	NOVO	-	Parcial	ORF=369aa
TF00263	954	4	2q33.2	NOVO	-	Parcial	ORF=203aa "Serine-rich region profile" PS50324 E value=0,21

TF00267	781	6	1p13.3	NOVO	-	Parcial	ORF=189aa "IG-like domain profile" PS50835 E value=0,0021
TF00274	574	6	16	Conhecido AK094783	-	-	
TF00280	1302	7	10q24.32	Extensão NM_024789	3	-	
TF00283	1924	12	14q23.2	Conhecido NM_015180	-	-	
TF00285	2264	11	19p13.2	Conhecido BC037565	-	-	
TF00288	932	9	12q14.1	Conhecido AB017814	-	-	
TF00289	2342	6	17p13.1	NOVO	-	Total	ORF=125aa
TF00292	880	9	15q11.2	NOVO	-	Total	ORF=207aa
TF00294	1291	10	1p32.3	NOVO	-	Total	ORF=230aa
TF00295	1636	14	1p32.3	NOVO	-	Parcial	ORF=198aa "Serine-rich region profile" PS50324 E value=0,00021
TF00297	1235	12	1p32.3	NOVO	-	Parcial	ORF=398aa
TF00305	919	7	10q23.33	Conhecido NM_022451	-	-	
TF00309	1366	12	16q22.1	Conhecido AK074773	-	-	
TF00312	923	6	16p12.2	Conhecido BC038400	-	-	
TF00313	1383	6	20q13.13	Conhecido NM_080829	-	-	
TF00314	739	4	15q15.3	Conhecido AL832683	-	-	
TF00315	961	5	9q21.11	Conhecido AL832333	-	-	
TF00318	1296	5	14q32.13	NOVO	-	Total	ORF=418aa Serpina ("serine protease inhibitor") IPR000215 E value=$1,7e^{-158}$
TF00327	1090	12	5p15.33	Conhecido NM_024830	-	-	
TF00350	2098	5	9q12	NOVO	-	Parcial	ORF=130aa
TF00351	1059	6	3p21.31	NOVO	-	Total	ORF=330aa
TF00355	777	4	5q14.3	Conhecido NM_153354	-	-	

TF00359	1008	9	12p13.31	Conhecido AL832139	-	-	
TF00363	944	4	22q11.21	Conhecido NM_153334	-	-	
TF00364	912	5	17q23.3	Conhecido NM_153335	-	-	
TF00371	655	4	19q13.2	NOVO	-	Total	ORF=130aa
TF00379	457	4	17p13.1	Conhecido AB076580	-	-	
TF00380	949	6	Xq23	Conhecido AF286598	-	-	
TF00393	811	8	9p24.3	Conhecido AK093572	-	-	
TF00394	1010	9	1q42.3	Conhecido NM_152490	-	-	
TF00395	1986	6	6p21.32	Conhecido AL713634	-	-	
TF00396	616	4	12q21.1	NOVO	-	Total	ORF=116aa
TF00398	1216	5	1p36.12	Conhecido AB007947	-	-	
TF00404	1124	7	4q34.2	Conhecido BC038536	-	-	
TF00411	1441	3	11p13	Extensão NM_024081	1	-	
TF00501	1013	8	14q23.2	Conhecido (NM_015180)	-	-	
TF00502	1612	2	16p12.2	Conhecido (NM_145865)	-	-	
TF00513	998	5	15q21.3	NOVO	-	Não predito	ORF=126aa
TF00517	1411	3	16p13.3	Conhecido (NM_032444)	-	-	
TF00518	920	8	2q33.2	Extensão (AK096293)	3	-	
TF00519	536	3	8p21.3	Conhecido (AK092034)	-	-	
TF00523	1058	6	16p13.3	Conhecido (AL833717)	-	-	
TF00525	974	4	19q13.43	Conhecido (BC033149)	-	-	
TF01004	943	6	Xp11.22	Conhecido (BC036767)	-	-	
TF01007	1128	6	16p13.3	Conhecido (NM_145294)	-	-	

TF01008	1032	6	16p13.3	Conhecido (NM_153239)	-	-	
TF01009	1287	4	22q13.1	Conhecido (BC031099)	-	-	
TF01013	1282	11	Xq22.3	NOVO	-	Total	ORF=122aa "Protein Kinase Domain" E value=2,1e^{-19}
TF01016	686	1	6p24.1	Conhecido (BC001646)	-	-	
TF01018	1079	6	Xp11.3	Conhecido (NM_032591)	-	-	
TF01024	739	4	21q22.3	Conhecido (BC038504)	-	-	
TF01034	1054	7	21q11.2	Conhecido (BC036510)	-	-	
TF01036	484	4	6p22.1	Conhecido (NM_032507)	-	-	
TF01052	1085	8	2p22.3	Conhecido (NM_016252)	-	-	
TF01054	1229	6	2p22.3	Conhecido (NM_016252)	-	-	
TF01057	1204	9	6p22.3	Extensão (NM_153042)	3	-	
TF01058	863	7	20q13.2	NOVO	-	Não predito	Não contém ORF
TF01061	1175	9	1p36.33	Conhecido (NM_023018)	-	-	
TF01074	769	6	10q23.1	NOVO	-	Parcial	Não contém ORF
TF01081	1331	5	9q34.3	Conhecido (AK092639)	-	-	
TF01083	968	6	4q21.22	Conhecido (AK091412)	-	-	
TF01087	1875	2	5q14.1	Conhecido (NM_152405?)	-	-	
TF01089	727	6	6q25.3	Conhecido (NM_032861)	-	-	
TF01092	1201	10	16q22.1	AL832446	-	-	
TF01100	1539	7	2p13.1	Conhecido (NM_133637)	-	-	
TF01102	960	6	7q22.3	Extensão (NM_152750)	4	-	
TF01105	1090	8	17q23.2	Conhecido (NM_022070)	-	-	

TF01112	1158	7	12q23.3	NOVO	-	Total	ORF=355aa "Ankirin repeat region circular profile" E value=2,1e^{-18}
TF01125	884	7	11q13.1	Extensão (AF001543)	4	-	
TF01130	1078	3	2q32.1	Conhecido (AL832632)	-	-	
TF01132	707	5	14q21.2	Conhecido (BC036056)	-	-	
TF01140	1488	7	1q43	Conhecido (AK095692)	-	-	
TF01145	754	4	7q11.23	Conhecido (BC022886)	-	-	

ANEXO 3

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo