CARACTERIZAÇÃO DE UM NOVO ANTÍGENO TUMORAL: CTSP-1

RAPHAEL BESSA PARMIGIANI

Tese apresentada à Fundação Antônio Prudente para obtenção do título de Doutor em Ciências

Área de concentração: Oncologia

Orientadora: Dra. Anamaria Aranha Camargo

Co-orientador: Dr. Sandro José de Souza

São Paulo

2005

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A.C. Camargo

Parmigiani, Raphael Bessa
Caracterização de um novo antígeno tumoral: CTSP-1 / Raphael Bessa Parmigiani -- São Paulo, 2005.
140p.
Tese(doutorado)-Fundação Antônio Prudente.
Curso de Pós-Graduação em Ciências-Área de concentração: Oncologia.
Orientador: Anamaria Aranha Camargo

Descritores: 1. ANTÍGENOS. 2. ANTIGENOS DE TUMORES/classificação. 3. GENOMA HUMANO.

"Feliz aquele que transfere o que sabe e aprende o que ensina."

Cora Coralina

DEDICATÓRIA

Aos meus pais, Haroldo e Luzia, e meus irmãos, Guilherme, Aline, Fred e Alice,
por mais uma vez me incentivarem a não desistir dos meus sonhos,
principalmente daqueles mais difíceis de serem alcançados. AMO VOCÊS!!!

À minha avó Flávia que, do seu jeito, sempre acreditou em mim. Tenho certeza de que ela está vendo e compartilhando comigo a conquista desta vitória.

Aos meus amigos, que se tornaram uma segunda família desde
o tempo da faculdade. Uma família que cresce cada vez mais
e que sabe o valor de uma verdadeira amizade.

AGRADECIMENTOS

À minha orientadora, **Dra Anamaria Aranha Carmargo**, por me dar o privilégio de realizar este trabalho sob sua orientação. Mais do que um exemplo profissional, você é a minha referência de paciência e humildade. Muito obrigado pela amizade, pelos conselhos em momentos difíceis e pela confiança em mim depositada.

À **Fabiana Bettoni e Lílian Pires**, popularmente conhecidas como Bu e Paia, pela amizade, discussões científicas, correção da tese e troca de conselhos dentro e fora do laboratório. Muito obrigado por me ouvirem e compreenderem o que eu estava sentindo em momentos que só quem trabalha com pesquisa entende. Os bons e os ruins!

À **Natanja Kirschbaum**, pela amizade, pelas importantes aulas de inglês, pelas caronas às aulas da sete e pela ajuda no laboratório. Vou sentir saudades!

À **Maria Vibranovski**, pela amizade e pelas análises de bioinformática.

Ao meu co-orientador **Dr. Sandro de Souza**, pelas sugestões dadas e pelos momentos de descontração fora do laboratório.

À **Ana Helena Perosa**, também conhecida como Saliva, pela amizade e por me incentivar sempre. Te admiro cada dia mais!

Aos meus amigos **Rodrigo, Gustavo, Marcelo, João Paulo e Gabriel** por me incentivarem a estudar e ir pra balada, pois existe vida fora do laboratório!!!

Á gênia aos 25, **Juliana Cruz**, pela amizade e por entender desde o começo o que significa: "Eu preciso estudar!". Com certeza você me ajudou muito!

Á **Mariana Brait**, vulga Miss, pela amizade, momentaneamente internacional, pela troca de experiências e por me ceder seu apartamento durante este último ano. "Se essa casa, se essa casa fosse minha..."

À **Marilene Lopes, Patrícia Sávio, Waleska Martins e Susana Diniz**, pela ajuda e ensinamentos em momentos em que NADA funcionava! Muito obrigado!

À **Dra Dirce Carraro**, pela ajuda e amizade dentro e fora do laboratório.

A todos os amigos do **Laboratório de Biologia Molecular**, em especial à **Elis, Anna Chris, Fernando, Elisa e Mari Granato**, pela amizade e ajuda no dia a dia.

Ao **Prof. Flávio Henrique da UFSCar** e todo o pessoal do seu laboratório, em especial à **Daniela, ao Luís e ao César**, por me ajudarem em um momento muito importante da realização deste trabalho.

Ao **Profº. Dr. Fernando Augusto Soares e Dra Isabela Cunha** pela ajuda nas análises anátomo-patológicas.

À **Ana Paula Lepique e Marilene Lopes** por me ajudarem muito na correção da tese.

Aos meus tios, **Cláudia e Cícero e primos, Mari, Edu e Pim** por continuarem me auxiliando e incentivando e por constituírem minha família paulista.

Ao meu amigo **Geovane**, pela amizade antiga e que mesmo estando sumido nos últimos meses, continua sendo um irmão.

Aos funcionários do Hospital do Câncer, em especial **Carlos Nascimento, Severino Silva e Miyuki Fukuda**, pela confecção das lâminas de material fresco e parafinado. Agradeço muito também à **Suely Nonogaki** pelas reações de imunohistoquímica. Obrigado pela paciência e disposição sempre que precisei.

Aos funcionários do Instituto Ludwig **Izabel Carneiro, Sirléia Miranda, Roseli Mendonça, Francisco Sampaio, Wanderley Lourenço e Ricardo Lima**, pela ajuda e por facilitarem a realização do meu trabalho.

À todos que me ajudaram de uma maneira ou de outra e que me mostraram que um bom trabalho nunca pode ser feito sem a ajuda e troca de informações entre colegas de trabalho.

Á **Márcia Hiratori e Ana Maria Kuninari**, pelos auxílios prestados durante estes anos de "amolação".

A todos os **funcionários da biblioteca** do Hospital A.C. Camargo, em especial à **Suely Francisco** pela ajuda prestada na correção da tese.

Ao **Dr. Ricardo Renzo Brentani**, pela direção do Instituto Ludwig e por permitir a realização deste trabalho.

Ao **Dr. Luís Fernando Lima Reis** pela direção da pós-graduação da Fundação Antônio Prudente.

A **Comissão de Aperfeiçoamento de Pessoal de Nível Superior** (CAPES) pela bolsa concedida.

RESUMO

Parmigiani RB. **Caracterização de um novo antígeno tumoral: CTSP-1.** São Paulo; 2005. [Tese de Doutorado-Fundação Antônio Prudente]

A necessidade de identificar novos antígenos tumorais, que possam ser utilizados no tratamento e diagnóstico do câncer, tem levado ao desenvolvimento de técnicas eficientes para essa finalidade. Antígenos de diferentes categorias foram identificados e caracterizados e, entre eles, os antígenos *cancer-testis* (*CT*) e os antígenos de diferenciação (CD) são os de maior importância clínica dado seu restrito padrão de expressão. Utilizando alinhamentos entre seqüências expressas e a seqüência do genoma humano, identificamos um novo gene localizado no cromossomo 21, denominado CTSP-1. Este gene apresenta alta similaridade com o antígeno tumoral NY-BR-1, o qual codifica um fator de transcrição tecido específico e é alvo potencial para imunoterapia de câncer. Para verificar se o gene CTSP-1 realmente corresponde a um novo antígeno tumoral, nós realizamos a completa caracterização do mesmo. Através de diferentes técnicas obtivemos a seqüência completa do gene CTSP-1 e identificamos diferentes formas de poliadenilação e de *splicing* alternativas. Além disso, avaliamos o padrão de expressão do CTSP-1 em tecidos normais, linhagens celulares tumorais e amostras de tumores. O gene CTSP-1 apresentou um padrão de expressão restrito, sendo expresso apenas em testículo dentre os tecidos normais, em diferentes linhagens celulares tumorais (9/22) e em diferentes tipos de tumores (74/178), o que condiz com o padrão de expressão dos

antígenos *Cancer-Testis*. Posteriormente, a proteína CTSP-1 recombinante foi expressa em sistema heterólogo e utilizada na produção de anticorpo policlonal em camundongo. Este anticorpo foi utilizado em experimentos de *immunoblotting* e imunohistoquímica para a detecção da proteína CTSP-1 em testículo normal e amostras normais e tumorais pareadas de mama e próstata. Através de *immunoblotting*, uma banda específica de 22kDa foi identificada em extrato total de testículo normal, correspondente ao peso esperado da proteína CTSP-1. Através de imunohistoquímica, verificamos uma marcação preferencial em células germinativas e células de Leydig de testículo normal. Nos tecidos com amostras pareadas normal/tumor, apenas as amostras tumorais foram fortemente marcadas. A proteína CTSP-1 recombinante também foi utilizada na investigação de anticorpos específicos em plasma de pacientes com câncer. Aproximadamente 150 amostras foram analisadas, das quais 20% apresentaram resposta imune humoral contra a proteína CTSP-1. Em conjunto, estes resultados confirmam que o gene CTSP-1 é um novo antígeno tumoral da categoria dos antígenos *Cancer-Testis*, com expressão restrita a tumor e testículo e com alta imunogenicidade em pacientes com câncer.

SUMMARY

Parmigiani RB. **[Characterization of a new tumor antigen: CTSP-1].** São Paulo; 2005. [Tese de Doutorado-Fundação Antônio Prudente]

The need to identify new tumor antigens to be used in cancer treatment and diagnosis has lead to the development of efficient techniques for this purpose. Antigens from different categories have been identified and characterized and, among those, the cancer-testis (CT) and the cancer differentiation (CD) antigens are of the greatest clinical interest due to their restricted expression pattern. Using alignments between expressed sequences and the Human Genome Sequence, we identified a new gene located on chromosome 21, named CTSP-1. This gene has a high similarity to the tumor antigen NY-BR-1, which encodes for a tissue specific transcription factor and is a potential target for cancer immunotherapy. In order to verify if the CTSP-1 gene is really a new tumor antigen, we performed its complete characterization. Using different techniques, we were able to obtain the complete sequence of the CTSP-1 gene and to identify different alternative polyadenilation and splicing forms. Moreover, we analyzed the CTSP-1 expression pattern in normal tissues, tumor cell lines and tumor samples. CTSP-1 showed a restricted expression pattern, being expressed only in testis among normal tissues, in different tumor cell lines (9/22) and in different tumor types (74/178), which matches with the expression pattern of Cancer-Testis antigens. Afterwards, the recombinant CTSP-1 protein was expressed in a heterologous system and used for the generation of polyclonal antibody in mice.

This antiboby was used in immunoblotting and immunohistochemistry experiments for the detection of CTSP-1 protein in normal testis and paired normal and tumor samples from breast and prostate. Using immunoblotting, a 22kDa specific band was identified in testis total protein extract, corresponding to the expected molecular weight of the CTSP-1 protein. Using immunohistochemistry, we verified the preferential staining of germ cells and Leydig cells in normal testis. Among tissues with paired normal/tumor samples, only tumor samples were strongly stained. The CTSP-1 recombinant protein was also used in the search for specific antibodies in plasma from cancer patients. Approximately 150 samples were analyzed, of which 20% showed a humoral immune response against the CTSP-1 protein. Taken together, these results confirm that the CTSP-1 gene is a new tumor antigen from the Cancer-Testis category, with restricted expression in testis and tumors and with high immunogenicity in cancer patients.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

Ω = Ohm

µF = Microfarad

µg = Micrograma

µl = Microlitro

µm = Micrômetro

µM = Micromolar

ADAM = *A Desintegrin and Metalloproteinase*

AFP = Alfa-fetoproteína

ATCC = *American Type Culture Collection*

BLAST = *Basic Local Alignment Search Tool*

BLAT = *BLAST-Like Alignment Tool*

BSA = Albumina Sérica Bovina (do inglês *Bovine Serum Albumin*)

cDNA = DNA complementar

CEA = Antígeno carcinoembrionário (do inglês *Carcinoembrionary Antigen*)

CT = *Cancer Testis*

CTL = *Cytolytic T lymphocyte*

DNA = Ácido desoxirribonucléico

DNA-PK = Proteína Quinase Dependente de DNA (do inglês *DNA-dependent protein kinase*)

dNTP = Deoxinucleotídeo

DO = Densidade Óptica

EBV = *Epstein-Barr virus*

EDTA = Ácido etilenodiaminotetracético disódico

EST = Etiqueta de seqüência expressa (do inglês *Expressed Seqüence Tags*)

FDA = *Food and Drugs Administration*

g = Grama

GAPDH = Gliceraldeído 3-fosfato desidrogenase

hMLH-1 = *Human mut-L homologue-1*

HPV = *Human papilloma virus*

HLA = Antígeno leucocitário humano (do inglês *Human Leucocyte Antigen*)

IFN-γ = *Interferon- γ*

Ig = Imunoglobulina

IHQ = Imunohistoquímica

IMAGE = *Integrated Molecular Analysis of Genomes and their Expression*

kb = Kilobases

kDa = Kilodaltons

M = Molar

mA = Miliamper

MCA = Metil-colantreno

MHC = Complexo principal de histocompatibilidade (do inglês *Major histocompatibility complex*)

ml = Mililitro

mM = Milimolar

mRNA = RNA mensageiro

ng = Nanograma

NK = *Natural Killer*

nm = Nanômetro

nM = Nanomolar

ORESTES = *Open Reading Frame ESTs*

ORF = Fase aberta de leitura (do inglês *Open Reading Frame)*

pb = Pares de base

PBS = *Phosphate Buffered Saline*

PBST = *Phosphate Buffered Saline Tween*

PCR = Reação em cadeia da polimerase (do inglês *Polimerase Chain Reaction*)

RAG = *Recombination activating gene*

RNA = Ácido ribonucléico

RT = Transcrição reversa (do inglês *Reverse Transcriptase*)

SAGE = *Serial Analysis of Gene Expression*

SCID = *Severe-combined immunodeficiency*

SDS = *Sodium lauryl sulfate*

SI = Sistema immune

SSC = *Saline Sodium Citrate*

TAE = Tris Acetato EDTA (do inglês *Tris Acetate EDTA*)

TGF-β = *Transforming growth factor β*

U = Unidade

V = Volume

ÍNDICE

Anexo 2: Ferreira EN, Pires LC, Parmigiani RB et al. Identification and complete sequencing of novel human transcripts through the use of mouse orthologs and testis cDNA sequences. **Genet Mol Res** 2004; 3(4):493-511.

Anexo 3: Sogayar MC, Camargo AA, Bettoni F et al. Ludwig-FAPESP Transcript Finishing Initiative. A transcript finishing initiative for closing gaps in the human transcriptome. **Genome Res** 2004; 14(7):1413-23.

Anexo 4: Reymond A, Camargo AA, Deutsch S et al. Nineteen additional unpredicted transcripts from human chromosome 21. **Genomics** 2002; 79(6):824-32.

Anexo 5: Kirschbaum NS, Parmigiani RB, Camargo AA, Souza SJ. Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data. **Physiological Genomics** 2005. *In press.*

Anexo 6: Parmigiani RB e Camargo AA. **In: Oncologia Molecular**. Ed Atheneu São Paulo, 2004; O genoma humano e o câncer. pp3-11.

INTRODUÇÃO

1 INTRODUÇÃO

O câncer é uma das mais importantes causas de morbidade e mortalidade entre crianças e adultos, o que o torna um dos maiores problemas de saúde mundial. Segundo estimativas do Ministério da Saúde as neoplasias constituem a segunda maior causa de morte por doença no Brasil (Ministério da Saúde 2002). O Instituto Nacional de Câncer (INCA) estimou que, em 2003, ocorreram mais de 400.000 novos casos e aproximadamente 127.000 óbitos por câncer em todo o país (Ministério da Saúde 2002).

Uma característica marcante do câncer é o descontrole das vias que regulam a proliferação e a morte celular. Na maioria dos casos (95%), alterações genéticas em células somáticas que alteram o controle do ciclo celular representam o agente causal. Estas alterações levam ao aumento da capacidade proliferativa das células que, ao sofrerem alterações em suas propriedades adesivas, adquirem a habilidade de invasão dos tecidos adjacentes, bem como de formação de metástases (LODISH et al. 2000).

Embora se saiba que existam inúmeros tipos de câncer, estudos têm mostrado a existência de importantes similaridades entre neoplasias aparentemente distintas. Sendo assim, HANAHAN e WEINBERG (2000) propuseram que as células transformadas devem adquirir seis alterações para o desenvolvimento e crescimento de um tumor (*the hallmarks of cancer*). São elas: 1) capacidade de crescimento autônomo; 2) insensibilidade a sinais inibitórios de crescimento; 3) evasão de sinais apoptóticos intrínsecos; 4) potencial proliferativo ilimitado; 5) capacidade de

promover angiogênese; 6) competência para invasão tecidual e formação de metástases. Recentemente uma outra característica tem sido considerada como um outro ponto em comum entre todas as neoplasias: a capacidade da célula maligna de evadir as funções supressoras de tumor do sistema imune (SI) do organismo (DUNN et al. 2004).

1.1 IMUNOLOGIA DE TUMORES – UM BREVE HISTÓRICO

O conceito de que o SI pode reconhecer e eliminar tumores primários em desenvolvimento, na ausência de intervenção terapêutica externa, existe há aproximadamente 100 anos (Ehrlich 1909, citado por DUNN et al. 2004). Porém, este conceito não pôde ser experimentalmente testado pois, na época em que foi proposto, pouco se sabia sobre as bases moleculares e celulares da imunologia.

Somente em meados dos anos 50, com a produção dos camundongos singênicos (geneticamente idênticos), é que esta hipótese começou a ser experimentalmente testada. Assim, a existência de antígenos tumor-específicos foi estabelecida através de experimentos nos quais camundongos foram imunizados contra transplantes singênicos de tumores induzidos por carcinógenos químicos, vírus ou outro agente externo. Estes estudos indicaram a existência da imunovigilância, uma vez que se não houvessem alterações nas células tumorais que fossem reconhecidas pelo SI, não haveria imunovigilância (BURNET 1970; STUTMAN 1974). As moléculas alvo capazes de desencadear a resposta imune foram denominadas antígenos tumor-específicos. Deste modo, THOMAS (1959) e BURNET (1970) propuseram formalmente a hipótese da imunovigilância (*The*

cancer immunosurveillance hypothesis). Ambos acreditavam que os linfócitos eram responsáveis por eliminar continuamente células transformadas recém formadas.

No entanto, contrariando todas expectativas, experimentos iniciais feitos em camundongos *nude*, o modelo experimental que melhor representava a imunodeficiência congênita naquela época, não forneceram nenhuma evidência de que esta hipótese fosse verdadeira (STUTMAN 1974 e 1979). Mais especificamente, os camundongos CBA/H *nude* não apresentaram uma maior incidência de tumores espontâneos ou quimicamente induzidos quando comparados aos animais controle (imunocompetentes), e nem sequer apresentaram um período menor de latência antes do aparecimento do tumor. Estes resultados fizeram com que a hipótese da imunovigilância fosse abandonada durante muito tempo.

Entretanto, hoje se sabe que existiam muitas falhas nestes experimentos as quais, naquele momento, não podiam ser previstas. Primeiro, o camundongo *nude* é reconhecido atualmente como um modelo imperfeito de imunodeficiência. Isto porque estes animais, mesmo possuindo um timo rudimentar, podem desenvolver algum grau de imunidade adquirida devido à produção de baixos, porém detectáveis, níveis de populações de células αβ T (HUNIG 1983; IKEHARA et al. 1984; MALECKAR e SHERMAN 1987). Segundo, a existência das células NK (*Natural Killer cells*) ainda não havia sido estabelecida, sendo as mesmas presentes e funcionais nos camundongos *nude*, revelando assim, a presença de uma imunidade inata muito importante no combate às células transformadas (SMYTH et al. 2002a). Desta forma, um resquício de imunidade adaptativa aliado a uma imunidade inata completamente funcional poderia fornecer ao camundongo *nude* alguma capacidade de imunovigilância. Por último, a linhagem de camundongo CBA/H é muito sensível

ao carcinógeno metil-colantreno (MCA), utilizado nestes experimentos, o que poderia mascarar qualquer tipo de proteção conferida pelo SI dos animais controle.

Nos anos 90, dois achados muito importantes incitaram um novo interesse pelo processo de imunovigilância contra o câncer. Primeiro, foi demonstrado o efeito protetor do interferon-γ (IFN-γ) endógeno contra o crescimento de tumores transplantados e formação de tumores primários espontâneos ou quimicamente induzidos (DIGHE et al. 1994; KAPLAN et al. 1998; STREET et al. 2001 e 2002). Estes estudos foram feitos com a injeção de anticorpos neutralizantes anti-IFN-γ em camundongos transplantados com tumores e em modelos experimentais de formação de tumores primários. Neste último caso, os camundongos insensíveis ao IFN-γ mostraram-se de 10 a 20 vezes mais sensíveis à formação de tumores induzidos por MCA do que os camundongos controle, nos quais a via de sinalização do IFN-γ estava preservada (KAPLAN et al. 1998).

O outro achado importante foi a observação de que camundongos C57BL/6 que não continham o gene da perforina (perforina -/-) eram mais susceptíveis à formação de tumores induzidos por MCA (VAN DEN BROEK et al. 1996; SMYTH et al. 2000a e b). A perforina é um componente dos grânulos citolíticos de linfócitos T citotóxicos e de células NK que desempenham um papel importante mediando a citotoxicidade dependente de linfócito (RUSSELL e LEY 2002). Além de tumores quimicamente induzidos, posteriormente verificou-se que os camundongos perforina -/- também apresentavam maior incidência de linfomas espontâneos quando comparados com os animais controle (SMYTH et al. 2000b).

O trabalho que demonstrou definitivamente a existência de um processo de imunovigilância dependente de IFN-γ e linfócitos foi baseado em experimentos com

camundongos que não continham os genes RAG-1 ou RAG-2 (*recombinase activating gene*). As enzimas codificadas por estes genes estão envolvidas no rearranjo de DNA e são expressas apenas nas linhagens celulares linfóides. Deste modo, os camundongos RAG-1-/- ou RAG-2-/- não conseguem fazer o rearranjo dos genes que codificam os receptores de antígeno dos linfócitos e, por isso, não possuem as células T, B e NKT (SHINKAI et al. 1992). Uma vez que a expressão destes genes está restrita às células do SI, o uso destes camundongos fornece um modelo apropriado para o estudo dos efeitos da imunodeficiência do hospedeiro no desenvolvimento de tumores pois, diferentemente de outros modelos de imunodeficiência [como os camundongos SCID (*Severe-combined Immunodeficiency*) que não possuem a *DNA-dependent protein kinase*], a ausência destes genes não resulta na perda da capacidade de reparo do DNA em células não linfóides que poderiam estar em processo de transformação.

Assim, pela primeira vez, os pesquisadores tiveram em mãos camundongos que podiam ser utilizados em experimentos de carcinogênese e cujos resultados poderiam ser corretamente interpretados. Após a injeção de MCA, camundongos RAG-2-/- desenvolveram sarcomas mais rapidamente e em maior freqüência que os animais tipo selvagem (SHANKARAN et al. 2001). Além disso, estes camundongos também apresentaram uma maior incidência de tumores epiteliais espontâneos.

Diferentes estudos que utilizaram outras linhagens singênicas de camundongos contendo disrupções em genes-alvo de componentes do SI (IL-12, cadeias β e δ do receptor TCR) ou mesmo que utilizaram anticorpos específicos contra os mesmos, forneceram resultados suficientes para comprovar o efeito protetor do SI do hospedeiro murino contra o desenvolvimento de tumores induzidos

e/ou espontâneos (DIGHE et al. 1994; KAPLAN et al. 1998; FALLARINO e GAJEWSKI 1999; KACHA et al. 2000). Mais do que isso, estes estudos também demonstraram o envolvimento de ambas as respostas imune, inata e adaptativa, na imunovigilância do câncer (GIRARDI et al. 2001; SMYTH et al. 2001).

Todos estes estudos evidentemente só puderam ser feitos em modelos experimentais (essencialmente camundongos), sugerindo que em humanos também exista uma imunovigilância contra o câncer, a qual precisava ser comprovada. Os estudos iniciais de pacientes transplantados que foram tratados com imunossupressores ou de indivíduos com imunodeficiência primária mostraram que os mesmos possuíam um risco relativo significativamente maior para o desenvolvimento de câncer quando comparados a indivíduos controle, indicando a existência de um mecanismo de imunovigilância em humanos (GATTI e GOOD 1971). No entanto, este risco aumentado estava limitado basicamente a tumores de origem viral, como linfomas não-Hodgkin, sarcoma de Karposi e carcinomas ano-geniturinários, associados respectivamente ao EBV (*Epstein-Barr virus*), HPV (*Human papilloma virus*) e herpes vírus 8 (PENN 1999). Cabe ressaltar que o aumento da susceptibilidade a infecções virais nestes pacientes não permite inferir de forma inequívoca o papel da deficiência imunológica no aumento da incidência de tumores.

Evidentemente experimentos de transplante de tumor heterólogo, como são realizados em camundongos, não podem ser feitos em humanos. Sendo assim, diferentes técnicas foram desenvolvidas para a investigação *in vitro* da presença de resposta imune celular e humoral em pacientes com câncer (CAREY et al. 1976; KNUTH et al. 1984; VAN DER BRUGGEN et al. 1991; TRAVERSARI et al. 1992;

SAHIN et al. 1995). Técnicas como *autologous typing* e *SEREX* (*Serological Analysis of Recombinant cDNA Expression Libraries of Human Tumors with Autologous Serum*) mostraram-se bastante eficientes na identificação da resposta anti-tumoral, através da utilização de soro e linfócitos de pacientes contra tumores autólogos. Além disso, a presença de linfócitos no interior de tumores tem se mostrado um bom indicador prognóstico (DUNN et al. 2004).

Apesar destas evidências suportarem de maneira indireta a existência de um processo de imunovigilância funcional, indivíduos imunocompetentes desenvolvem câncer. Isto acontece porque existe uma falha no processo imunológico que favorece o crescimento de tumores com baixa imunogenicidade em indivíduos imunocompetentes. Uma vez que é alta a instabilidade genética nas células transformadas, alterações que favorecem o não reconhecimento destas células pelo SI são selecionadas e permitem o escape do tumor (ABBAS 2000). Considerando-se que estas alterações são positivamente selecionadas pelo SI, pode-se afirmar que o mesmo favorece o desenvolvimento de tumores menos imunogênicos que dificilmente serão eliminados.

Sendo assim, acreditando neste duplo papel do SI, a hipótese de imunovigilância vem sendo reformulada. Originalmente esta hipótese preconizava que a imunovigilância contra o câncer fosse essencialmente uma função protetora exercida pela resposta imune adaptativa nos momentos iniciais da transformação celular. Entretanto, hoje sabe-se que ambas as respostas, inata e adaptativa, participam deste processo e servem não apenas para proteger o hospedeiro do desenvolvimento do tumor, mas também para "esculpir" ou editar a imunogenicidade dos tumores que eventualmente se formam. Assim, o uso do termo imunoedição do

câncer (*cancer immunoediting*) tem sido proposto para melhor representar este duplo papel do SI (DUNN et al. 2002).

1.2 A TEORIA DA IMUNOEDIÇÃO DO CÂNCER

O processo de imunoedição do câncer pode ser dividido em 3 fases, denominadas "os três 'Es' da imunoedição": eliminação, equilíbrio e escape (DUNN et al. 2004). A fase de eliminação consiste no conceito original da imunovigilância e se a mesma for bem sucedida, ou seja, se a célula tumoral for eliminada, ela representa todo o processo de imunoedição sem que ocorra a progressão para as fases subseqüentes. Já na fase de equilíbrio, o SI do hospedeiro e as células tumorais que sobreviveram à fase de eliminação entram em um equilíbrio dinâmico no qual os componentes do SI exercem uma potente e contínua pressão seletiva sobre as células tumorais geneticamente instáveis e mutantes, suficiente para contê-las, porém não o bastante para eliminá-las completamente. Esta é a fase mais longa de todo o processo, podendo durar muitos anos.

No final da fase de equilíbrio o que se encontra é uma nova população de clones tumorais com imunogenicidade reduzida, originada a partir de uma população heterogênea inicial que foi "esculpida" pelo SI. Assim, se inicia a fase de escape, na qual as variantes tumorais selecionadas na fase de equilíbrio crescem em um ambiente imunologicamente intacto.

Toda célula tumoral possui múltiplos mecanismos de escape para fugir da ação integrada das respostas imune inata e adaptativa do SI já existentes contra os clones imunogênicos que lhe deram origem. Provavelmente, um dos mecanismos

mais importantes é a diminuição ou perda da expressão de moléculas de HLA (*Human Leucocyte Antigen*) classe I nas células tumorais, o que dificulta o seu reconhecimento pelos linfócitos T citotóxicos (*CTL = Cytolytic T Lymphocyte*) (ABBAS 2000). Além disso, também pode haver diminuição da expressão de β_2-microglobulina, ou ainda, de componentes da maquinaria de processamento de antígenos (CHEEVER et al. 1995).

Outro mecanismo de escape verificado em tumores é a perda da expressão de antígenos que desencadeiam resposta imune, a qual é mais freqüente em tumores de crescimento muito acelerado. Dada a alta taxa de mitose das células tumorais e sua instabilidade genética, mutações ou deleções de genes que codificam proteínas reconhecidas pelo SI são freqüentes. Se estas proteínas não forem necessárias para o crescimento do tumor ou para a manutenção do fenótipo transformado, esta perda pode representar uma vantagem para o crescimento das células proteína específica-negativas.

Além destes mecanismos, pode-se citar também: a secreção de citocinas imunossupressoras pelo tumor (como TGF-β e IL-10), o não reconhecimento das células tumorais pelos *CTLs*, devido à ausência de expressão de moléculas co-estimulatórias (como B7) ou de HLA classe II (necessárias para a ativação de linfócitos T *Helper*) e ainda, mecanismos de escape da destruição imunológica (como falhas na via de morte por apoptose e insensibilidade à ação do IFN-γ). Estes mecanismos certamente permitem que as células tumorais escapem da fase de eliminação e de equilíbrio, sendo de fato encontrados em tumores (KHONG e RESTIFO 2002).

Deste modo, fica evidente que uma melhor compreensão do processo de imunoedição do câncer terá implicações importantes na imunoterapia de neoplasias. A possibilidade de explorar estas reações do SI, na tentativa de aumentá-las e assim eliminar o tumor ou mesmo impedir que o mesmo se desenvolva, tem se tornado o principal objetivo da imunologia de tumores.

Sendo assim, a identificação de antígenos tumorais capazes de induzir resposta imune em pacientes com câncer torna-se um importante passo para o desenvolvimento de vacinas e a aplicação da imunoterapia no tratamento de tumores. Neste sentido, um grande esforço tem sido feito na tentativa de identificar um grande número de antígenos que sejam freqüentemente encontrados em diferentes tipos de tumor.

1.3 IDENTIFICAÇÃO DE ANTÍGENOS TUMORAIS

Nos últimos anos, uma grande variedade de antígenos tumorais reconhecidos por linfócitos T e B, tem sido identificada (SAHIN et al. 1997). A identificação de tais antígenos iniciou-se na década de 60 com a análise de soro heteroimune, obtido a partir de coelhos e outros animais imunizados com tumores humanos. O desafio, geralmente não alcançado, era remover os anticorpos que reagiam contra antígenos de tecidos normais que diferiam entre as espécies. Mesmo com estas limitações, dois antígenos importantes foram identificados com esta metodologia: a alfa-fetoproteína (AFP) e o antígeno carcinoembrionário (CEA) (GOLD e FREEDMAN 1965).

Pouco tempo depois, técnicas que envolviam aloanticorpos gerados em camundongos singênicos foram utilizadas na identificação de antígenos tumorais de

células linfóides malignas. Desta maneira, antígenos de diferenciação da superfície celular de linfócitos, tais como TL (*thymus leukemia*) e CD8, foram identificados (BOYSE e OLD 1969; OLD e STOCKERT 1977). Mesmo com o desenvolvimento da tecnologia de hibridomas e a produção de anticorpos monoclonais, que potencialmente poderiam ser a solução para a identificação de antígenos tumorais de superfície, a comprovação da existência de um real antígeno tumoral humano ainda não era tão fácil. Isto porque mesmo o antígeno tumoral com padrão de expressão mais restrito geralmente acabava sendo um antígeno de diferenciação, também presente no tecido normal.

Na tentativa de solucionar estes problemas, uma técnica imunológica denominada *autologous typing* foi desenvolvida. O objetivo desta técnica era restringir a análise aos reagentes autólogos (células tumorais, soro e linfócitos do mesmo paciente) para eliminar a contribuição dos aloantígenos nas reações observadas. Para tanto, o cultivo de células tumorais era necessário, o que limitou esta análise a tipos de tumores que poderiam ser adaptados à cultura *in vitro* com maior facilidade, tais como: leucemia, linfomas, melanoma, tumor renal e cerebral (PFREUNDSCHUH et al. 1978; UEDA et al. 1979). Embora muito promissora, a técnica revelou que apenas uma pequena parte dos pacientes devenvolve anticorpos ou linfócitos T específicos para antígenos de superfície celular do tumor autólogo. Após o desenvolvimento de técnicas de clonagem gênica e sistemas de expressão, a caracterização dos alvos moleculares que estavam sendo reconhecidos pelas células T CD8 tornou-se mais fácil (TRAVERSARI et al. 1992). Por outro lado, os baixos títulos de anticorpos encontrados nos pacientes muitas vezes eram insuficientes para a caracterização dos antígenos identificados.

Somente em meados da década de 90, com o desenvolvimento de uma técnica inovadora, a identificação de antígenos tumorais reconhecidos por anticorpos tornou-se mais produtiva. A técnica, denominada *SEREX*, consiste em isolar antígenos tumorais que elicitam resposta imune com altos títulos de imunoglobulina G (IgG) em pacientes com câncer (SAHIN et al. 1995). Para tal, é feita a biópsia do tumor, com posterior extração de RNA e síntese de cDNA (DNA complementar). Em seguida, este cDNA é clonado em vetor de expressão e a biblioteca resultante é rastreada com o soro do paciente em busca de clones reativos. Os clones positivos são seqüenciados e verifica-se se há similaridade com algum gene já descrito. Paralelamente, é feita também uma avaliação do padrão de expressão destes candidatos em tecidos normais e tumorais e, ainda, a medida da imunogenicidade dos mesmos, através da determinação da freqüência de anticorpos no soro de indivíduos sadios e de pacientes com o mesmo tipo de tumor. Logo no início do emprego da técnica, SAHIN et al. (1997) identificaram os antígenos MAGE-A1 e tirosinase, dois antígenos originalmente identificados como indutores de resposta imune celular, indicando que existem antígenos tumorais que elicitam os dois tipos de resposta imune, celular e humoral.

1.4 ANTÍGENOS TUMORAIS

Pode-se definir os antígenos tumorais como sendo proteínas de restrito padrão de expressão e que, necessariamente, são reconhecidas pelo SI de pacientes com câncer. Baseando-se no reconhecimento específico das células tumorais por linfócitos T citotóxicos (KNUTH et al. 1984) ou anticorpos autólogos (SAHIN et al.

1995; OLD e CHEN 1998) de pacientes com câncer, um grande número de antígenos tumorais tem sido identificado, os quais podem ser classificados em 5 grupos:

1. Antígenos *cancer/testis* (*CT*) os quais são expressos em diferentes tipos de tumores mas que estão silenciados em tecidos normais, exceto testículo. Ex: MAGE-A1, NY-ESO-1 e SSX-1 (VAN DER BRUGGEN et al. 1991; CHEN et al. 1997; GURE et al. 1997);
2. Antígenos de diferenciação, expressos em tipos celulares específicos e em seus respectivos tumores. Ex: Melan A, gp100 e Tirosinase de melanócitos (BRICHARD et al. 1993; KAWAKAMI et al. 1994);
3. Antígenos codificados por genes mutados. Ex: p53, CDK4 e MUM-1 (COULIE et al. 1995; WOLFEL et al. 1995; SCANLAN et al. 1998);
4. Antígenos superexpressos em tecidos tumorais. Ex: HER2/neu, anidrase carbônica e eIF-4gamma (FISK et al. 1995; SAHIN et al. 1995; BRASS et al. 1997);
5. Antígenos virais. Ex: oncoproteínas de HPV e EBV (LENNETTE et al. 1995; TINDLE et al. 1996);

Dentre estas classes, os antígenos com restrito padrão de expressão, tanto os antígenos *CT* quanto os antígenos de diferenciação, têm sido muito estudados por apresentarem grande potencial terapêutico.

1.4.1 Antígenos de diferenciação (*Cluster differentiation* = CD)

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). Outra característica destes antígenos é

que sua expressão geralmente é preservada nas células tumorais, o que os torna importantes marcadores no diagnóstico imunopatológico diferencial do câncer. Por exemplo, pode-se saber qual a origem de uma metástase através da expressão de um antígeno de diferenciação específico de um determinado tecido. Esta categoria de antígenos foi melhor estudada em linfócitos, sendo a maioria dos primeiros antígenos de diferenciação específica de linfócitos e outras células hematopoiéticas, como por exemplo CD4 e CD8 (CANTOR e BOYSE 1975; SHIKU et al. 1975).

Estes antígenos também podem ser utilizados como alvo para imunoterapia específica, considerando-se que apenas o tecido normal do qual o tumor se originou será afetado. Este procedimento possui uma grande vantagem sobre a quimio e radioterapia, ambas completamente inespecíficas. Um bom exemplo é o anticorpo anti-CD20, que reconhece um antígeno de diferenciação de linfócito, o primeiro anticorpo monoclonal aprovado pelo *Food and Drugs Administration* (FDA) para imunoterapia de linfoma (GRILLO-LOPEZ et al. 1999).

1.4.2 Antígenos *Cancer/testis* (*CT*)

A principal característica dos antígenos *CT* é a sua restrita expressão em testículo, dentre os tecidos normais, e em tumores de diferentes tipos histológicos (CHEN et al. 1998). Além disso, estes antígenos geralmente fazem parte de famílias gênicas e, curiosamente, muitos deles (60%) estão localizados no cromossomo X. Em termos evolutivos, famílias gênicas com organização exon/íntron semelhante, como as famílias MAGE e SSX, sugerem que os antígenos *CT* provavelmente tenham surgido por retrotransposição, conversão gênica ou duplicação gênica (SCANLAN et al. 2002a).

Existem exceções para alguns membros desta categoria que também são expressos em outros tecidos normais, no entanto, sempre em um nível mais baixo do que em testículo e nos tumores (LETHE et al. 1998). Após a identificação de vários membros desta categoria, sendo mais de 89 genes no total, algumas observações quanto à expressão dos mesmos puderam ser feitas. Primeiramente, a freqüência de expressão de um determinado antígeno *CT* é muito variável entre os diferentes tipos de tumor. O antígeno NY-ESO-1 e alguns membros da família MAGE são os que apresentam uma maior freqüência de expressão variando de 20 a 50% dos casos dependendo do tipo tumoral (SCANLAN et al. 2002a). Para os outros antígenos desta categoria essa freqüência é bem mais baixa, muitas vezes não ultrapassando 10% das amostras. Considerando-se a origem tecidual do tumor, geralmente os melanomas têm a freqüência mais alta de expressão de um determinado antígeno *CT*, seguido por tumores de bexiga, pulmão, mama e próstata. Além disso, estima-se que mais de 95% das lesões metastáticas de melanoma expressem pelo menos um antígeno *CT* (ZENDMAN et al. 2003).

Além disso, a expressão dos antígenos *CT* parece estar associada à progressão tumoral e a tumores com maior potencial de malignidade (SCANLAN et al. 2002a). Por exemplo, a expressão dos genes MAGE-1, 2, 3 e 4 foi detectada, respectivamente, em 16%, 41%, 36% e 11% dos casos de melanoma primário (n=100) e em 48%, 70%, 76% e 22% de melanomas metastáticos (n=145) (BRASSEUR et al. 1995). Do mesmo modo, a expressão do NY-ESO-1 foi detectada em 10% das amostras de melanoma primário (n=20) e em 47% das lesões metastáticas (n=32) (GOYDOS et al. 2001).

Por último, há uma tendência para que a expressão destes antígenos esteja agrupada, isto é, certos tumores apresentam expressão de múltiplos antígenos *CT* simultaneamente. Este "agrupamento" pode estar relacionado aos mecanismos de controle da expressão destes antígenos, principalmente mecanismos como a metilação de regiões promotoras e a deacetilação de histonas, nos quais o controle da expressão gênica não é restrito a apenas um gene, mas abrange vários genes localizados em uma determinada região genômica (SHICHIJO et al. 1996; WEISER et al. 2001). A constatação de que muitos destes genes estão localizados no cromossomo X e portanto, possivelmente próximos, fortalece esta hipótese.

Nos estudos com imunohistoquímica (IHQ), nos quais pode-se detectar o padrão de expressão da proteína com anticorpos monoclonais, outra característica muito relevante destes antígenos pode ser observada. A expressão da proteína em um mesmo tumor é muito variável, podendo estar restrita a uma pequena quantidade de células, de maneira dispersa ou em pequenos focos celulares, mas também presente em todo o tumor (JUNGBLUTH et al. 2000; JUNGBLUTH et al. 2001a). Esta expressão heterogênea é bastante diferente do que se verifica para membros de outras categorias de antígenos, como os antígenos de diferenciação, os quais possuem expressão bastante homogênea. Esta heterogeneidade na expressão dos antígenos *CT* representa um grande obstáculo para o estabelecimento de uma imunoterapia eficiente para o tratamento do câncer. Mesmo assim, acredita-se que o desenvolvimento de vacinas polivalentes (contendo diferentes antígenos) seja uma alternativa razoável para contornar este obstáculo.

O fato dos antígenos *CT* serem expressos apenas em tumores, gametas e trofoblastos sugere um mecanismo regulatório comum para estas células

(JUNGBLUTH et al. 2001b). Além da regulação da expressão por processos como a metilação, tem se postulado que a expressão dos antígenos *CT* em tumores possa ser conseqüência da indução de um programa gametogênico no câncer (OLD 2001). Sustentando esta hipótese, existem inúmeras características compartilhadas entre as células tumorais e as gametogênicas, tais como: (a) imortalização; (b) capacidade de invasão; (c) crescimento independente de adesão; (d) capacidade de promover angiogênese; (e) hipometilação genômica; e (f) baixa expressão de moléculas de HLA.

Pouco se sabe sobre a função da maioria dos antígenos *CT*, com exceção do SCP-1, OY-TES-1 e CT15/Fertilina-β. O SCP-1 faz parte do complexo sinaptoneal, envolvido no pareamento dos cromossomos homólogos durante a meiose (MEUWISSEN et al. 1992). O CT15 é um membro da família das proteínas ADAM (A *Desintegrin and Metalloproteinase protein*), que são proteases e/ou moléculas de adesão de superfície celular (VIDAEUS et al. 1997). Neste caso, o CT15 está envolvido na interação entre o óvulo e o espermatozóide, via ligação à integrina $\alpha_6\beta_1$. Já o OY-TES-1 é o precursor da proteína sp32, envolvida no empacotamento da acrosina na cabeça do espermatozóide (BABA et al. 2001). Fora estes antígenos *CT*, os outros só possuem função associada a domínios protéicos conhecidos, encontrados em suas seqüências de aminoácidos. Além disso, estudos funcionais têm sido feitos com membros da família MAGE que são ubiquamente expressos, como por exemplo a necdina e o MAGE-D1 (SALEHI et al. 2000; MATSUMOTO et al. 2001). Os resultados destes trabalhos sugerem que certos membros da família MAGE estejam envolvidos no controle do ciclo celular e apoptose, funções importantes no processo

de formação e desenvolvimento do câncer (TANIURA et al. 1999; JORDAN et al. 2001).

Independentemente de suas funções, dado seu padrão de expressão restrito às espermatogônias, os antígenos *CT* tornam-se um alvo ideal para imunoterapia do câncer. Isto porque as células do testículo não expressam moléculas de HLA classe I e, portanto, não apresentam antígenos para os *CTLs* (FISZER e KURPISZ 1998). Tal característica somada ao bloqueio físico imposto pela barreira hemato-testicular, que impede a passagem de anticorpos para o testículo, torna o mesmo um tecido imunologicamente protegido (DE WIT et al. 2002). Assim, a princípio, vacinas que tenham como alvo estes antígenos não devem reagir com a proteína expressa no testículo, garantindo, pelo menos aparentemente, que apenas as células tumorais sejam destruídas.

Com base nestes dados, diferentes ensaios clínicos têm sido realizados, nos quais antígenos *CT* são utilizados como alvo para o desencadeamento da resposta imune anti-tumoral. Merecem destaque os antígenos NY-ESO-1 e MAGE-3 normalmente utilizados em ensaios clínicos de pacientes com melanoma avançado. Diferentes abordagens terapêuticas têm sido utilizadas, dentre as quais destacam-se: vacinação com peptídeos antigênicos ou com a proteína recombinante, transferência de células T ativadas *in vitro* e transferência de células dendríticas "pulsadas" com o antígeno de interesse (NESTLE et al. 1998; ROSENBERG et al. 1998; COULIE et al. 2002; DUDLEY et al. 2002).

Alguns destes trabalhos mostraram pacientes que, além da ativação do SI, também apresentaram regressão tumoral associada à resposta clínica (MUKHERJI et al. 1995; TITZER et al. 2000; LONCHAY et al. 2004). Deste modo, estes resultados

confirmam o potencial terapêutico destes antígenos e motivam a realização de novos trabalhos que tenham como objetivo a identificação dos mesmos.

1.5 TÉCNICAS NÃO IMUNOLÓGICAS UTILIZADAS NA IDENTIFICAÇÃO DE NOVOS ANTÍGENOS *CT*

Dada a restrita expressão dos antígenos *CT* a testículo normal e tumores, uma busca alternativa por novos antígenos desta categoria tem sido utilizada através de técnicas de bancada e computacionais que permitem avaliar o padrão de expressão gênica em larga escala.

Técnicas de bancada capazes de avaliar a expressão diferencial de transcritos, como, por exemplo, *RDA* (*Representational Difference Analysis*), *differential display* e *cDNA oligonucleotide array*, estão sendo utilizadas e têm gerado bons resultados. Entretanto, técnicas complementares, tais como *RT-PCR (Reverse Transcriptase – Polimerase Chain Reaction)* e *Northern-blot* são necessárias para a confirmação do padrão de expressão dos genes encontrados. Com a aplicação destas técnicas, alguns antígenos *CT* foram identificados, tais como LAGE-1, CTp11 e MMA-1A (LETHE et al. 1998; ZENDMAN et al. 1999; DE WIT et al. 2002).

Dada a grande quantidade de informação gerada pela produção de *ESTs* (*Expressed Sequence Tags*), hoje é possível utilizar os dados de bioinformática para avaliar o padrão de expressão de um determinado gene. *ESTs* são seqüências parciais de transcritos provenientes de uma biblioteca de cDNA construída a partir de um determinado tecido. Desta maneira, em um banco de dados contendo seqüências de diferentes bibliotecas, um mesmo gene pode estar representado por várias *ESTs*.

Conhecendo a origem tecidual das *ESTs*, é possível prever o padrão de expressão de um determinado gene. Desta forma, a seleção de genes que possuem apenas *ESTs* de testículo normal e de qualquer tipo de tumor revelaria candidatos a novos antígenos *CT*. Devido às limitações inerentes à biologia computacional, tais como o número insuficiente de *ESTs* de diferentes tecidos, problemas de clusterização e de anotação de bibliotecas de cDNA, a confirmação do padrão de expressão por técnicas de bancada é necessária para a validação dos candidatos. Utilizando esta abordagem, alguns antígenos foram identificados, como CT9, CT15, CT16 e CT17 (SCANLAN et al. 2002a e b).

Além dos dados de *ESTs*, outra fonte de informação que também pode ser utilizada para a identificação de novos antígenos *CT* são as bibliotecas de SAGE (*Serial Analysis of Gene Expression*). Esta técnica permite a avaliação da expressão gênica através da identificação de uma etiqueta (*tag*) em cada transcrito. A *tag* corresponde a uma seqüência de 10 nucleotídeos 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 *tags* correspondentes aos diferentes transcritos que são expressos no tecido em estudo. O nível de expressão de cada transcrito está diretamente associado ao número de vezes que sua *tag* específica é identificada em uma determinada biblioteca. Para a identificação de antígenos *CT*, pode-se procurar transcritos cujas *tags* estejam presentes apenas em bibliotecas de testículo normal e de tumores. Com o uso desta abordagem, DONG et al. (2003) identificaram 2 novos antígenos *CT*, que inicialmente estavam descritos como genes de expressão restrita a testículo.

Após a descoberta de que a maioria dos antígenos *CT* faz parte de famílias gênicas e dada a disponibilidade da seqüência do genoma humano, outra abordagem pôde ser empregada: a busca de novos candidatos por similaridade de seqüências. Assim, utilizando-se a seqüência de um antígeno já conhecido como "sonda", é possível fazer uma busca em todo o genoma, procurando regiões de alta similaridade. Assim como os dados de *ESTs*, estes resultados precisam ser confirmados experimentalmente. Primeiramente é necessário demonstrar que a região encontrada é transcrita e, portanto, um novo gene. Em seguida, é necessário avaliar o padrão de expressão do transcrito em diferentes tecidos, pois alguns membros da categoria dos antígenos *CT*, como por exemplo NY-ESO-1 e MAGE-A1, possuem genes parálogos com expressão ubíqua em tecidos normais (NY-ESO-3, e MAGE-D respectivamente) (LUCAS et al. 1999; ALPEN et al. 2002).

Embora muito úteis e promissoras na identificação de novos antígenos *CT*, nenhuma destas técnicas não imunológicas comprova que de fato os genes identificados sejam antígenos tumorais. Mesmo apresentando um restrito padrão de expressão, como esperado para um antígeno tumoral, os candidatos só poderão ser considerados como novos antígenos *CT* após a identificação de resposta imune específica contra os mesmos em pacientes com câncer.

1.6 BANCO DE DADOS DO TRANSCRIPTOMA HUMANO E IDENTIFICAÇÃO DE NOVOS GENES NO CROMOSSOMO 21

O projeto ***Transcript Finishing Initiative*** iniciado em fevereiro de 2001 teve como objetivo a identificação de novos genes através do alinhamento entre

seqüências expressas [*ESTs, Open Reading Frame ESTs (ORESTES)* e seqüências completas de mRNAs] e a seqüência do genoma humano (SOGAYAR et al. 2004) (artigo em Anexo). Todos os dados utilizados neste projeto foram armazenados em um banco de dados relacional do Laboratório de Biologia Computacional do Instituto Ludwig de São Paulo. Além disso, foram geradas etiquetas 3' através dos cromatogramas de *ESTs* fornecidos pela *Washington University*. Estas etiquetas também foram mapeadas na seqüência genômica e serviram como âncoras na delimitação entre os genes e na orientação dos transcritos.

Em colaboração com um grupo da Universidade de Genebra-Suíça, o banco do transcriptoma foi utilizado na identificação de 23 novos genes localizados no cromossomo 21 (REYMOND et al. 2002) (artigo em Anexo). A confirmação da existência de tais genes, a obtenção da seqüência completa, caracterização do padrão de expressão dos mesmos e a presença de genes ortólogos em camundongo, foram realizadas neste trabalho.

Um dos novos genes identificados neste trabalho, denominado C21orf99 (AF427490), apresentou um padrão de expressão bastante restrito, sendo abundantemente expresso apenas em testículo dentre 22 tecidos normais analisados. Além disso, este gene revelou uma alta similaridade (87%) com os genes NY-BR-1 e NY-BR-1.1. O NY-BR-1 é um antígeno de diferenciação tumoral, o que nos chamou a atenção para a possibilidade de termos encontrado um terceiro membro desta família gênica.

1.7 O ANTÍGENO NY-BR-1

JAGER et al. (2001) utilizaram a técnica de *SEREX* com o objetivo de identificar novos antígenos de câncer de mama. Dentre os candidatos isolados, um novo gene denominado NY-BR-1 foi identificado (AF269087). Este gene está localizado no cromossomo 10, organizado em 37 *exons* e possui uma região codificadora de 4125 pares de base (pb). A análise da seqüência de aminoácidos mostrou a presença de diferentes domínios protéicos: (a) sinal de localização nuclear; (b) cinco repetições anquirina; e (c) sítio de ligação ao *DNA* seguido de um *leucine zipper motif* (sugerindo se tratar de um fator de transcrição). Além disso, foram identificados três elementos repetitivos com função desconhecida. Formas de *splicing* alternativo também foram encontradas, mas sem modificar a fase de leitura ou afetar os domínios protéicos.

O padrão de expressão gênica do NY-BR-1 foi verificado por *RT-PCR* e, dentre vários tecidos normais, o gene apresentou-se fortemente expresso apenas em mama e testículo. Em amostras tumorais, a expressão do antígeno pôde ser identificada em 21 das 25 (84%) amostras de tumor de mama e, dentre 82 amostras de outros tipos de tumor, apenas 2 amostras de melanoma apresentaram alguma expressão do transcrito. Devido ao seu restrito padrão de expressão, o NY-BR-1 pôde ser classificado como um antígeno de diferenciação de mama. No mesmo trabalho, foi identificado um outro gene, com alta similaridade com o gene descrito, sendo denominado NY-BR-1.1 (AF269088). Entretanto, este outro gene que está localizado no cromossomo 9 apresentou padrão de expressão bastante diferente. Por

ser expresso em vários tecidos normais, o NY-BR-1.1 não foi considerado um antígeno de diferenciação.

Devido à grande importância desta categoria de antígenos tumorais, a identificação do C21orf99 que possui alta similaridade com o NY-BR-1 e o NY-BR-1.1 nos chamou bastante atenção. Assim, este projeto teve como objetivo a caracterização completa deste novo gene, agora denominado CTSP-1. Para tanto, foi necessário avaliar seu padrão de expressão em amostras normais e tumorais, bem como obter sua seqüência completa. Além disso, também foi feita a expressão da proteína recombinante para a produção de anticorpo policlonal em camundongo. Posteriormente, este anticorpo foi utilizado em experimentos de imunohistoquímica e *immunoblotting* para a identificação e avaliação da expressão da proteína CTSP-1 em tecidos humanos. Por último, foi feita a avaliação da presença de resposta imune anti-CTSP-1, através da identificação de anticorpos específicos em plasma de pacientes com câncer.

OBJETIVOS

2 OBJETIVOS

2.1 OBJETIVO PRINCIPAL

- Caracterização do gene CTSP-1 como um novo antígeno tumoral.

2.2 OBJETIVOS SECUNDÁRIOS

- Obtenção da seqüência completa do gene CTSP-1;
- Avaliação do seu padrão de expressão em diferentes tecidos normais, linhagens celulares tumorais e amostras de tumores provenientes de diferentes tecidos;
- Verificação da existência de transcritos alternativos para o gene CTSP-1 e determinação dos respectivos padrões de expressão nos mesmos tecidos citados anteriormente;
- Expressão e purificação da proteína recombinante CTSP-1 em sistema heterólogo (*E. coli*);
- Produção de anticorpo policlonal contra a proteína recombinante CTSP-1 em camundongos;
- Detecção da proteína CTSP-1 por *immunoblotting*;
- Avaliação do padrão de expressão da proteína CTSP-1 em diferentes tecidos normais e tumorais através da técnica de imunohistoquímica;
- Avaliação da presença de anticorpos específicos contra a proteína CTSP-1 em plasma de pacientes com diferentes tipos de câncer.

MATERIAL E MÉTODOS

3 MATERIAL E MÉTODOS

3.1 OBTENÇÃO DA SEQÜÊNCIA COMPLETA DO GENE CTSP-1

3.1.1 Seqüenciamento completo de clones de cDNA

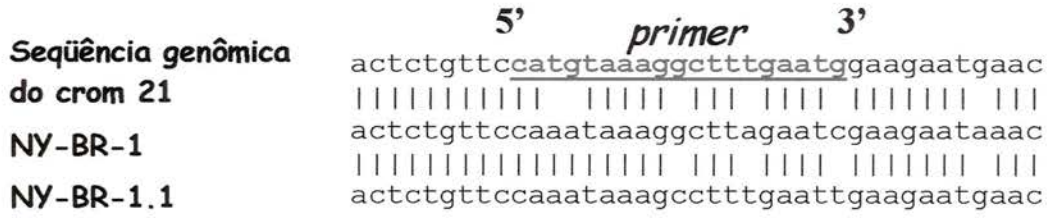
A estratégia inicial utilizada foi o seqüenciamento completo de clones de cDNA correspondentes às *EST*s utilizadas para a identificação de novos genes localizados no cromossomo 21. Para tanto, tais clones foram escolhidos através de ferramentas de buscas disponibilizadas na internet (http://www.rzpd.de/dist/html/clones). Posteriormente, os clones foram solicitados ao RZPD (*Resource Center German Human Genome Project*) e completamente seqüenciados. Inicialmente os insertos foram seqüenciados com *primers* que alinham no vetor (**T7**: 5' TAA TAC GAC TCA CTA TAG GGA GA 3' e **T3**: 5' AAT TAA CCC TCA CTA AAG GGA GA 3') e, quando necessário, foram desenhados *primers* internos e específicos para cada inserto.

3.1.2 Alinhamento dos Genes Parálogos seguido de *RT-PCR*

Através do alinhamento das seqüências dos genes NY-BR-1 e NY-BR-1.1 com a seqüência do clone genômico do cromossomo 21 no qual o CTSP-1 está localizado regiões de alta similaridade puderam ser identificadas. Estes alinhamentos foram feitos através do programa *BLAST*. Desta maneira, foram identificadas regiões de alta similaridade que não estavam cobertas pela seqüência obtida a partir dos clones de cDNA e que, possivelmente, faziam parte do gene CTSP-1. Para a

validação das mesmas, foram feitas *RT-PCRs* com 2 pares de *primers*, que cobriam as regiões de alta similaridade.

As regiões escolhidas para a construção dos *primers* foram as regiões que apresentaram alta similaridade com os dois genes. No entanto, devido à alta similaridade existente entre os genes, os *primers* foram construídos de forma que a extremidade 3' dos mesmos conferisse especificidade ao gene localizado no cromossomo 21 (Figura 1). Além disso, os pares de *primers* foram desenhados de maneira que os fragmentos obtidos apresentassem uma região de sobreposição, facilitando assim a montagem das seqüências. Assim, foram desenhados os seguintes pares de *primers*: **CTSP-F1**: 5' ATA TCT AAA AAT TCT CAA AAT AG 3'; **CTSP-R1**: 5' GCT TTC CCC AAA CAT TGA AC 3'; **CTSP-F2**: 5' AAG ACT GAA TGA GTG GCA G 3'; **CTSP-R2**: 5' CTG ATT CAA ATT ACT TCT TAC AG 3'.



Legenda: A construção foi baseada no alinhamento entre a seqüência genômica do cromossomo 21 e os genes NY-BR-1 e NY-BR-1.1, sendo o nucleotídeo da extremidade 3' específico para a seqüência genômica do cromossomo 21.

Figura 1 - Representação da construção dos *primers* utilizados nas *RT-PCRs* do gene CTSP-1.

Para o fragmento esperado na *RT-PCR* com o primeiro par de *primers* (**CTSP-F1** e **CTSP-R1**), foi necessário a construção de *primers* mais internos para a

realização de *Nested-PCR*, devido à dificuldade de amplificação do mesmo. Assim, foram construídos os seguintes *primers*: **CTSP-F1N**: 5' TGA AAG CTT GGT GGA AAG 3' e **CTSP-R1N**: 5' GTT CCT TCT TCC AAA ACT TC 3'

3.1.2.1 *RT-PCR*

Para a síntese de cDNA foram utilizados 2µg de RNA total de testículo normal e ainda: 0,025µg/µl oligo dT, 0,5mM dNTPs, em 12µl de reação. Esta solução foi incubada a 65°C durante 5 minutos e armazenada em gelo. Em seguida, foram adicionados 1X *First strand buffer*, 0,01M DTT, 200U de *RNAse out-inhibitor* (*Invitrogen®*) e 40U de *SuperScript II* (*Invitrogen®*), em volume final de 20µl. A reação seguiu-se com incubação a 42°C por 1 hora e depois a 70°C por 15 minutos para inativação da enzima.

Antes do cDNA ser utilizado na *PCR* de interesse, foi realizado um teste para avaliar a qualidade do mesmo, no qual fez-se uma *PCR* utilizando-se *primers* específicos para os exons 6 e 7 do gene GAPDH (gliceraldeído desidrogenase): **GAPDH-F**: 5' CTG CAC CAC CAA CTG CTT A 3' e **GAPDH-R**: 5' CAT GAC GGC AGG TCA GGT C 3'. Esta reação foi feita nas seguintes condições: 0,5µl de cDNA, 1,0mM $MgCl_2$, 0,1mM dNTPs, 0,4µM de cada primer e 1U de Taq *DNA* polimerase (*Invitrogen®*), em tampão apropriado e volume final de 20µl. A amplificação foi iniciada com desnaturação a 94°C por 5 minutos, seguida de 22 ciclos de: 1 minuto a 94°C, 45 segundos a 60°C e 1 minuto a 72°C, e extensão final de 10 minutos a 72°C.

Em seguida, a *PCR* para o CTSP-1 foi feita utilizando-se 1µl de cDNA, 1mM $MgCl_2$, 0,1mM dNTPs, 0,5µM de cada *primer* (**CTSP-F1** e **CTSP-R1** ou **CTSP-F2**

e **CTSP-R2**) e 1U de Taq *DNA* polimerase (*Invitrogen®*), em tampão apropriado e volume final de 25µl. A reação iniciou-se com desnaturação a 94°C por 5 minutos, seguida de 40 ciclos de: 1 minuto a 94°C, 45 segundos a 60°C e 1 minuto a 72°C, e extensão final de 10 minutos a 72°C. Na reação de *Nested-PCR* foram utilizados os *primers* citados anteriormente (**CTSP-F1N** e **CTSP-R1N**) e, ao invés de cDNA, foi utilizado como molde 1µl do produto da reação inicial. Ao final, os produtos de amplificação foram visualizados em gel de 8% poliacrilamida, corado com prata (SANGUINETTI et al. 1994).

3.1.2.2 Clonagem e Seqüenciamento

Após a amplificação dos fragmentos desejados, estes foram clonados no vetor PCR® 2.1, utilizando-se o *TA cloning kit* (*Invitrogen®*), seguindo-se as instruções do fabricante. A presença de inserto foi verificada através de *PCR* direto da colônia, com *primers* específicos do vetor: **Forward**: 5' CGC CAG GGT TTT CCC AGT CAC GAC 3' e **Reverse**: 5' TCA CAC AGG AAA CAG CTA TGA C 3'.

As condições de *PCR* foram as mesmas utilizadas na amplificação do gene CTSP-1, entretanto com 30 ciclos de amplificação. O seqüenciamento foi feito a partir do produto desta reação, utilizando-se o *ET-Terminator kit* (*Amersham®*) em seqüenciador automático *ABI 3100* (*Applied Biosystems®*), segundo especificações do fornecedor. Após o seqüenciamento, foi feita uma montagem das seqüências com o programa *PhredPhrapConsed* (GORDON et al. 1998). Para a confirmação da especificidade da amplificação obtida, a seqüência consenso final foi alinhada contra a seqüência genômica humana através da ferramenta *BLAST*.

3.1.3 *RACE (Rapid Amplification of cDNA Ends)*

A técnica de *RACE* permite a amplificação das extremidades 5' ou 3' de seqüências transcritas e também foi utilizada na obtenção da seqüência completa do CTSP-1. Para tanto, foi utilizado o *Marathon™ Amplification cDNA kit* (*Clontech®*) que contém bibliotecas de cDNA fita dupla ligado a adaptadores em ambas extremidades. Primeiramente, foi feita uma *PCR* com o *primer* do adaptador e o *primer* específico do transcrito de interesse (CTSP-1). A reação foi feita nas seguintes condições: 5µl de cDNA, 0,2mM dNTPs, 0,2µM de cada *primer* e 1U de *Advantage Taq DNA polimerase* (*Clontech®*), em tampão apropriado. A reação iniciou-se com desnaturação a 94°C por 1 minuto, seguida de 5 ciclos de: 5 segundos a 94°C e 4 minutos a 70°C, 5 ciclos de 5 segundos a 94°C e 4 minutos a 68°C e, por último, 25 ciclos de 5 segundos a 94°C, 30 segundos a 65°C e 4 minutos a 68°C. Em seguida, foi feita uma reação de *Nested-PCR*, na qual foi utilizada 1µl da primeira reação como molde, nas mesmas condições, mas com *primers* internos aos utilizados na primeira, tanto para o adaptador quanto para a seqüência de interesse. Os *primers* específicos para o CTSP-1, direcionados para a extremidade 3', foram: **SP-RACE3**: 5' CCA TGG CTC ACA CCT GTA ATC TCA TCA C 3' e **SP-RACE3N**: 5' CCA AGG CGG GCA GAT CAT GAC 3'.

Após a amplificação, os produtos foram clonados e seqüenciados como descrito anteriormente. Novamente, foi feita a montagem destas seqüências com o programa *PhredPhrapConsed* e a seqüência consenso foi alinhada contra a seqüência genômica humana através da ferramenta *BLAST*.

3.2 AVALIAÇÃO DA EXISTÊNCIA DE POLIADENILAÇÃO ALTERNATIVA

Esta avaliação foi feita através de *Northern-blot* e devido à necessidade de uma grande quantidade de RNA e à falta de material suficiente para a extração do mesmo, foi realizada a amplificação *in vitro* de RNA mensageiro. A amplificação foi feita seguindo-se o protocolo descrito por WANG et al. (2003), com algumas modificações. Para a reação inicial, de síntese da primeira fita de cDNA, foram utilizados 3µg de RNA total de testículo normal, 0,025µg/µl de oligo dT-T7primer (5' AAA CGA CGG CCA GTG AAT TGT AAT ACG ACT CAC TAT AGG CGC TTT TTT TTT TTT TTT 3'), 0,01M DTT, em 9µl de volume inicial de reação. Esta solução foi incubada a 65°C durante 5 minutos e armazenada em gelo. Em seguida, foram adicionados 1X *First strand buffer*, 1mM dNTPs, 200U de *RNAse out* (*Invitrogen®*), 0,05µg/µl TS primer (5'AAG CAG TGG TAA CAA CGC AGA GTA CGC GGG 3') e 40U de *SuperScript II* (*Invitrogen®*). A reação seguiu-se com incubação a 42° C por 2 horas e denaturação a 70° C por 15 minutos.

A síntese da segunda fita foi feita adicionando-se ao produto da primeira reação, 1X *Advantage cDNA polimerase Mix* (*BD Biosciences®*), 2U de *RNAse H* (*Invitrogen®*) e 0,2mM dNTP, em tampão apropriado e em volume final de 150µl. A reação iniciou-se com a digestão do mRNA a 37°C por 10 minutos, seguindo-se com denaturação a 94°C por 2 minutos, ligação dos *primers* a 65°C por 2 minutos e extensão a 75°C por 30 minutos. A reação foi inativada com 7,5µl da solução 1M NaOH e 2mM EDTA e incubação a 65°C por 10 minutos.

Finalmente, o cDNA dupla fita foi submetido a uma extração por fenol/clorofórmio. A solução aquosa resultante foi transferida para coluna Microcon-YM30 (*Millipore®*) e centrifugada a 7000xg por 6 minutos a temperatura ambiente. A coluna foi lavada 3 vezes com 400µl de água DEPC (Dietil-pirocarbonato) e, ao final, a mesma foi transferida para um tubo novo e invertida, seguindo-se com centrifugação a 1000xg por 3 minutos. Ao final, o volume restante foi ajustado a 20µl em *Speed Vac*.

Todo o cDNA obtido foi utilizado na transcrição *in vitro*, ao qual foram adicionados 7,5mM rNTP, 1X *React Buffer* e 5µl de *Enxime Mix* (*RNA polimerase, Recombinant RNasin ribonuclease inhibitor e Yeast Inorganic Pyrophosphstase - RiboMAX™ Large Scale*) (*Promega®*), em 50µl de reação. A transcrição foi feita a 37°C por 6 horas. Para a purificação do mRNA, foi utilizado o reagente Trizol (*Invitrogen®*), seguindo-se as recomendações do fabricante. Ao final, o mRNA foi ressuspendido em água DEPC e dosado em espectrofotômetro.

Para a realização do *Northern-blot* foi utilizado todo o mRNA amplificado de testículo normal (aproximadamente 6µg). Inicialmente, o gel de agarose foi preparado dissolvendo-se 1g de agarose (*Invitrogen®*) em 87ml de água DEPC e incubado a 55°C em banho maria. Em seguida, foram adicionados 10ml de 10X MOPS [0,2M MOPS (Ácido 3-N-morfolino-propanosulfônico), 50mM Acetato de Sódio, 10mM EDTA, pH 7,0) e 5,1ml de Formaldeído saturado com Carbonato de Cálcio.

O mRNA amplificado foi concentrado em 5µl de água DEPC para aplicação no gel, sendo adicionado ao mesmo 25µl de tampão de amostra (50% Formamida deionizada, 1X MOPS, 6% Formaldeído, 6% Glicerol e traços de Azul de

Bromofenol). Este volume final foi incubado a 65°C por 15 minutos para denaturação do mRNA e posteriormente armazenado em gelo. Antes da aplicação no gel, foi adicionado à amostra 1µl de Brometo de Etídio (1mg/ml).

O gel foi submetido à eletroforese em tampão 1X MOPS, sob a voltagem constante de 60V, por aproximadamente 6 horas. Após a corrida o gel foi lavado 2 vezes em 10X SSC (1,5M NaCl e 0,15M Citrato de Sódio, pH 7,0) durante 20 minutos para a remoção do Formaldeído. Ao final, o gel foi colocado para tranferência em ponte contendo tampão 10X SSC, durante 16 horas. Para a imobilização do mRNA foi utilizada membrana de nylon (*Hybond®*). Após a transferência, a membrana foi lavada em tampão apropriado (2X SSC: 0,6M NaCl e 60mM Citrato de Sódio, pH 7,0) e, em seguida, foi feito o *cross-link* em luz ultravioleta (UV) (120mJ/cm^2).

Foi utilizado como sonda um produto de *RT-PCR*, previamente seqüenciado, que reconhece as 3 formas de poliadenilação do gene CTSP-1 (Figura 8D). A sonda foi marcada com fósforo radioativo através da incorporação de nucleotídeos [α-^{32}P]dCTP (*Amersham®*). A marcação foi feita com *Random Primers DNA Labeling System* (*Invitrogen®*), seguindo as instruções do fabricante. Depois de marcada, a sonda foi purificada em coluna de Sepharose 50 (*Invitrogen®*) para a eliminação dos nucleotídeos não incorporados. Posteriormente, a sonda foi hibridada com as membranas, a 65°C durante 16 horas em solução apropriada (0,25M Na_2HPO_4, 7% SDS, 1% BSA, 1mM EDTA). Após a incubação, as membranas foram lavadas em solução de lavagem (0,25M Na_2HPO_4, 1% SDS, 1mM EDTA) durante 30 minutos a 65°C, por duas vezes. Finalmente, as mesmas foram expostas a filme fotográfico (Kodak®) a -80°C por diferentes períodos de tempo.

3.3 IDENTIFICAÇÃO DA INSERÇÃO DE ELEMENTOS REPETITIVOS NA SEQÜÊNCIA DO GENE CTSP-1

A seqüência nucleotídica final do CTSP-1 foi analisada pelo programa *RepeatMasker* (http://www.repeatmasker.org), o qual identifica elementos repetitivos conhecidos em diferentes espécies.

3.4 ANÁLISE *IN SILICO* DOS DOMÍNIOS PROTÉICOS DA PROTEÍNA CTSP-1

Para a obtenção da seqüência de aminoácidos da proteína codificada pelo CTSP-1, foi feita a tradução da seqüência nucleotídica através de uma ferramenta disponível no site do *Expasy* (http://www.expasy.com). Posteriormente, foi feita uma busca por domínios protéicos na seqüência de aminoácidos através de outra ferramenta, o *PROSITE*, também disponível no site do *Expasy*. Esta ferramenta dispõe de um banco de seqüências de domínios protéicos conhecidos, os quais podem ser identificados na seqüência analisada.

3.5 AVALIAÇÃO DA FUNCIONALIDADE DO GENE CTSP-1

3.5.1 Análises de Mutações Sinônimas e não Sinônimas

Para verificar se o CTSP-1 era um gene funcional, foi feita a comparação entre a seqüência codificante do mesmo e a do NY-BR-1, na qual foram avaliadas as taxas de substituição sinônima (dS) e não sinônima (dN). Os alinhamentos foram

feitos pelo programa *ClustalW* (THOMPSON et al. 1994) e as taxas de substituição estimadas pelo método Nei-Gojobori modificado com distância Jukes-Cantor (NEI e KUMAR 2000). Este método corrige os efeitos de substituições múltiplas e de diferentes taxas de transversão e de transição. Em seguida, o Teste Exato de Fisher foi utilizado para verificar se a razão dN/dS era estatisticamente diferente de 1. Todos os cálculos e testes foram realizados no programa MEGA (versão 2.1) (KUMAR et al. 2001).

3.5.2 Identificação de Gene Ortólogo em Chimpanzé

Inicialmente, o programa *ClustalW* (THOMPSON et al. 1994) foi utilizado para o alinhamento múltiplo entre a *EST* de chimpanzé (CB298911) e as seqüências completas dos genes CTSP-1, NY-BR-1 e NY-BR-1.1, no qual puderam ser identificadas as alterações nucleotídicas entre as seqüências. Em seguida, foram feitos alinhamentos globais pelo programa *Nap e Gap* do pacote **AAT** (HUANG 1994) entre as seqüências dos 3 genes e a seqüência genômica de chimpanzé. Este programa produz alinhamentos de nucleotídeo *versus* nucleotídeo considerando-se os sítios de *splicing* GT-AG e calcula a porcentagem de nucleotídeos idênticos entre as seqüências alinhadas.

3.6 AVALIAÇÃO DO PADRÃO DE EXPRESSÃO DO GENE CTSP-1

3.6.1 Tecidos Normais

Devido à dificuldade de obtenção de tecidos normais, foi obtido comercialmente um painel de RNA de tecidos normais (*Clontech*®). O painel possui

RNA total diluído em água dos seguintes tecidos: testículo, mama, cérebro, próstata, cólon, timo, intestino delgado, pulmão, coração, medula espinhal, baço, cérebro fetal, fígado, fígado fetal, medula óssea, glândula salivar, mama, útero, glândula adrenal, músculo, placenta e rim.

3.6.2 Linhagens Celulares Tumorais

Linhagens celulares tumorais, disponibilizadas pela ATCC® (*American Type Culture Collection*), foram utilizadas para avaliar o padrão de expressão do CTSP-1. Para tanto, foi obtido o maior número de linhagens tumorais possível, dos mais diferentes tipos de tecido. Foram elas: CaSki (carcinoma epidermóide da cérvix uterina), Hela (adenocarcinoma de útero), A172 (glioblastoma), T98G (glioblastoma multiforme), HL-60 (leucemia aguda mielocítica), K562 (leucemia mielóide crônica), IM9 (linfoblasto B, transformado com EBV), H358 (adenocarcinoma de pulmão), H1155 (adenocarcinoma de pulmão), Du145 (carcinoma de próstata), PC3 (adenocarcinoma de próstata), SCABER (carcinoma de células escamosas da bexiga), FaDu (carcinoma de células escamosas da faringe), MDA-436 (adenocarcinoma de mama), MCF-7 (adenocarcinoma de mama) e ZR75.3A (carcinoma ductal mamário), SW-480 (adenocarcinoma coloretal), SaOS-2 (osteossarcoma), A2058 (melanoma), SKmel-28 (melanoma maligno) e HepG2 (carcinoma hepatocelular). Cada linhagem foi cultivada segundo as especificações do fornecedor.

As linhagens celulares foram cultivadas em meio apropriado até obtenção de confluência (aproximadamente $4x10^4$ células/cm^2) e submetidas à extração de RNA pelo método de sedimentação em Cloreto de Césio (CHIRGWIN et al. 1979).

Inicialmente, o meio de cultura foi aspirado e 9ml da solução de lise (4M Isotiocianato de Guanidina, 25mM Citrato de Sódio – pH 7.0, 0.1M β-mercaptoetanol) foram adicionados à garrafa de cultura ($75cm^2$). Em seguida, o lisado celular foi homogeneizado e transferido para um tubo de ultracentrífuga contendo 4ml de solução de Cloreto de Césio (5.7M CsCl e 25mM NaAc) e então centrifugado a 150000xg por 17 horas a 20° C (rotor SW40Ti, *Beckman*®). Após a centrifugação, formou-se um precipitado de RNA e o sobrenadante contendo proteínas e DNA foi descartado. Finalmente, a parede interna do tubo foi limpa e o RNA solubilizado em 50 a 200µl de água DEPC. A dosagem foi feita em espectrofotômetro apropriado a 260nm de comprimento de onda. Além disso, também foi avaliada a razão entre as leituras a 260 e 280nm, a qual indica a pureza do material obtido.

3.6.3 Amostras de Tumores

Amostras de diferentes tipos de tumor foram solicitadas junto ao Banco de Tumores do Hospital A. C. Camargo. Foram utilizadas apenas amostras cedidas ao Banco de Tumores após o consentimento informado dos pacientes e todas as precauções pertinentes para manter o sigilo e a confidencialidade dos dados dos pacientes foram adotadas. Além destes tumores, foram utilizadas também 13 amostras de RNA de glioblastoma gentilmente cedidas pelo Dr. Gregory Riggins da Universidade de Dude – Estados Unidos. Estes glioblastomas são *xenografts* originados a partir da injeção de células tumorais humanas em camundongos. Sendo assim, no total foram obtidas 177 amostras de 15 tumores diferentes conforme descrito na Tabela 1.

Tabela 1 - Amostras tumorais utilizadas para avaliação da expressão do CTSP-1.

Tecido	**Número de amostras**
Cólon	18
Esôfago	5
Estômago	9
Glioblastoma	13
Mama	25
Melanoma	18
Próstata	25
Pulmão	14
Tireóide	24
Útero	20
Hemangioma	2
Linfangioma	1
Ovário	1
Bexiga	1
Rim	1
Total	**177**

A extração de RNA dos tumores foi feita com o reagente Trizol (Invitrogen®). Para tanto, os tumores, ainda congelados, foram imersos em 1ml do reagente Trizol (sobre uma placa de petri estéril) e cortados em pequenos fragmentos com o auxílio de um bisturi estéril. Posteriormente, estes fragmentos foram transferidos para tubos cônicos de poliestireno de 5ml (*Falcon®*) aos quais foi acrescido 1ml do reagente Trizol. Com auxílio de um *Polytron* (*Kinematica® AG*), as amostras foram completamente homogeneizadas e o protocolo foi seguido conforme as especificações do fabricante.

3.6.4 Avaliação da Qualidade dos RNAs Extraídos

A qualidade dos RNAs foi avaliada de duas maneiras: verificou-se se houve degradação do material; e se o mesmo estava contaminado com DNA genômico. A integridade dos RNAs foi visualizada aplicando-se 1μg de RNA total em gel de 1% agarose. Antes de ser aplicado, o RNA foi desnaturado a 65°C por 5 minutos, sendo mantido em condição desnaturante em tampão de amostra contendo Uréia (2X TAE, 30% Glicerol, 7M Uréia, traços de Azul de Bromofenol). A coloração do material foi feita com brometo de etídio e o gel foi visualizado em luz UV. Foram considerados íntegros os *RNA*s que apresentaram as bandas correspondentes aos *RNA*s ribossômicos 28S e 18S bem evidentes e na razão 2:1.

A contaminação com DNA genômico foi verificada através do teste de *hMLH1* (*human mut-L homologue 1*). Este teste consiste em uma *PCR*, na qual utilizam-se 200ng de *RNA* total e *primers* desenhados nos íntrons 12 e 13 do gene *hMLH1*, de maneira que, qualquer amplificação obtida deve-se à contaminação com DNA (tamanho esperado do fragmento: 250pb). Os *primers* utilizados foram: **HMLH-F**: 5' TGG TGT CTC TAG TTC TGG 3' e **HMLH-R**: 5' CAT TGT TGT AGT AGC TCT GC 3'.

As seguintes condições de amplificação foram utilizadas: 1,5mM $MgCl_2$, 0,1mM dNTPs, 0,4μM de cada *primer*, 1U de *Taq DNA polimerase* (*Invitrogen®*), em tampão apropriado. A reação foi iniciada com desnaturação a 94°C por 5 minutos, seguida de 35 ciclos de: 45 segundos a 94°C, 45 segundos a 55°C e 1 minuto a 72°C, e extensão final de 6 minutos a 72°C. Os produtos de amplificação foram visualizados em gel de 8% poliacrilamida corado em prata. As amostras que apresentaram alguma amplificação e portanto, estavam contaminadas com DNA

genômico, foram tratadas com *DNAse* (*Invitrogen®*), segundo instruções do fabricante. Para confirmar se a contaminação havia sido eliminada, o teste de *hMLH1* foi repetido.

3.6.5 *RT-PCR*

A síntese de cDNA e as condições da reação de *PCR* foram feitas conforme descrito no item **3.1.2.1**. Os *primers* utilizados na amplificação do CTSP-1 foram desenhados em exons diferentes para possibilitar a avaliação da expressão do mesmo em amostras de RNA que não foram tratadas com *DNAse* e que, eventualmente, poderiam estar contaminadas com DNA genômico. Além disso, os *primers* também foram construídos considerando-se as formas de *splicing* alternativo que foram identificadas durante a obtenção da seqüência completa do transcrito. Assim, os *primers* foram construídos nos exons 3 e 8, os quais são flanqueadores dos exons que sofrem *splicing* alternativo. Os *primers* contruídos foram: **SPEXP-F**: 5' GCT GTC CAT TAT GCT GTT AAC 3' e **SPEXP-R**: 5' TTT TGA GAA TTT TTA GAT ATC 3'.

Os produtos destas reações foram analisados inicialmente em gel de 8% poliacrilamida corados em prata. Entretanto, para obter-se uma maior sensibilidade e especifidade em nossa análise, os mesmos foram reanalisados por *Southern-blot*.

3.6.6 *Southern-blot*

Inicialmente, os produtos de *RT-PCR* foram fracionados em gel de agarose 2% corado com brometo de etídio e depois transferidos para membranas de nylon (*Hybond®*). Antes da transferência, o gel foi lavado em solução desnaturante (1,5M

NaCl e 1M NaOH) por 45 minutos e, depois, duas vezes em solução de neutralização (0,5M Tris pH 7,5 e 1,5M NaCl) por 30 minutos, sob leve agitação. Após a transferência por capilaridade com tampão 10X SSC, as membranas foram lavadas em tampão apropriado (2X SSC) e em seguida, foi feito o *cross-link* em luz UV ($120mJ/cm^2$).

Como sonda foi utilizado um produto de *RT-PCR* previamente seqüenciado e que cobria todas as formas de *splicing* encontradas (Figura 16). A sonda foi marcada com fósforo radioativo através da incorporação de nucleotídeos [α-^{32}P]dCTP (*Amersham®*). A marcação foi feita com *Random Primers DNA Labeling System* (*Invitrogen®*), seguindo-se as instruções do fabricante. Depois de marcada, a sonda foi purificada e hibridada conforme descrito no item **3.2**.

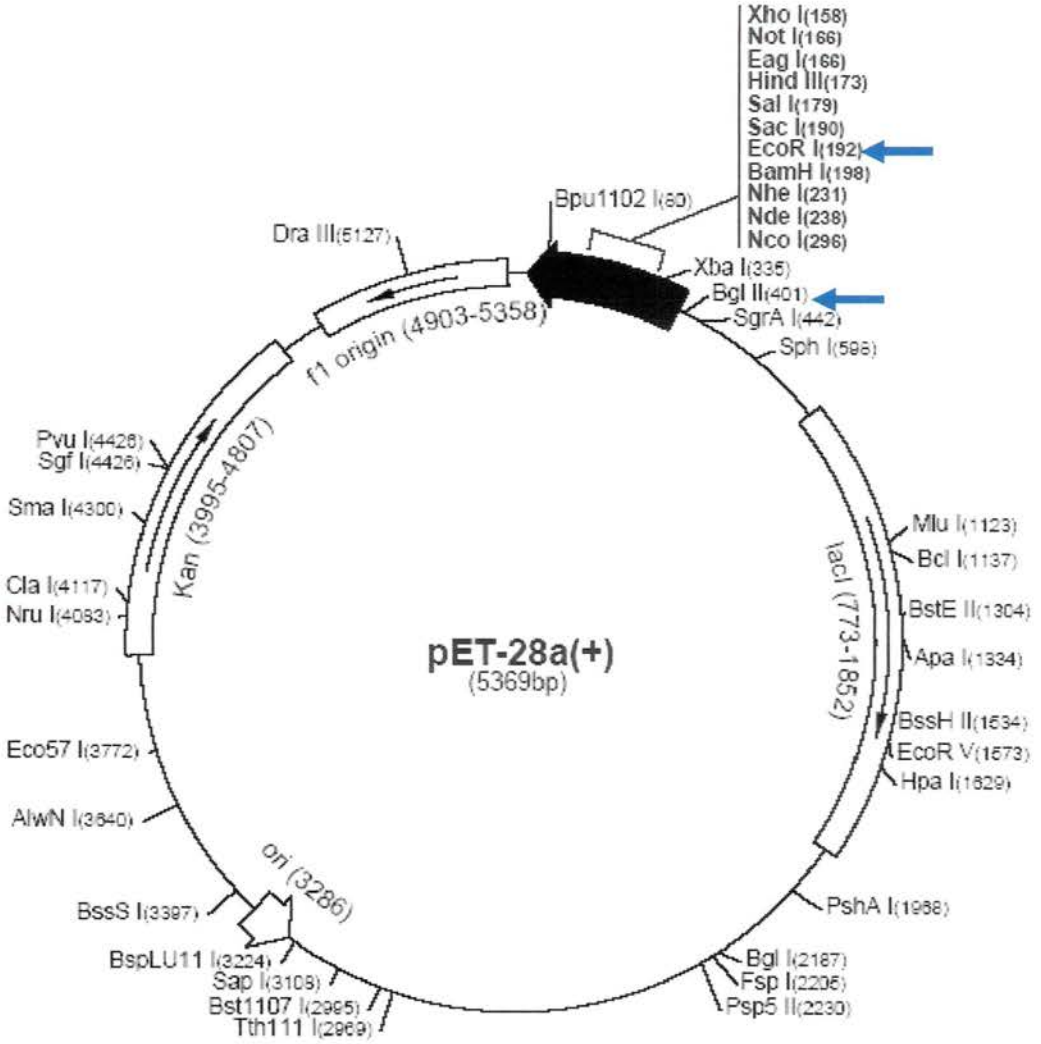
3.7 AVALIAÇÃO DO ENVOLVIMENTO DA METILAÇÃO NO CONTROLE DA EXPRESSÃO DO GENE CTSP-1

A linhagem MCF-7 (derivada de adenocarcinoma de mama) foi submetida ao tratamento com 5'aza-2'deoxicitidina (*Sigma®*). Foram plaqueadas 10^6 células em placa de 75mm e, após 24 horas, foi adicionado meio de cultura contendo 30µM de 5'aza. As células foram tratadas por 48 horas e, ao final, foi feita extração do RNA com o reagente Trizol (*Invitrogen®*), seguindo-se recomendações do fabricante. A *RT-PCR* foi feita conforme descrito no item **3.1.2.1**. Os *primers* utilizados nesta reação foram os mesmos utilizados para a avaliação do padrão de expressão do transcrito (**SPEXP-F** e **SPEXP-R**).

3.8 EXPRESSÃO DA PROTEÍNA RECOMBINANTE CTSP-1

3.8.1 Clonagem da Fase Aberta de Leitura do Gene CTSP-1 em Vetor de Expressão

O vetor de expressão pET28a (*Novagen®*), cujo mapa está esquematizado na Figura 2, foi utilizado para a clonagem do inserto. As enzimas selecionadas para a clonagem sítio dirigida foram: *Bgl* II e *EcoR* I, as quais não possuem sítios de restrição na seqüência do fragmento a ser clonado. Os *primers* utilizados na amplificação do inserto foram construídos de maneira que possuíssem os sítios das enzimas nas posições necessárias (em vermelho), para que, quando o inserto fosse inserido no vetor, este estivesse na fase aberta de leitura correta. São eles: **SPCLONE-F**: 5' CGA GAT CTA TGA AGA AGA CGA CAA TG 3' e **SPCLONE-R**: 5' CGG AAT TCC TAT TCG TCA GGT GTT CT 3'.



Legenda: As setas indicam os sítios de clonagem utilizados.

Figura 2 - Mapa do vetor pET28a.

A síntese de cDNA foi feita conforme descrito no item **3.1.2.1**. A reação de *PCR* para amplificação do inserto de interesse foi feita utilizando-se 1µl de *cDNA*, 1,4mM $MgSO_4$, 0,1mM dNTPs, 0,4µM de cada *primer* e 1U de Taq *Platinum Hi-Fidelity* (*Invitrogen*®), em tampão apropriado. A reação iniciou-se com desnaturação a 94°C por 2 minutos, seguida de 35 ciclos de: 30 segundos a 94°C, 30 segundos a

68°C e 1 minuto e 30 segundos a 68°C, e extensão final de 10 minutos a 68°C. Os produtos de amplificação foram visualizados em gel de 8% poliacrilamida corado em prata.

Após a amplificação, o produto de *PCR* foi purificado a partir de gel de agarose através do *QIAquick Gel Extraction Kit* (*QIAGEN®*), segundo as instruções do fabricante. Em seguida, o fragmento de interesse foi digerido com as enzimas selecionadas para a clonagem, permitindo assim a clonagem sítio dirigida. Nesta reação foram utilizadas 5 unidades de cada enzima, *Bgl* II e *EcoR* I (*New England-Biolabs®*), em tampão apropriado, durante 16 horas a 37°C. Ao final, o produto da digestão foi purificado com o *QIAquick PCR purification Kit* (*QIAGEN®*) seguindo as recomendações do fabricante. Paralelamente, o vetor pET28a também foi digerido com as mesmas enzimas sob as mesmas condições e, posteriormente purificado a partir de gel de agarose.

Cinquenta nanogramas do vetor e 200ng do inserto (CTSP-1), ambos duplamente digeridos com *Bgl* II e *EcoR* I, foram adicionados à uma reação de ligação contendo 200U *T4 DNA ligase* (*New England-Biolabs®*) em tampão apropriado e volume final de 20µl. A reação foi incubada durante 16 horas a 16°C.

Após a ligação, o produto da reação foi dialisado em filtro de nitrocelulose (*Millipore®*) para a remoção do excesso de sal. Em seguida, 2µl do produto final foram utilizados para transformação em bactéria *E. coli* cepa DH10B (*Stratagene®*) eletrocompetente. Após o choque elétrico (resistência de 200Ω, capacitância de 25mF e voltagem de 1.8Kv, em eletroporador *Gene Pulser - BioRad®*), as bactérias foram incubadas durante 1 hora em 1ml de meio LB (1% de triptona; 0,5% de extrato de levedura; 1% de NaCl; pH 7,5), a 37°C sob agitação constante (200rpm). Em

seguida, as bactérias foram sedimentadas através de centrifugação, ressuspensas em 100µl de meio LB e semeadas em placa de LB-ágar (1,5% de ágar), contendo 25µg/ml Kanamicina. Ao final, a placa foi incubada por 16 horas a 37°C em estufa seca.

Algumas colônias isoladas foram coletadas da placa e crescidas em 5ml de LB acrescido de 25µg/ml Kanamicina, por 16 horas a 37°C, para posterior purificação de DNA plasmidial em pequena escala. Os plasmídeos foram purificados através do *Wizard*^{TM} *Mini Preps DNA Purification System* (*Promega*®) e digeridos com as enzimas *Bgl* II e *EcoR* I nas mesmas condições citadas anteriormente, para verificação da liberação do inserto. Os clones que apresentaram um inserto com o tamanho esperado foram então seqüenciados com os *primers* **SPEXP-F** e **SPEXP-R** para a confirmação de sua inserção em fase no vetor. Nas reações de seqüenciamento foi utilizado o *Dynamic*^{TM} *ET terminator cycle sequencing kit* (*Amersham*®), conforme instruções do fabricante e o processamento das mesmas foi realizado no seqüenciador automático *ABI Prism 3100 DNA Sequencer* (*Perkin Elmer*®).

3.8.2 Indução da Expressão da Proteína Recombinante

O plasmídeo contendo a seqüência confirmada foi utilizado para transformação de bactérias *E. coli*, cepa *BL-21 A494* (*Stratagene*®) quimiocompetentes. Inicialmente, 50ng do plasmídeo foram adicionados às bactérias e incubados em gelo durante 20 minutos. Após o choque térmico a 42°C por 1 minuto, as bactérias foram incubadas por 1 hora em 1ml de meio LB a 37°C. Em seguida, foram semeadas em placa de LB-ágar (1,5% de ágar), contendo 25µg/ml Kanamicina, e incubadas por 16 horas a 37°C em estufa seca.

Três colônias transformantes foram inoculadas e incubadas durante toda a noite a 37°C em 5ml de meio LB contendo 25µg/ml Kanamicina. A partir deste pré-inóculo, as bactérias foram crescidas em 10ml de meio LB, contendo Kanamicina, até atingirem a $D.O._{600} = 0.6$. Neste momento, a expressão da proteína recombinante foi induzida através da adição de 0,4mM IPTG (β-D-tiogalactopiranosídeo de Isopropila), por 4 horas a 37°C. Um volume de 3ml de meio contendo as bactérias foi centrifugado a 4000xg, por 10 minutos, a 4°C. O sobrenadante foi descartado e o *pellet* solubilizado em 300µl de tampão de amostra (240mM Tris pH 6,8, 0,8% SDS, 200mM β-mercaptoetanol, 40% Glicerol e traços de Azul de Bromofenol). Os extratos brutos foram então fracionados em gel de 12% poliacrilamida com SDS, posteriormente corado em solução de Coomassie blue [50% metanol, 10% ácido acético e 0,25% Coomassie R250 (*Amersham®*)].

3.8.3 Detecção da Expressão da Proteína Recombinante através de *Immunoblotting*

As proteínas do extrato bruto de bactéria foram separadas através de eletroforese em gel de 12% poliacrilamida com SDS e, em seguida, transferidas para membranas de nitrocelulose (0,45µm - *Schleicher & Schuell*) durante 1 hora sob corrente constante de $0,8mA/cm^2$ no sistema *Nova blot* (*Amersham®*) em tampão de transferência (39mM Glicina, 48mM Tris pH 7,4, 0,037% de SDS e 20% de metanol). Para avaliação da eficiência de transferência, a membrana de nitrocelulose foi corada com *Ponceau* (*Sigma®*) para a visualização das proteínas.

As membranas foram então bloqueadas durante 1 hora a temperatura ambiente com PBST (137mM NaCl, 2,5mM KCl, 10mM Na_2PO_4, 2mM KH_2PO_4 e

0,05% de Tween 20) contendo 5% de leite desnatado liofilizado (*Molico®, Nestlè*). Posteriormente as membranas foram incubadas com anticorpo primário anti-His, conjugado com peroxidase (*Invitrogen®*) na diluição de 1:5000 em PBST contendo 5% de leite, por 1 hora a temperatura ambiente. Após 3 lavagens com PBST, de 10 minutos cada, a reação foi revelada com a solução do *ECL Western blotting analysis system* (*Amersham®*) e as membranas expostas a filmes radiográficos (*Kodak®*).

3.8.4 Teste de solubilidade da Proteína Recombinante

Para o teste de solubilidade da proteína recombinante foram utilizados 5ml da cultura de células após a indução da expressão da proteína recombinante, conforme descrito anteriormente. As bactérias foram sedimentadas por centrifugação a 10000xg, por 2 minutos a temperatura ambiente. O *pellet* de células foi então solubilizado em 10ml de tampão de sonicação (100mM NaCl, 10mM Tris pH 8,0, 50mM NaH_2PO_4).

Posteriormente, as células foram lisadas com o auxílio de um sonicador, através de 3 pulsos de 20mA de 1 minuto cada. Todo este procedimento foi feito em gelo. Após a sonicação, o lisado celular foi centrifugado a 16000xg, por 5 minutos a 4°C. O sobrenadante, contendo as proteínas solúveis, e o *pellet* contendo as proteínas insolúveis foram separados e reservados. Aos mesmos foi adicionado tampão de amostra (240mM Tris pH 6,8, 0,8% SDS, 200mM β-mercaptoetanol, 40% Glicerol e traços de Azul de Bromofenol) e os produtos finais foram fracionados em gel de 12% poliacrilamida com SDS e corado em solução de Coomassie blue.

3.8.5 Purificação da Proteína Recombinante sob Condições Desnaturantes

Para a produção de uma grande quantidade de proteína recombinante, o clone de bactéria que apresentou uma melhor expressão dentre os 3 analisados, foi crescido em 250ml de meio LB e a expressão da proteína foi induzida nas mesmas condições descritas anteriormente. Em seguida, a cultura de células foi centrifugada a 16000xg, por 10 minutos a 4°C. O *pellet* de células foi solubilizado em 30ml de tampão de sonicação. Posteriormente, as células foram lisadas com o auxílio de um sonicador, através de 5 pulsos de 20 mA de 1 minuto cada. Todo este procedimento foi feito em gelo. Após a sonicação, o lisado foi centrifugado a 16000xg, por 10 minutos a 4°C. Como a proteína estava expressa na forma insolúvel, o *pellet*, contendo os corpos de inclusão, foi reservado e o sobrenadante descartado. Para a solubilização das proteínas insolúveis, o *pellet* foi ressuspendido em 10ml de tampão de sonicação contendo 6M uréia. A ressuspensão foi feita com o auxílio de agitador magnético (sob baixa rotação) em um béquer mantido em gelo, durante 2 horas.

Em seguida, a solução, já bem homogênea, foi centrifugada em tubo corex a 16000xg, por 10 minutos a 4°C. Após a centrifugação, o sobrenadante foi reservado para posterior purificação da proteína recombinante em resina de agarose carregada com Níquel (*Ni-NTA resin – Invitrogen®*). Assim, 1ml de resina foi empacotado em uma seringa de 5ml contendo lã de vidro. A coluna foi equilibrada com tampão de sonicação-uréia e, posteriormente, o sobrenadante contendo a proteína recombinante foi aplicado na mesma. Em seguida, a coluna foi lavada com 20ml do mesmo tampão. Após esta lavagem, seguiram-se eluições seriadas utilizando diferentes concentrações de Imidazol (*Sigma®*). As eluições foram feitas com 5ml de tampão de sonicação-uréia contendo 10, 25, 50, 100 e 250mM Imidazol. Para a remoção

completa da proteína recombinante da resina, ao final foram adicionados 10ml de tampão Imidazol 250mM.

As frações foram coletadas separadamente e, em seguida, analisadas em gel de 12% poliacrilamida com SDS corado com Coomassie blue. Após verificar quais frações continham a maior parte da proteína recombinante, as mesmas foram reunidas e dialisadas para a remoção parcial da Uréia e total do Imidazol. A diálise foi feita em membrana de celulose (*Dialysis Tubing-12KDa-Sigma®*) contra o mesmo tampão de sonicação-uréia, mas reduzindo-se gradativamente a concentração de uréia (4M, 2M e 1M) a cada 6 horas de diálise. A diálise foi feita a 4°C e a concentração final de úreia alcançada foi de 1M, sem qu houvesse precipitação da proteína.

Ao final da diálise foi feita a dosagem da proteína com o auxílio do reagente de Bradford (*BioRad®*). A dosagem foi feita a 595nm em leitor de ELISA (*Microplate Reader Benchmark - BioRad®*).

3.9 PRODUÇÃO DE ANTICORPOS ANTI-CTSP-1 EM CAMUNDONGOS

3.9.1 Animais

Foram utilizadas fêmeas das linhagens C57 e Swiss, sendo 3 de cada, com média de 3 meses de idade, mantidas no biotério do Instituto Ludwig.

3.9.2 Imunização

Em cada animal foram injetados intraperitonealmente 25ug de proteína recombinante CTSP-1 diluída em 80µl de água, acrescidos de 100µl de adjuvante completo de *Freunds* (*Sigma®*) e 20µl Hidróxido de Alumínio (60mg/ml, *EMS®*). Após 3 semanas, foi feito o "reforço" com a mesma quantidade de proteína e de Hidróxido de Alumínio, mas desta vez com adjuvante incompleto de *Freunds* (*Sigma®*). Um segundo reforço foi feito após 3 semanas.

O sangramento dos animais foi realizado após 2 semanas do último reforço, via bulbo ocular com auxílio de pipeta *Pasteur* de vidro. Foram coletados, em média, 100µl de sangue por animal. O soro foi obtido após coagulação a 37ºC por 30 minutos seguida de centrifugação de 10000xg, por 1 minuto à temperatura ambiente.

3.9.3 Titulação de Anticorpos Presentes no Soro dos Animais Imunizados

3.9.3.1 ELISA (*Enzime-linked immunosorbent assay*)

Inicialmente, a proteína recombinante foi adsorvida em placa de 96 wells (*DYNEX Immulon® 2HB*), sendo utilizado 1µg de proteína recombinante por poço. A adsorção foi feita a 4ºC por 16 horas, com a proteína diluída em tampão Carbonato (0,015M Carbonato de Sódio e 0,031M Bicarbonato de Sódio, pH 9,6), sendo aplicado 100µl por poço. Em seguida, a placa foi lavada com PBST (1X PBS, 0,05% Tween 20) e bloqueada com solução de PBST contendo 5% BSA (150µl por poço), por 1 hora à temperatura ambiente.

Após o bloqueio, a placa foi lavada novamente com PBST e 100µl do soro dos camundongos já diluído em PBST contendo 0,1% BSA foi adicionado aos poços. Foram feitas 7 diluições seriadas para cada soro: 1:2000, 1:4000, 1:8000, 1:16000,

1:32000. 1:64000 e 1:128000. A incubação do anticorpo com a proteína recombinante foi feita por 1 hora à temperatura ambiente. A placa foi então submetida a extensivas lavagens com PBST. O anticorpo secundário, anti-mouse IgG, conjugado com peroxidase (*Anti-mouse IgG, HRP-linked whole antibody from sheep - Amersham®*), foi adicionado na diluição 1:5000 em PBST contendo 0,1% BSA (100µl por poço) e incubado por 1 hora à temperatura ambiente.

Após uma última lavagem com PBST, a reação foi revelada com 100µl de solução reveladora por poço (40mM Na_2HPO_4, 27mM $C_6H_8O_7$, 0,35mg/ml OPD, 0,006% H_2O_2, pH 5,0), incubando-se por aproximadamente 5 minutos no escuro. A reação foi interrompida com solução de 0,2N Ácido Sulfúrico (*Merck®*). A dosagem foi feita em leitor de ELISA (*Microplate Reader Benchmark - BioRad®*) a 492nm.

3.9.3.2 *Immunoblotting*

A proteína recombinante purificada foi fracionada através de eletroforese em gel de 12% poliacrilamida com SDS, em seguida, transferida e imobilizada em membrana de nitrocelulose como descrito no item **3.8.3**. As membranas foram bloqueadas durante 1 hora à temperatura ambiente com PBST contendo 5% de leite. Posteriormente, as membranas foram incubadas por 1 hora à temperatura ambiente com o soro dos camundongos nas diluições 1:10000, 1:20000, 1:40000, 1:80000 e 1:160000 em PBST contendo 5% de leite. Após 3 lavagens com PBST, de 10 minutos cada, o anticorpo secundário anti-mouse IgG, conjugado com peroxidase (*Amersham®*) foi adicionado na diluição 1:5000 em PBST e incubado por 1 hora a temperatura ambiente. Novamente, após 3 lavagens com PBST, a reação foi revelada e a membrana exposta a filmes radiográficos, como descrito no item **3.8.3**.

3.10 DETECÇÃO DA PROTEÍNA CTSP-1 EM AMOSTRAS DE TECIDO HUMANO

3.10.1 *Immunoblotting*

Uma amostra de testículo normal, congelada a -70°C, foi utilizada para a detecção da proteína CTSP-1. Com auxílio de um bisturi estéril, a mesma foi fragmentada em pequenas partes, às quais foram adicionados 200µl de tampão de amostra (240mM Tris pH 6,8, 0,8% SDS, 200mM β-mercaptoetanol, 40% Glicerol e traços de Azul de Bromofenol). Após homogeneização quase completa, a amostra foi incubada em banho seco a 95°C por 10 minutos. Em seguida, os fragmentos de tecido remanescentes foram sedimentados por centrifugação e o extrato protéico em solução foi fracionado através de eletroforese em gel de 12% poliacrilamida com SDS e, em seguida, transferido e imobilizado em membrana de nitrocelulose, como descrito no item **3.8.3.**

A dosagem da massa de proteína aplicada não pôde ser feita com o auxílio do reagente de Bradford devido ao fato do tampão utilizado na lise celular conter o corante Azul de Bromofenol e ainda uma grande quantidade de detergente. A estimativa da massa utilizada foi feita após a transferência e a membrana corada com Ponceau (*Sigma®*). Assim, a massa do extrato de interesse foi comparada com a massa de outro extrato protéico do mesmo tecido com concentração conhecida, obtido com tampão de lise 0,05% NP40. Desta maneira, estima-se que a massa aproximada de proteína total fracionada foi de 100µg.

A membrana foi então bloqueada por 1 hora a temperatura ambiente, com PBST contendo 5% de leite. Posteriormente, a membrana foi incubada com o soro de

camundongo na diluição 1:10000 em PBST, por 1 hora a temperatura ambiente. Após 3 lavagens com PBST, de 10 minutos cada, o anticorpo secundário anti-mouse IgG, conjugado com peroxidase (*Amersham®*) foi adicionado na diluição 1:5000 em PBST contendo 5% de leite e incubado por 1 hora à temperatura ambiente. Ao final, após 3 lavagens com PBST, a reação foi revelada com a solução do *ECL Western Blotting Analysis System* (*Amersham®*) e as membranas expostas a filmes radiográficos *Hyperfilm™ ECL™* (*Amersham®*).

3.10.2 Imunohistoquímica

Para a padronização e análise inicial das reações de imunohistoquímica foram utilizadas as seguintes amostras: uma amostra de testículo normal, duas amostras de tumor de próstata com os respectivos tecidos normais pareados e duas amostras de tumores de mama com seus respectivos tecidos normais pareados. A partir dos blocos de parafina das amostras selecionadas foram feitos cortes de 3-4μm em micrótomo, os quais foram depositados em lâminas de vidro previamente tratadas com 3-aminopropyltriethoxysilano (*ERV-SFPLUS, Erviegas®*) e deixados por 24 horas em estufa a 60°C. As lâminas foram tratadas com xilol a 60°C por 20 minutos e à temperatura ambiente por mais 20 minutos. Em seguida, as lâminas foram incubadas por 30 segundos em concentrações decrescentes de etanol (100%, 95%, 80% e 70%) e lavadas em água destilada. A seguir, as lâminas foram tratadas com solução fervente de Ácido Cítrico 10mM pH 6,0 em panela de pressão. As lâminas foram lavadas em água destilada e em seguida tratadas com 3% H_2O_2 por 5 minutos, repetindo-se o processo por 4 vezes. As lâminas foram novamente lavadas com água destilada e posteriormente com solução de PBS pH 7,4 por 5 minutos. As lâminas

foram incubadas por 18 horas a 4°C em câmara úmida com o anticorpo primário diluído (1:2000) em PBS contendo 1% BSA e 0,1% Azida Sódica (NaN3). O soro utilizado como anticorpo primário foi do camundongo C57 que apresentou o maior título de anticorpos específicos. Como controle negativo da reação foram utilizados soro irrelevante de camundongo C57 e o soro imune depletado de IgG anti-CTSP-1, conforme será descrito no item a seguir.

Após a incubação, as lâminas foram lavadas 3 vezes com PBS, por 5 minutos cada. Posteriormente, as lâminas foram incubadas a 37°C por 30 minutos com anticorpo secundário biotinilado anti-IgG de camundongo (*StreptABComplex/HRP Duet Kit - DAKO®*, Dinamarca), diluído 1:200 em PBS. As lâminas foram lavadas 3 vezes em PBS por 5 minutos cada, e incubadas por 30 minutos a 37°C com peroxidase acoplada a estreptavidina, diluída 1:200 em PBS. Após lavagem em PBS, as lâminas foram incubadas por 5 minutos a 37°C no escuro com solução contendo o substrato 3,3'-Diaminobenzidina Tetrahidrocloreto (DAB) (0,6mg/ml) (*Sigma®*), 1,4M Dimetilsulfóxido (DMSO), 1% H_2O_2 (20 volumes) em PBS pH 7,4. Após a reação, as lâminas foram lavadas com água destilada por 3 minutos, contracoradas com Hematoxilina de Harris por 1 minuto, e novamente lavadas com água destilada. As lâminas foram imersas em 0,5% Hidróxido de Amônio e lavadas com água destilada. Em seguida, as lâminas foram desidratadas em concentrações crescentes de etanol (80%, 95% e 100%), por 30 segundos em cada concentração e tratadas com xilol 4 vezes, durante 30 segundos cada.

3.10.3 Depleção de IgG Anti-CTSP-1 do Soro de Camundongo

Inicialmente, 10µL do soro de camundongo utilizado na IHQ foram diluídos em PBS em volume final de 1ml. Em seguida, incubou-se este soro diluído por 2 horas à temperatura ambiente sob agitação constante com 500µL da resina NiNTA (*Invitrogen®*) carregada com 60µg de proteína recombinante CTSP-1. Ao final da incubação, a solução foi centrifugada e o sobrenadante armazenado. No precipitado restou a resina ligada ao complexo proteína recombinante-anticorpo específico, os quais também foram armazenados e utilizados em experimentos de *immunoblotting* para a confirmação da ligação do anticorpo à proteína recombinante.

Em seguida, imobilizou-se a proteína recombinante CTSP-1 em membrana de nitrocelulose, a qual foi incubada com o soro depletado para verificar se ainda restava anticorpos específicos anti-CTSP-1 no mesmo. Para tanto, a reação de *immunoblotting* foi feita conforme descrito no item **3.9.3.2**. Além disso, o complexo proteína recombinante-anticorpo foi removido da resina com tampão de amostra SDS-βmercaptoetanol e, em seguida, imobilizado em membrana de nitrocelulose. Assim, esta membrana foi incubada com anticorpo anti-IgG de camundongo (*Amersham®*) para a confirmação da ligação dos anticorpos específicos presentes no soro à recombinante conjugada na resina. Para confirmar a especificidade da reação de IHQ feita anteriormente, o soro depletado também foi utilizado em experimentos com os mesmos tecidos utilizados.

3.11 DETECÇÃO DE ANTICORPOS ANTI-CTSP-1 EM PLASMA DE PACIENTES

3.11.1 Amostras Utilizadas

Amostras de plasma de pacientes com diferentes tipos de tumores foram obtidas junto ao banco de tumores do Hospital A.C. Camargo. Foram utilizadas 148 amostras (Tabela 2) das quais 126 também tiveram a expressão do gene CTSP-1 analisada por *RT-PCR*.

Tabela 2 - Amostras de plasma de pacientes utilizadas na detecção de anticorpo anti-CTSP-1.

Tecido	**Número de amostras**
Cólon	20
Esôfago	4
Estômago	8
Mama	18
Melanoma	23
Próstata	24
Pulmão	13
Tireóide	10
Útero	22
Hemangioma	2
Linfangioma	1
Ovário	1
Bexiga	1
Rim	1
Total	**148**

Foram utilizados como controle negativo plasma de 50 indivíduos sadios, obtidos a partir de doadores do banco de sangue do Hospital A.C.Camargo.

3.11.2 ELISA

A proteína recombinante foi adsorvida em placa de 96 wells (*DYNEX Immulon® 2HB*), na qual foram utilizados 250ng da proteína por poço. A adsorção foi feita a 4°C por 16 horas, com a proteína diluída em tampão Carbonato (0,015M Carbonato de Sódio e 0,031M Bicarbonato de Sódio, pH 9,6), sendo aplicado 100µl de tampão por poço. Em seguida, a placa foi lavada com PBST e bloqueada por 1 hora à temperatura ambiente com solução de PBST contendo 5% BSA (150µl por poço).

Após o bloqueio, a placa foi lavada com PBST e 100µl do plasma dos pacientes diluído em PBST contendo 0,1% BSA foram adicionados aos poços. Foram feitas 4 diluições seriadas para cada plasma: 1:100, 1:400, 1:1600 e 1:6400. A incubação do anticorpo com a proteína recombinante foi feita por 1 hora à temperatura ambiente. A placa foi então submetida a extensivas lavagens com PBST. O anticorpo secundário, anti-IgG humana, conjugado com peroxidase (*Amersham®*), foi adicionado na diluição 1:10000 em PBST contendo 0,1% BSA (100µl por poço) e incubado por 1 hora à temperatura ambiente.

Após uma última lavagem com PBST, a reação foi incubada no escuro com 100µl de solução reveladora por poço (40mM Na_2HPO_4, 27mM $C_6H_8O_7$, 0,35mg/ml OPD, 0,006% H_2O_2, pH 5,0). Após aproximadamente 5 minutos, a reação foi interrompida com solução de 0,2N Ácido Sulfúrico. A dosagem foi feita em leitor de ELISA (*Microplate Reader Benchmark - BioRad®*) a 492nm.

3.11.3 *Immunoblotting*

A proteína recombinante purificada foi fracionada através de eletroforese em gel de 12% poliacrilamida com SDS, em seguida, transferida e imobilizada em membrana de nitrocelulose, como descrito no item **3.8.3**. Em cada canaleta do gel foi aplicado 0,5μg da proteína recombinante.

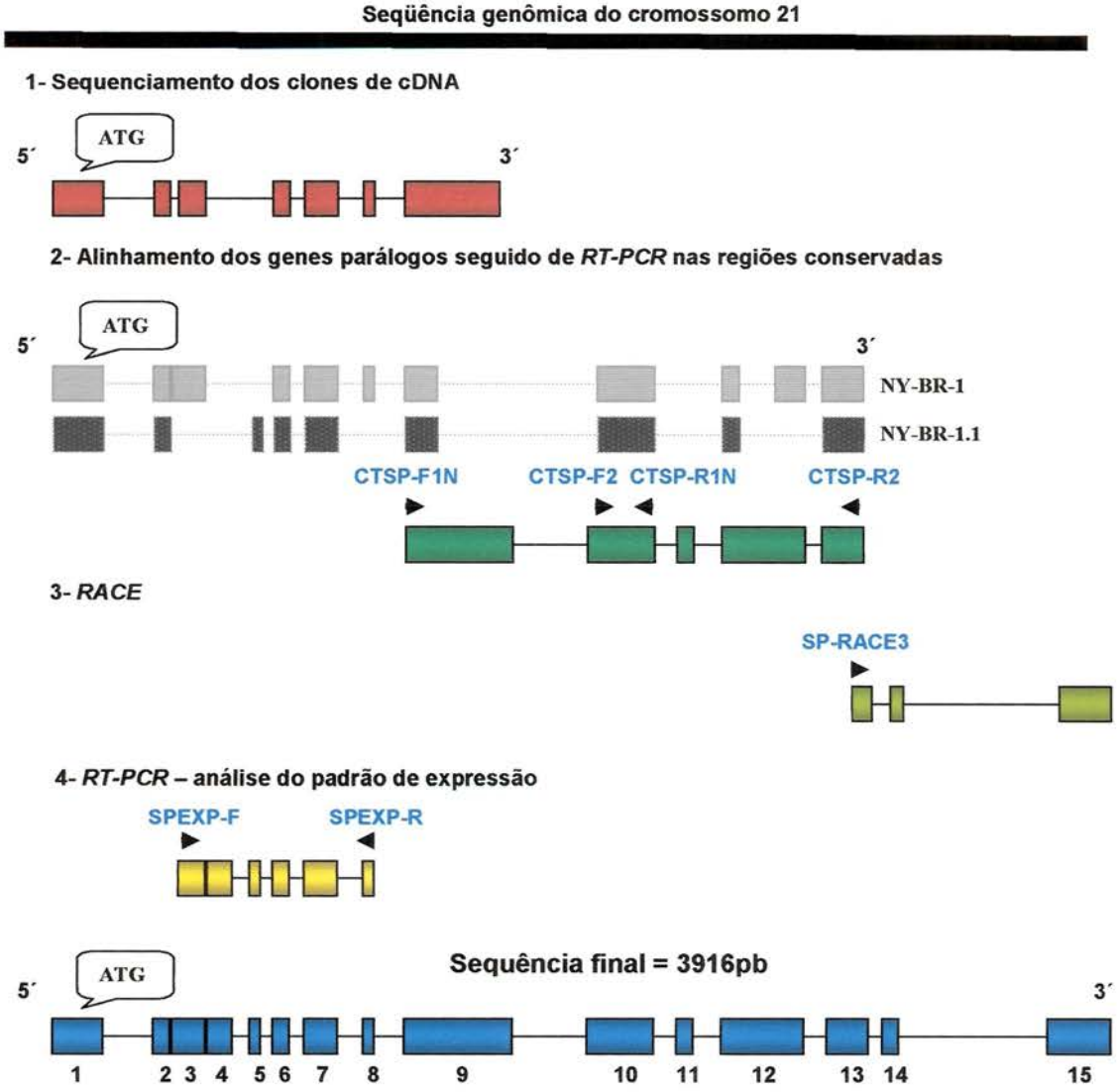
As membranas foram bloqueadas durante 1 hora à temperatura ambiente com PBST contendo 5% leite. Posteriormente, as membranas foram incubadas por 1 hora à temperatura ambiente com o plasma dos pacientes na diluição 1:25 em PBST. Após 3 lavagens de 10 minutos cada com PBST, o anticorpo secundário anti-IgG humana conjugado com peroxidase (*Amersham®*) foi adicionado na diluição 1:10000 em PBST contendo 5% de leite e incubado por 1 hora à temperatura ambiente. Ao final, após 3 lavagens com PBST, a reação foi revelada e as membranas expostas a filmes radiográficos *Hyperfilm™ ECL™* (*Amersham®*).

RESULTADOS E DISCUSSÃO

4 RESULTADOS E DISCUSSÃO

4.1 OBTENÇÃO DA SEQÜÊNCIA COMPLETA DO GENE CTSP-1

Diferentes metodologias foram empregadas para a obtenção da seqüência completa do CTSP-1. Após a identificação do transcrito C21orf99 (posteriormente denominado CTSP-1) no banco de dados do transcriptoma, foram analisadas todas as *ESTs* correspondentes ao mesmo. Assim, 2 clones de cDNA que deram origem às *ESTs* foram selecionados e solicitados ao centro RZPD. Após o seqüenciamento completo dos clones de cDNA, foi obtida uma seqüência consenso de 1272pb. Quando esta foi alinhada à seqüência genômica do cromossomo 21, verificamos que a mesma estava distribuída em 7 exons (Figura 3).



Legenda: As setas indicam os *primers* utilizados na amplificação dos fragmentos. Estão representados apenas os maiores fragmentos obtidos em cada amplificação.

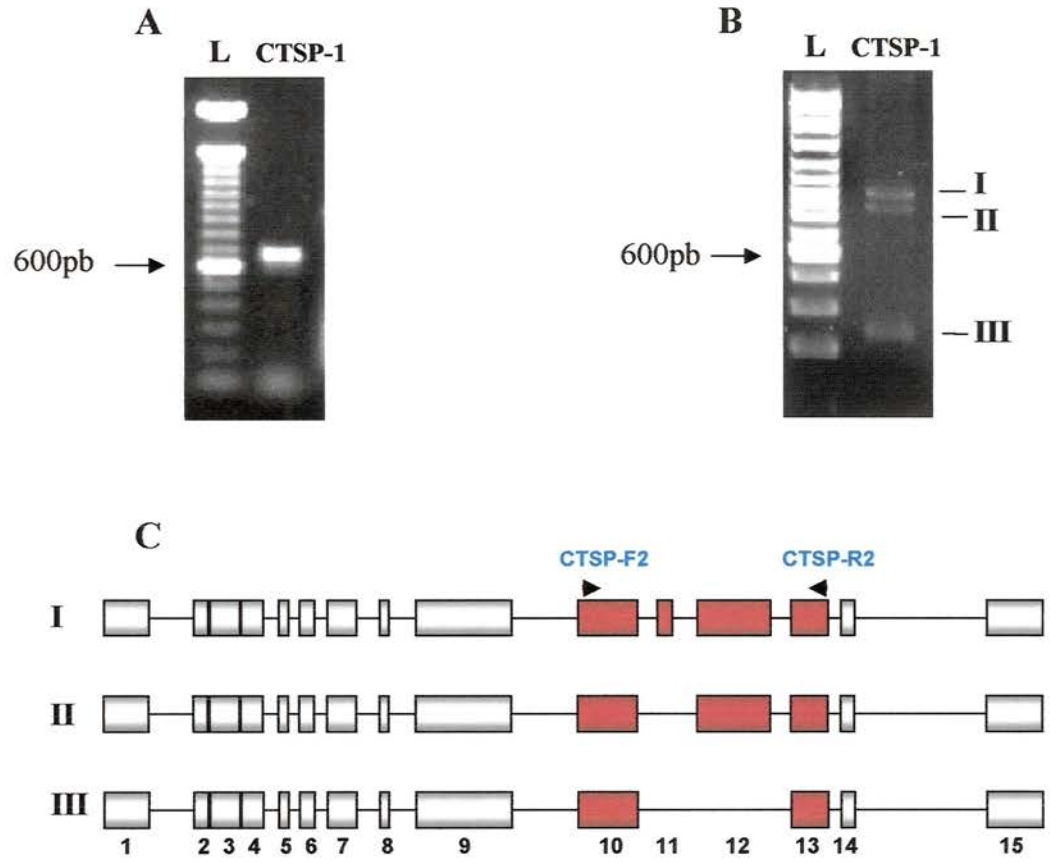
Figura 3 - Estratégias utilizadas para a obtenção da seqüência completa do gene CTSP-1.

Quando a seqüência consenso dos clones de cDNA foi comparada à seqüência dos genes parálogos NY-BR-1 e NY-BR-1.1, verificou-se que a mesma correspondia apenas à porção 5' dos parálogos (Figura 3). Além disso, quando os genes NY-BR-1 e NY-BR-1.1 foram alinhados à seqüência genômica do

cromossomo 21, algumas regiões de alta similaridade, que não estavam cobertas pela seqüência consenso dos clones de cDNA, foram identificadas, sugerindo que a seqüência consenso fosse uma seqüência parcial do gene CTSP-1 (Figura 3). Assim, visando extender a seqüência consenso dos clones de cDNA, dois pares de *primers* (**CTSP-F1** e **CTSP-R1**; **CTSP-F2** e **CTSP-R2**) foram desenhados nas regiões de alta similaridade entre a seqüência genômica do cromossomo 21 e os 2 parálogos. Conforme descrito em material e métodos, os *primers* foram desenhados de maneira que a extremidade 3' fosse específica para a seqüência genômica do cromossomo 21, evitando assim a amplificação dos genes parálogos (Figura 1).

O fragmento amplificado com o primeiro par de *primers* (**CTSP-F1** e **CTSP-R1**) foi de difícil obtenção, sendo necessária a realização de *Nested-PCR* com os *primers* **CTSP-F1N** e **CTSP-R1N**. Ao final, um produto específico de 700pb foi obtido (Figura 4A). Já para o segundo par de *primers* (**CTSP-F2** e **CTSP-R2**) a amplificação foi feita com apenas uma reação, sempre com a presença de 3 bandas evidentes, de aproximadamente 300pb, 1000pb e 1100pb (Figura 4B).

Os fragmentos obtidos em ambas reações foram clonados e seqüenciados. Para a confirmação da especificidade dos mesmos, suas seqüências foram alinhadas à seqüência genômica humana, onde verificamos uma maior similaridade com o clone genômico do cromossomo 21 (AL163202). Os três fragmentos amplificados com os *primers* **CTSP1-F2** e **CTSP1-R2** tiveram sua especificidade confirmada e correspondem a formas de *splicing* alternativo do gene (Figura 4C).



Legenda: A- Produto da *Nested-PCR* com os *primers* **CTSP-F1N** e **CTSP-R1N**. B- Produto da *RT-PCR* com os *primers* **CTSP-F2** e **CTSP-R2**. O padrão de peso molecular utilizado foi o 100bp (L). C- Variantes de *splicing* identificadas após o seqüenciamento dos fragmentos obtidos na amplificação com os *primers* **CTSP-F2** e **CTSP-R2**.

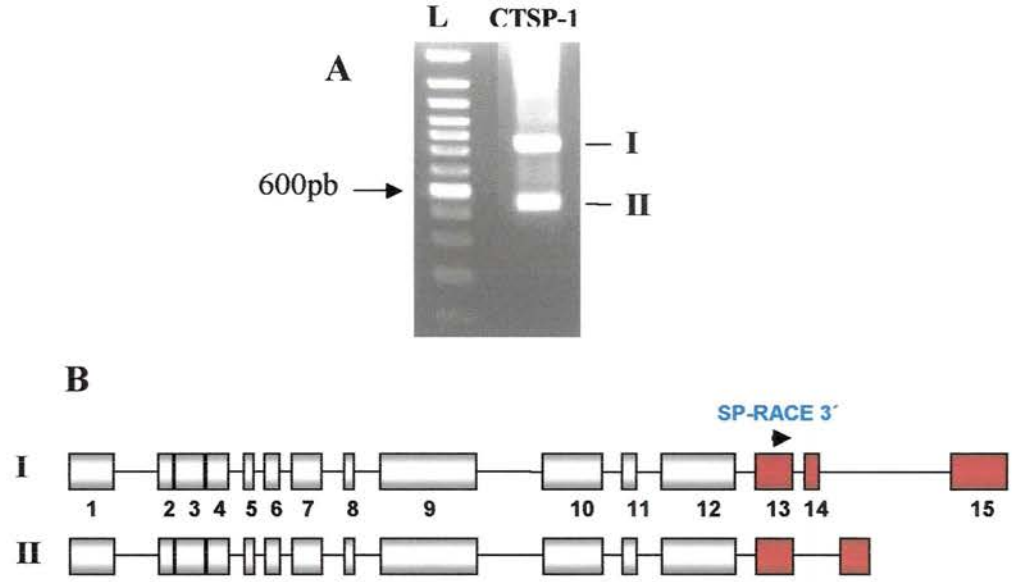
Figura 4 - *RT-PCR* do gene CTSP-1 com os *primers* construídos nas regiões conservadas entre os genes parálogos e a seqüência genômica do cromossomo 21.

Assim, foi gerada a seqüência consenso formada pela seqüência obtida com o seqüenciamento dos clones de cDNA e as seqüências obtidas com o seqüenciamento dos fragmentos gerados por *RT-PCR*. Ao final, foi gerada uma seqüência consenso de 3207 pb, distribuída em 11 exons quando alinhada à seqüência genômica.

Posteriormente, a técnica de *RACE* foi utilizada para extender apenas a porção 3' do transcrito, uma vez que o alinhamento entre as extremidades 5' dos genes parálogos e a seqüência consenso de 3207pb indicava que a região 5' do gene

estava completa (Figura 3). Além disso, a tradução da seqüência consenso revelou a presença de um códon de terminação em fase, localizado *upstream* ao primeiro códon de iniciação da tradução (ATG), tornando desnecessária a realização do *RACE 5'*.

Deste modo, através da realização do *RACE 3'* dois produtos distintos foram obtidos (Figura 5A), os quais, depois de seqüenciados, foram confirmados como seqüências específicas e representando formas de poliadenilação alternativa do gene CTSP-1 (Figura 5B).



Legenda: A- Produto obtido na amplificação. O padrão de peso molecular utilizado foi o 100bp (L). B- Formas de poliadenilação alternativa do gene CTSP-1, identificadas após o seqüênciamento dos fragmentos amplificados.

Figura 5 - *RACE 3'* do gene CTSP-1.

Para análise do padrão de expressão do gene CTSP-1 foram utilizados *primers* localizados nos exons 3 e 8 (**SPEXP-F** e **SPEXP-R**). Em diferentes

amplificações foram obtidos fragmentos que após o seqüenciamento foram confirmados como sendo novas formas de *splicing* do CTSP-1 (Figura 6). As seqüências destas variantes também foram utilizadas na geração da seqüência consenso final. Desta maneira, a seqüência protótipo do gene CTSP-1 possui 3916pb distribuídos em 15 exons (Figura 3).

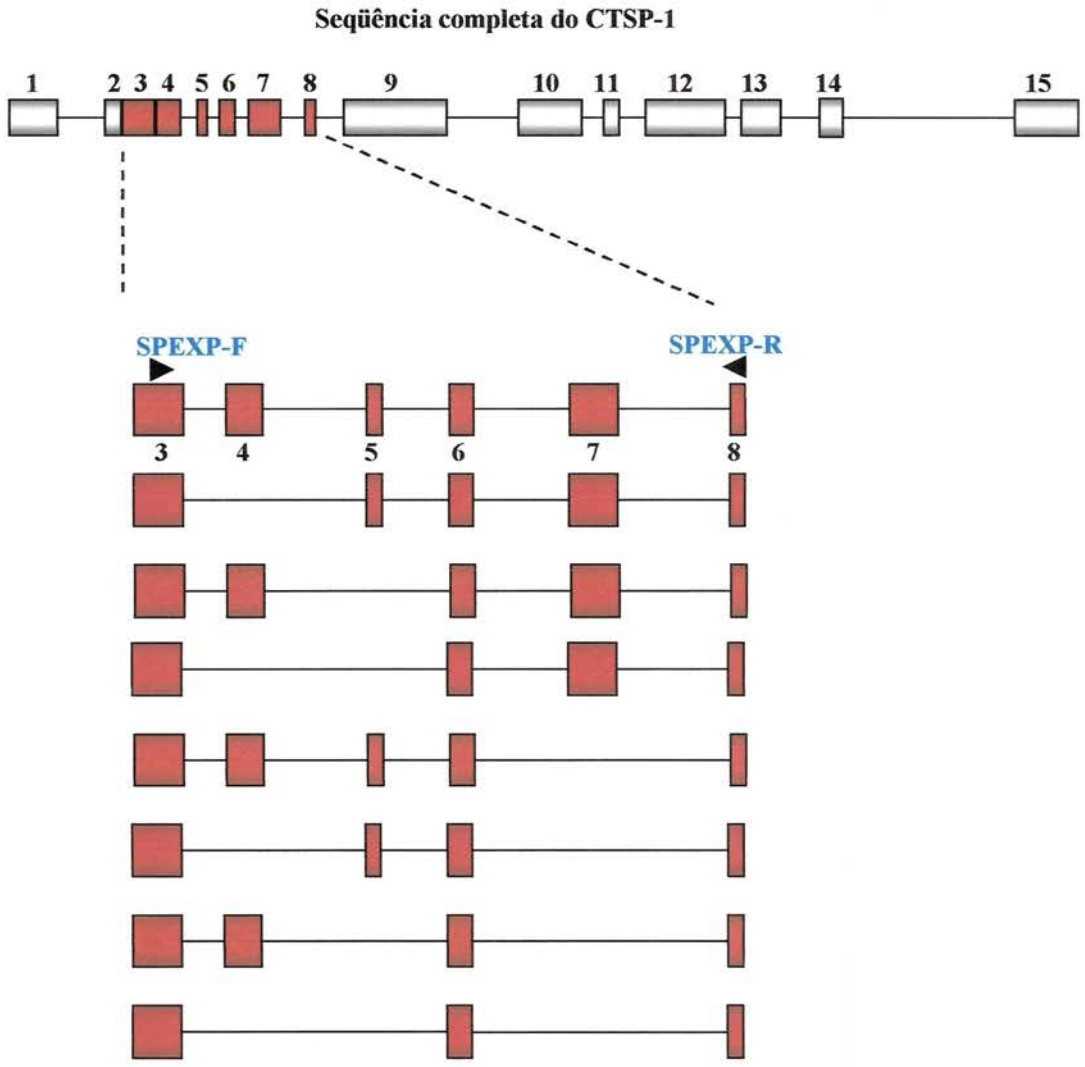


Figura 6 - Formas de *splicing* alternativo do gene CTSP-1 identificadas durante a análise do padrão de expressão.

4.2 ANÁLISE *IN SILICO* DOS DOMÍNIOS PROTÉICOS DA PROTEÍNA CTSP-1

Após verificarmos que o gene CTSP-1 possui diferentes formas de *splicing* alternativo, foram geradas diferentes seqüências consenso correspondentes a cada uma das variantes identificadas. Posteriormente estas seqüências foram traduzidas para a identificação de uma fase aberta de leitura, através da ferramenta *TRANSLATE DNA* disponibilizada no *site* do *Expasy*. Assim, verificamos que a variante que não possui os exons 4, 5 e 7, possui a maior fase aberta de leitura, formada por 202 aminoácidos (Figura 7C).

Esta fase aberta de leitura extende-se apenas até o nono exon, a partir do qual verificam-se vários códons de terminação em todas as fases de leitura. Dada a grande diferença de tamanho entre as regiões codificantes dos genes CTSP-1 (606pb) e NY-BR-1 (4125pb), a qualidade do nosso seqüenciamento foi novamente certificada através da comparação da seqüência consenso do CTSP-1 com a seqüência do genoma humano disponível no *GenBank*. Nenhuma alteração nucleotídica entre as seqüências que causasse mudança de fase ou a criação de um códon de terminação foi identificada, confirmando assim a curta fase aberta de leitura do CTSP-1.

Outra ferramenta disponível no *site* do *Expasy*, o *PROSITE*, permitiu a verificação da presença de domínios protéicos importantes na seqüência de aminoácidos codificada pelo gene CTSP-1. Devido à alta similaridade entre os genes CTSP-1 e NY-BR-1, já esperávamos encontrar determinados domínios protéicos semelhantes aos que foram identificados no NY-BR-1, tais como: domínio de localização nuclear, repetições anquirina e sítio de ligação ao DNA. Foram

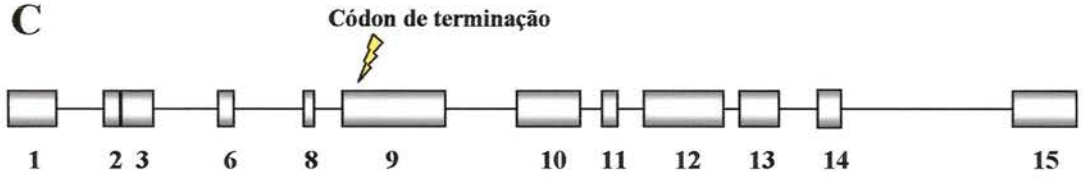
identificados quase todos os domínios citados, excetuando-se o sítio de ligação ao DNA, uma vez que, no gene CTSP-1, a seqüência nucleotídica que codificaria este domínio encontra-se *downstream* ao primeiro códon de terminação (Figura 7A e B).

A

MMKKTTMDLNIRDAKKRTALHWACANGHAEVVTLLVDRKCQLDVLDGENRTTL
MKALQCQREACANILIDSGADPNIVDVYGNTAVHYAVNSENLSVVAKLLSCGT
DIKVKNKAGHTPLLLAIRKRSEQIVEFLLTKNANANGVDKFKCIHQQLLEYKQ
KISKNSQNSNPEGTSEGTPDEAAPLAERTPDTAESLVERTPDE-

B

C

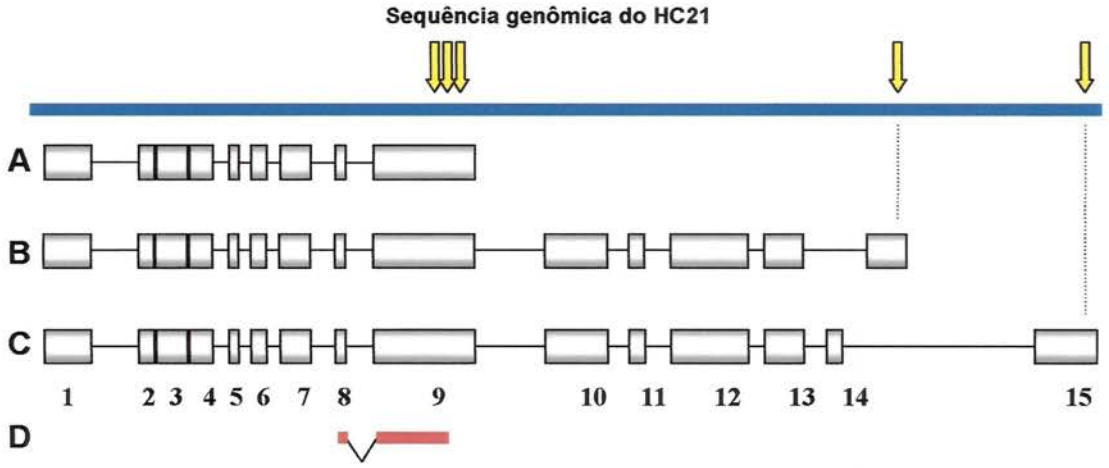


Legenda: A- Em destaque, os domínios protéicos encontrados: em vermelho o domínio de localização nuclear e, em azul, as repetições anquirina. B- Esquema gerado pelo *PROSITE* no qual se verifica a distribuição dos domínios na seqüência protéica. C- Esquema da variante de *splicing* que codifica esta fase aberta de leitura.

Figura 7 - Seqüência de aminoácidos obtida a partir da tradução da variante do gene CTSP-1 que apresentou a maior fase aberta de leitura.

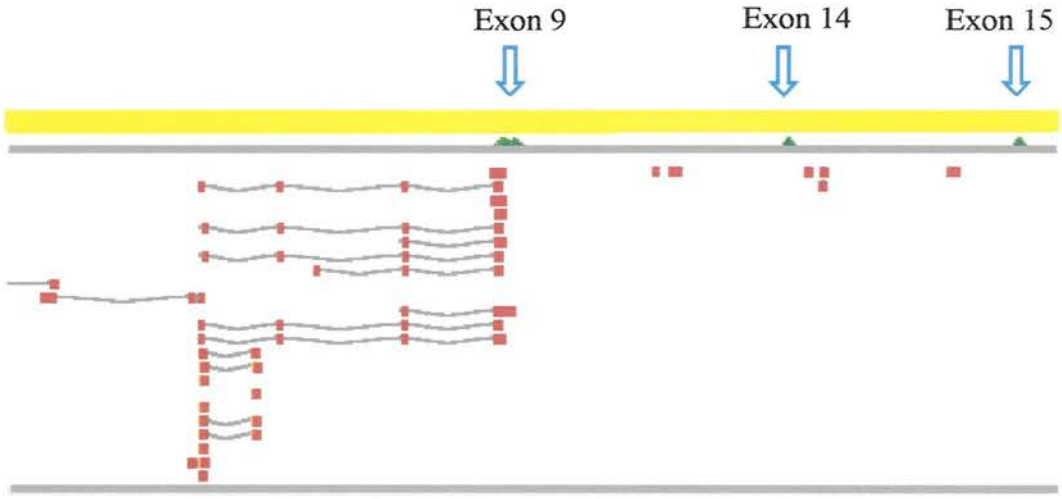
4.3 AVALIAÇÃO DA EXISTÊNCIA DE POLIADENILAÇÃO ALTERNATIVA

Durante a obtenção da seqüência completa do gene CTSP-1 foram identificadas 3 formas de poliadenilação alternativa (Figura 8). A primeira e de menor tamanho corresponde à seqüência representada pelos clones de cDNA. As demais variantes foram identificadas através dos experimentos de *RACE 3'*. A avaliação da distribuição das *ESTs* correspondentes a este transcrito revelou uma grande quantidade de *ESTs* 3' no exon 9, reforçando a existência da variante menor e sugerindo que essa variante seja mais abundante (Figura 9).



Legenda: As setas indicam os sítios de poliadenilação presentes nas seqüências do transcrito (representada em azul). A- Variante menor, representada pelos clones de cDNA. B- Variante intermediária. C- Variante maior, na qual o sítio de poliadenilação utilizado está localizado no exon 15. D- Sonda utilizada no *Northern-blot* para a identificação das variantes.

Figura 8 - Variantes do gene CTSP-1 segundo o sítio de poliadenilação.

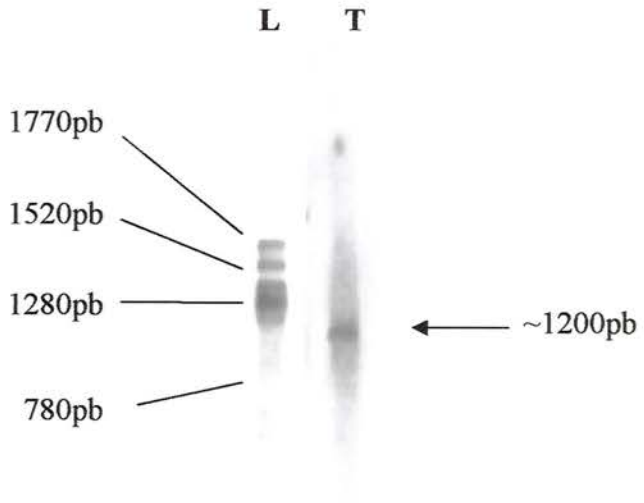


Legenda: Os triângulos em verde representam os sítios de poliadenilação. A barra amarela representa a seqüência genômica do cromossomo 21 e as *ESTs* estão representadas pelas caixas em vermelho (exons), geralmente ligadas por linhas que representam os íntrons.

Figura 9 - Visualização das *ESTs* correspondentes ao gene CTSP-1 na interface gráfica do Projeto *Transcript Finishig Initiative*.

Para confirmar a existência das formas de poliadenilação alternativa foram realizados experimentos de *Northern-blot* utilizando RNA total de testículo normal. Este experimento foi realizado repetidas vezes com uma grande massa de RNA total (100-200µg), mas sem sucesso, provavelmente devido ao baixo nível de expressão do transcrito. Utilizando-se a técnica de amplificação de RNA através de transcrição *in vitro*, uma grande quantidade de RNA mensageiro (aproximadamente 6µg) foi produzida e, em seguida utilizada em experimento de *Northern-blot*.

Na Figura 10, uma banda de aproximadamente 1200pb pode ser observada, correspondendo ao tamanho esperado para a variante menor do gene CTSP-1, representada pelos clones de cDNA. Entretanto, as outras variantes não puderam ser identificadas. O resultado obtido confirma a existência da variante menor e seu maior nível de expressão.



Legenda: L - RNA *ladder*; T - mRNA de testículo normal transcrito *in vitro*.

Figura 10 - Identificação das formas de poliadenilação alternativa do gene CTSP-1 por *Northern-blot*.

A dificuldade de detecção das outras duas variantes do gene CTSP-1 pode ser explicada por um processo específico de degradação de mRNA chamado NMD (*nonsense-mediated mRNA decay*) (SCHELL et al 2002). Este processo permite que células eucariontes eliminem mRNAs que contenham códons de terminação da tradução prematuros. Tais mRNAs codificam polipeptídeos truncados que poderiam exercer efeitos negativos. Assim, torna-se vantajoso para as células minimizar a tradução destes mRNAs. A questão central sobre este mecanismo é como as células identificam tais moléculas de mRNA.

Acredita-se que as junções exon-exon sejam "marcadas" por um complexo chamado EJC (*exon-junction complex*), o que permite que a posição do íntron removido possa ser "detectada" pela maquinaria de tradução (ZHANG et al 1998).

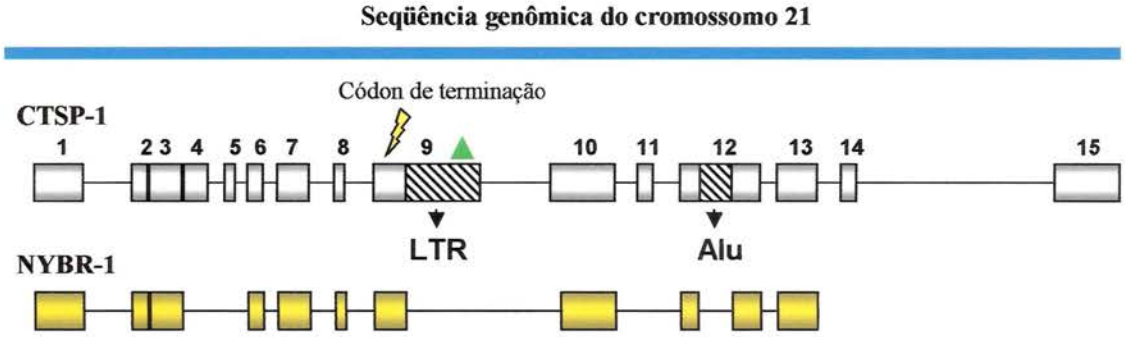
Após o processamento do mRNA, este complexo é depositado 20 pares de base *upstream* às junções exon-exon, e parte dele migra juntamente com o mRNA para o citoplasma. Em seguida, este complexo é removido pela maquinaria de síntese protéica durante o primeiro ciclo de tradução do mRNA (ISHIGAKI et al. 2001). Caso exista uma junção exon-exon a uma distância superior a 50 pares de bases *downstream* ao códon de terminação, o complexo EJC não é removido. Deste modo, esta molécula de mRNA permanece "marcada", podendo ser reconhecida e posteriormente destruída pelo processo de NMD.

No gene CTSP-1, o códon de terminação da tradução está localizado no exon 9. Considerando-se as três formas de poliadenilação alternativa do gene, apenas a variante que contém os 9 exons iniciais (variante A - Figura 7), não estaria sujeita ao processo de NMD. Por outro lado, as outras duas variantes que possuem exons *downstream* ao códon de terminação (variantes B e C - Figura 7), podem ser degradadas por este processo, explicando assim a ausência das mesmas nos experimentos de *Northern-blot*.

4.4 IDENTIFICAÇÃO DA INSERÇÃO DE ELEMENTOS REPETITIVOS NA SEQÜÊNCIA DO GENE CTSP-1 E SUPOSIÇÕES SOBRE SUA EVOLUÇÃO

Durante a análise da seqüência completa do gene CTSP-1 foi identificada a inserção de 2 tipos de elementos repetitivos: um elemento *LTR* (*Long Terminal Repeat*) localizado no exon 9 (nt 1386-2125) e uma seqüência *Alu* localizada no exon 12 (nt 2741-3047) (Figura 11). Podemos especular que a inserção do *LTR* foi

responsável pela criação do códon de terminação prematuro e dos sítios alternativos de poliadenilação presentes no exon 9, e conseqüentemente pela origem da variante menor do gene CTSP-1.



Legenda: Neste esquema verifica-se que os mesmos elementos não estão presentes na seqüência do NY-BR-1. LTR = *Long Terminal Repeat*. O triângulo em verde representa os sítios de poliadenilação presentes no exon 9.

Figura 11 - Esquema demonstrando a inserção de elementos repetitivos na seqüência do gene CTSP-1.

Através de alinhamentos entre as seqüências dos genes CTSP-1, NY-BR-1 e NY-BR-1.1, verificamos que a similaridade é bem maior entre os genes NY-BR-1 e NY-BR-1.1 do que entre o CTSP-1 e os mesmos. Isto sugere que a divergência entre o NY-BR-1 e o NY-BR-1.1 é evolutivamente mais recente do que o surgimento do CTSP-1. Desta forma podemos especular que o gene CTSP-1 se originou a partir da duplicação de um ancestral comum ao NY-BR-1 e NY-BR-1.1. A identificação dos elementos repetitivos apenas na seqüência do CTSP-1 reforça esta hipótese.

A presença do códon de terminação prematuro localizado no exon 9 revelou que a região codificante do gene CTSP-1 é muito menor que a do NY-BR-1. Este fato, juntamente com a inserção dos elementos repetitivos, gerou a possibilidade de que o CTSP-1 fosse um pseudogene. Neste caso, a curta fase aberta de leitura

encontrada, embora estivesse preservada, não corresponderia a uma proteína real. Antes da produção de anticorpos policlonais para a identificação da proteína CTSP-1 por experimentos de *immunoblotting* ou imunohistoquímica, foram feitas algumas análises *in silico* das seqüências disponíveis, com o objetivo de avaliar se o gene apresentava características que pudessem sugerir sua funcionalidade.

4.5 AVALIAÇÃO DA FUNCIONALIDADE DO GENE CTSP-1

4.5.1 Análise de Mutações Sinônimas e não Sinônimas

Segundo a teoria neutra de evolução molecular, pseudogenes evoluem como seqüências neutras, isto é, as taxas de substituição de nucleotídeos devem ser uniformes ao longo de toda seqüência (KIMURA 1991; LI et al. 1981). Para avaliar a possibilidade do CTSP-1 ser um pseudogene, foi feita a avaliação da taxa de substituição sinônima (dS) e não sinônima (dN) entre as seqüências codificantes dos genes CTSP-1 e NY-BR-1. Em seguida, estas taxas foram utilizadas para o cálculo da razão dN/dS. Uma razão dN/dS maior que 1 indicaria seleção positiva, isto é, uma seleção a favor da mudança dos aminoácidos, enquanto que uma razão dN/dS menor que 1 indicaria seleção purificadora, ou seja, uma seleção no sentido da conservação da proteína. Em pseudogenes espera-se que esta razão seja igual ou próxima de 1, pois não há nenhum tipo de seleção atuando sobre o mesmo (KIMURA 1991). As análises entre as seqüências do CTSP-1 e NY-BR-1 revelaram uma razão dN/dS de aproximadamente 0,6 (P = 0,035). Este resultado sugere fortemente a existência de uma seleção purificadora, no sentido de conservação da proteína codificada pelo CTSP-1. Desta maneira, mesmo sendo uma proteína "truncada" em relação à

proteína NY-BR-1, a proteína codificada pelo CTSP-1 deve possuir uma função importante, justificando assim esta conservação.

Uma possível explicação para a preservação desta proteína "truncada" seria a presença de um mecanismo regulatório entre as proteínas CTSP-1, NY-BR-1 e NY-BR-1.1. Como, aparentemente, o NY-BR-1 e o NY-BR-1.1 são fatores de transcrição, a atividade dos mesmos poderia ser regulada através da competição de sua ligação a seus sítios alvo. Semelhantemente aos mesmos, a proteína CTSP-1 também possui repetições anquirina, o que sugere que a mesma seja capaz de se ligar às mesmas proteínas alvo. Entretanto, a transcrição dos genes alvo não deve ser ativada, uma vez que a proteína CTSP-1 não possui o domínio de ligação ao DNA. Um mecanismo semelhante de competição por sítios alvo entre uma forma ativa de proteína (com todos os domínios) e outra regulatória (com falta do domínio importante para o desempenho da função) foi verificado entre variantes de *splicing* das DNA metil transferases (ROBERTSON et al. 1999).

4.5.2 Identificação de Gene Ortólogo em Chimpanzé

Outra maneira de confirmar a funcionalidade de um gene é a conservação do mesmo em diferentes espécies. Durante as análises da seqüência consenso do CTSP-1, verificamos que a mesma possui alta similaridade com um clone genômico (AC125393) e uma *EST* (CB298911) de chimpanzé (*Pan troglodytes*), sugerindo a presença de um gene ortólogo ao CTSP-1 neste organismo. Para a confirmação de que este possível gene era um ortólogo do CTSP-1 e não do NY-BR-1 ou NY-BR-1.1, foi feito um alinhamento múltiplo entre a seqüência da *EST* de chimpanzé e as seqüências dos 3 genes parálogos (Figura 12). Este alinhamento revelou que o

possível gene de chimpanzé é um ortólogo do CTSP-1, uma vez que foi encontrada apenas a inserção de um nucleotídeo entre a seqüência do CTSP-1 e a *EST* de chimpanzé, sendo esta *upstream* ao códon de iniciação da tradução. Quando comparados à *EST* de chimpanzé, os genes NY-BR-1 e NY-BR-1.1 apresentaram 3 inserções e uma grande quantidade de substituições (Tabela 3).

```
EST_chimp          ----------------------------------------------------------TCTC 4
CTSP-1_mRNA        CCCTTCAGTCAGCTGGTCTACACCAACAACGACTCTTACGTGATTCACCATGGGGATCTC 270
NY-BR-1_mRNA       ------------CTAGTCTATACCAGCAACGACTCCTACATCGTCCACTCTGGGGATCTT 48
NY-BR-1.1_mRNA     CCCTTCAGCGAACGGGTCTACACTGAGAAGGACTACGGGACCATCTACTTCGGGGATCTA 297
                                                                        ***

EST_chimp          AGGAAGATCCACAAAGCTGCCTCCCGGG-CCAAGCCTGGAAGCTGGAGAGGATGAT---G 60
CTSP-1_mRNA        AGGAAGATCCACAAAGCTGCCTCCCGGGGCCAAGCCTGGAAGCTGGAGAGGATGAT---G 327
NY-BR-1_mRNA       AGAAAGATCCATAAAGCTGCCTCCCGGGGACAAGTCCGGAAGCTGGAGAAGATGACAAAG 108
NY-BR-1.1_mRNA     GGGAAGATCCATACAGCTGCCTCCCGGGGCCAAGTCCAGAAGCTGGAGAAGATGACAGTA 357
                    * ******** * ************** **** *  ********** *****

EST_chimp          AAGAAGACGACAATGGACCTGAACATAAGAGATGCGAAGAAGAGGACTGCTCTACACTGG 120
CTSP-1_mRNA        AAGAAGACGACAATGGACCTGAACATAAGAGATGCGAAGAAGAGGACTGCTCTACACTGG 387
NY-BR-1_mRNA       AGGAAGAAGACCATCAACCTTAATATACAAGACGCCCAGAAGAGGACTGCTCTACACTGG 168
NY-BR-1.1_mRNA     GGGAAGAAGCCCGTCAACCTGAACAAAAGAGATATGAAGAAGAGGACTGCTCTACACTGG 417
                    ***** * *  *  **** ** * *  *** ***********************

EST_chimp          GCCTGTGCCAATGGCCATGCAGAAGTAGTAACACTTCTGGTAGATAGAAAGTGCCAGCTT 180
CTSP-1_mRNA        GCCTGTGCCAATGGCCATGCAGAAGTAGTAACACTTCTGGTAGATAGAAAGTGCCAGCTT 447
NY-BR-1_mRNA       GCCTGTGTCAATGGCCATGAGGAAGTAGTAACATTTCTGGTAGACAGAAAGTGCCAGCTT 228
NY-BR-1.1_mRNA     GCCTGTGTCAATGGCCATGCANAAGTAGTAACATTTCTGGTAGACAGAAAGTGCCNGCTT 477
                   ******* ********** *********** ********** ********** ****

EST_chimp          GACGTCCTTGATG---------------------------------------------- 193
CTSP-1_mRNA        GACGTCCTTGATGGCGAAAACAGGACAACTCTGATGAAGGCTCTGCAATGCCAGAGGGAG 507
NY-BR-1_mRNA       GACGTCCTTGATGGCGAACACAGGACACCTCTGATGAAGGCTCTACAATGCCATCAGGAG 288
NY-BR-1.1_mRNA     AATGTCCTTGATGGCGAAGGGAGGACACCTCTGATGAAGGCTCTACAATGCGAGAGGGAA 537
                    * **********
```

Legenda: Em vermelho está representado o códon de iniciação da tradução e, em amarelo, os sítios que apresentam nucleotídeos variáveis.

Figura 12 - Alinhamento múltiplo entre a *EST* de chimpanzé e as seqüências dos genes CTSP-1, NY-BR-1, NY-BR-1.1.

Tabela 3 - Alterações nucleotídicas entre a *EST* de chimpanzé e as seqüências dos genes CTSP-1, NY-BR-1 e NY-BR-1.1.

Gene humano	Alteração nucleotídica em relação à *EST* de chimpanzé		
	Substituição	Deleção	Inserção
CTSP-1	0	0	1
NY-BR-1	24	0	3
NY-BR-1.1	27	0	3

Posteriormente, foram feitos alinhamentos através do programa *Nap e Gap* (HUANG 1994) entre as seqüências do CTSP-1, NY-BR-1 e NY-BR-1.1 e a seqüência genômica do chimpanzé. A análise destes alinhamentos revelou que a similaridade entre a seqüência genômica do chimpanzé e o CTSP-1 (95%) é bem maior do que a similaridade entre a mesma e o NY-BR-1 (46%) ou o NY-BR-1.1 (60%). Além disso, o códon de terminação prematuro presente no exon 9 do gene CTSP-1 também foi encontrado na seqüência genômica do chimpanzé (Figura 13). Cabe ressaltar que este clone genômico de chimpanzé está localizado no cromossomo 22, o qual é sintênico ao cromossomo 21 humano, corroborando com nossos achados.

Além disso, os possíveis genes ortólogos do NY-BR-1 e do NY-BR-1.1 foram encontrados em chimpanzé nos cromossomos 8 e 17, respectivamente. Estes cromossomos são sintênicos aos cromossomos humanos nos quais estão localizados os genes NY-BR-1 e NY-BR-1.1, cromossomo 10 e 18 respectivamente. Toda esta família de genes parece ser exclusiva de primatas pois não encontramos seqüências ortólogas em camundongo e rato, cujos genomas já estão completamente seqüenciados (http://www.ensembl.org).

Assim, foi confirmado que o possível gene presente em chimpanzé é um ortólogo do CTSP-1. Mais do que isso, este resultado sugere que a duplicação gênica ou qualquer que seja o evento que levou ao aparecimento dos parálogos, tenha ocorrido antes da diferenciação das duas espécies. Também de grande importância evolutiva é o fato de que os mesmos elementos repetitivos (*LTR* e *Alu*) encontrados na seqüência do CTSP-1, também foram encontrados na seqüência genômica de chimpanzé, o que demonstra que a inserção destes elementos também ocorreu antes da diferenciação das duas espécies.

A conservação da proteína codificada pelo CTSP-1 em outra espécie corrobora com os achados da análise das taxas de substituição entre os parálogos, e contribui para a exclusão da hipótese de que o CTSP-1 seja um pseudogene.

Legenda: A seqüência superior representa a seqüência de chimpanzé e a inferior a região codificante do gene CTSP-1. Os íntrons estão representados pelos hífens e, em vermelho, a conservação do códon de terminação prematuro.

Figura 13 - Alinhamento global entre a seqüência genômica de chimpanzé e o gene CTSP-1.

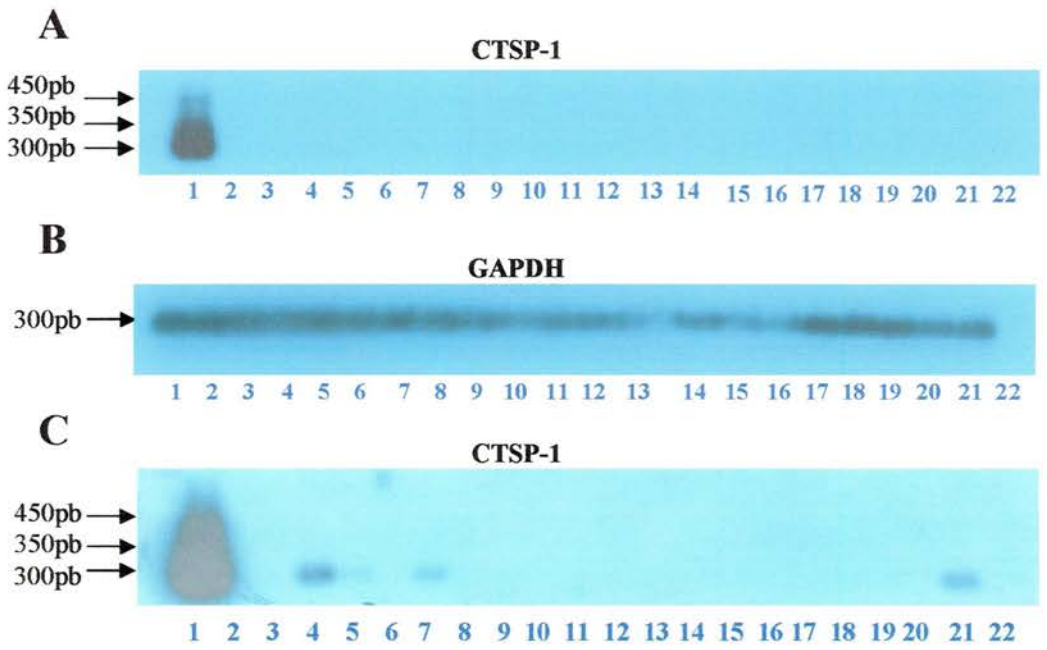
4.6 AVALIAÇÃO DO PADRÃO DE EXPRESSÃO DO GENE CTSP-1

Durante o trabalho de identificação de novos genes localizados no cromossomo 21, foi feita a análise do padrão de expressão em diferentes tecidos normais de todos os transcritos identificados. Dentre os transcritos, o CTSP-1 apresentou-se abundantemente expresso apenas em testículo. Dada a sua similaridade com o antígeno NY-BR-1 e considerando-se que a expressão dos antígenos *CT* é restrita a testículo e a tumores, faltava ainda a avaliação da expressão do CTSP-1 em tecidos tumorais. Assim, decidimos reavaliar seu padrão de expressão em tecidos normais para posteriormente seguirmos com as linhagens celulares tumorais e amostras de tumores de pacientes.

A reação de *RT-PCR* para avaliação do padrão de expressão do CTSP-1 foi inicialmente padronizada com cDNA de testículo normal e posteriormente realizada com os outros tecidos. Os produtos das amplificações foram transferidos para membranas de nylon e então hibridados com sonda específica. Desta maneira, foi possível aumentar a especificidade e sensibilidade da visualização da amplificação. Ao final de 4 horas de exposição, uma forte expressão do transcrito pôde ser observada apenas em testículo (Figura 14A). Foram identificadas diferentes bandas específicas que, após o seqüenciamento, foram caracterizadas como formas de *splicing* alternativo. Posteriormente foi feita uma análise mais detalhada da expressão destas variantes de *splicing*, cujos resultados serão apresentados a seguir.

Para verificarmos se havia uma expressão muito baixa do CTSP-1 em outros tecidos normais, foram feitas outras exposições por um maior período de tempo. Após 16 horas de exposição foi possível detectar um sinal de baixa intensidade em

outros tecidos normais: pulmão, placenta e glândula salivar (Figura 14C). Entretanto, somente a variante de *splicing* menor foi detectada.



Legenda: A- Expressão do CTSP-1 após 4 horas de exposição; B- Expressão do gene GAPDH como controle da qualidade da síntese de cDNA, após 4 horas de exposição; C- Expressão do CTSP-1 após 16 horas de exposição. 1- testículo, 2- cérebro, 3- cólon, 4- pulmão, 5- útero, 6-medula óssea, 7- placenta, 8- fígado, 9- próstata, 10- timo, 11- rim, 12- cérebro fetal, 13- intestino delgado, 14- mama, 15- baço, 16- fígado fetal, 17- músculo, 18- coração, 19- medula espinhal, 20- glândula adrenal, 21- glândula salivar, 22- controle negativo. As três bandas verificadas, indicadas pelas setas, foram confirmadas como formas alternativas de *splicing*.

Figura 14 - Avaliação do padrão de expressão do gene CTSP-1 em tecidos normais por *RT-PCR* seguido por *Southern-blot.*

O fato dos produtos amplificados nas amostras normais serem identificados apenas após um longo período de exposição, revela que o gene é muito pouco expresso nestes tecidos. Cabe ressaltar que os trabalhos mais recentes de identificação de antígenos *CT* utilizaram apenas 30 ou 35 ciclos de amplificação, sendo o produto da reação visualizado em gel de agarose, o que torna a detecção da

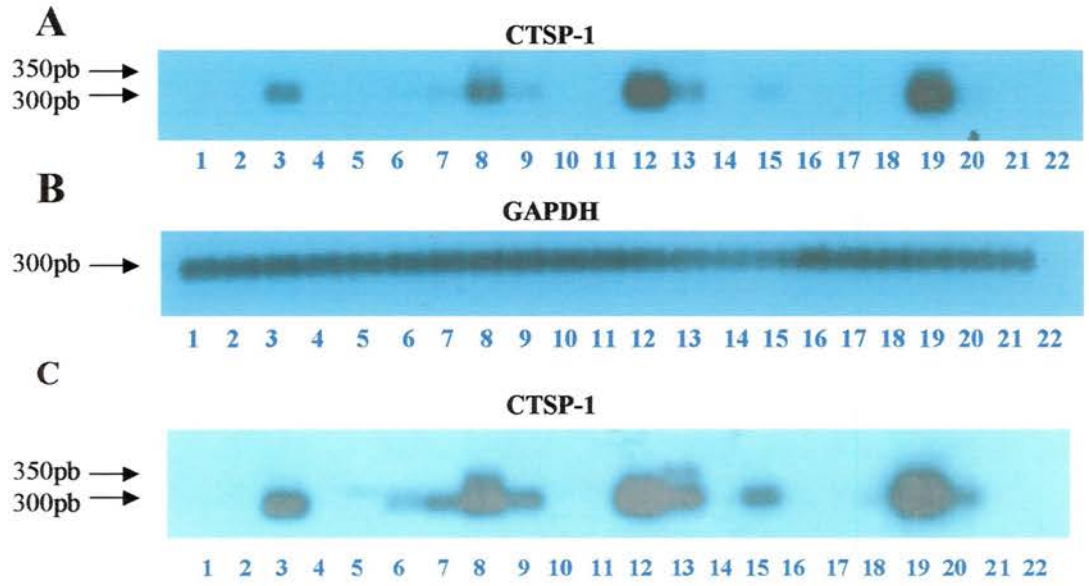
expressão bem menos sensível (DE WIT et al. 2002; SCANLAN et al. 2002b; ZENDMAN et al. 2002).

Em um recente estudo, SCANLAN et al. (2004) procuraram padronizar a análise do padrão de expressão dos antígenos *CT* em tecidos normais. Foi feita a análise de 43 antígenos por *RT-PCR*, utilizando-se 35 ciclos de amplificação. Considerando-se apenas a expressão em tecidos normais, os antígenos puderam ser divididos em 4 grupos: 1) transcritos restritos a testículo; 2) transcritos expressos em 2 ou menos tecidos não gametogênicos (testículo, ovário e placenta); 3) transcritos expressos entre 3 e 6 tecidos não gametogênicos; e 4) transcritos com expressão ubíqua. Dos 43 antígenos estudados, apenas 19 apresentaram expressão restrita a testículo, correspondente ao padrão esperado para os antígenos *CT*. Os outros 24 antígenos apresentaram expressão em outros tecidos normais, sendo 5 deles expressos na maioria dos tecidos normais testados. Se a avaliação do padrão de expressão do CTSP-1 fosse feita nas mesmas condições deste trabalho, certamente ele pertenceria ao grupo dos antígenos restritos a testículo. Para efeitos comparativos, o NY-ESO-1 (provavelmente o antígeno tumoral mais estudado no momento) além dos tecidos gametogênicos, apresentou-se expresso também em pâncreas e fígado.

Dando continuidade à avaliação do padrão de expressão do CTSP-1, foi feita a análise em linhagens celulares tumorais. Assim como para os tecidos normais, o produto das amplificações foi analisado por *Southern-blot*. Ao final de 4 horas de exposição, a expressão do CTSP-1 pôde ser verificada em 9 das 21 linhagens testadas (40%), sendo estas derivadas de glioblastoma (A172), tumor de pulmão (H358 e H1155), tumor de próstata (Du145), linfoblasto B (IM9), tumor de faringe (FADu), mama (MDA-436) e melanoma (A2058) (Figura 15). Assim como em

testículo normal, em algumas linhagens também foi possível identificar formas de *splicing* alternativo.

Conforme feito para os tecidos normais, a membrana foi exposta por um período maior de tempo para verificar se haviam outras linhagens que expressavam o CTSP-1 em menor intensidade (Figura 15). Desta maneira, a expressão do transcrito foi detectada em mais uma linhagem de leucemia (K562). Além disso, verificamos também que as linhagens de tumor de pulmão (H1155), leucemia (IM9), tumor de faringe (FADu) e melanoma (A2058) apresentaram expressão de diferentes variantes de *splicing*.



Legenda: A- Expressão do CTSP-1 após 4 horas de exposição; B- Expressão do gene GAPDH como controle da qualidade da síntese de cDNA, após 4 horas de exposição; C- Expressão do CTSP-1 após 16 horas de exposição.1- Caski; 2- Hela; 3- A172; 4- T98G; 5- HL-60; 6-K562; 7- H358; 8- H1155; 9- Du145; 10- PC3; 11- SCABER; 12- IM9; 13- FADu; 14- MCF-7; 15- MDA-436; 16- ZR75.3A; 17- SW-480; 18- SAOS-2; 19- A2058; 20- SKmel-25; 21- HEPG2; 22- controle negativo

Figura 15: Avaliação do padrão de expressão do CTSP-1 em linhagens celulares tumorais por *RT-PCR*.

Após a verificação da expressão do CTSP-1 em linhagens celulares tumorais, foi feita a avaliação em amostras de tumores de pacientes do Banco de Tumores do Hospital A.C. Camargo. Devido ao grande número de amostras, o produto das amplificações foi analisado apenas em gel de acrilamida corado em prata. A sensibilidade desta análise é bem menor do que a feita por *Southern-blot*, sugerindo que a freqüência de expressão do CTSP-1 em tumores seja maior do que a que encontramos.

Foram analisadas 177 amostras de tumores, derivadas de 15 tecidos diferentes (Tabela 4). No total, 42% das amostras apresentaram expressão do gene CTSP-1. Esta porcentagem pode ser considerada alta frente aos resultados encontrados na literatura para outros antígenos da categoria dos antígenos *CT*, cuja freqüência de positividade varia de 10 a 40% (DE PLAEN et al. 1994; SCANLAN et al. 2002a; ZENDMAN et al. 2002). Considerando-se apenas os tumores que tiveram mais de 10 amostras analisadas, os que apresentaram as maiores porcentagens de expressão foram: tumor de pulmão (57%), melanoma (55,5%), tumor de próstata (48%) e glioblastoma (46%) (Tabela 4).

Tabela 4 - Padrão de expressão do CTSP-1 nas amostras tumorais testadas.

Tecido	**Expressão**
Pulmão	8/14 (57%)
Melanoma	10/18 (55%)
Próstata	12/25 (48%)
Glioblastoma	6/13 (46%)
Estômago	4/9 (44%)
Útero	8/20 (40%)
Esôfago	2/5 (40%)
Mama	9/25 (36%)
Cólon	6/18 (33%)
Tireóide	6/24 (25%)
Hemangioma	2/2
Linfangioma	0/1
Ovário	0/1
Bexiga	1/1
Rim	0/1
Total	**74/177 (42%)**

No trabalho citado anteriormente, de SCANLAN et al. (2004), foi feita também uma análise da expressão de 41 antígenos *CT* em amostras de tumores segundo dados da literatura. Esta análise revelou que o perfil de expressão destes antígenos varia bastante conforme o tipo de tumor estudado. Embora os autores ressaltem que as condições de amplificação utilizadas nos diferentes trabalhos foram muito diferentes, algumas comparações generalizadas puderam ser feitas. Considerando-se o número de antígenos expressos e a sua freqüência de expressão, os tumores puderam ser divididos em 3 grupos: 1) *High CT expressors*: tumores que expressam mais de 50% dos antígenos *CT* avaliados com uma freqüência maior que 20% (por exemplo melanoma, câncer de pulmão e de bexiga); 2) *Moderate CT*

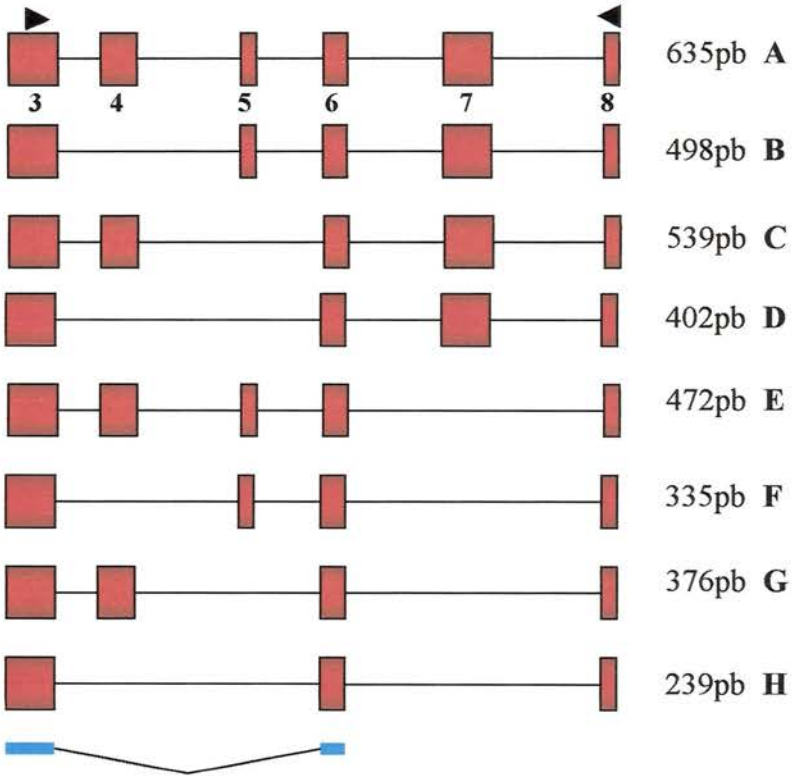
expressors: tumores que expressam entre 30 e 50% dos antígenos analisados com uma freqüência maior que 20% (por exemplo câncer de mama e de próstata); 3) *Low CT expressors*: expressam menos de 30% dos antígenos com freqüência maior que 20% (câncer renal e de cólon). Diferentes tipos de tumores, incluindo tumores de cérebro e de pâncreas, não puderam ser analisados devido ao baixo número de trabalhos sobre expressão de antígenos *CT* nos mesmos. Desta maneira, considerando-se os tipos tumorais em que foi feita a avaliação da expressão do CTSP-1, verifica-se que o perfil de expressão do CTSP-1 nestes tumores é condizente com os outros membros da categoria dos antígenos *CT*. Nos chamou atenção a freqüência de expressão em tumor de cólon (33%), um pouco maior do que o encontrado na literatura, uma vez que este tipo de tumor é considerado um *low CT expressor*.

A alta freqüência de expressão do gene CTSP-1 em amostras tumorais de diferentes tecidos o torna um excelente candidato para imunoterapia em pacientes com câncer. Como nenhum antígeno tumoral foi encontrado em 100% das amostras analisadas até o momento, acredita-se que um tratamento eficiente poderá ser obtido somente através do uso de vacinas polivalentes. Certamente, o desenvolvimento e a eficiência destas vacinas dependem da identificação de novos antígenos com alta freqüência de expressão em diferentes tipos de tumores.

Uma vez que as diferentes formas de *splicing* alternativo do gene CTSP-1 apresentaram alteração da fase aberta de leitura, nos pareceu interessante avaliar seu padrão de expressão em algumas amostras tumorais. Originalmente estas variantes foram identificadas em testículo normal, um tipo de tecido germinativo que apresenta expressão de uma grande quantidade de genes, bem como de suas diferentes formas

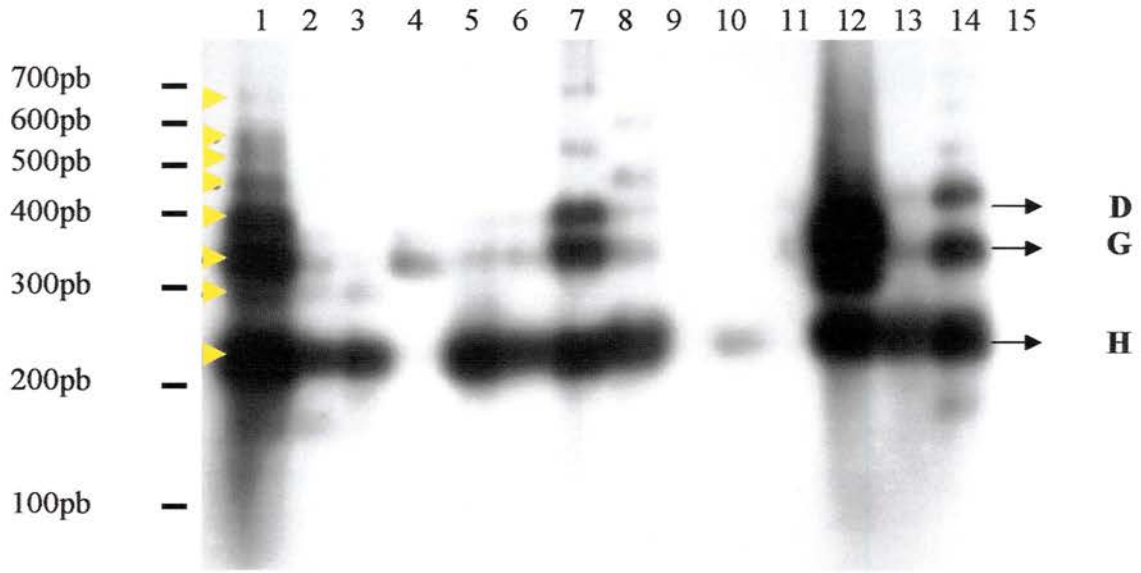
de *splicing* alternativo (EDDY 2002). Desta maneira, foi feita uma análise para verificarmos se as variantes do CTSP-1 encontradas em testículo, em especial a variante que possui a maior fase aberta de leitura, também eram expressas nas amostras tumorais.

Na Figura 16 estão representadas todas as variantes existentes na região codificante do CTSP-1. Foram selecionadas algumas linhagens tumorais e amostras de tumores para avaliação da expressão destas variantes do CTSP-1, através de *RT-PCR* analisado por *Southern-blot*. Nesta reação foram utilizados *primers* desenhados nos exons 3 e 8 na tentativa de identificarmos todas as variantes da região codificante. Como pode ser verificado na Figura 17, foi difícil o isolamento de cada variante, mesmo quando o produto foi fracionado em um grande e concentrado gel de agarose (2%). Entretanto, fica evidente que a variante menor (variante H - Figura 16) é a mais freqüentemente expressa nas amostras analisadas. É exatamente esta variante que possui a maior fase aberta de leitura. Além disso, verifica-se que outras variantes também são expressas nas amostras de tumores, podendo destacar a forte expressão das variantes D e G em algumas amostras.



Legenda: Do lado direito a denominação da variante e o tamanho aproximado do produto amplificado por *RT-PCR* utilizando-se *primers* (**SPEXP-F** e **SPEXP-R**) desenhados no exons 3 e 8 (setas). Em azul, a sonda que foi utilizada no *Southern-blot* para a identificação das variantes.

Figura 16 - Variantes de *splicing* da porção codificante do gene CTSP-1.



Legenda: As amplificações foram feitas com *primers* que flanqueiam todas as variantes da região codificante do gene. As setas amarelas indicam as variantes confirmadas pelo seqüenciamento. 1- Testículo normal, 2- A172 (linhagem de glioblastoma), 3- A2058 (linhagem de melanoma), 4- H1155 (linhagem de tumor de pulmão), 5- Amostra A de tumor de mama, 6- Amostra B de tumor de mama, 7- Amostra A de tumor de próstata, 8- Amostra B de tumor de próstata, 9- Amostra A de melanoma, 10- Amostra B de melanoma, 11- Amostra A de tumor de útero, 12- Amostra B de tumor de útero, 13- Amostra A de tumor de tireóide, 14- Amostra B de tumor de tireóide, 15- Controle negativo (sem cDNA). O peso molecular aproximado está indicado no lado esquerdo da figura.

Figura 17 - Avaliação das variantes de *splicing* do gene CTSP-1 por *RT-PCR* seguido de *Southern-blot*.

4.7 AVALIAÇÃO DO ENVOLVIMENTO DA METILAÇÃO NO CONTROLE DA EXPRESSÃO DO GENE CTSP-1

A metilação é um evento epigenético que consiste na adição de um radical metil (CH_3) na porção 5' de citosinas que compõe um dinucleotídeo CG (BIRD 1992). A distribuição destes dinucleotídeos no genoma não é aleatória. Existem no genoma regiões que apresentam uma freqüência de CG maior do que o esperado, as chamadas ilhas de CpG. Na maioria das vezes, as ilhas de CpG estão localizadas na

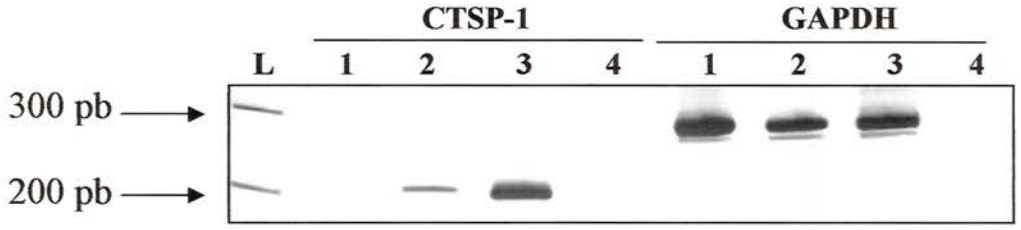
região promotora dos genes, onde freqüentemente não estão metiladas. Entretanto, quando uma ilha de CpG está metilada, verifica-se uma repressão da expressão do gene ao qual ela está associada (JONES e TAKAI 2001).

Uma característica comum dos antígenos *CT* é a indução de sua expressão, através de tratamento com inibidores da DNA metiltransferase-1 ou inibidores das histona deacetilases. Esta indução foi verificada em todos os antígenos *CT* testados até o momento, tais como membros da família MAGE e SSX e o NY-ESO-1 (SCANLAN et al. 2002a). Deste modo, pode-se inferir que a metilação do DNA e a modificação da cromatina possuem um papel importante no controle da expressão gênica destes antígenos. Provavelmente, a região promotora destes genes encontra-se metilada em tecidos normais, e a re-expressão dos mesmos em amostras tumorais se deve à diminuição global da metilação do genoma observada durante o processo de tumorigênese.

Neste contexto, nos pareceu interessante, como parte da caracterização do gene CTSP-1, avaliar seu padrão de expressão em linhagens celulares tumorais após o tratamento com agentes desmetilantes. Estes agentes são análogos da citosina e, quando incorporados ao DNA, se ligam irreversivelmente às DNA metiltransferases inibindo o processo de metilação durante a replicação celular. Desta maneira, a cada divisão celular os níveis de metilação são reduzidos, havendo uma re-expressão de genes cuja expressão é regulada pela metilação.

Para tanto a linhagem celular MCF-7, que não expressa o gene CTSP-1 (Figura 15), foi tratada com 5'aza-2'deoxicitidina por 48 horas. Após o tratamento, a expressão do gene CTSP-1 foi avaliada por *RT-PCR*, na qual observou-se uma expressão significativa do gene nas células tratadas (Figura 18). Cabe ressaltar que

apenas a variante de *splicing* que apresenta a maior fase aberta de leitura foi verificada. Este resultado revela que, a exemplo do que ocorre com outros membros da categoria dos antígenos *CT*, a metilação possui um papel importante no controle da expressão do gene CTSP-1.



Legenda: (1) MCF-7 Mock, (2) MCF-7 tratada com 5'aza 30uM durante 48 horas, (3) Testículo normal, (4) Controle negativo (sem cDNA).

Figura 18 - *RT-PCR* dos genes CTSP-1 e GAPDH após o tratamento da linhagem MCF-7 com 5'aza-2'deoxicitidina.

4.8 EXPRESSÃO E PURIFICAÇÃO DA PROTEÍNA RECOMBINANTE CTSP-1

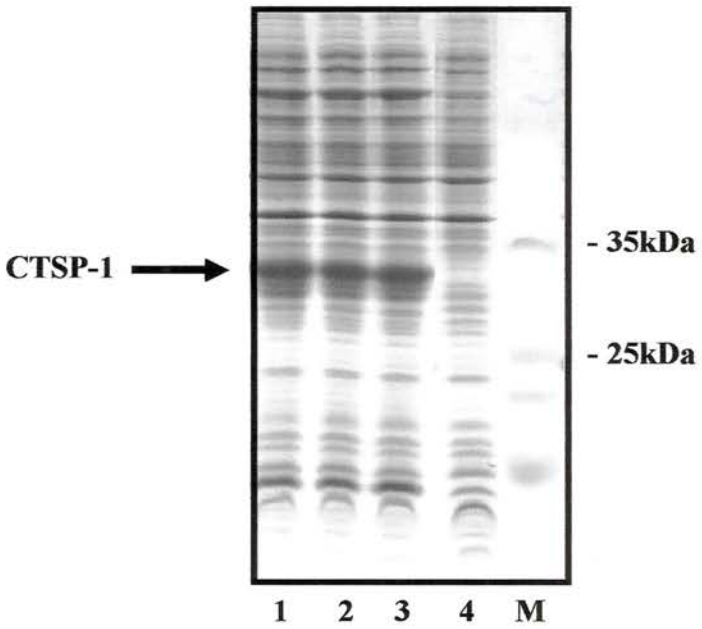
A expressão da proteína CTSP-1 sob a forma de proteína recombinante em sistema heterólogo (*E. coli*) foi realizada com duas finalidades distintas. Em um primeiro momento, a proteína recombinante foi utilizada para a produção de anticorpo policlonal em camundongos, o qual foi posteriormente utilizado para a detecção da proteína CTSP-1 em tecidos humanos através de experimentos de imunohistoquímica e *immunoblotting*. Posteriormente, a proteína recombinante foi utilizada para detectar a presença de anticorpos específicos contra a proteína CTSP-1

em plasma de pacientes com câncer através de experimentos de ELISA e *immunoblotting*.

A maior fase aberta de leitura do gene CTSP-1 foi amplificada por *RT-PCR* e clonada no vetor pET28a. Este vetor possui um sistema de fusão da proteína recombinante à uma seqüência formada por 6 resíduos de histidinas consecutivos (*His-tag*), que permite sua rápida e eficiente purificação em coluna de agarose carregada com níquel. Além disso, este vetor também possui um controle eficiente da expressão do gene de interesse, diminuindo a expressão basal da proteínas de interesse antes da indução por IPTG.

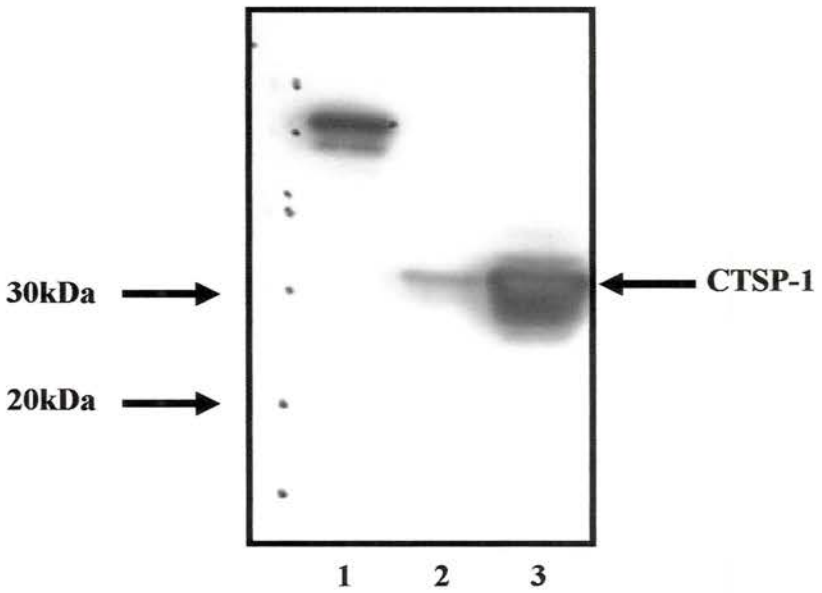
Após a indução das colônias transformantes com 0,4mM de IPTG a 37° C por 16 horas, uma forte expressão da proteína recombinante foi obtida (Figura 19). A banda que aparece fortemente expressa apenas no extrato total de bactérias induzidas corresponde ao tamanho esperado da proteína recombinante (32 KDa).

Para a confirmação de que a banda predominante no extrato bruto de bactéria realmente correspondia à proteína recombinante CTSP-1, o extrato foi analisado por *immunoblotting*, utilizando-se anticorpo anti-*His-tag*. Como esperado, o anticorpo anti-*His-tag* (Figura 20) reconheceu apenas a banda predominante do extrato bruto corado com *Coomassie Blue* (Figura 19). No controle negativo, feito com bactérias não induzidas com IPTG, observa-se apenas uma fraca marcação correspondente à expressão basal da proteína (Figura 20).



Legenda: 1, 2 e 3- Extratos de 3 clones de bactéria induzidos por 16 horas a 37° C com 0,4mM IPTG. 4- Extrato de bactéria não induzida. M- Marcador de peso molecular.

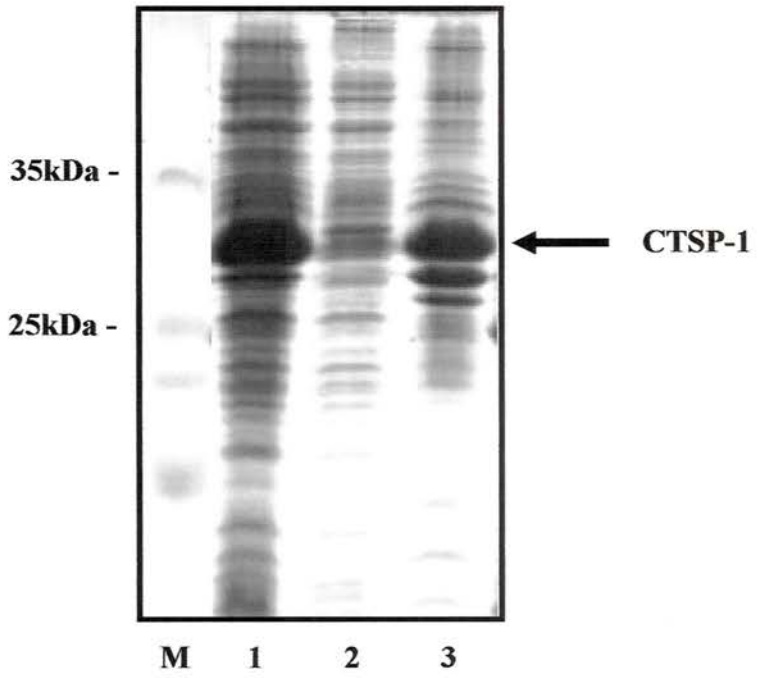
Figura 19 - Extrato bruto de bactéria BL21 expressando a proteína recombinante CTSP-1.



Legenda: Os extratos de bactéria foram transferidos para uma membrana de nitrocelulose e posteriormente incubados com anticorpo anti-*His-tag*. 1- Controle positivo: outra proteína recombinante expressa no mesmo vetor e que também contém a cauda de histidina. 2- Extrato de bactéria não induzida. 3- Extrato de bactéria induzida por 8 horas a 37° C com 0,4mM IPTG. As indicações do padrão de peso molecular aparecem no lado esquerdo da figura.

Figura 20 - *Immunoblotting* com extrato bruto de bactéria BL21 expressando a proteína recombinante CTSP-1.

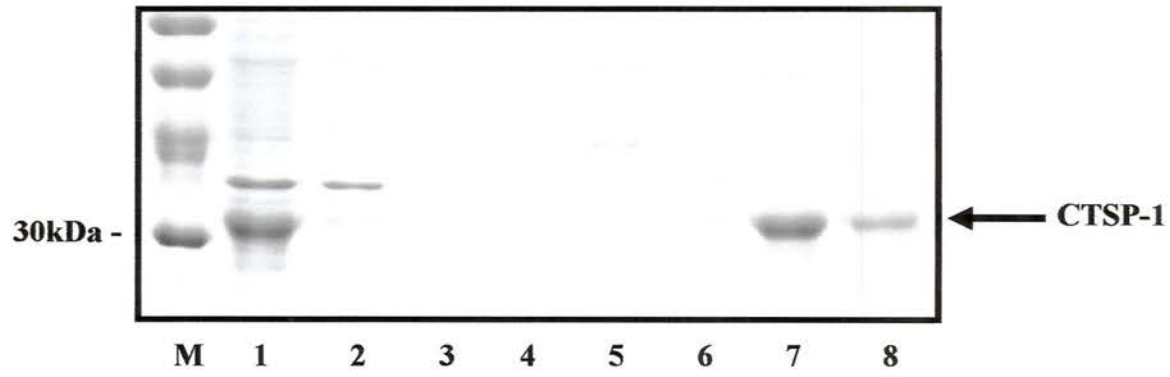
Para a padronização dos protocolos de purificação da proteína recombinante é importante obter informações sobre a solubilidade da mesma. Para tanto, a indução da expressão da proteína recombinante foi realizada em pequena escala e as bactérias induzidas foram lisadas por sonicação em tampão apropriado. Deste modo, as proteínas insolúveis, contidas nos corpos de inclusão, foram sedimentadas por centrifugação e separadas das proteínas solúveis que permaneceram no sobrenadante. As duas frações de proteínas foram desnaturadas em tampão de amostra e fracionadas por eletroforese em gel de SDS-acrilamida como descrito em materiais e métodos. Como pode ser verificado na Figura 21, a maior parte da proteína CTSP-1 recombinante encontra-se expressa na forma insolúvel.



Legenda: 1- Extrato bruto induzido por 16 horas. 2- Extrato de proteínas solúveis. 3- Extrato de proteínas insolúveis. M- Marcador de peso molecular.

Figura 21 - Teste de solubilidade da proteína recombinante CTSP-1.

Uma vez determinada a solubilidade da proteína recombinante, a purificação da mesma foi feita em larga escala a partir de uma cultura de 250mL de meio, utilizando as mesmas condições de indução descritas anteriormente. No entanto, após a lise das bactérias, apenas o *pellet* de proteínas insolúveis foi utilizado para a purificação em coluna acoplada a níquel. Inicialmente o *pellet* de proteínas foi dissolvido em tampão contendo uréia (em condição desnaturante) e, em seguida, aplicado em uma coluna de agarose carregada com níquel. Devido à alta afinidade da cauda de histidina pelo níquel, a proteína recombinante ficou retida na coluna. Após a completa lavagem da coluna, para a retirada de proteínas inespecíficas, iniciou-se a eluição da proteína recombinante com tampão contendo diferentes concentrações de imidazol, o qual compete com o grupo imidazol presente nos resíduos de histidina pela ligação ao níquel. As frações eluídas da coluna foram coletadas separadamente e analisadas em gel de SDS-acrilamida, no qual verificou-se que a maior parte da proteína recombinante foi eluída nas frações de 100 e 250mM de imidazol (Figura 22). Estas frações foram reunidas e submetidas à diálise para a remoção parcial da uréia e total do imidazol. Ao final da diálise, a proteína foi armazenada em tampão sódio-fosfato contendo 1M de uréia e em concentração aproximada de 0,3μg/μL.



Legenda: 1- Extrato protéico antes da purificação; 2- Extrato protéico após a passagem pela coluna; 3- Produto da lavagem da coluna com tampão de sonicação 4- Produto da eluição com tampão 10mM Imidazol; 5, 6, 7 e 8- Idem com tampão 25mM, 50mM, 100mM, 250mM Imidazol, respectivamente; M- Marcador de peso molecular.

Figura 22 - Purificação da proteína recombinante CTSP-1 em coluna de agarose.

4.9 PRODUÇÃO DE ANTICORPOS ANTI-CTSP-1 EM CAMUNDONGOS

Anticorpos policlonais anti-proteína recombinante CTSP-1 foram produzidos em camundongos de duas linhagens: C57 e Swiss. Foram utilizados 3 animais de cada linhagem, com 3 meses de idade em média. Na imunização foi utilizado o adjuvante de *Freund* para potencializar a resposta ao antígeno. Foram feitas 3 imunizações por animal com intervalo de 21 dias e o soro dos mesmos foi testado ao final de 2 semanas após a última imunização. O título dos anticorpos anti-CTSP-1 foi verificado por ELISA e *immunoblotting* (Figuras 23, 24 e 25). Em todos os experimentos foram feitos controles com soro de camundongo não imune da mesma linhagem (C57 ou Swiss), os quais não reagiram com a proteína CTSP-1. Foram

feitos também os seguintes controles negativos: BSA (*Bovine Serum Albumin*) no ELISA (considerado como "branco" das reações); proteína irrelevante produzida no mesmo sistema de expressão no *Western-blot*.

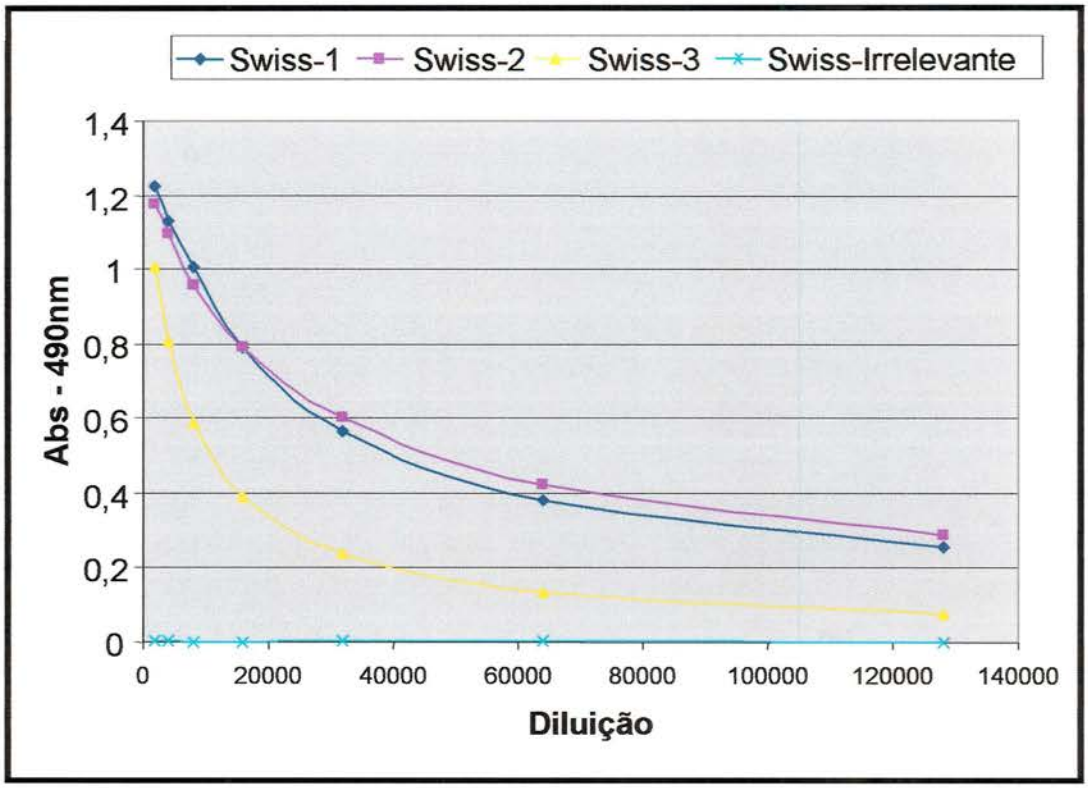


Figura 23 - Titulação do anticorpo anti-CTSP-1 proveniente de camundongos Swiss através de ensaio de ELISA.

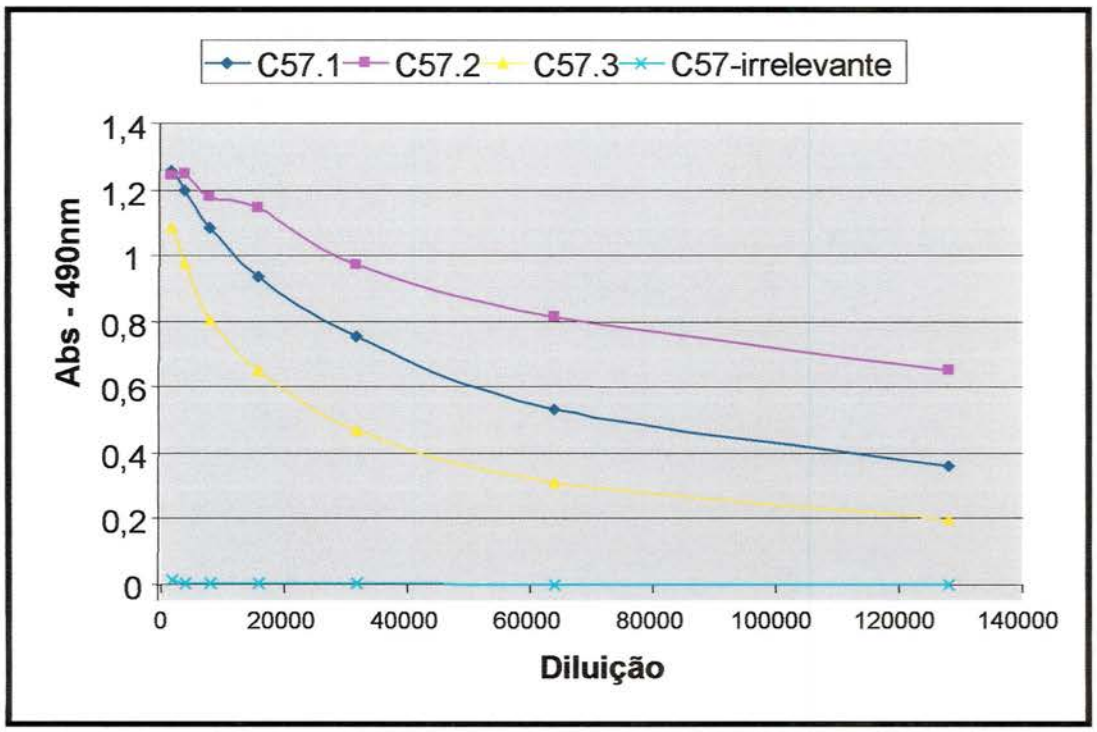
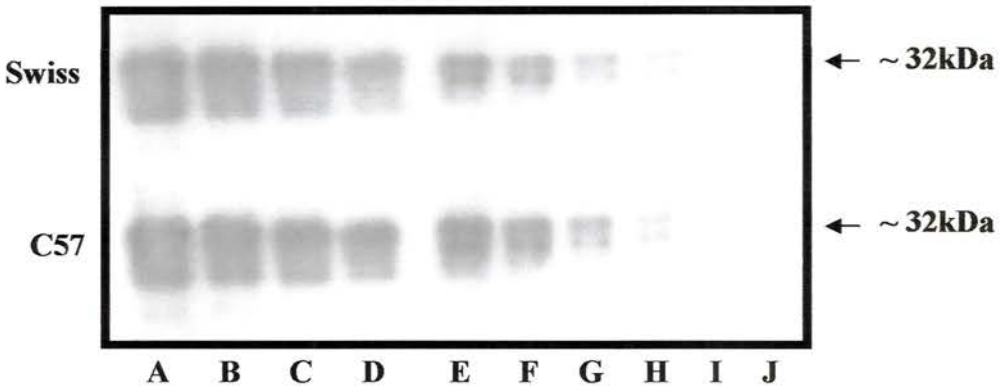


Figura 24 - Titulação do anticorpo anti-CTSP-1 proveniente de camundongos C57 através de ensaio de ELISA.



Legenda: Foi testado apenas o soro do camundongo que apresentou o título mais alto no ELISA. Na membrana foi imobilizada 0,5ug de proteína recombinante CTSP-1 por canaleta. Diluições: A- 1:1000; B- 1:2500; C- 1:5000; D- 1:10000; E- 1:20000; F- 1:40000; G- 1:80000; H- 1:160000; I- 1:320000; J- Controle negativo: proteína irrelevante produzida no mesmo sistema de expressão.

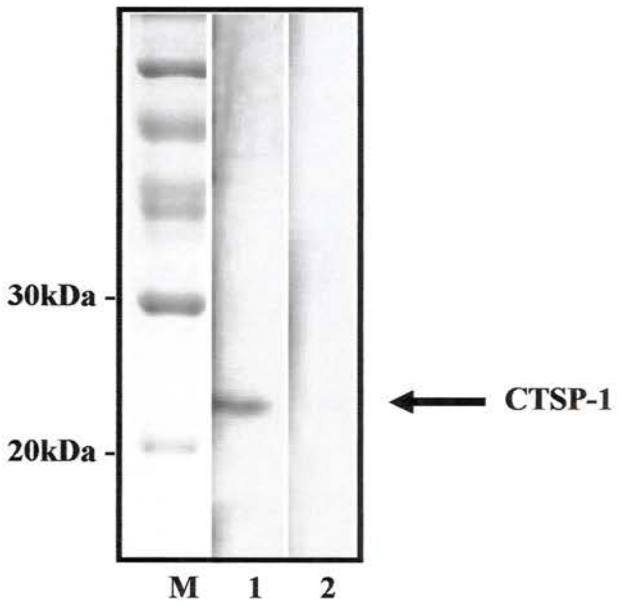
Figura 25 - Titulação do anticorpo anti-CTSP-1 proveniente de camundongos C57 e Swiss através de *immunoblotting*.

4.10 DETECÇÃO DA PROTEÍNA CTSP-1 EM AMOSTRAS DE TECIDO HUMANO

4.10.1 *Immunoblotting*

Para a detecção inicial da proteína CTSP-1 em tecido humano, o tecido escolhido foi testículo normal, uma vez que o mesmo apresentou a expressão mais forte dentre os tecidos analisados por *RT-PCR*. O tecido foi lisado em tampão desnaturante e fracionado em gel de SDS-acrilamida como descrito em materiais e métodos. Em seguida, o extrato total de proteínas foi transferido e imobilizado em membrana de nitrocelulose para posterior incubação com o anticorpo anti-CTSP-1 policlonal de camundongo. Neste caso optou-se por usar o soro do camundongo C57 que apresentou o título de anticorpo mais alto nos ensaios de ELISA e

immunoblotting. Uma banda específica de aproximadamente 22kDa pôde ser identificada, correspondendo ao tamanho esperado (202aa) (Figura 26). No controle feito com o soro irrelevante de camundongo C57, nenhuma banda foi visualizada.



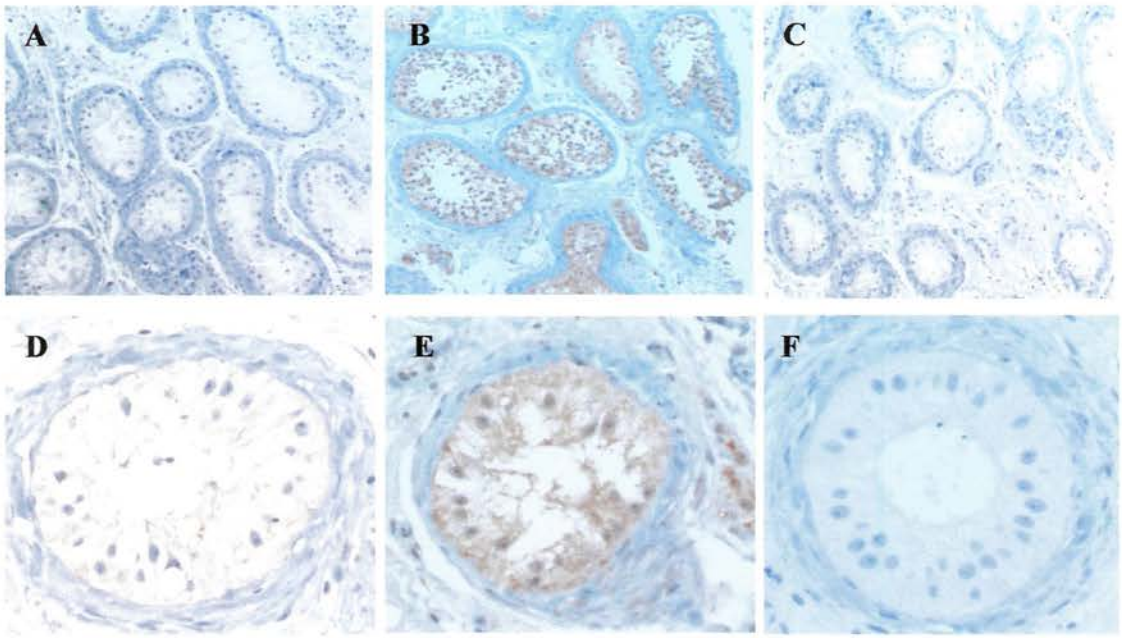
Legenda: 1- Soro do camundongo C57-2; 2- Soro irrelevante de camundongo C57. Ambos foram dilu[idos 1:10000. M- Marcador de peso molecular.

Figura 26 - Detecção da proteína CTSP-1 em extrato protéico de testículo normal por *immunoblotting*.

Deste modo, a hipótese de que o CTSP-1 pudesse ser um pseudogene foi completamente afastada. Entretanto, para a caracterização do mesmo como um antígeno tumoral, ainda restava realizar a análise do padrão de expressão da proteína em amostras de tecidos tumorais e a detecção de anticorpos específicos contra a proteína CTSP-1 em amostras de plasma de pacientes com câncer.

4.10.2 Imunohistoquímica

Para avaliarmos o padrão de expressão da proteína CTSP-1 nos diferentes tipos celulares que compõem um tecido e ainda determinar sua localização sub-celular, foram feitos experimentos de imunohistoquímica (IHQ) a partir de amostras emblocadas em parafina. Para tanto, foram utilizados o mesmo anticorpo utilizado em *immunoblotting* e a amostra de testículo normal como controle positivo. Na Figura 27 verifica-se uma forte marcação nas células germinativas (no interior dos túbulos seminíferos) e também nas células de Leydig (grupos celulares específicos fora dos túbulos). Dependendo do campo do tecido analisado observa-se uma forte marcação nuclear compatível com a presença do domínio de localização nuclear na seqüência de aminoácidos da proteína CTSP-1. Entretanto, a maior parte da marcação é predominantemente citoplasmática. Já no controle feito com soro irrelevante de camundongo C57, nenhuma marcação foi verificada.



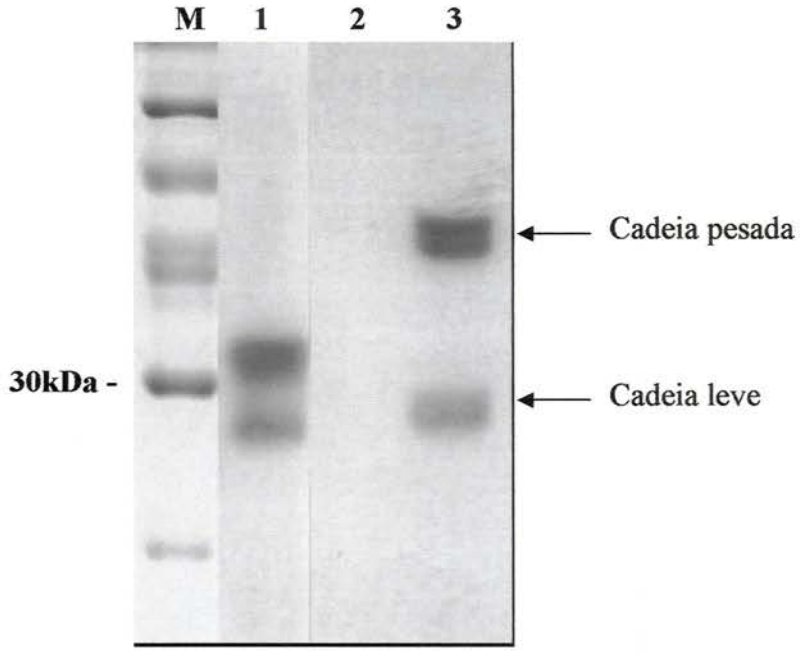
Anticorpo irrelevante (1:2000) | **Anticorpo anti-CTSP-1 (1:2000)** | **Anticorpo anti-CTSP-1 depletado (1:500)**

Legenda: A- Anticorpo irrelevante, diluição. 1:2000; B- Anticorpo anti-CTSP-1, diluição 1:2000; C- Anticorpo anti-CTSP-1 depletado, dil. 1:500; A, B e C - aumento de 200X. D, E e F- idem em aumento de 400X.

Figura 27: Cortes histológicos de testículo normal reagidos contra diferentes anticorpos para a detecção da proteína CTSP-1.

Para a confirmação da especificidade do anticorpo policlonal, foi feito um teste de depleção do soro do camundongo, no qual a proteína recombinante CTSP-1 foi utilizada para a remoção dos anticorpos específicos contra mesma. Assim, o soro depletado foi utilizado inicialmente em experimentos de *immunoblotting* contra a proteína recombinante para a confirmação da depleção total dos anticorpos anti-CTSP-1. Como pode ser verificado na Figura 28, o soro depletado não foi capaz de reconhecer a proteína recombinante, nem mesmo quando testado em uma concentração maior (diluição de 1:10000) do que a originalmente utilizada para a detecção da proteína (1:160000). Além disso, o complexo protéico ligado à resina de agarose foi eluído e transferido para uma membrana de nitrocelulose, a qual foi

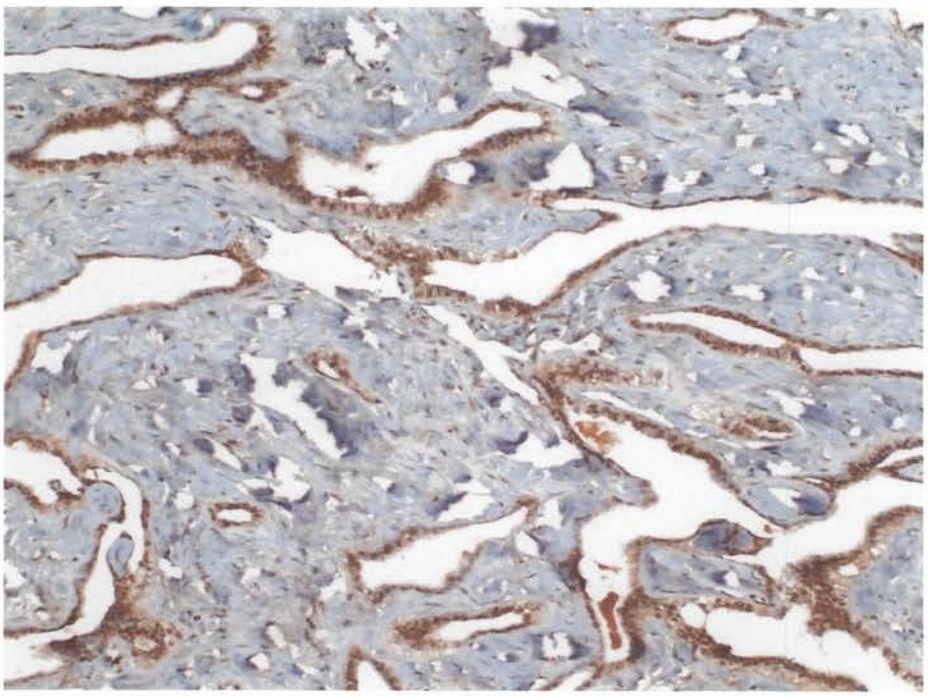
incubada com anticorpo anti-imunoglobulina G de camundongo. Assim, observamos as bandas correspondentes às cadeias leve e pesada de imunoglobulina demonstrando a ligação do anticorpo policlonal à proteína recombinante fixada na resina (Figura 28). Confirmada a eliminação dos anticorpos anti-CTSP-1 por *immunoblotting*, o soro depletado foi utilizado em experimentos de imunohistoquímica na amostra de testículo normal, para avaliar a especificidade da forte marcação encontrada nas células germinativas. Como esperado, o soro depletado não apresentou nenhuma marcação, mesmo quando utilizado em um título mais concentrado (1:500) do que o original (1:2000) (Figura 27).



Legenda: 1 e 2- Proteína recombinante CTSP-1 fracionada em gel e imobilizada na membrana, posteriormente incubada com o soro anti-CTSP-1 antes da depleção (1) (diluição 1:100000) e após a depleção do soro (2) (diluição 1:10000); 3- O eluato de proteínas ligadas à resina (previamente carregada com a recombinante), após a depleção do soro, foi imobilizado na membrana. Posteriormente, a membrana foi incubada com anticorpo anti-IgG de camundongo. M- Marcador de peso molecular.

Figura 28 - *Immunoblotting* para verificação da depleção de IgG anti-CTSP-1 do soro murino.

A expressão preferencial em células germinativas está de acordo com o esperado para os antígenos *CT*, o que reforça a caracterização da proteína CTSP-1 como um novo membro desta categoria. No entanto, também foi encontrada uma forte marcação para a proteína CTSP-1 nas células epiteliais do epidídimo, o que ainda não foi descrito na literatura para os demais antígenos *CT* (Figura 29). No controle negativo, feito com anticorpo irrelevante, nenhuma marcação foi verificada (dado não mostrado).

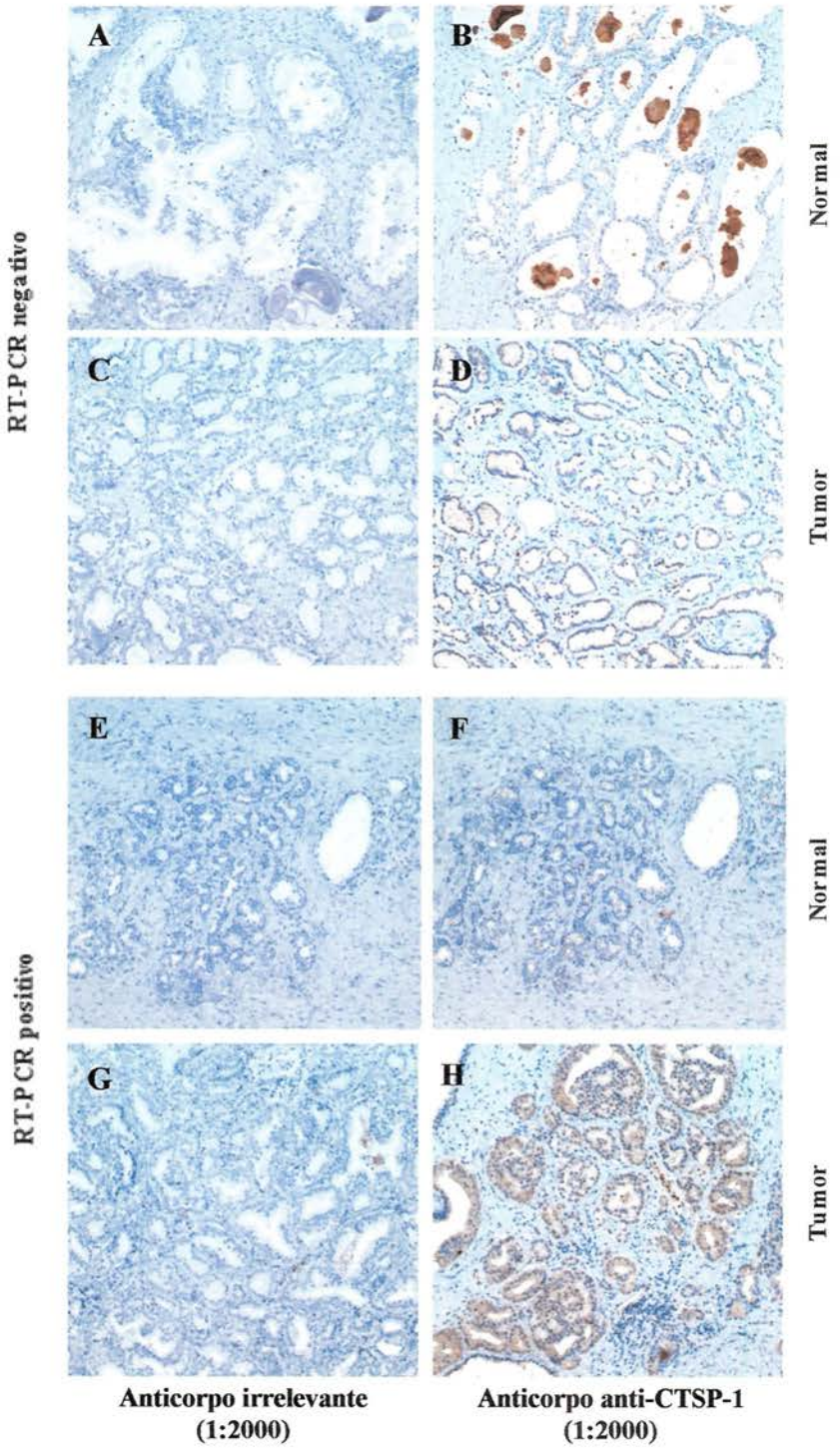


Legenda: Aumento de 200X. Diluição de 1:2000.

Figura 29 - Corte histológico de epidídimo normal reagido contra anticorpo anti-CTSP-1.

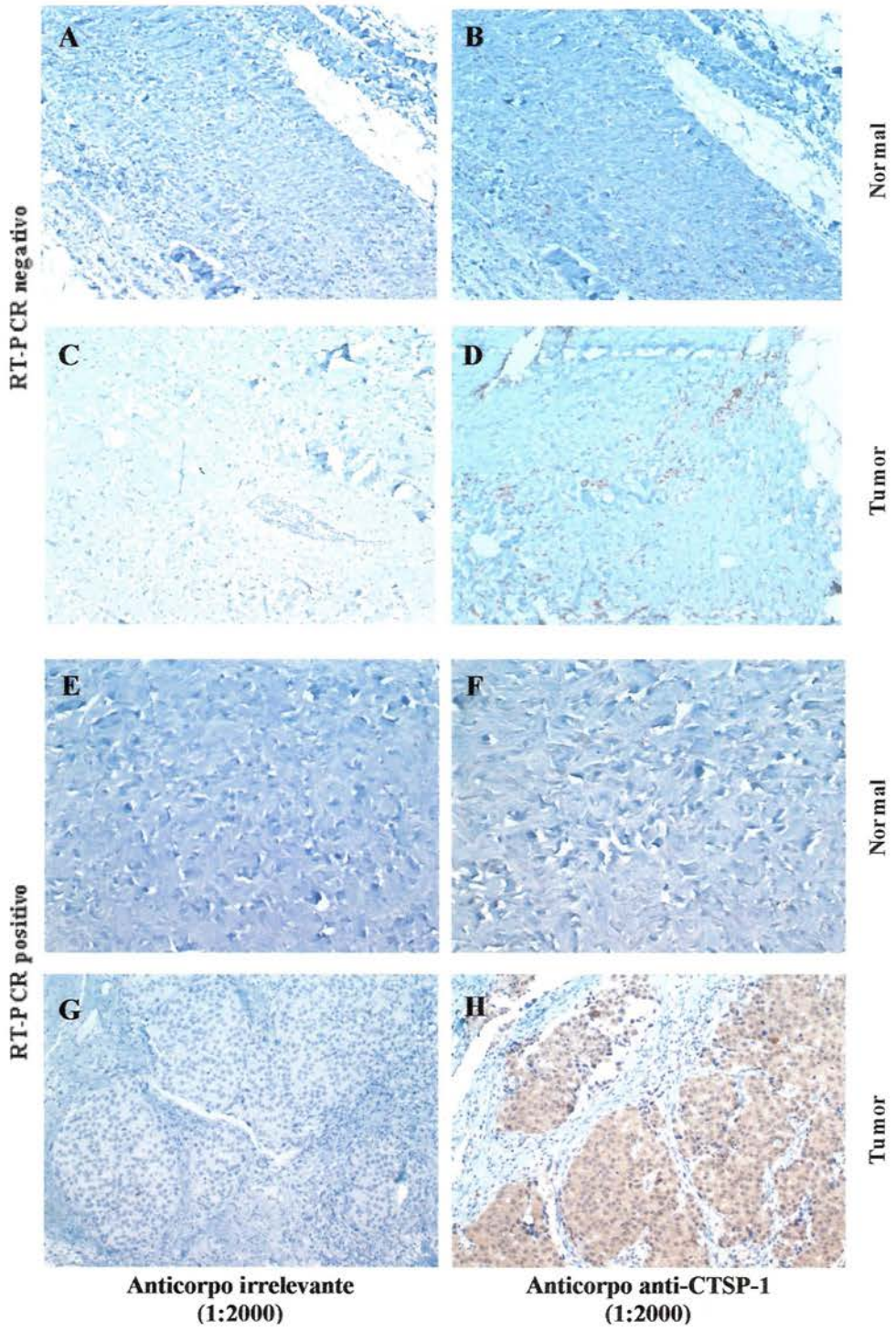
Após a confirmação da especificidade do anticorpo anti-CTSP-1, amostras de outros tecidos também foram analisadas por IHQ, para a confirmação da expressão diferencial da proteína CTSP-1 em tumores. Para tanto, inicialmente foram selecionados 2 casos de tumor de próstata e 2 de tumor de mama, todos com seus

tecidos normais pareados. As amostras de tumores de mama e próstata escolhidas já haviam sido testadas por *RT-PCR* sendo uma delas positiva e a outra negativa para a expressão do gene CTSP-1. De forma interessante, todas as amostras tumorais apresentaram marcação específica, independente do resultado obtido na *RT-PCR* (Figuras 30 e 31). Já dentre as amostras normais, nenhuma apresentou marcação significativa, confirmando a restrita expressão da proteína a tumores. Nos controles feitos com o soro irrelevante (Figuras 30 e 31) e com o soro depletado (dado não mostrado), nenhuma marcação foi verificada em nenhuma das amostras normais ou tumorais.



Legenda: A- tecido normal correspondente ao tumor *RT-PCR* negativo, incubado com anticorpo irrelevante; B- tecido normal correspondente ao tumor *RT-PCR* negativo, incubado com anticorpo anti-CTSP-1; C- tecido tumoral com *RT-PCR* negativo, incubado com anticorpo irrelevante; D- tecido tumoral com *RT-PCR* negativo, incubado com anticorpo anti-CTSP-1; E, F, G e H- Idem, com as amostras correspondentes ao tumor *RT-PCR* positivo. Todos os painéis com aumento de 100X.

Figura 30 - Cortes histológicos de amostras de próstata reagidos contra diferentes anticorpos para detecção da proteína CTSP-1.



Legenda: A- tecido normal correspondente ao tumor *RT-PCR* negativo, incubado com anticorpo irrelevante; B- tecido normal correspondente ao tumor *RT-PCR* negativo, incubado com anticorpo anti-CTSP-1; C- tecido tumoral com *RT-PCR* negativo, incubado com anticorpo irrelevante; D- tecido tumoral com *RT-PCR* negativo, incubado com anticorpo anti-CTSP-1; E, F, G e H- Idem, com as amostras correspondentes ao tumor com *RT-PCR* positivo. Todos os painéis com aumento de 100X.

Figura 31 - Cortes histológicos de amostras de mama reagidos contra diferentes anticorpos para detecção da proteína CTSP-1.

Em vista destes resultados discrepantes, a *RT-PCR* e a IHQ das amostras tumorais selecionadas foram repetidas. Como esperado, na IHQ os resultados foram os mesmos obtidos anteriormente, sendo observada marcação do anticorpo em todas as amostras tumorais. Entretanto, na *RT-PCR*, a amostra de tumor de próstata que inicialmente foi considerada como negativa para a expressão do transcrito, revelou uma fraca expressão, que corrobora com a fraca marcação do anticorpo anti-CTSP-1 encontrada na IHQ. Já para as amostras de tumor de mama os resultados obtidos na primeira *RT-PCR* foram confirmados e a discrepância mantida.

A aparente ausência de correlação entre a expressão do transcrito e da proteína CTSP-1 pode ser explicada pelo baixo nível de expressão e uma eventual instabilidade do transcrito. Este fato explicaria a dificuldade encontrada na detecção do transcrito em experimentos de *Northern-blot*, assim como a necessidade de um elevado número de ciclos na amplificação do transcrito por *RT-PCR*. Além disso, discrepâncias entre a expressão de proteína e RNA mensageiro podem refletir a heterogeneidade do tumor e os diferentes níveis de sensibilidade das técnicas utilizadas na detecção dos mesmos. Por último, a degradação parcial do RNA extraído de amostras tumorais devido à presença de necroses e às condições de coleta do material, é muito freqüente.

Diferentes trabalhos na literatura já descreveram esta aparente discrepância para outros genes, como o receptor de estrógeno e *cerbB2* e a principal justificativa dada é a heterogeneidade da massa tumoral (ONODY et al. 2001; OMOTO et al. 2002). Outro trabalho muito interessante fez uma análise em larga escala da correlação entre o nível de RNA mensageiro (avaliado por *cDNA microarray*) e a proteína correspondente (avaliada por *tissue-array*) em amostras de pulmão de rato, e

os resultados encontrados foram muito discrepantes, sendo a maior meia-vida das proteínas a principal explicação dada pelos autores do trabalho (IZZOTTI et al. 2004).

De qualquer maneira, será de grande importância a análise da expressão do CTSP-1 em um maior número de amostras de tumores, permitindo assim a confimação dos resultados de *RT-PCR* e ainda a realização de estudos de associação entre a expressão do antígeno e os dados anátomo-clínicos do paciente. Estudos anteriores revelaram que a expressão dos antígenos *CT* pode estar associada à progressão tumoral e a tumores com maior potencial de malignidade (SCANLAN et al. 2002a). Por exemplo, a expressão dos genes MAGE-1, 2, 3 e 4 foi detectada, respectivamente, em 16%, 41%, 36% e 11% dos casos de melanoma primário (n=100) e em 48%, 70%, 76% e 22% de melanomas metastáticos (n=145) (BRASSEUR et al. 1995). Do mesmo modo, a expressão do NY-ESO-1 foi detectada em 10% das amostras de melanoma primário (n=20) e em 47% das lesões metastáticas (n=32) (GOYDOS et al. 2001).

4.11 DETECÇÃO DE ANTICORPOS ANTI-CTSP-1 EM PLASMA DE PACIENTES COM CÂNCER

Para finalizar a caracterização da proteína CTSP-1 como um novo antígeno tumoral, era necessário demonstrar a existência de resposta imune específica contra essa proteína em pacientes com câncer. Devido à disponibilidade de amostras de plasma de pacientes no Banco de Tumores do Hospital A.C. Camargo, optou-se por investigar a existência de resposta imune humoral, ou seja, a presença de anticorpos

circulantes anti-CTSP-1. Assim, foram analisados plasmas de pacientes com diferentes tipos de tumor, sendo que, para a maioria dos casos, a detecção do transcrito na amostra tumoral correspondente já havia sido verificada na parte inicial deste projeto. Além disso, plasmas de 50 indivíduos sadios também foram analisados e utilizados como controles negativos.

A maioria dos trabalhos que avaliam a presença de anticorpos contra antígenos tumorais da categoria dos antígenos *CT* utiliza a técnica de ELISA para tal finalidade (JAGER et al. 1998 e 1999; DONG et al. 2003). Nestes trabalhos, são analisados soros de indivíduos sadios que servem de referência para determinar a positividade das amostras dos pacientes com câncer. Assim, para determinar o valor de *cut-off*, é feita a média das leituras das amostras dos indivíduos sadios e calculado o desvio padrão das leituras. Considera-se positivo todo soro de paciente cuja leitura ultrapasse a média obtida entre os soros sadios somada de 3 vezes o desvio padrão determinado.

Deste modo, inicialmente foram feitos experimentos controle com plasma de indivíduos sadios para o cálculo do *cut-off* a ser utilizado na determinação da positividade para os pacientes com resposta específica anti-CTSP-1. De maneira geral, foi verificado um alto *background* nas amostras do grupo controle e ainda, uma grande variabilidade entre as mesmas. Com isso, segundo o *cut-off* estabelecido pela análise dos plasmas de indivíduos sadios, ao analisarmos as amostras dos pacientes, apenas um foi considerado positivo para a presença de anticorpos anti-CTSP-1.

O alto *background* e a variabilidade entre as amostras podem ser explicados pela utilização de plasma ao invés de soro na realização dos experimentos de ELISA.

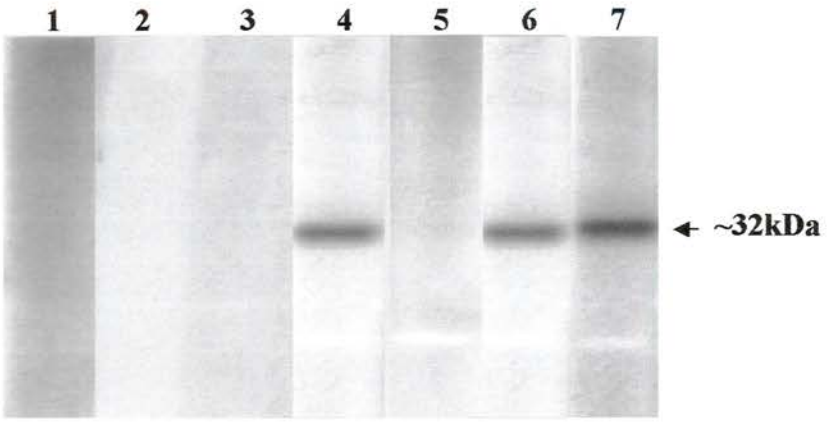
Isto se deve à presença de uma grande quantidade de proteínas séricas no plasma, que interferem na especificidade do experimento. Além disso, espera-se que o título de anticorpos (quando presentes) seja baixo e, por isso a amostra não pode ser muito diluída no momento da análise, aumentando a inespecificidade da reação. Embora o resultado fosse importante para revelar a existência de resposta imune contra a proteína CTSP-1, acreditamos que o uso da técnica de ELISA não seja apropriado para a investigação de anticorpos em plasma de pacientes.

A presença de anticorpos anti CTSP-1 no plasma de pacientes com câncer passou então a ser avaliada por *immunoblotting*. Neste caso, o resultado obtido foi bastante promissor, pois amostras de indivíduos sadios que, quando analisadas por ELISA, apresentaram a mesma leitura que algumas amostras de pacientes, mostraram-se negativas quanto à presença de anticorpos específicos. Por outro lado, o número de amostras de pacientes que apresentaram anticorpos anti-CTSP-1 aumentou significativamente. Cabe ressaltar que, embora mais trabalhosa, a técnica de *immunoblotting* é bem mais específica do que o ELISA (comumente utilizado na literatura), uma vez que a positividade se dá pelo reconhecimento da proteína recombinante no tamanho esperado após seu fracionamento em gel. Neste caso, o *background* causado pela alta concentração das amostras pode ser facilmente diferenciado de um sinal positivo específico (Figura 32).

A freqüência encontrada de anticorpos anti-CTSP-1 no plasma dos pacientes foi de 20,0% (Tabela 5). Os portadores de câncer de próstata foram os pacientes que apresentaram a maior freqüência de anticorpos (40%). Esta freqüência encontrada é extremamente significativa quando comparada aos resultados de imunogenicidade de outros antígenos *CT* já estudados (SCANLAN et al. 2002a). O NY-ESO-1 é o

antígeno mais imunogênico da categoria dos antígenos *CT* já descrito até o momento, sendo encontrados anticorpos específicos em pacientes com diferentes tipos de tumor: câncer de ovário (12%), melanoma (9%), câncer de mama (8%) e câncer de pulmão (4%) (STOCKERT et al. 1998). Esta porcentagem pode subir para 25-50% dos casos, se forem considerados apenas pacientes cujos tumores apresentem expressão do transcrito correspondente. Já para os outros antígenos *CT* a presença de anticorpos em pacientes com câncer é pouco freqüente, nunca ultrapassando 10% dos casos.

No futuro será interessante analisar um maior número de amostras de um mesmo tipo de tumor, na tentativa de avaliar o efeito prognóstico da presença de resposta imune humoral anti-CTSP-1. Em um estudo com pacientes portadores de carcinoma de células transicionais de bexiga, a presença de anticorpos anti-NY-ESO-1 foi detectada apenas nos pacientes com lesões de alto grau (G3), mostrando assim que, além da expressão do transcrito no tumor, a presença de anticorpos específicos no soro do paciente também está associada ao grau avançado do tumor (KURASHI



Legenda: A proteína recombinante CTSP-1 foi fracionada e imobilizada na membrana, a qual foi incubada com: 1, 2 e 3- plasmas de indivíduos sadios; 4, 5 e 6- plasmas de pacientes com câncer; 7- Controle positivo: soro murino anti-CTSP-1.

Figura 32 - Identificação de anticorpos específicos em plasma de pacientes com câncer por *Immunoblotting*.

Tabela 5 - Freqüência de anticorpos anti-CTSP-1 em plasma de pacientes com diferentes tipos de tumor.

Tecido	**Anticorpo**
Próstata	8/24 (33%)
Mama	6/18 (33%)
Tireóide	3/10 (30%)
Cólon	4/20 (20%)
Útero	3/22 (14%)
Estômago	1/8 (12%)
Melanoma	2/23 (9%)
Pulmão	1/13 (8%)
Esôfago	0/4
Hemangioma	0/2
Linfangioma	0/1
Ovário	1/1
Bexiga	0/1
Rim	0/1
Total	**29/148 (20,0%)**

Para a maioria dos casos analisados (126 amostras) foi possível fazer o pareamento entre o estudo da expressão do transcrito CTSP-1 no tumor e a investigação da presença de anticorpos específicos no plasma do paciente (Tabela 6). Dentre as amostras em que foram realizadas análises simultâneas de expressão do transcrito e de presença de anticorpos, 59 (47%) não apresentaram expressão do transcrito no tumor nem a presença de anticorpos específicos no plasma. Para 11 amostras (9%) foi possível detectar como esperado tanto a presença do transcrito no tumor como a de anticorpos específicos no plasma do paciente. No entanto, para o restante das amostras não foi possível estabelecer uma correlação direta entre a expressão do transcrito e a presença de anticorpos. Para 43 amostras (34%) foi possível detectar a expressão do transcrito mas não a presença de anticorpos. De maneira inversa, para 13 amostras (10%) a presença de anticorpos no plasma do paciente foi detectada mesmo na ausência de expressão do transcrito no tumor correspondente.

Tabela 6 - Casos de pacientes em que foram analisadas a expressão do gene CTSP-1 no tumor por *RT-PCR* e a presença de anticorpos específicos (Ac) no plasma.

Tecido	*RT* positivo Ac positivo	*RT* positivo Ac negativo	*RT* negativo Ac positivo	*RT* negativo Ac negativo	Total
Cólon	2	4	2	9	**17**
Esôfago	0	2	0	2	**4**
Estômago	0	4	1	3	**8**
Mama	2	4	3	6	**15**
Melanoma	1	8	0	8	**15**
Próstata	4	4	4	9	**21**
Pulmão	1	7	0	5	**13**
Tireóide	0	3	2	3	**8**
Útero	1	6	0	12	**19**
Hemangioma	0	2	0	0	**2**
Linfangioma	0	0	0	1	**1**
Ovário	0	0	1	0	**1**
Bexiga	0	1	0	0	**1**
Rim	0	0	0	1	**1**
Total	**11**	**43**	**13**	**59**	**126**

Trabalhos semelhantes com outros antígenos tumorais, incluindo o NY-ESO-1, também encontraram a presença de anticorpos em pacientes com tumores sem expressão dos respectivos antígenos (MAIO et al. 2003; AKCAKANAT et al. 2004). Possíveis explicações para esses resultados discrepantes seriam: 1) o baixo nível de expressão e a instabilidade do transcrito (como discutido anteriormente); 2) a expressão heterogênea do transcrito na amostra tumoral que também dificultaria sua detecção por *RT-PCR*; 3) presença de resposta imune anti-CTSP-1 no início do desenvolvimento do tumor com posterior seleção de clones de células tumorais que não expressam o antígeno; 4) expressão do transcrito apenas na metástase do tumor.

De forma interessante, dentre os 50 plasmas de indivíduos sadios analisados, um apresentou sinal positivo na análise feita por *immunoblotting*. Casos como este já foram descritos na literatura para outros antígenos *CT*, como o MAGE-A1 e outros membros da família MAGE, e acredita-se que seja resultado de reação cruzada em indivíduos que têm Vitiligo e Lupus Eritematoso ou ainda resposta imune efetiva em indivíduos que se submeteram à vasectomia (HOON et al. 1995; MCCURDY et al. 1998; LEA et al. 1997; ROCHA et al. 2000). Além disso, não podemos descartar a possibilidade de termos identificado uma resposta imune a um tumor ainda não diagnosticado. Este seria um exemplo no qual a presença de anticorpos específicos poderia funcionar como um marcador para câncer.

De qualquer maneira, a detecção de anticorpos anti-CTSP-1 no plasma de pacientes com câncer confirmou sua antigenicidade e a caracterização do mesmo como um novo antígeno tumoral da categoria dos antígenos *CT*. Como dito anteriormente, estes antígenos possuem grande potencial terapêutico, principalmente devido ao seu restrito padrão de expressão. Entretanto, excetuando-se o NY-ESO-1 e alguns membros da família MAGE, a maioria dos antígenos desta categoria estão presentes em menos de 10% das amostras de tumores analisadas, o que dificulta a utilização dos mesmos no tratamento do câncer (SCANLAN et al. 2002a).

A freqüência de expressão e a espontânea imunogenicidade são critérios importantes para alvos potenciais de vacinas para o câncer. O perfil de expressão dos antígenos tumorais é um pré-requisito para o desenvolvimento da vacina, enquanto que a pré-existência de imunidade autóloga demonstra o potencial do sistema imune em reconhecer o antígeno, revelando que provavelmente a vacinação resultará em resposta imune. O NY-ESO-1 é o antígeno mais imunogênico desta categoria, sendo

encontrados anticorpos específicos em aproximadamente 8% dos pacientes com câncer. Esta aparente pequena porcentagem faz com que o NY-ESO-1 seja o antígeno tumoral mais estudado atualmente e com o maior número de ensaios clínicos em desenvolvimento (SCANLAN et al. 2004; CHEN et al. 2004).

Quando comparado aos antígenos *CT* descritos na literatura, o CTSP-1 destaca-se pela alta freqüência de expressão em diferentes tipos tumorais e mais ainda pela alta imunogenicidade encontrada nestes pacientes. Isto o torna um excelente candidato para imunoterapia no tratamento do câncer. Para que ensaios clínicos futuros sejam viabilizados outros estudos *in vitro* serão necessários para uma melhor compreensão da resposta imune espontânea anti-CTSP-1 em pacientes com câncer. Um importante passo neste sentido será a verificação da presença de células T anti-CTSP-1 nestes pacientes. Para a maioria dos antígenos *CT* a presença de resposta imune celular está associada à presença da resposta humoral, o que sugere fortemente a existência de tal resposta contra o CTSP-1. Além da identificação de linfócitos T CD8, alguns trabalhos também têm investigado a presença de células T CD4 *helper* específicas contra estes antígenos, visto que a presença das mesmas é imprescindível para a manutenção da resposta citotóxica e da produção de anticorpos (CHEN et al. 2004; DAVIS et al. 2004).

Outro importante passo a ser dado será a identificação dos epítopos antigênicos do CTSP-1 que são reconhecidos pelo SI dos pacientes com câncer. Vários peptídeos antigênicos de diferentes antígenos *CT* reconhecidos por células T citotóxicas já foram identificados e alguns deles estão sendo testados como candidatos a imunoterapia em pacientes com câncer (GNJATIC et al. 2000 e 2002; CHEN et al. 2004). Alguns ensaios clínicos em pacientes com melanoma têm sido

feitos utilizando-se um peptídeo derivado do MAGE-3 e evidências de regressão do tumor têm sido verificadas em 20% dos casos (MARCHAND et al. 1999).

Assim, verifica-se que este trabalho gera perspectivas para muitos outros, nos quais a função, a imunogenicidade, e o real potencial terapêutico do CTSP-1 deverão ser avaliados. Certamente, os resultados obtidos neste trabalho contribuirão de maneira significativa para a caracterização completa deste promissor antígeno tumoral.

CONCLUSÕES

5 CONCLUSÕES

Os resultados obtidos neste trabalho nos permitem concluir que:

- O gene CTSP-1 está organizado em 15 exons no cromossomo 21 humano e possui diferentes formas de poliadenilação e *splicing* alternativas.
- A proteína codificada pela isoforma mais abundante do gene CTSP-1 possui 202 aminoácidos e contém alguns dos domínios encontrados na proteína codificada pelo gene NY-BR-1 (repetições anquirina e sítio de localização nuclear).
- Análises comparativas entre as seqüências dos genes CTSP-1, NY-BR-1 e NY-BR-1.1 revelaram a inserção de elementos repetitivos na seqüência do CTSP-1. Estas análises também revelaram que há uma pressão seletiva no sentido da conservação da proteína codificada pelo CTSP-1, sugerindo a funcionalidade deste gene.
- O gene CTSP-1 possui um gene ortólogo em chimpanzé, no qual também foi identificada a inserção dos elementos repetitivos. Aparentemente esta família gênica é específica de primatas.
- O gene CTSP-1 possui um padrão de expressão restrito a testículo normal, a diferentes tipos de linhagens celulares tumorais e a tumores.
- A metilação possui um papel importante no controle da expressão do gene CTSP-1.
- A proteína CTSP-1 foi detectada em amostras de tecido humano, descartando a hipótese do gene CTSP-1 ser um pseudogene.
- Em testículo normal, a proteína CTSP-1 é expressa no citoplasma e está restrita às células germinativas e células de Leydig.

- A proteína CTSP-1 também é expressa em amostras tumorais de próstata e de mama e não foi detectada nas respectivas amostras normais pareadas.
- A presença de anticorpos específicos em plasma de pacientes com câncer revelou a imunogenicidade da proteína CTSP-1.
- O CTSP-1 é um novo membro da categoria dos antígenos *cancer-testis*.
- A alta freqüência de expressão em tumores e a alta imunogenicidade em pacientes com câncer sugerem que o gene CTSP-1 é um candidato promissor a alvo para a imunoterapia do câncer.

REFERÊNCIAS BIBLIOGRÁFICAS

6 REFERËNCIAS BIBLIOGRÁFICAS

Abbas AK, Lichtman AH, Pober JS. **Celular and molecular immunology**. 4th ed. Philadelphia: W.B. Saunders; 2000. Immunity to tumors; p.384-403.

Akcakanat A, Kanda T, Koyama Y, et al. NY-ESO-1 expression and its serum immunoreactivity in esophageal cancer. **Cancer Chemother Pharmacol** 2004; 54:95-100.

Alpen B, Gure AO, Scanlan MJ, Old LJ, Chen YT. A new member of the NY-ESO-1 gene family is ubiquitously expressed in somatic tissues and evolutionarily conserved. **Gene** 2002; 297:141-9.

Baba T, Koizumi M, Suzuki T, Yamanaka I, Yamashita S, Kudo R. Cloning and characterization of a tumor-associated antigen, beta-casein-like protein. **Biochem Biophys Res Commun** 2001; 284:340-5.

Bird A. The essentials of DNA methylation. **Cell** 1992; 70:5-8.

Boyse EA, Old LJ. Some aspects of normal and abnormal cell surface genetics. **Annu Rev Genet** 1969; 3:269-90.

Brass N, Heckel D, Sahin U, Pfreundschuh M, Sybrecht GW, Meese E. Translation initiation factor eIF-4gamma is encoded by an amplified gene and induces an immune response in squamous cell lung carcinoma. **Hum Mol Genet** 1997; 6:33-9.

Brasseur F, Rimoldi D, Lienard D, et al. Expression of mage genes in primary and metastatic cutaneous melanoma. **Int J Cancer** 1995; 63:375-80.

Brichard V, Vanpel A, Wolfel T, et al. The tyrosinase gene codes for an antigen recognized by autologous cytolytic T-lymphocytes on Hla-A2 melanomas. **J Exp Med** 1993; 178:489-95.

Burnet FM. The Concept of immunological surveillance. **Prog Exp Tumor Res** 1970; 13:1-27.

Cantor H, Boyse EA. Functional subclasses of T-lymphocytes bearing different Ly antigens. I. The generation of functionally distinct T-cell subclasses is a differentiative process independent of antigen. **J Exp Med** 1975; 141:1376-89.

Carey TE, Takahashi T, Resnick LA, Oettgen HF, Old LJ. Cell surface antigens of human malignant melanoma: mixed hemadsorption assays for humoral immunity to cultured autologous melanoma cells. **Proc Natl Acad Sci U.S.A** 1976; 73:3278-82.

Cheever MA, Disis ML, Bernhard H, et al. Immunity to oncogenic proteins. **Immunol Rev** 1995; 145:33-59.

Chen YT, Scanlan MJ, Sahin U, et al. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. **Proc Natl Acad Sci U.S.A** 1997; 94:1914-8.

Chen YT, Gure AO, Tsang S, et al. Identification of multiple cancer/testis antigens by allogeneic antibody screening of a melanoma cell line library. **Proc Natl Acad Sci U.S.A** 1998; 95:6919-23.

Chen QY, Jackson H, Parente P, et al. Immunodominant CD4(+) responses identified in a patient vaccinated with full-length NY-ESO-1 formulated with ISCOMATRIX adjuvant. **Proc Natl Acad Sci U.S.A** 2004; 101:9363-8.

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

Coulie PG, Lehmann F, Lethe B et al. A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T-lymphocytes on a human-melanoma. **Proc Natl Acad Sci U.S.A** 1995; 92:7976-80.

Coulie PG, Karanikas V, Lurquin C, et al. Cytolitic T-cell responses of cancer patients vaccinated with a MAGE antigen. **Immunol Rev** 2002; 188:33-42.

De Plaen E, Arden K, Traversari C, et al. Structure, chromosomal localization, and expression of 12 genes of the MAGE family. **Immunogenetics** 1994; 40:360-9.

Davis ID, Chen WS, Jackson H, et al. Recombinant NY-ESO-1 protein with ISCOMATRIX adjuvant induces broad integrated antibody and CD4+ and CD8+ T cell responses in humans. **Proc Natl Acad Sci U.S.A** 2004; 101:10697-702.

de Wit NJ, Weidle UH, Ruiter DJ, van Muijen GN. Expression profiling of MMA-1a and splice variant MMA-1b: new cancer/testis antigens identified in human melanoma. **Int J Cancer** 2002; 98:547-53.

Dighe AS, Richards E, Old LJ, Schreiber RD. Enhanced in-vivo growth and resistance to rejection of tumor-cells expressing dominant-negative IFN-gamma receptors. **Immunity** 1994; 1:447-56.

Dong XY, Su YR, Qian XP, et al. Identification of two novel CT antigens and their capacity to elicit antibody response in hepatocellular carcinoma patients. **Br J Cancer** 2003; 89:291-7.

Dudley ME, Wunderlich JR, Robbins PF, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. **Science** 2002; 298:850-4.

Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. **Nat Immunol** 2002; 3:991-8.

Dunn GP, Old LJ, Schreiber RD. The three Es of cancer immunoediting. **Ann Rev Immunol** 2004; 22:329-60.

Eddy EM. Male germ cell gene expression. **Recent Prog Horm Res** 2002; 57:103-28.

Fallarino F, Gajewski TF. Cutting edge: Differentiation of antitumor CTL in vivo requires host expression of Stat1. **J Immunol** 1999; 163:4109-13.

Fisk B, Blevins TL, Wharton JT, Ioannides CG. Identification of an immunodominant peptide of her-2/neu protooncogene recognized by ovarian tumor-specific cytotoxic T-lymphocyte lines. **J Exp Med** 1995; 181:2109-17.

Fiszer D, Kurpisz M. Major histocompatibility complex expression on human, male germ cells: A review. **Am J Reprod Immunol** 1998; 40:172-6.

Gatti RA, Good RA. Occurrence of malignancy in immunodeficiency diseases - literature review. **Cancer** 1971; 28:89-98.

Girardi M, Oppenheim DE, Steele CR, et al. Regulation of cutaneous malignancy by gamma delta T cells. **Science** 2001; 294:605-9.

Gnjatic S, Nagata Y, Jager E, et al. Strategy for monitoring T cell responses to NY-ESO-1 in patients with any HLA class I allele. **Proc Natl Acad Sci U.S.A** 2000; 97:10917-22.

Gnjatic S, Jager E, Chen W, et al. CD8(+) T cell responses against a dominant cryptic HLA-A2 epitope after NY-ESO-1 peptide immunization of cancer patients. **Proc Natl Acad Sci U.S.A** 2002; 99:11813-8.

Gold P, Freedman SO. Specific carcinoembryonic antigens of the human digestive system. **J Exp Med** 1965; 122:467-81.

Gordon D, Alajian C, Green P. Consed: a graphical tool for sequence finishing. **Genome Res** 1998; 8:195-202.

Goydos JS, Patel M, Shih WC. NY-ESO-1 and CTp11 expression may correlate with stage of progression in melanoma. **J Surg Res** 2001; 98:76-80.

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.

Gure AO, Tureci O, Sahin U, et al. SSX: a multigene family with several members transcribed in normal testis and human cancer. **Int J Cancer** 1997; 72:965-71.

Hanahan D, Weinberg RA. The hallmarks of cancer. **Cell** 2000; 100:57-70.

Hoon DSB, Yuzuki D, Hayashida M, Morton DL. Melanoma patients immunized with melanoma cell vaccine induce antibody-responses to recombinant mage-1 antigen. **J Immunol** 1995; 154:730-7.

Huang X. On global sequence alignment. **Comput Appl Biosci** 1994; 10:227-35.

Hunig T. T-Cell function and specificity in athymic mice. **Immunol Today** 1983; 4:84-7.

Ikehara S, Pahwa RN, Fernandes G, Hansen CT, Good RA. Functional T-cells in athymic nude-mice. **Proc Natl Acad Sci U.S.A** 1984; 81:886-8.

Ishigaki Y, Li X, Serin G, Maquat LE. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. **Cell** 2001;106:607-17.

Izzotti A, Bagnasco M, Cartiglia C, Longobardi M, De Flora S. Proteomic analysis as related to transcriptome data in the lung of chromium(VI)-treated rats. **Int J Oncol** 2004; 24:1513-22.

Jager E, Chen YT, Drijfhout JW, et al. Simultaneous humoral and cellular immune response against cancer-testis antigen NY-ESO-1: definition of human histocompatibility leukocyte antigen (HLA)-A2-binding peptide epitopes. **J Exp Med** 1998; 187:265-70.

Jager E, Stockert E, Zidianakis Z, et al. Humoral immune responses of cancer patients against 'Cancer-Testis' antigen NY-ESO-1: correlation with clinical events. **Eur J Cancer** 1999; 35:S353-S4.

Jager D, Stockert E, Gure AO, et al. Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. **Cancer Res** 2001; 61:2055-61.

Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. **Science** 2001; 293:1068-70.

Jordan BW, Dinev D, LeMellay V, et al. Neurotrophin receptor-interacting mage homologue is an inducible inhibitor of apoptosis protein-interacting protein that augments cell death. **J Biol Chem** 2001; 276:39985-9.

Jungbluth AA, Busam KJ, Kolb D, et al. Expression of MAGE-antigens in normal tissues and cancer. **Int J Cancer** 2000; 85:460-5.

Jungbluth AA, Chen YT, Stockert E, et al. Immunohistochemical analysis of NY-ESO-1 antigen expression in normal and malignant human tissues. **Int J Cancer** 2001a; 92:856-60.

Jungbluth AA, Antonescu CR, Busam KJ, et al. Monophasic and biphasic synovial sarcomas abundantly express cancer/testis antigen NY-ESO-1 but not MAGE-A1 or CT7. **Int J Cancer** 2001b; 94:252-6.

Kacha AK, Fallarino F, Markiewicz MA, Gajewski TF. Cutting edge: spontaneous rejection of poorly immunogenic P1.HTR tumors by Stat6-deficient mice. **J Immunol** 2000; 165:6024-8.

Kaplan DH, Shankaran V, Dighe AS, et al. Demonstration of an interferon gamma-dependent tumor surveillance system in immunocompetent mice. **Proc Natl Acad Sci U.S.A.** 1998; 95:7556-61.

Kawakami Y, Eliyahu S, Delgado CH, et al. Cloning of the gene coding for a shared human-melanoma antigen recognized by autologous t-cells infiltrating into tumor. **Proc Natl Acad Sci U.S.A.** 1994; 91:3515-9.

Khong HT, Restifo NP. Natural selection of tumor variants in the generation of "tumor escape" phenotypes. **Nat Immunol** 2002; 3:999-1005.

Kimura M. The neutral theory of molecular evolution: a review of recent evidence. **Jpn J Genet** 1991; 66:367-86.

Knuth A, Danowski B, Oettgen HF, Old LJ. T-cell-mediated cyto-toxicity against autologous malignant-melanoma: analysis with interleukin 2-dependent t-cell cultures. **Proc Natl Acad Sci U.S.A** 1984; 81:3511-5.

Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. **Bioinformatics** 2001; 17:1244-5.

Kurashige T, Noguchi Y, Saika T, et al. NY-ESO-1 expression and immunogenicity associated with transitional cell carcinoma: Correlation with tumor grade. **Cancer Res** 2001; 61:4671-4.

Lea IA, Adoyo P, O'Rand MG. Autoimmunogenicity of the human sperm ptotein Sp17 in vasectomized men and identification of linear B cell pitopes. **Fertil Steril** 1997; 67:355-61.

Lennette ET, Winberg G, Yadav M, Enblad G, Klein G. Antibodies to Lmp2A/2B in Ebv-Carrying Malignancies. **Eur J Cancer** 1995; 31A:1875-8.

Lethe B, Lucas S, Michaux L, et al. LAGE-1, a new gene with tumor specificity. **Int J Cancer** 1998; 76:903-8.

Li WH, Gojobori T, Nei M. Pseudogenes as a paradigm of neutral evolution. **Nature** 1981; 292:237-9.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell JE. **Molecular cell biology.** 4th ed. New York: W. H. Freeman; 2000.Cancer; p.1054-84.

Lonchay C, van der Bruggen P, Connerotte T, et al. Correlation between tumor regression and T cell responses in melanoma patients vaccinated with a MAGE antigen. **Proc Natl Acad Sci U.S.A** 2004; 101:14631-8.

Lucas S, Brasseur F, Boon T. A new MAGE gene with ubiquitous expression does not code for known MAGE antigens recognized by T cells. **Cancer Res** 1999; 59:4100-3.

Maio M, Coral S, Sigalotti L, et al. Analysis of cancer/testis antigens in sporadic medullary thyroid carcinoma: Expression and humoral response to NY-ESO-1. **J Clin Endocrinol Metabolism** 2003; 88:748-54.

Maleckar JR, Sherman LA. The composition of the T-cell receptor repertoire in nude-mice. **J Immunol** 1987; 138:3873-6.

Marchand M, van Baren N, Weynants P, et al. Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA- A1. **Int J Cancer** 1999; 80:219-30.

Matsumoto K, Taniura H, Uetsuki T, Yoshikawa K. Necdin acts as a transcriptional repressor that interacts with multiple guanosine clusters. **Gene** 2001; 272:173-9.

McCurdy DK, Tai LQ, Nguyen J, et al. MAGE Xp-2: A member of the MAGE gene family isolated from an expression library using systemic lupus erythematosus sera. **Mol Genet Metabolism** 1998; 63:3-13.

Meuwissen RL, Offenberg HH, Dietrich AJ, Riesewijk A, van Iersel M, Heyting C. A coiled-coil related protein specific for synapsed regions of meiotic prophase chromosomes. **EMBO J** 1992; 11:5091-100.

Ministério da Saúde. Instituto Nacional de Câncer. **Estimativa 2003: incidência de câncer no Brasil.** Rio de Janeiro: INCA; 2002.

Mukherji B, Chakraborty NG, Yamasaki S, et al. Induction of antigen-specific cytolytic t-cells in-situ in human-melanoma by immunization with synthetic peptide-pulsed autologous antigen-presenting cells. **Proc Natl Acad Sci U.S.A** 1995; 92:8078-82.

Nei M, Kumar S. **Molecular evolution and phylogenetics**. Oxford: Oxford University Press; 2000.

Nestle FO, Alijagic S, Gilliet M, et al. Vaccination of melanoma patients with peptide- or tumor lysate-pulsed dendritic cells. **Nat Med** 1998; 4:328-32.

Old LJ, Stockert E. Immunogenetics of cell surface antigens of mouse leukemia. **Annu Rev Genet** 1977; 11:127-60.

Old LJ, Chen YT. New paths in human cancer serology. **J Exp Med** 1998; 187:1163-7.

Old LJ. Cancer/testis (CT) antigens - a new link between gametogenesis and cancer. **Cancer Immun** 2001; 1:1-7.

Omoto Y, Kobayashi S, Inoue S, et al. Evaluation of oestrogen receptor beta wild-type and variant protein expression, and relationship with clinicopathological factors in breast cancers. **Eur J Cancer** 2002; 38:380-6.

Onody P, Bertrand F, Muzeau F, Bieche I, Lidereau R. Fluorescence in situ hybridization and immunohistochemical assays for HER-2/neu status determination - Application to node-negative breast cancer. **Arch Pathol Lab Med** 2001; 125:746-50.

Penn I. Posttransplant malignancies. **Transplant Proc** 1999; 31:1260-2.

Pfreundschuh M, Shiku H, Takahashi T, et al. Serological analysis of cell surface antigens of malignant human brain tumors. **Proc Natl Acad Sci U.S.A** 1978; 75:5122-6.

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

Reymond A, Camargo AA, Deutsch S, et al. Nineteen additional unpredicted transcripts from human chromosome 21. **Genomics** 2002; 79:824-32.

Robertson KD, Uzvolgyi E, Liang G, et al. The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. **Nucleic Acids Res** 1999; 27:2291-8.

Rocha IM, Oliveira LJN, de Castro LCM, et al. Recognition of melanoma cell antigens with antibodies present in sera from patients with vitiligo. **Int J Dermatol** 2000; 39:840-3.

Rosenberg SA, Yang JC, Schwartzentruber DJ, et al. Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma. **Nat Med** 1998; 4:321-7.

Russell JH, Ley TJ. Lymphocyte-mediated cytotoxicity. **Ann Rev Immunol** 2002; 20:323-70.

Sahin U, Tureci O, Pfreundschuh M. Serological identification of human tumor antigens. **Curr Opin Immunol** 1997; 9:709-16.

Sahin U, Tureci O, Schmitt H, et al. Human neoplasms elicit multiple specific immune responses in the autologous host. **Proc Natl Acad Sci U.S.A** 1995; 92:11810-3.

Salehi AH, Roux PP, Kubu CJ, et al. NRAGE, a novel MAGE protein, interacts with the p75 neurotrophin receptor and facilitates nerve growth factor-dependent apoptosis. **Neuron** 2000; 27:279-88.

Sanguinetti CJ, Dias NE, Simpson AJ. Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. **Biotechniques** 1994; 17:914-21.

Scanlan MJ, Chen YT, Williamson B, et al. Characterization of human colon cancer antigens recognized by autologous antibodies. **Int J Cancer** 1998; 76:652-8.

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

Scanlan MJ, Gordon CM, Williamson B, et al. Identification of cancer/testis genes by database mining and mrna expression analysis. **Int J Cancer** 2002b; 98:485-92.

Scanlan MJ, Simpson AJ, Old LJ. The cancer/testis genes: Review, standardization, and commentary. **Cancer Immun** 2004; 4:1-15.

Schell T, Kulozik AE, Hentze MW. Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. **Genome Biol** 2002; 3: 1006.1-1006.6.

Shankaran V, Ikeda H, Bruce AT, et al. IFN gamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. **Nature** 2001; 410:1107-11.

Shichijo S, Yamada A, Sagawa K, et al. Induction of MAGE genes in lymphoid cells by the demethylating agent 5-aza-2'-deoxycytidine. **Jpn J Cancer Res** 1996; 87:751-6.

Shiku H, Kisielow P, Bean MA, et al. Expression of T-cell differentiation antigens on effector cells in cell- mediated cytotoxicity in vitro: evidence for functional heterogeneity related to the surface phenotype of T cells. **J Exp Med** 1975; 141:227-41.

Shinkai Y, Rathbun G, Lam KP, et al. Rag-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. **Cell** 1992; 68:855-67.

Smyth MJ, Thia KYT, Street SEA, et al. Differential tumor surveillance by natural killer (NK) and NKT cells. **J Exp Med** 2000a; 191:661-8.

Smyth MJ, Thia KYT, Street SEA, MacGregor D, Godfrey DI, Trapani JA. Perforin-mediated cytotoxicity is critical for surveillance of spontaneous lymphoma. **J Exp Med** 2000b; 192:755-60.

Smyth MJ, Crowe NY, Godfrey DI. NK cells and NKT cells collaborate in host protection from methylcholanthrene-induced fibrosarcoma. **Int Immunol** 2001; 13:459-63.

Smyth MJ, Crowe NY, Hayakawa Y, Takeda K, Yagita H, Godfrey DI. NKT cells - conductors of tumor immunity? **Curr Opin in Immunol** 2002a; 14:165-71.

Smyth MJ, Hayakawa Y, Takeda K, Yagita H. New aspects of natural-killer-cell surveillance and therapy of cancer. **Nat Rev Cancer** 2002b; 2:850-61.

Sogayar MC, Camargo AA, Bettoni F, et al. A transcript finishing initiative for closing gaps in the human transcriptome. **Genome Res** 2004; 14:1413-23.

Stockert E, Jager E, Chen YT, et al. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens**. J Exp Med** 1998; 187:1349-54.

Street SEA, Cretney E, Smyth MJ. Perforin and interferon-gamma activities independently control tumor initiation, growth, and metastasis**. Blood** 2001; 97:192-7.

Street SEA, Trapani JA, MacGregor D, Smyth MJ. Suppression of lymphoma and epithelial malignancies effected by interferon gamma. **J Exp Med** 2002; 196:129-34.

Stutman O. Tumor Development After 3-Methylcholanthrene in Immunologically Deficient Athymic Nude Mice. **Science** 1974; 183:534-6.

Stutman O. Chemical carcinogenesis in nude-mice - comparison between nude-mice from homozygous matings and heterozygous matings and effect of age and carcinogen dose. **J Natl Cancer Inst** 1979; 62:353-8.

Taniura H, Matsumoto K, Yoshikawa K. Physical and functional interactions of neuronal growth suppressor necdin with p53. **J Biol Chem** 1999; 274:16242-8.

Thomas L. Discussion. In: Lawrence HS, editor. **Cellular and humoral aspects of the hypersensitive states.** New York: Hoeber-Harper; 1959. p.529-32.

Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res** 1994; 22:4673-80.

Tindle RW. Human papillomavirus vaccines for cervical cancer. **Curr Opin Immunol** 1996; 8:643-50.

Titzer S, Christensen O, Manzke O, et al. Vaccination of multiple myeloma patients with idiotype-pulsed dendritic cells: immunological and clinical aspects. **Br J Haematol** 2000; 108:805-16.

Traversari C, van der Bruggen P, van den EB, et al. Transfection and expression of a gene coding for a human melanoma antigen recognized by autologous cytolytic T lymphocytes. **Immunogenetics** 1992; 35:145-52.

Ueda R, Shiku H, Pfreundschuh M, et al. Cell surface antigens of human renal cancer defined by autologous typing. **J Exp Med** 1979; 150:564-79.

van der Bruggen P, Traversari C, Chomez P, et al. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. **Science** 1991; 254:1643-7.

van den Broek MF, Kagi D, Ossendorp F, et al. Decreased tumor surveillance in perforin-deficient mice. **J Exp Med** 1996; 184:1781-90.

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

Vidaeus CM, Kapp-Herr C, Golden WL, Eddy RL, Shows TB, Herr JC. Human fertilin beta: identification, characterization, and chromosomal mapping of an ADAM gene family member. **Mol Reprod Dev** 1997; 46:363-9.

Wang J, Hu L, Hamilton SR, Coombes KR, Zhang W. RNA amplification strategies for cDNA microarray experiments. **Biotechniques** 2003; 34:394-400.

Weiser TS, Guo ZS, Ohnmacht GA, et al. Sequential 5-Aza-2'-deoxycytidine-depsipeptide FR901228 treatment induces apoptosis preferentially in cancer cells and facilitates their recognition by cytolytic T lymphocytes specific for NY-ESO-1. **J Immunother** 2001; 24:151-61.

Wolfel T, Hauer M, Schneider J, et al. A P16(Ink4A)-insensitive cdk4 mutant targeted by cytolytic T-lymphocytes in a human-melanoma. **Science** 1995; 269:1281-4.

Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. **Mol Cell Biol** 1998; 18: 5272-83.

Zendman AJ, Cornelissen IM, Weidle UH, Ruiter DJ, van Muijen GN. CTp11, a novel member of the family of human cancer/testis antigens. **Cancer Res** 1999; 59:6223-9.

Zendman AJ, Van Kraats AA, Weidle UH, Ruiter DJ, van Muijen GN. The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma. **Int J Cancer** 2002; 99:361-9.

Zendman AJ, Ruiter DJ, Van Muijen GNP. Cancer/testis-associated genes: Identification, expression profile, and putative function. **J Cell Physiol** 2003; 194:272-88.

ANEXO 1

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A novel human G protein-coupled receptor is over-expressed in prostate cancer

Raphael B. Parmigiani[2], Geraldo S. Magalhães[1], Pedro A.F. Galante[1,2], Carina V.B. Manzini[1], Anamaria A. Camargo[2] and Bettina Malnic[1]

[1]Departamento de Bioquímica, Universidade de São Paulo, São Paulo, SP, Brasil
[2]Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente, 109, 4th floor, São Paulo, SP, Brasil
Corresponding author: B. Malnic
E-mail: bmalnic@iq.usp.br

Genet. Mol. Res. 3 (4): 521-531 (2004)
Received October 4, 2004
Accepted December 10, 2004
Published December 30, 2004

ABSTRACT. G protein-coupled receptors (GPCRs) are involved in a large variety of physiological functions. The number of known members that belong to this large family of receptors has been rapidly increasing. Now, with the availability of the human genome sequence databases, further family members are being identified. We describe the identification of a novel GPCR that shows no significant amino acid identity to any one of the known members of the GPCR superfamily. The gene expression pattern of this receptor is restricted: in normal tissues it is confined to the nervous system and testis, but we also detected gene expression in several tumor types, most notably prostate cancer, suggesting a potential role for this gene as a marker for this disease.

Key words: GPCR, Brain, Testis, Prostate cancer, Orphan receptor, Bioinformatics

INTRODUCTION

G protein-coupled receptors (GPCRs) comprise the largest family of surface molecules involved in signal transduction. They are activated by a large variety of ligands, including hormones, growth factors, light, peptides, neurotransmitters, nucleotides and odorants, and are involved in the regulation of a range of cellular responses (Bockaert and Pin, 1999; Pierce et al., 2002). All GPCRs have a common central core domain consisting of seven transmembrane helices connected by three intracellular loops and three extracellular loops. The lengths of the N-terminal and C-terminal domains are variable in different GPCRs.

The large number of GPCRs and their involvement in human diseases has made these receptors attractive drug targets. According to a recent analysis, over 30% of marketed drugs are active against this receptor super-family and yet, only a small fraction of the known GPCRs are at present drug targets (Drews, 2000). Over the last 10 years, a large number of novel GPCRs have been identified, mostly through homology cloning. The ligands remain unknown for most of these receptors (Marchese et al., 1999; Howard et al., 2001; Lee et al., 2002). Many GPCR genes do not have introns, facilitating their identification within the recently available human genome sequences (I.H.G.S.C., 2001; Venter et al., 2001). Some of the newly discovered genes are more closely related to previously known GPCRs and are likely to be activated by similar ligands. Others are more distantly related to known GPCRs and may have quite different functions. The identification of novel GPCRs, and their respective ligands, will continue to contribute to the understanding and identification of new signaling pathways involved in different aspects of human physiology. Of particular interest are receptors with recognized expression in the central nervous system, given that many psychiatric and neurodegenerative disorders are mediated by unknown mechanisms.

We describe a novel putative GPCR from the human genome sequence database and characterize its expression pattern in normal human tissues as well as in tumor cell lines and tissues. The receptor was named brain testis restricted (BTR), due to its restricted expression in normal tissues. We also detected expression of the BTR gene in tumor cell lines and tissues, predominantly in prostate cancer. There is a significant correlation between BTR gene expression and prostate cancer, suggesting a potential role for BTR as a marker for detection of this type of cancer.

MATERIAL AND METHODS

Sequence analysis

The human genome database (htgs, NCBI) was searched with consensus motifs common to members of odorant receptors (MAYDRYVAIC and KAFSTCASH) using TBLASTN. A genomic sequence containing an open reading frame (ORF) of 1044 bp was identified. The amino acid sequence was analyzed using tools in the protein fingerprint database (PRINTS) (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS), PROSITE (http://www.expasy.org/prosite/), and PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). Transmembrane regions were predicted with TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) (Krogh et al., 2001).

Phylogenetic tree

We selected 31 GPCR amino acid sequences belonging to different families. The extremely variable amino-terminal and carboxy-terminal regions were deleted from all of these receptor sequences. The sequences were then aligned using ClustalW version 1.8. The resulting multiple alignment was used as input to Mega2 (Kumar et al., 2001) to construct a neighbor-joining tree from 1000 replicates of the interior branch test.

RT-PCR

A panel of total RNA from 20 different human tissues was purchased from Clontech (Palo Alto, CA, USA). Additionally, RNA was prepared from tumor cell lines and tumor tissues by the guanidinium thiocyanate method. RNAs were treated with DNAse (Promega) according to the manufacturer's instructions. Tumor samples were obtained from the A.C. Camargo Hospital tumor collection. First, 1 µg total RNA, plus 100 ng oligo (dT) in 13.5 µl DEPC-treated water, was incubated for 2 min at 70°C. The reaction was rapidly chilled on ice and used to synthesize cDNA in 20 µl 1X Superscript II first strand buffer containing 0.5 mM dNTP, 3 mM $MgCl_2$, 20 U RNAse inhibitor (RNaseOUT; Invitrogen Life Technologies) and 200 U Superscript II reverse transcriptase at 42°C for 60 min. Control templates for checking amplification of genomic DNA were prepared without reverse transcriptase. The product was diluted to 100 µl and 2.5 µl was used for PCR to amplify BTR.

Primers were synthesized by Operon Technologies Inc. The forward and reverse primers match putative transmembrane regions IV and VII, respectively, and should produce a 450-bp long PCR product (BTR1F: 5' GCTACCTCTCCTTCATGTCC; BTR4R: 5' GCATC ATGAGTACCTCACTG).

Twenty-five microliter PCR reactions containing 2.5 µl cDNA, 0.2 mM dNTP, 1.5 mM $MgCl_2$, 0.5 µM of each forward and reverse primers, 1.25 U Platinum Taq DNA polymerase (Invitrogen Life Technologies) were heated to 95°C for 2 min, followed by 35 thermal cycles of 95°C for 45 s, 55°C for 45 s, 72°C for 1 min, and a final incubation at 72°C for 6 min. PCR products were analyzed in 1.5% agarose gels stained with ethidium bromide or in Southern blots using a purified BTR PCR product as a probe. The identity of the PCR products was also confirmed by DNA sequencing.

The relative intensities of bands in Figures 2A, 3 and 4 were determined by densitometry using the program LabWorks (UVP). The densities for the different bands were normalized with their corresponding GAPDH band densities.

The primers BTRA: 5' ATGGGGGATGAGCTGGCAC and BTRB: 5' CTAGGAAA TGGTAAAGATGGC were used to amplify the whole coding region of BTR from human brain cDNA. The sequence of the obtained PCR was deposited in Genbank (accession number: AY280965).

RESULTS

Identification of a novel GPCR

In previous experiments we used peptide motifs commonly seen in odorant receptors

(OR) as queries to scan the human genome for new OR genes (Malnic et al., 2004). OR genes, like many other GPCR genes, have no introns in their coding sequence (Buck and Axel, 1991), which makes it easier to search for related genes by scanning the human genome databases. Among the many OR gene sequences identified, we found a putative GPCR that did not show any of the common motifs which characterize the OR family members (see Methods). BLASTN and TBLASTN searches, using this translated gene as query, were performed to browse for homologous proteins in the NCBI database (nr and htgs). No highly similar sequences were found: the highest identities were to the human serotonin 1D receptor (5HT1D) (24% AI; 49/202) and alpha-A1-adrenergic receptor (25% AI; 49/190). Although the identity scores to these two GPCRs are very low, a protein motif search with TMHMM (a tool for prediction of transmembrane helices in proteins) indicated seven putative transmembrane regions, suggesting that this gene must code for a new member of the GPCR super family. The fact that we did not find other human proteins closely related to this new putative GPCR indicates that it does not belong to a new family of genes, but instead it is a solitary gene. Chromosomal mapping using Entrez map viewer assigned this gene to chromosome 2q21.

During this research the identification of 109 novel human GPCRs out of the human genome sequence (including BTR) was reported by another group (Takeda et al., 2002). A second group also recently reported the identification of 367 total GPCRs in humans, and BTR is also included in this list (Vassilatis et al., 2003). Nevertheless, neither of the two groups described the tissue distribution of BTR.

We also scanned the mouse genome sequence from Celera and NCBI (build 30) databases (I.H.G.S.C., 2001; Venter et al., 2001) for a putative homolog for this gene. Neither of the two databases contained sequences related to the BTR gene. We compared the human genomic region containing the human BTR gene (on chromosome 2q21.3) to its corresponding synthenic region on the mouse chromosome 1 (1B). We found mouse orthologs for two human genes flanking the BTR gene in this region, but we found no gene similar to BTR located between these genes. Additionally, no mouse ESTs could be found matching the BTR gene. Thus, it seems that there is no BTR gene in the mouse genome, although we cannot exclude the possibility that gaps in the mouse genome sequences might have precluded us from identifying the mouse counterpart for this gene.

Comparison with other GPCRs

GPCRs comprise very large and diverse groups of gene families that recognize distinct ligands. Members of the GPCR superfamily can be grouped in different families based on their sequence similarities. The families characterized best so far are: the rhodopsin family (family A), the secretin receptor family (family B), and the metabotropic glutamate receptor family (family C) (Attwood and Findlay, 1994; Kolakowski, 1994; Strader et al., 1995; Bockaert and Pin, 1999; Pierce et al., 2002). In addition to rhodopsin, family A includes the β-adrenergic receptors, serotonin receptors, OR, adenosine receptors, and many others. Family B includes receptors for polypeptide hormones, such as glucagon, secretin and calcitonin. Family C includes all mGluR types, Ca^{2+} sensing and $GABA_B$ receptors, a group of pheromones (termed VRs or G_0VN; reviewed in Bargmann, 1997), and a small group of taste receptors, the T1Rs). Other families include the pheromone receptors (VNRs; Dulac and Axel, 1995) and the taste receptors T2Rs (for a review on taste receptors, see Montmayeur and Matsunami, 2002).

Sequence identities can be very low among the most distant GPCRs (<20% ASI), but it is expected that receptors with related functions share conserved sequence motifs. In order to determine if BTR can be assigned to one of the known families, we generated a multiple alignment containing 31 GPCR amino acid sequences representing different families and we constructed a phylogenetic tree from 1000 interior branch test replicates. BTR constituted a separate branch, more closely related to family A, although with an interior branch test value of only 38% (Figure 1). This indicates that BTR is not closely related to any of the known families.

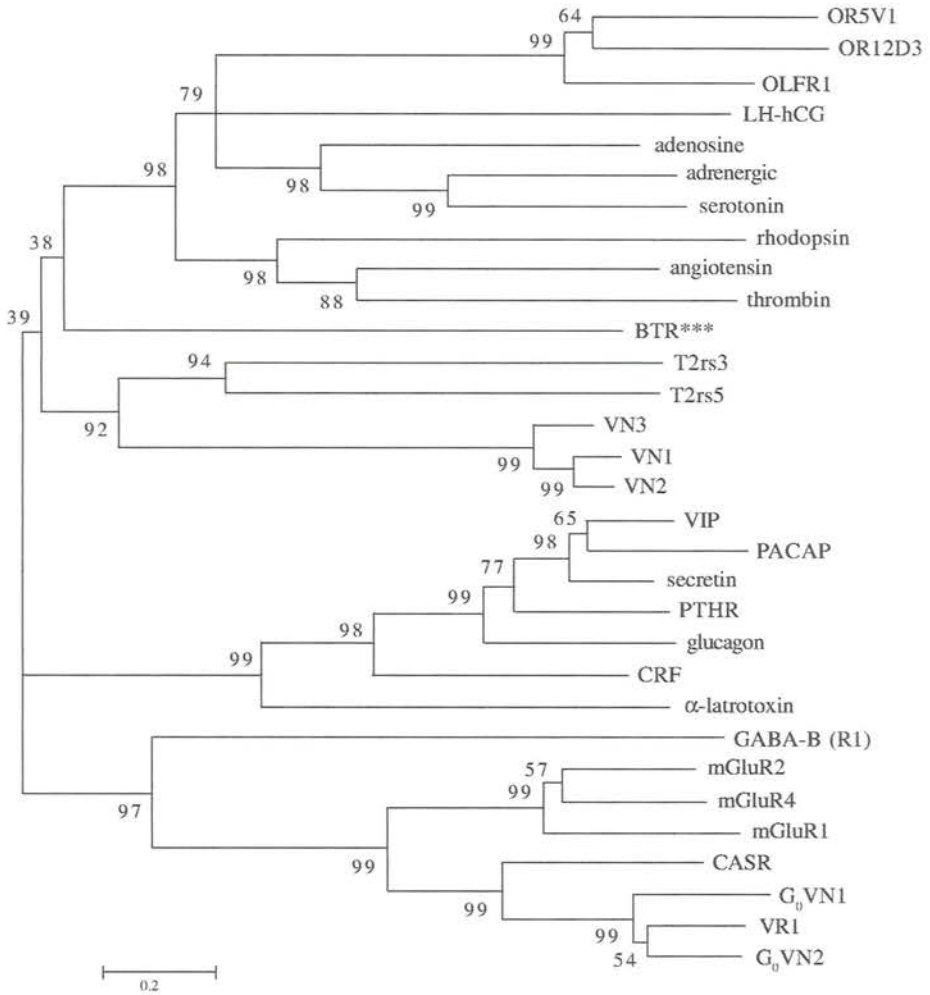


Figure 1. Phylogenetic tree showing the comparison of BTR with other GPCRs. Family A sequences are: odorant (OR5V1, OR12D3, OLFR1), luteinizing hormone-human choriogonadotropic hormone (LH-hCG), adenosine, adrenergic, serotonin, rhodopsin, angiotensin, and thrombin receptors. Family B sequences are: vasoactive intestinal polypeptide (VIP), pituitary adenylate cyclase-activating polypeptide (PACAP), secretin, parathyroid hormone receptor (PTHR), glucagon, corticotropin-releasing factor (CRF) and α-latrotoxin. Family C sequences are: GABA-B, metabotropic glutamate receptors (mGluR) 1, 2 and 4, calcium-sensing receptor (CASR) and the VRs or G_0VN pheromone receptors (G_0VN1, VR1, G_0VN2). The T2R taste receptors (T2rs3 and T2rs5) and the VNR or Gi2VN pheromone receptors (VN1, VN2 and VN3) were also included. Brain testis restricted (BTR) is indicated by asterisks. The numbers indicate the interior branch test values for 1000 replicates. The bar indicates the number of amino acid substitutions per site.

BTR gene expression in normal tissues

To investigate the tissue expression pattern of BTR gene, we conducted RT-PCR experiments using a panel of total RNA from 20 different normal human tissues. Expression was detected only in testis, fetal brain, whole brain, and spinal cord (Figure 2A). The gene expression levels in fetal brain, whole brain and spinal cord tissues were detected at 4, 27, and 19%, respectively, of the level detected in the testis (see Methods). When human genomic DNA was used as a template with the same set of primers, a PCR product containing a sequence identical to the cDNA PCR product was obtained, confirming that BTR is an intronless gene. We only detected PCR products when reverse transcriptase (RT) was added to the reaction, which confirmed the absence of contaminating genomic DNA in our RNA samples (Figure 2B). The complete BTR coding region was also amplified from brain cDNA, and the sequence was deposited in Genbank (accession number AY280965). These results indicate that the BTR gene is preferentially expressed in testis and in the nervous system.

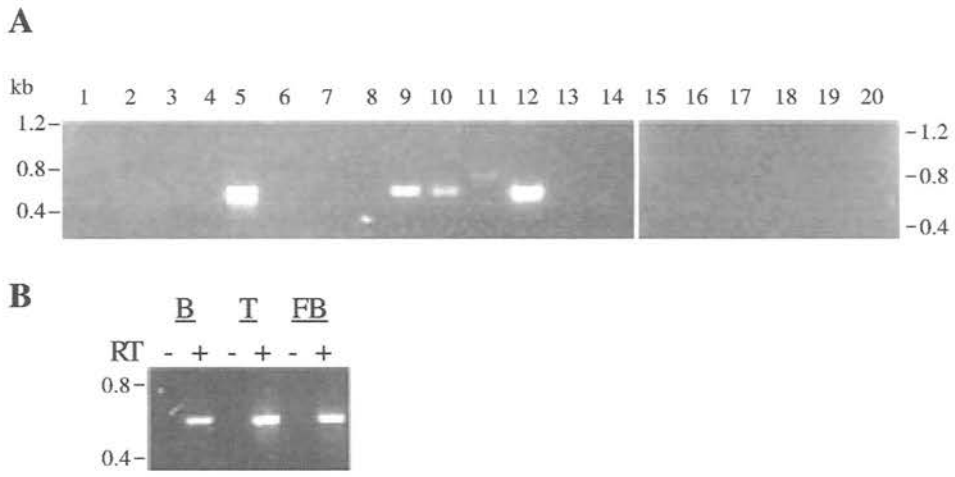


Figure 2. RT-PCR analysis of BTR gene expression in normal tissues. **A**. RT-PCR was conducted using 1 µg total RNA and 35 PCR cycles from the following human tissues: lanes 1, kidney; 2, lung; 3, uterus; 4, colon; 5, testis; 6, liver; 7, spleen; 8, thymus; 9, whole brain; 10, spinal cord; 11, adrenal gland; 12, fetal brain; 13, heart; 14, trachea; 15, bone marrow; 16, prostate; 17, skeletal muscle; 18, fetal liver; 19, placenta, and 20, salivary gland. **B**. A control RT-PCR was conducted in the absence (-) or presence (+) of reverse transcriptase (RT) from 1 µg of brain (B), testis (T) or fetal brain (FB) total RNA. BTR PCR products (450 bp) were observed only when RT was added to the reaction.

BTR gene expression in tumor cell lines and tissues

There are some examples of highly tissue-restricted gene products that are also expressed in cancer. A group of proteins, called cancer/testis/brain antigens (CTB antigens), expressed in normal testis and brain and in different types of tumors has been described. These proteins were originally identified as the target molecules recognized by autoantibodies in patients with paraneoplastic syndromes (Dropcho et al., 1987; Dalmau et al., 1999; Voltz et al., 1999; Scanlan et al., 2002a). We checked whether the BTR gene is also expressed in malignant tissues. We performed RT-PCR using RNA extracted from different cell lines as templates

(Figure 3). BTR was expressed in the prostate tumor cell lines Dul45 and PC3, in the H1155 lung tumor cell line, in the Skmel melanoma cell line, and in the MDA436 mammary cell line. We also detected weak BTR gene expression after 35 cycles in normal prostate tissue, which we had not detected before (compare Figures 2A and 3A). This is probably because we used Southern blot to detect the PCR products, which is a more sensitive technique. BTR gene expression in the prostate cell lines was 20X (Dul45) or 7X (PC3) stronger than the expression in normal prostate tissue after 35 cycles.

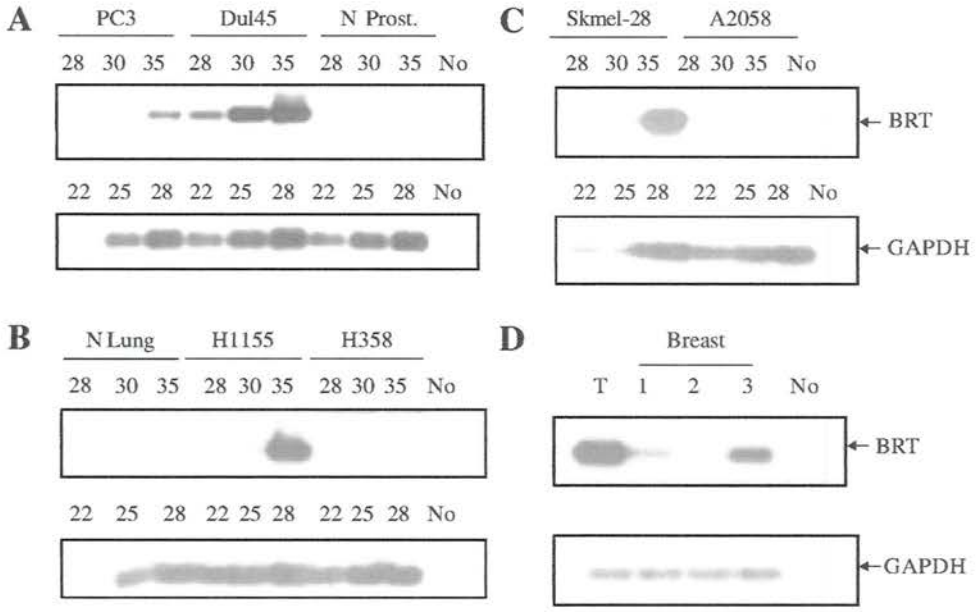


Figure 3. RT-PCR analysis of BTR gene expression in tumor cell lines. Expression of brain testis restriction (BTR) mRNA transcripts in: **A**, normal prostate (N Prost.) and the Dul45 and PC3 prostate tumor cell lines; **B**, normal lung (N Lung) and the H1155 and H358 lung tumor cell lines, and **C**, the Skmel-28 and A2058 melanoma cell lines, following 28, 30 and 35 cycles of PCR. The PCR products corresponding to BTR gene (450 bp) are indicated. No: negative control (no DNA added). The bottom panels show expression of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) mRNA from the same samples following 22, 25 and 28 cycles of PCR. **D**, Expression of BTR gene in the ZR75.3A (lane 1), MCF-7 (lane 2) and MDA436 (lane 3) mammary tumor cell lines after 35 cycles of PCR. T: normal testis tissue. The bottom panel shows expression of GAPDH mRNA from the same samples following 35 cycles of PCR. PCR products were detected through Southern blotting using a BTR probe.

We next assessed whether prostate tumor samples also express the BTR gene. Expression was detected in all prostate tumor samples tested (25/25; Figure 4). Cancer/testis antigen expression may be associated with tumor progression and with tumors of higher malignant potential (Scanlan et al., 2002b); however, this correlation was not observed in our analysis since expression was detected in all samples independently of the progression stage. We compared the relative levels of BTR gene expression in normal prostate tissue and in the prostate tumor samples. Of the 25 samples, only three showed expression levels similar or lower to the one observed for normal prostate tissue (samples 20, 21 and 24). The remaining 22 samples showed higher levels of expression, varying from three to thirty times the amount of BTR

expression observed in the normal prostate tissue. These results indicate that BTR gene expression may constitute a consistent marker for prostate cancer.

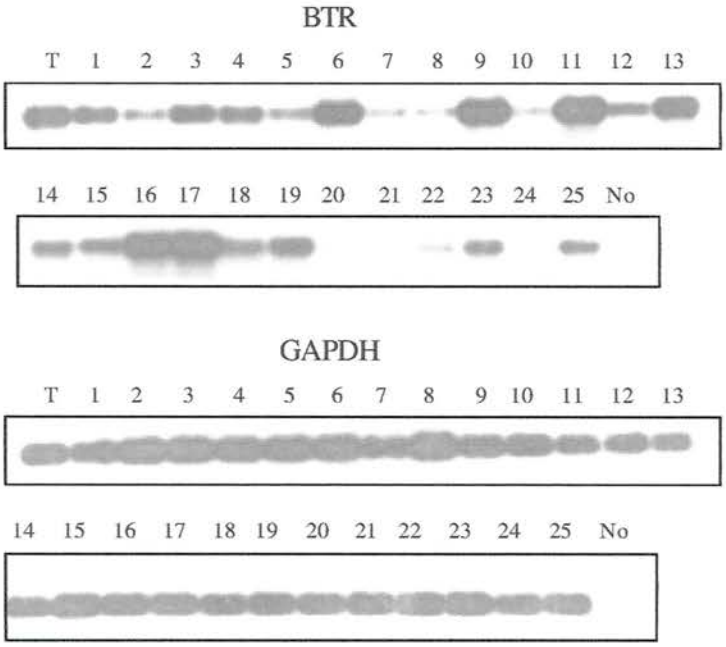


Figure 4. RT-PCR analysis of BTR gene expression in prostate tumor tissues. The upper panels show the expression of brain testis restricted (BTR) mRNA in prostate tumor samples from 25 different patients (1-25), following 35 cycles of PCR. The bottom panels show the expression of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) mRNA from the same samples following 35 cycles of PCR. PCR products were detected through Southern blotting using a BTR probe. T: normal testis tissue; No: no DNA added.

DISCUSSION

GPCRs comprise the largest family of proteins in many species. There are hundreds of GPCRs in humans, many of which have no known ligand or function (called "orphan GPCRs"; Howard et al., 2001). Analysis of the human genome sequence indicates that there are many other unknown GPCRs (I.H.G.S.C., 2001; Venter et al., 2001). We describe the identification and characterization of a novel putative member of this large family of receptors. Comparison of BTR with representative members of the GPCR super family indicates that this novel member constitutes a separate subtype of receptor and may have a quite distinct function, yet to be determined.

We found no evidence that there is a mouse BTR gene. It is possible that other GPCRs are also absent in rodents. Indeed, it was previously demonstrated that GPR8, a human orphan GPCR related to opioid and somatostatin receptors, was not found in rodents (Lee et al., 1999). It is believed that most mouse genes have corresponding orthologs in humans, with only a small fraction of the genes (less than 1%) showing no homolog in the other species (Mouse Genome

Sequencing Consortium, 2002). These genes must be either involved in specific functions, inherent to one species but not the other, or just be redundant or even non-functional. The identification of all of the species-specific genes through the comparison of the mouse and human genome sequences will clarify if some of these genes play important roles in the determination of characteristics that are specific to human or mouse.

The function of the BTR gene product is unknown. The fact that this gene is preferentially expressed in the brain and spinal cord, besides the testis, indicates that its function must be related to physiological aspects of the nervous system. We do not know if this gene is expressed throughout the brain, or if expression is confined to some particular regions within the brain. The identification of the exact regions of BTR expression in the nervous system will probably contribute to the understanding of this receptor's function.

GPCRs have only been recently identified as mediators of cellular growth and differentiation (Dhanasekaran et al., 1995; Luttrell et al., 1999). A large number of studies have also demonstrated that GPCRs can promote tumor formation (for a review, see Whitehead et al., 2001). Recently, a prostate specific GPCR (termed *PSGR*), which is expressed at higher levels in prostate cancer, was identified (Xu et al., 2000; Xia et al., 2001). Tumor-associated overexpression of *PSGR* was identified in 62% of the prostate specimens analyzed (32 of 52) (Xu et al., 2000). Like the CTB antigens (Dropcho et al., 1987; Dalmau et al., 1999; Voltz et al., 1999; Scanlan et al., 2002a), the BTR gene is expressed in cancer cell lines and tumors. Whether the BTR gene product is able to elicit an autoimmune response in patients with prostate cancer, still needs to be determined. If this turns out to be the case, the detection of anti-BTR antibodies could be useful for the diagnosis of prostate cancer.

In conclusion, we identified a new human GPCR that is expressed preferentially in the normal nervous system and testis. Further studies are required to determine the precise regions within the brain where the BTR gene is expressed, which would contribute to the understanding of BTRs function. We also found a correlation between prostate tumor and elevated levels of BTR gene expression, suggesting a potential role for the gene as a marker for the disease.

ACKNOWLEDGMENTS

We thank Jean Pierre Montmayeur, Sandro José de Souza and Andrew Simpson for reviewing the manuscript. We also thank the Hospital A.C. Camargo (São Paulo, Brazil) for providing the tumor samples, Dr. Simone Treiger Sredni for help with the description of the clinical stages of the different prostate tumors analyzed in this study and Dr. Ronaldo B. Quaggio for help with the Labworks software. Research supported by FAPESP, by the São Paulo branch of the Ludwig Institute for Cancer Research, and by grants from FAPESP to P.A.F. Galante and from CAPES to R.B. Parmigiani.

REFERENCES

Attwood, T. and **Findlay, J.** (1994). Fingerprinting G-protein-coupled receptors. *Protein Eng. 7*: 195-203.

Bargmann, C. (1997). Olfactory receptors, vomeronasal receptors and the organization of olfactory information. *Cell 90*: 585-587.

Bockaert, J. and **Pin, J.** (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J. 18*: 1723-1729.

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

Dalmau, J., Gultekin, S., Voltz, R., Hoard, R., DesChamps, T., Balmaceda, C., Batchelor, T., Gerstner, E., Eichen, J., Frennier, J., Posner, J. and **Rosenfeld, M.** (1999). Ma 1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. *Brain 122*: 27-39.

Dhanasekaran, N., Heasley, L. and **Johnson, G.** (1995). G protein-coupled receptor systems involved in cell growth and oncogenesis. *Endocrinol. Rev. 16*: 259-270.

Drews, J. (2000). Drug discovery: a historical perspective. *Science 287*: 1960-1964.

Dropcho, E., Chen, Y., Posner, J. and **Old, L.** (1987). Cloning of a brain protein identified by autoantibodies from a patient with paraneoplastic cerebellar degeneration. *Proc. Natl. Acad. Sci. USA 84*: 4552-4556.

Dulac, C. and **Axel, R.** (1995). A novel family of genes encoding putative pheromone receptors. *Cell 83*: 195-206.

Howard, A., McAllister, G., Feighner, S., Liu, Q., Nargund, R., Van de Ploeg, L. and **Patchett, A.** (2001). Orphan G protein-coupled receptors and natural ligand discovery. *Trends in Pharmacol. Sci. 22*: 132-140.

I.H.G.S.C. (2001). Initial sequencing and analysis of the human genome. *Nature 409*: 860-921.

Kolakowski, L.F. (1994). GCRDb: a G-protein-coupled receptor database. *Receptors Channels 2*: 1-7.

Krogh, A., Larsson, B., Von Heijne, G. and **Sonnhammer, E.** (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol. 305*: 567-580.

Kumar, S., Tamura, K., Jakobsen, I. and **Nei, M.** (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics 17*: 1244-1245.

Lee, D., Nguyen, T., Porter, C., Cheng, R., George, S. and **O'Dowd, B.** (1999). Two related G protein-coupled receptors: the distribution of GPR7 in rat brain and the absence of GPR8 in rodents. *Mol. Brain Res. 71*: 96-103.

Lee, D., George, S. and **O'Dowd, B.** (2002). Novel G protein-coupled receptor genes expressed in the brain: continued discovery of important therapeutic targets. *Expert Opin. Ther. Targets 6*: 185-202.

Luttrell, L., Daaka, Y. and **Lefkowitz, R.** (1999). Regulation of tyrosine kinase cascades by G-protein coupled receptors. *Curr. Opin. Cell Biol. 11*: 177-183.

Malnic, B., Godfrey, P. and **Buck, L.** (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. 101*: 2584-2589.

Marchese, A., George, S., Kolakowski Jr., L., Lynch, K. and **O'Dowd, B.** (1999). Novel GPCRs and their endogenous ligand: expanding the boundaries of physiology and pharmacology. *Trends Pharmacol. Sci. 20*: 370-375.

Montmayeur, J. and **Matsunami, H.** (2002). Receptors for bitter and sweet taste. *Curr. Opin. Neurobiol. 10*: 519-527.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature 420*: 520-562.

Pierce, K., Premont, R. and **Lefkowitz, R.** (2002). Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol. 3*: 639-650.

Scanlan, M., Gordon, C., Williamson, B., Lee, S., Chen, Y., Stockert, E., Jungbluth, A., Ritter, G., Jager, D., Jager, E., Knuth, A. and **Old, L.** (2002a). Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int. J. Cancer 98*: 485-492.

Scanlan, M., Gure, A., Jungbluth, A., Old, L. and **Chen, Y.** (2002b). Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol. Rev. 188*: 22-32.

Strader, C., Ming Fong, T., Graziano, M. and **Tota, M.** (1995). The family of G-protein-coupled receptors. *FASEB J. 9*: 745-754.

Takeda, S., Kadowaki, S., Haga, T., Takaesu, H. and **Mitaku, S.** (2002). Identification of G protein-coupled receptor genes from the human genome sequence. *FEBS Let. 520*: 97-101.

Vassilatis, D., Hohmann, J., Zeng, H., Li, F., Ranchalis, J., Mortrud, M., Brown, A., Rodriguez, S., Weller, J., Wright, A., Bergmann, J. and **Gaitanaris, G.** (2003). The G protein-coupled receptor repertoires of human and mouse. *Proc. Natl. Acad. Sci. USA 100*: 4903-4908.

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

Voltz, R., Gultekin, S., Rosenfeld, M., Gerstner, E., Eichen, J., Posner, J. and **Dalmau, J.** (1999). A serologic marker of paraneoplastic limbic and brain-stem encephalitis in patients with testicular cancer. *N. Engl. J. Med. 340*: 1788-1795.

Whitehead, I., Zohn, I. and **Der, C.** (2001). Rho GTPase-dependent transformation by G protein-coupled receptors. *Oncogene 20*: 1547-1555.

Xia, C., Ma, W., Wang, F., Hua, S.-B. and **Liu, M.** (2001). Identification of a prostate-specific G-protein coupled receptor in prostate cancer. *Oncogene 20*: 5903-5907.

Xu, L., Stackhouse, B., Florence, K., Zhang, W., Shanmugam, N., Sesterhenn, I., Zou, Z., Srikantan, V., Augustus, M., Roschke, V., Carter, K., McLeod, D., Moul, J., Soppett, D. and **Srivastava, S.** (2000). PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor, is overexpressed in prostate cancer. *Cancer Res. 60*: 6568-6572.

ANEXO 2

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Identification and complete sequencing of novel human transcripts through the use of mouse orthologs and testis cDNA sequences

Elisa N. Ferreira[1*], Lilian C. Pires[1*], Raphael B. Parmigiani[1], Fabiana Bettoni[1], Renato D. Puga[3], Daniel G. Pinheiro[17], Luís Eduardo C. Andrade[4], Luciana O. Cruz[3], Theri L. Degaki[3], Milton Faria Jr.[7], Fernanda Festa[3], Daniel Giannella-Neto[20], Ricardo R. Giorgi[20], Gustavo H. Goldman[8], Fabiana Granja[9], Arthur Gruber[10], Christine Hackel[11], Flávio Henrique-Silva[12], Bettina Malnic[13], Carina V.B. Manzini[13], Suely K.N. Marie[14], Nilce M. Martinez-Rossi[15], Sueli M. Oba-Shinjo[14], Maria Ines M.C. Pardini[16], Paula Rahal[18], Cláudia A. Rainho[21], Silvia R. Rogatto[22], Camila M. Romano[10], Vanderlei Rodrigues[23], Magaly M. Sales[16], Marcela Savoldi[8], Ismael D.C.G. da Silva[5], Neusa P. da Silva[4], Sandro J. de Souza[6], Eloiza H. Tajara[19], Wilson A. Silva Jr.[17], Andrew J.G. Simpson[1,2], Mari C. Sogayar[3], Anamaria A. Camargo[1] and Dirce M. Carraro[1,24]

[1]Laboratory of Molecular Biology and Genomics,
Ludwig Institute for Cancer Research, São Paulo, SP, Brazil
[2]Ludwig Institute for Cancer Research, New York, NY, USA
[3]Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brasil
[4]Divisão de Reumatologia, Universidade de São Paulo, São Paulo, SP, Brasil
[5]Laboratório de Ginecologia Molecular, Departamento de Ginecologia,
Universidade Federal de São Paulo, São Paulo, SP, Brasil
[6]Laboratório de Biologia Computacional,
Instituto Ludwig de Pesquisa sobre o Câncer, São Paulo, SP, Brasil
[7]Departamento de Engenharia Química e de Informática, Bioinformática,
Universidade de Ribeirão Preto, Ribeirão Preto, SP, Brasil
[8]Faculdade de Ciências Farmacêuticas de Ribeirão Preto,
Universidade de São Paulo, Ribeirão Preto, SP, Brasil
[9]Laboratório de Genética Molecular do Câncer,
Departamento de Clínica Médica, Faculdade de Ciências Médicas,
Universidade Estadual de Campinas, Campinas, SP, Brasil
[10]Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia,
Universidade de São Paulo, São Paulo, SP, Brasil
[11]Departamento de Genética Médica, Faculdade de Ciências Médicas,
Universidade Estadual de Campinas, Campinas, SP, Brasil

[12]Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, SP, Brasil
[13]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brasil
[14]Departamento de Neurologia, Faculdade de Medicina, Universidade de São Paulo, São Paulo, SP, Brasil
[15]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil
[16]Laboratório de Biologia Molecular, Hemocentro, Faculdade de Medicina, Universidade Estadual Paulista, Botucatu, SP, Brasil
[17]Centro de Terapia Celular, Hemocentro e Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil
[18]Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São Jose do Rio Preto, SP, Brasil
[19]Departamento de Biologia Molecular, Faculdade de Medicina de São José do Rio Preto, São José do Rio Preto, SP, Brasil
[20]Laboratório de Endocrinologia Molecular e Celular (LIM-25), Hospital das Clínicas da Faculdade de Medicina, Universidade de São Paulo, São Paulo, SP, Brasil
[21]Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, Botucatu, SP, Brasil
[22]Laboratório NeoGene, Departamento de Urologia, Faculdade de Medicina, Universidade Estadual Paulista, Botucatu, SP, Brasil
[23]Departamento de Bioquímica e Imunologia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil
[24]Laboratory of Gene Expression Analysis, Ludwig Institute for Cancer Research, São Paulo, SP, Brazil
*These authors contributed equally to this study.
Corresponding author: D.M. Carraro
E-mail: dcarraro@ludwig.org.br

Genet. Mol. Res. 3 (4): 493-511 (2004)
Received October 4, 2004
Accepted December 7, 2004
Published December 30, 2004

ABSTRACT. The correct identification of all human genes, and their derived transcripts, has not yet been achieved, and it remains one of the major aims of the worldwide genomics community. Computational programs suggest the existence of 30,000 to 40,000 human genes. However, definitive gene identification can only be achieved by experimental approaches. We used two distinct methodologies, one based on the alignment of mouse orthologous sequences to the human genome, and an-

other based on the construction of a high-quality human testis cDNA library, in an attempt to identify new human transcripts within the human genome sequence. We generated 47 complete human transcript sequences, comprising 27 unannotated and 20 annotated sequences. Eight of these transcripts are variants of previously known genes. These transcripts were characterized according to size, number of exons, and chromosomal localization, and a search for protein domains was undertaken based on their putative open reading frames. *In silico* expression analysis suggests that some of these transcripts are expressed at low levels and in a restricted set of tissues.

Key words: Novel human transcripts, Mouse orthologous sequence, Testis cDNA

INTRODUCTION

Due to the complexity of the human genome, the main objective of the Human Genome Project, the correct identification of all human genes and their derived transcripts, has yet to be achieved. The human genome is estimated to contain 30,000 to 40,000 genes (Lander et al., 2001; Venter et al., 2001). This number of genes is only slightly higher than what is found in much simpler organisms, such as *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998), suggesting that the greater complexity found in vertebrates might be due to a variety of mechanisms involving regulation of gene expression. Such mechanisms might include, for example, inhibition of gene expression by antisense transcripts (Shendure and Church, 2002; Yelin et al., 2003) and alternative splicing (Modrek and Lee, 2002).

The coding sequences of the human genome, comprising only 3% of the entire sequence, are separated by large intergenic regions, and they are composed of short transcribed exons, normally interrupted by numerous, very long non-transcribed introns. Gene prediction programs have been used to recognize patterns and predict genes within the genome sequence. These programs are based not only on *ab initio* gene prediction, but they also make use of orthologous sequences, motif and domain structure, and the presence of polyadenylation sites and/or signals (Burge and Karlin, 1997; Rogic et al., 2001; Solovyev, 2001; Blanco et al., 2002). Despite the progress in this area, genomic structural complexity does not allow such programs to correctly predict all human genes without experimental confirmation. As a result, the correct identification of all human genes, and their derived transcripts, remains a real challenge.

Many large-scale sequencing projects have contributed to the global identification of human genes and their variants by generating expressed sequence tags (ESTs) (Adams et al., 1991; Houlgatte et al., 1995) and open reading frame (ORF) ESTs (ORESTES) (Dias-Neto et al., 2000; de Souza et al., 2000; Camargo et al., 2001; Brentani et al., 2003). ESTs are partial and single-pass sequences derived from the 5' or 3' extremities of cDNA clones (Adams et al., 1991), whereas the ORESTES approach is biased towards the central portion of the coding regions of transcripts (Dias-Neto et al., 2000; Camargo et al., 2001). Nevertheless, full-length transcript sequences are crucial for final confirmation of gene structure. Considerable effort

has been made in the generation of full-length transcript sequences (Strausberg et al., 1999; Wiemann et al., 2001; Kikuno et al., 2002; Nakajima et al., 2002; Strausberg et al., 2002) directly from high-quality cDNA libraries (Bonaldo et al., 1996; Carninci et al., 2000). The transcript finishing strategy, developed by Sogayar and collaborators (2004), utilized RT_PCR experiments to bridge gaps between EST clusters mapped to the human genome to achieve final confirmation of the structure of transcripts.

Our involvement with this latter project motivated us to further explore the complete characterization of new human transcripts through integrative approaches involving experimental and *in silico* strategies. We used mouse transcript sequences and cDNA molecules derived from a human testis library to identify new human transcripts. We completed the sequence of 47 transcripts, including 27 that were first annotated by us in the human genome.

MATERIAL AND METHODS

Testis cDNA library generation and clone selection

A unidirectional human testis cDNA library was constructed from polyA RNA using the Superscript™ Plasmid System Gateway™ Technology for double-strand cDNA synthesis and cloning. Cloned fragments were selected by size on Sepharose CL-2B (Pharmacia) columns (40 cm long, 1 mm ID) according to the protocol described by Vettore and collaborators (2001). Fractions containing cDNA molecules larger than 800 bp were ligated into pSPORT6 vectors (Invitrogen) at the *Sal*I-*Not*I site and the resulting plasmids were transformed in DH10B cells (Invitrogen) by electroporation (BioRad). The transformants were spread on LB agar medium containing ampicilin (100 µg/ml), IPTG (100 mM) and X-Gal (20 mg/ml), and plasmid DNA was purified using the alkaline lysis method (Sambrook et al., 1989). In order to estimate the frequency of full-length clones, putative new transcripts, and the level of redundancy in the library, 5' sequences from 192 clones were generated, resulting in 153 high-quality sequences (300 bp with Phred >20) suitable for further analysis.

The 5' sequences were aligned to the human genome using the BLAT search tool provided by the University of California, Santa Cruz (UCSC) (http://genome.ucsc.edu/cgi-bin/hgBlat, version Nov. 2002), and the annotation tracks corresponding to Known Genes, human mRNA and RefSeq genes were used for the comparison. The sequences that aligned with already identified full-length human mRNAs were used to estimate the frequency of full-length clones within the library. A sequence was considered a full-length clone when the 5' end aligned with, or upstream of, the start codon site of a corresponding CDS-annotated mRNA molecule. The sequences that did not align with any full-length human mRNA were used to assess the frequency of putative new human transcripts. The CAP3 assembler program with default parameters (Huang and Madan, 1999) was used to join sequences with a high-identity level into contigs. The number of contigs and singletons obtained were divided by the total number of reads, and the redundancy level was assumed as 1 minus the value obtained in the previous calculation.

cDNA evaluation and sequencing

Inserts were amplified by PCR using the primers (SP6 Promoter primer - 5' ATTTAG

GTGACACTATA 3' - and T7 Promoter primer - 5' CCCTATAGTGAGTCGTATTA 3') in a standard reaction containing 1X Taq polymerase buffer, 0.25 mM dNTP, 1.5 mM $MgCl_2$, 1.0 mM each primers and 2 U Taq DNA polymerase (Invitrogen) in a final volume of 25 µl. Reactions were carried out at 94°C for 30 s, 55°C for 45 s and 72°C for 4 min for 35 cycles. An initial cycle of 94°C for 4 min and a final extension at 72° for 6 min were used. For generation of the complete sequence, two strategies were used: primer walking for clones smaller than 1,500 bp and shotgun libraries for clones larger than 1,500 bp. For the primer walking strategy, direct sequencing was undertaken, using internal primers designed approximately 100 bases from the end of the available high-quality sequence. For the shotgun strategy, the inserts were amplified by PCR and randomly fragmented by sonication. Fragments of 500 bp to 1,000 bp were isolated from agarose gels (1%) and cloned using the TOPO Shotgun Cloning Kit (Invitrogen).

Identification of human transcripts using mouse cDNA sequences

Mouse mRNA sequences available from UniGene were analyzed by the Laboratory of Computational Biology, Ludwig Institute for Cancer Research, São Paulo Branch. All mouse sequences were mapped onto the human genome sequence using Blast, and those that did not correspond to a full-length human mRNA were selected for further analysis. A total of 672 sequences of the list were manually checked, and those sequences that presented a functional annotation and size less than 4.0 kb were selected for evaluation by RT_PCR.

RT_PCR amplification of human transcripts identified with mouse sequences

Following selection of the mouse orthologs of interest, their sequences were aligned to the human genome sequence (http://genome.ucsc.edu/cgi-bin/hgBlat, version Nov. 2002), and RT_PCR primers were designed from conserved regions using Primer3 with default parameters (www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi). Since most sequences were larger than 500 bp, the construction of one or more pairs of primers was necessary to cover the entire transcript. Each pair of primers delimited a fragment of up to 1,000 bp, with 100 bp overlapping between pairs of primers to facilitate assembly of a consensus sequence. In order to determine from which tissue the cDNA should be used as the RT_PCR template, mouse mRNA sequences were aligned against dbEST (http://www.ncbi.nlm.nih.gov/BLAST/). PCR was undertaken using a mixture that contained 1X Taq polymerase buffer, 0.20 mM dNTP, 1.5 mM $MgCl_2$, 1.0 mM primers, and 2 U Taq DNA polymerase (Invitrogen) in a final volume of 25 µl. Reactions were carried out with a basic cycle consisting of 94°C for 30 s for denaturating, primer annealing at calculated temperature for 45 s and extension for 1 min at 72°C for 35 cycles, together with an initial denaturating at 94°C for 4 min and a final extension at 72°C for 6 min. Modifications in the PCR conditions for candidates showing no specific amplification were attempted including the addition of betaine (1 M) (Henke et al., 1997), alteration of annealing temperature and adjustments of $MgCl_2$ concentration.

RNA extraction and cDNA synthesis

Total RNA was prepared from cultured cells using the cesium chloride cushion technique (Chirgwin et al., 1979) and subsequently treated with 100 U DNAse I (FPLC-pure; Amer-

sham). For cDNA synthesis, a reverse-transcriptase PCR was carried out, using 5 to 10 µg total RNA, oligo(dt)12-18, random primer and Superscript II (Invitrogen), according to the manufacturer's instructions. Following the synthesis, cDNA molecules were treated with RNAse H. Genomic DNA contamination and cDNA quality were evaluated through PCR amplifications using primers annealing to intronic sequences flanking exon 12 of the hMLH-1 gene (forward: 5'TGGTGTCTCTAGTTCTGG3' - reverse: 5' CATTGTTGTAGTAGCTCTCG3') and primers located at the 5' extremity of the Notch 2 transcript (11,433 bp) (forward: 5'ACTGTG GCCAACCAAGTTCTC3' - reverse: 5'CTCTCACAGGTGCTCCCTTC3'), respectively.

Template preparation and DNA sequencing

DNA templates were prepared for the testis cDNA clones in a 96-well plate, using the alkaline lysis method. Transcripts identified with mouse sequences were sequenced directly following purification of PCR products with the QIAquick PCR Purification Kit (QIAGEN). Sequencing reactions were carried out using ABI Prism BigDye Terminator v3.0 Cycle Sequencing Ready Reactions (Applied Biosystems) and an ABI Prism 3100 (Applied Biosystems).

Sequencing analysis and database update

Chromatograms were classified into testis cDNA sequences (TESTIS) or mouse-derived sequences (MO) and were processed with the PredPhrap package (including phred, phd2fasta and cross_match, in this order) with default parameters. Before the Phrap was called to assemble the data into one contig, the sequence reads were screened against the vector sequence. The contig sequences, and their respective chromatograms, were visually analyzed.

Transcript characterization

Consensus sequences were aligned against the May 2004 version of the human genome sequence available at UCSC Genome Browser (Kent et al., 2002), using the BLAT search tool (http://genome.ucsc.edu/cgi-bin/hgBlat). This allowed determination of overlap with known genes and gene predictions. A transcript was considered novel if its alignment coordinates did not match the coordinates of Known Genes, RefSeq Genes, MGC sequences, or human mRNA annotation tracks available through the BLAT search tool. A transcript was defined as a splicing variant if the alignments revealed intron retention and/or alternative exon usage when compared to sequences available in the databases cited above. The following tracks were used for comparison with prior gene prediction: Fgenesh++ (Solovyev, 2001), Geneid (Guigó et al., 1992) and Genscan (Burset and Guigó, 1996), available through the BLAT search tool in the July 2003 version. An exon was considered to be predicted if it aligned within the coordinates defined by any of the three gene prediction programs (not necessarily sharing borders) and a transcript was considered not predicted if none of the exons were predicted by any of the computer programs.

The consensus sequences corresponding to newly identified transcripts were translated into amino acid sequences using TRANSLATE (http://us.expasy.org/tools/dna.html). The ORF sequences were searched against Pfam and Prosite databases by the Hits tool (http://hits.isb-

sib.ch/) to determine putative protein domains. ORFs from the TESTIS cDNA clones were those with the longest amino acid sequences, containing at least 40 amino acids. In the case of the MO transcripts, ORFs were selected on the basis of their match with those in the mouse transcript.

SAGE

Virtual tags were assigned to the transcript sequences comprising the 10 bp downstream of the *Nla*III site most proximal to the 3' extremity. The extracted tags were then analyzed using serial analysis of gene expression (SAGE) genie (http://cgap.nci.nih.gov/SAGE) to generate a putative expression pattern for each transcript.

RESULTS

Library validation and selection of testis cDNA molecules

The testis cDNA library contained 2 x 10^6 cDNA clones. Approximately 35% (35 of 99) of cDNA molecules tested in an initial sample were judged to be full-length sequences, based on alignment with CDS-annotated human transcripts available in the public databases. This percentage was not as high as that achieved using special protocols for obtaining full-length cDNA sequences (Carninci et al., 2000, 2001), but it was superior to what was obtained with commercially available, high-quality cDNA libraries evaluated by our group (data not shown). From an initial set of 153 sequences, we found 14 in which the 5' extremity sequence did not match any known human transcript. This result indicated that up to 9% of our set might correspond to previously unidentified transcripts. Redundancy, as evaluated by CAP3 (Huang and Madan, 1999), was around 3.3%, with 145 singletons and three contigs. Based on these results the library was judged adequate to search for unknown transcripts.

We then generated 1,152 5' sequences, 835 of which were of high quality. Fifty-nine of these did not match known human transcripts, including 55 that were 500 bp or longer, based on PCR amplification. In order to detect chimeric clones, the sequences of both extremities were mapped to the Nov. 2002 version of the human genome sequence using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). Based on this, 4 clones were discarded, since their extremities did not align to the same genome region. Of the remaining 51 clones, 15 aligned discontinuously to the genomic sequence (revealing the presence of exons and introns), whereas 36 aligned continuously (indicating non-spliced structures). The presence of exons greatly increases the probability that the sequence is a bona fide transcript (Sorek and Safer, 2003). However, recent studies identified 3,500 single-exon human transcripts on the human genome sequence (approximately 10% of known genes) (Sakharkar and Kangueane, 2004), which appear to play an important role in transcript regulation and cell differentiation (Hickox et al., 2002; Sakharkar and Kangueane, 2004). We, thus, selected cDNA clones corresponding to putative single-exon transcripts where at least one corresponding EST was available in dbEST (http://www.ncbi.nlm.nih.gov/BLAST/). This was found to be the case for 31 sequences that mapped continuously to the genome. These and the 15 sequences corresponding to multi-exon transcripts were then submitted to complete sequencing (Figure 1).

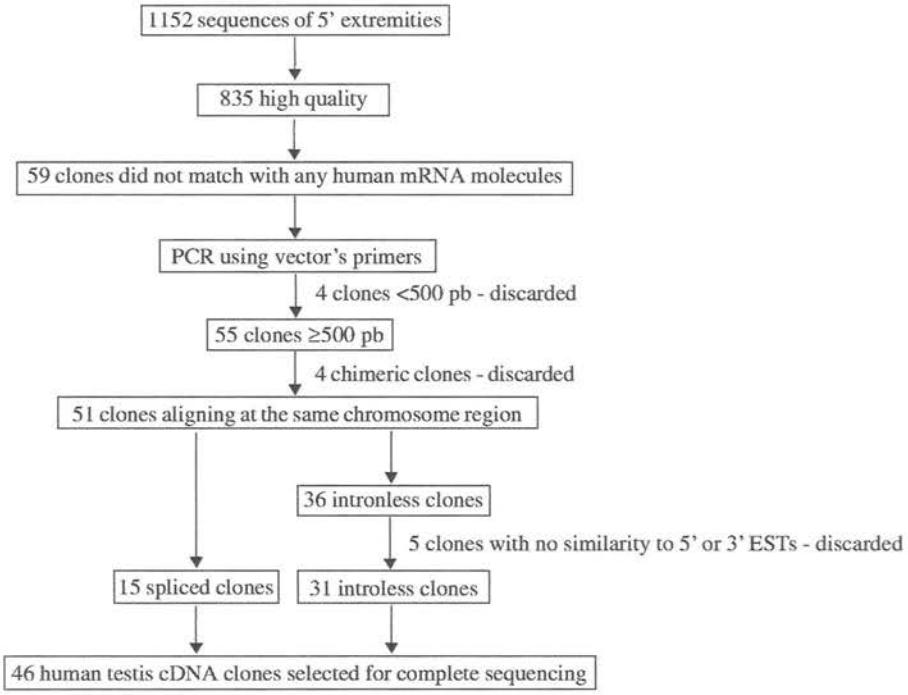


Figure 1. A general overview of the human testis cDNA library validation and the process of cDNA clone selection for complete sequencing. ESTs = expressed sequence tags.

The use of mouse sequences to identify human transcripts

One thousand four hundred and fourteen mouse sequences that had no match to any full-length human cDNA were considered. Manual inspection of 672 randomly selected sequences indicated that 217 had no alignment to any full-length human mRNA sequence in the public database. Several other criteria (See Methods) were then applied to reduce the dataset to a list of 29 regions in the human genome likely to contain a human transcript. An overview of the selection process is shown in Figure 2.

The source of ESTs matching those regions was used as a guide for tissue selection for PCR amplification. A pool of cDNA molecules from different tissues was used when no EST information was available. Among the 29 attempted amplifications, 28 generated at least one fragment of the expected size, indicating that as many as 96.5% of our candidates had a corresponding human ortholog.

Consensus sequence assembly

A total of 713 sequences were generated during the project, of which 305 were from TESTIS cDNA clones and 408 from the cDNA molecules identified with MO sequences; these were successfully assembled into 27 TESTIS cDNA molecules and 20 MO sequences. Fifteen TESTIS cDNA molecules and 6 MO sequences were abandoned once full-length human mRNA

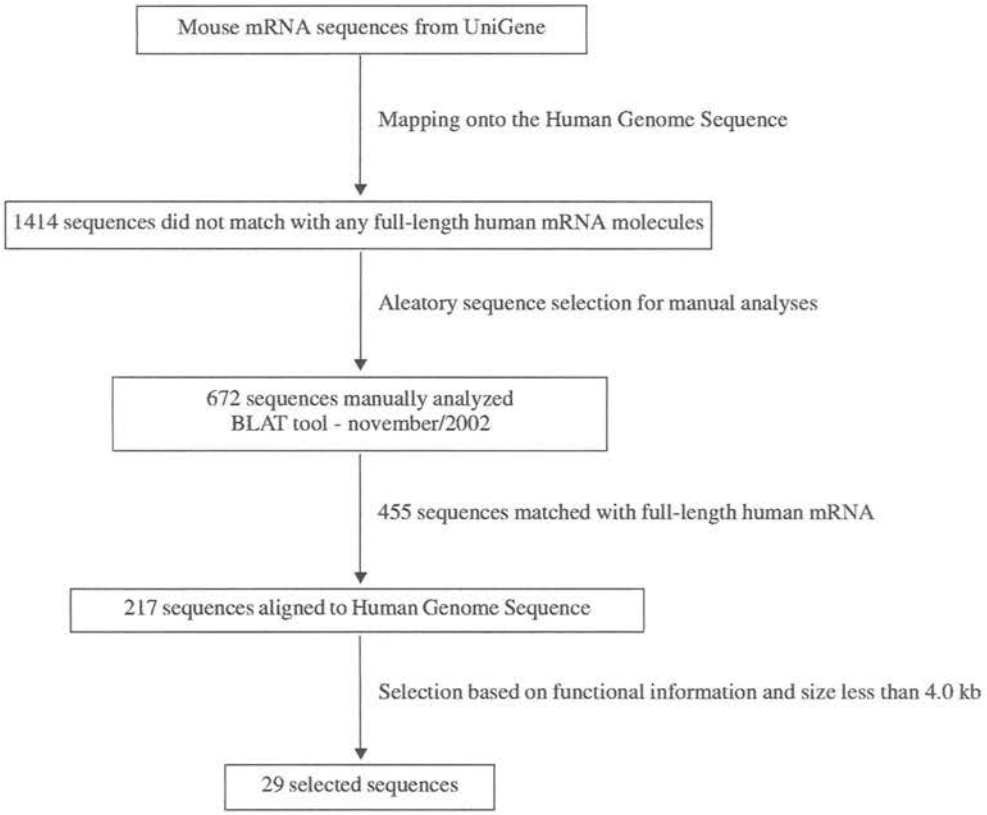


Figure 2. A general overview of the process of orthologous sequence selection for complete sequencing.

sequences were submitted to the GenBank by other groups. In the case of 4 TESTIS cDNA molecules and 3 MO sequences, it was impossible to generate a consensus sequence due to the presence of repetitive sequences (in the case of TESTIS cDNA molecules) or due to nonspecific amplification of fragments (in the case of MO sequences). Sequence discrepancies were manually corrected based on the genome sequence. Less than 0.9% of the nucleotides in the consensus assemblies required alteration. In the case of MO sequences, final confirmation came from RT_PCR, using primers annealing to the extremities of the consensus to demonstrate the existence of the complete transcript in the tissue (Figure 3). The 47 consensus sequences had an average size of 1,547 bp (1,721 bp for the TESTIS cDNA molecules and 1,312 bp for the MO sequences).

Analysis of the testis-derived transcripts

Alignment of the 27 TESTIS cDNA transcripts to the human genome sequence (UCSC - May, 2004) revealed that 19 (70%) did not match full-length human mRNA sequences. Furthermore, only 5% of these presented a structure predicted by *ab initio* gene prediction programs (Fgenesh++, Geneid and Genscan), thus representing totally unknown transcribed re-

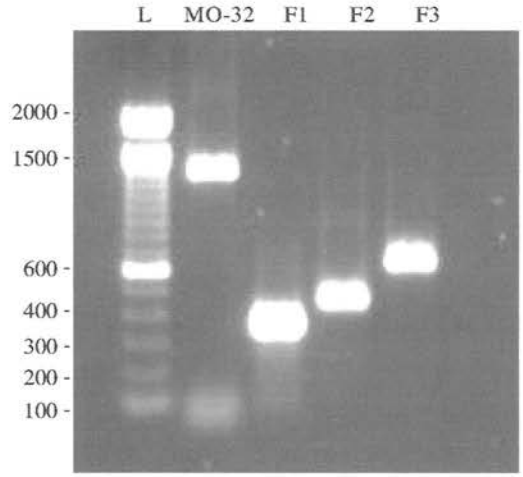


Figure 3. Final confirmation of the MO-32 (AY726601) consensus assembly. Agarose gel showing RT_PCR of MO-32 entire transcript and the three individual fragments using placental cDNA as a template. L: Ladder 100 bp; MO-32: RT_PCR using extremity primers (size: 1,287 bp); F1: fragment 1 (size: 368 bp); F2: fragment 2 (size: 458 bp); F3: fragment 3 (size: 628 bp).

gions in the human genome. Five TESTIS cDNA molecules comprised previously unknown splicing variants of known genes. Overall our data revealed 46 previously unidentified transcribed exons in the human genome, corresponding to 33,487 bp. The average size of the new exons was 727.9 bp. However, when single-exon transcripts were excluded the exon average size decreased to 438.0 bp (12 of 19 transcripts were single exon).

Analysis of transcripts identified using mouse sequences

A similar analysis for the 20 MO transcripts revealed that 8 did not match to full-length human mRNA sequences (40%) and one corresponded to a splicing variant of a known gene. Half of the novel transcripts had been predicted by *ab initio* gene prediction programs (Fgenesh++, Geneid and Genscan). Our data delimited 22 previously unknown exons in the human genome sequence comprising 9,060 bp. The average size of the new exons was 411.8 bp. When the six single-exon transcripts were excluded, the average exon size decreased to 231.8 bp.

Transcript annotation

Of the 27 full insert transcripts, 22 were found to contain an ORF of at least 40 amino acids, with the average ORF size being 129 amino acids. In 6 of these, the transcript contained an identifiable protein domain, such as a serine-rich region profile, a zinc finger C_2H_2 type domain and G-protein-coupled receptor family 1 profile. Amongst the MO transcripts, five of eight contained such profiles, while amongst the TESTIS transcripts the proportion was much lower, being 1 of 19. A complete listing of the characteristics of the TESTIS cDNAs is shown in Table 1 and another of the MO transcripts is shown in Table 2.

Table 1. Annotation of the testis cDNA sequences (TESTIS), including accession numbers, consensus size, number of exons, chromosomal localization, status related to new human transcript, putative open reading frame (ORF) size, and number of tags per 200,000. aa = amino acids.

Accession number	Size (bp)	Exon number	Chromosome location	Status	Number of new exons	Prediction	ORF Protein domain	Tags per 200,000
TESTIS-602 AY726558	1734	2	9q22.31	New	-	0	165 aa	9
TESTIS-603 AY726559	1149	2	1q24.1	New	-	0	151 aa	-
TESTIS-604 AY726560	2108	16	1p34.3	Alternative variant BC041360	2	0	-	38
TESTIS-607 AY726561	2167	3	19p13.3	Alternative variant AX775861	1 exon retention	0	-	191
TESTIS-608 AY726562	1669	5	2q21.1	Alternative variant BC064385	3	2	-	42
TESTIS-609 AY726563	2594	2	11p11.2	Alternative variant AK097878	1	1	-	70
TESTIS-612 AY726564	2334	1	11q12.311q13.1	New	-	0	52 aa	-
TESTIS-614 AY726565	1667	9	11p12	New	-	0	57 aa	1554
TESTIS-706 AY726566	1212	1	16p11.2	New	-	0	103 aa	-
TESTIS-713 AY726567	1921	3	18p11.22	New	-	1	75 aa	41
TESTIS-714 AY726568	2290	1	1q25.1	New	-	0	133 aa	12
TESTIS-721 AY726569	1774	1	13q32.1	New	-	0	113 aa	2
TESTIS-724 AY726570	2216	1	1q25.1	Known BX537597	-	-	-	450
TESTIS-725 AY726571	1677	1	1p22.2	Known BC053364	-	-	-	47
TESTIS-732 AY726572	1577	1	6p25.1	New	-	0	118 aa	23
TESTIS-735 AY726573	771	1	6q27	New	-	0	113 aa	07
TESTIS-738 AY726574	1334	1	16q24.3	Alternative variant BC025283	-	-	-	663
TESTIS-740 AY726575	855	3	19q12	New	-	0	88 aa	5
TESTIS-742 AY726576	1971	1	5q23.2	New	-	0	83 aa; serine-rich region profile	1343
TESTIS-744 AY726577	2813	1	2q14.1	Known AK124683	-	-	-	223
TESTIS-750 AY726578	2241	1	9q22.33	New	-	0	59 aa	5
TESTIS-809 AY726579	919	1	7q11.21	New	-	0	60 aa	54
TESTIS-814 AY726580	515	1	Xq25	New	-	0	41 aa	52
TESTIS-817 AY726581	2421	5	4q35.1	New	-	0	67 aa	1510
TESTIS-822 AY726582	1952	1	5p13.1	New	-	0	37 aa	47
TESTIS-823 AY726583	1560	3	11q14.1	New	-	0	83 aa	4
TESTIS-828 AY726584	1037	1	12p11.1	New	-	0	38 aa	5

Table 2. Annotation of the mouse-derived sequence (MO) transcripts, including accession numbers, consensus size, number of exons, chromosomal localization, status related to new human transcript, putative open reading frame (ORF) size and number of tags per 200,000. aa = amino acids.

Accession number	Size (bp)	Exon number	Chromosome location	Status	Number of new exons	Prediction	ORF/Protein domain	Tags per 200,000
MO-01 AY726585	1028	6	11q13.2	Known BC047953	-	-	-	489
MO-06 AY726586	2727	20	10q21.1	Extension AK026129	10	9	-	19
MO-07 AY726587	440	2	15q23	New	-	1	20 aa	1006
MO-09 AY726588	2026	4	7q31.2	New	-	4	463 aa; zinc finger C2H2 type domain profile	-
MO-13 AY726589	782	1	9p13.3	New	-	1	260 aa; G-protein-coupled receptor family 1 profile	-
MO-16 AY726590	840	1	11p15.4	New	-	0	279 aa; G-protein-coupled receptor family 1 profile	1
MO-18 AY726591	1321	8	3p21.1	Known BC047015	-	-	-	144
MO-21 AY726592	781	1	11q24.2	New	-	1	259 aa; 7 transmembrane receptor (rhodopsin family)	55
MO-22 AY726593	3212	8	3q13.2	Known AF506819 AB052098	-	-	-	3584
MO-23 AY726594	797	5	20q11.22	Known AB100261 AY329085	-	-	-	68
MO-24 AY726595	542	5	7q11.22	Known BC020200 AY007302	-	-	-	-
MO-25 AY726596	891	3	1p33	Known AF398527	-	-	-	44
MO-27 AY726597	2379	14	1q21.3	Known BC053562	-	-	-	338
MO-28 AY726598	852	1	14q11.2	Known OR4E2	-	-	-	2
MO-30 AY726599	790	1	15q21.3	New	-	0	42 aa	1049
MO-31 AY726600	834	6	5p13.2	Known BX538177 BX538178	-	-	-	66
MO-32 AY726601	1271	1	Xq22.1	New	-	0	309 aa; protein of unknown function (DUF634)	68
MO-33 AY726602	887	1	4q35.1	New	-	0	26 aa	122
MO-34 AY726603	1128	1	Xq21.1	Known BX648117 BC067294	-	-	-	524
MO-35 AY726604	2716	3	5q23.3	Known AJ504664	-	-	-	58

In silico analysis of transcript abundance

In order to have a general view of the transcript expression profile from the human transcripts identified here, we performed an *in silico* analysis based on SAGE. The virtual tags corresponded to the 10 nucleotides immediately downstream of the 3'most *Nla*III site in a given transcript. These virtual tags were then submitted to SAGE Genie (Boon et al., 2002) in order to establish transcript expression profiles (available at http://gdm.fmrp.usp.br/cgi-bin/tagmap/index.pl?template_file=view_sequence). We were able to obtain expression information for 41 of 47 sequences. Among these, 18 had less than 50 tags per 200,000, and were thus relatively lowly expressed (Figure 4A). Thirteen of the 41 transcripts were expressed in 5 or less tissues, as judged by the source of SAGE tags (Figure 4B). Amongst the TESTIS cDNA molecules analyzed, 14 of 24 had less than 50 tags per 200,000, whereas amongst the MO transcripts, only 4 of 17 presented this profile, and these were judged as rare messages. Similar results were found in tissue distribution analysis where 10 of 24 TESTIS cDNAs, and 3 of 17 MO transcripts were expressed in only 5 or less tissues.

DISCUSSION

There is currently a worldwide effort to complete the catalogue of human genes and derived transcripts, involving several independent initiatives and distinct approaches.

Full-length cDNA sequences are the gold standard for the definition of transcripts. Progress has been made in full-length sequence generation, using standard and full-length enriched cDNA library from many human tissues (Bonaldo et al., 1996; Carninci et al., 2000; Strausberg et al., 1999, 2002; Nakajima et al., 2002). A total of 28,256 UniGene clusters currently include at least one full-length cDNA sequence (UniGene Build #171 - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene).

To contribute to the definition of the human transcript catalogue, we have used two alternative strategies: the construction of a unidirectional human testis cDNA library and the alignment of mouse sequences to the human genome. Testis is a highly specialized tissue that expresses a large number of transcript species, which makes it suitable as a potential source for unidentified transcripts (Warrington et al., 2000; Yao et al., 2004). The strategy using mouse sequences is powerful, due to the high degree of conservation between human and mouse observed both in the coding sequence (Makalowski et al., 1996) and in 3' and 5' UTR (Makalowski et al., 1996; Shabalina et al., 2004).

We identified and completely sequenced 47 previously unknown human transcripts, of which 27 had still not been annotated in the May 2004 version of the UCSC genome browser. The use of a cDNA testis library was found to be more effective (19 novel transcripts) than the use of mouse mRNA sequences (8 novel transcripts) for the identification of unknown human transcripts.

Intronless transcripts have increasingly been perceived as playing an important role in the regulation of transcription (Hickox et al., 2002; Sakharkar and Kangueane, 2004), and they represent a significant proportion of the human gene catalogue (Sakharkar et al., 2002; Sakharkar and Kangueane, 2004). A significant proportion of our candidates were intronless transcripts (24 of 47). Many of them had still not been reported by others by the end of the project (18 of 24 intronless transcripts and 9 of 23 transcripts with multiple exons). The high number of novel

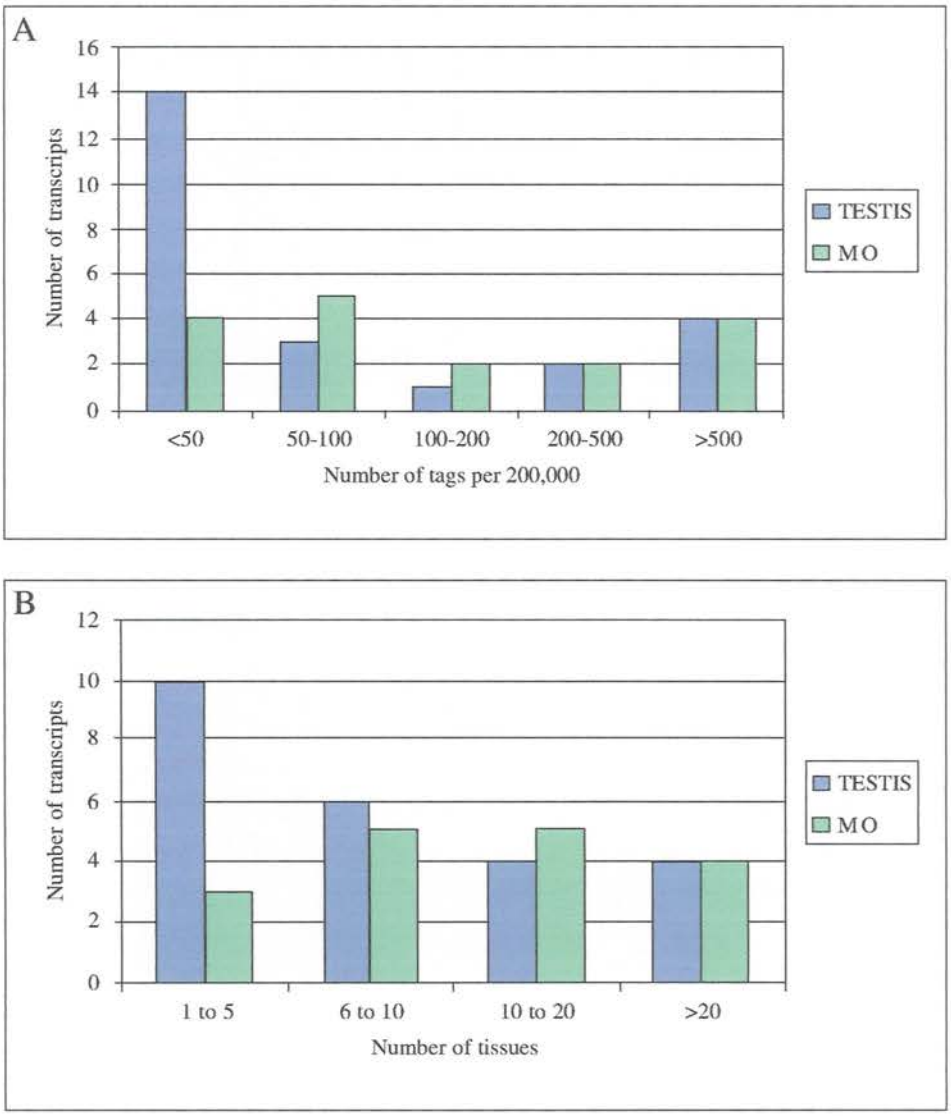


Figure 4. Distribution of extracted tags of the 41 transcripts that have expression information in the SAGE Genie web site. TESTIS - testis cDNA clones; MO - mouse human orthologs. **A**, The number of transcripts per number of tags. **B**, Number of transcripts per different tissue types.

intronless transcripts identified in this project may be due to the fact that many groups have been using strategies for transcript identification that exclude intronless sequences (Sogayar et al, 2004).

Our analysis of the novel transcripts revealed that 22 of 27 were not identified by any of the commonly used *ab initio* gene prediction programs. Amongst the TESTIS transcripts, 18 of 19 were not predicted, while in the case of MO transcripts 4 of 8 were not predicted. This leads us to suggest that most of the unidentified transcripts have atypical structures that are difficult to

identify using *ab initio* computational prediction programs.

Amongst the novel human transcripts, only 9 of 27 had an ORF of more than 100 amino acids, precluding extensive analysis of predicted protein structure and function. Short ORF transcripts may be more difficult to predict by computational programs, and consequently their experimental characterization might have been delayed. Moreover, only 6 transcripts exhibited recognizable protein domains, 5 of which were MO transcripts. Whether the transcripts containing short ORFs without the presence of a known protein domain produce a functional protein remains to be determined.

Recent technological advances in large-scale gene expression analysis have been made, including SAGE (Velculescu et al., 1995). Currently there are around 15 million tags available in the SAGE Genie database. These tags can be associated with a human transcript, providing a global expression portrait of the human genome, and the use of bioinformatics tools allows a general view of individual transcript distribution. The TESTIS cDNA molecules had a higher frequency of transcripts expressed at a low level and a restricted number of tissue types, when compared to MO transcripts. This result is supported by the higher proportion of novel human transcripts identified by the testis library approach, since this expression pattern could have made their previous identification more difficult.

We conclude that there are still unidentified human transcripts, many of which might be found using the testis as a tissue source. Surprisingly, we also found that even genes fully annotated in the mouse genome remained cryptic in the human genome, although many of these transcripts either may not encode proteins or they could produce rather short polypeptides. Nevertheless, continued efforts at human gene identification would appear to be worthwhile.

ACKNOWLEDGMENTS

We thank Dr. Ricardo R. Brentani (Director of the Ludwig Institute-São Paulo Branch and of the A.C. Camargo Hospital) for valuable support. We also thank Anna Christina de Matos Salim, Elisangela Monteiro, Jane Kaiano, Dr. Maria Rita Passos Bueno, Guilherme M. Orabona, and Elisson Campos Osório for technical and computational assistance and Natanja Slager for critical reading and important comments on this manuscript. Research 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., 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.

Blanco, E., Parra, G. and **Guigó, R.** (2002). Finding genes. In: *Current Protocols in Bioinformatics* (Baxevanis, A., ed.). John Wiley & Sons Ltd., New York (in press).

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

Boon, K., Osorio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Buetow, K.H., Strausberg, R.L., De Souza, S.J. and **Riggins, G.J.** (2002). An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA 99*: 11287-11292.

Brentani, H., Caballero, O.L., Camargo, A.A., da Silva, A.M., da Silva Jr., W.A., Dias Neto, E., Grivet, M., Gruber, A., Guimarães, P.E., Hide, W., Iseli, C., Jongeneel, C.V., Kelso, J., Nagai, M.A., Ojopi, E.P., Osório, E.C., Reis, E.M., Riggins, G.J., Simpson, A.J., de Souza, S., Stevenson, B.J., Strausberg,

R.L., Tajara, E.H., Verjovski-Almeida, S., Acencio, M.L., Bengtson, M.H., Bettoni, F., Bodmer, W.F., Briones, M.R., Camargo, L.P., Cavenee, W., Cerutti, J.M., Coelho Andrade, L.E., Costa dos Santos, P.C., Ramos Costa, M.C., da Silva, I.T., Estecio, M.R., Sa Ferreira, K., Furnari, F.B., Faria Jr., M., Galante, P.A., Guimarães, G.S., Holanda, A.J., Kimura, E.T., Leerkes, M.R., Lu, X., Maciel, R.M., Martins, E.A., Massirer, K.B., Melo, A.S., Mestriner, C.A., Miracca, E.C., Miranda, L.L., Nóbrega, F.G., Oliveira, P.S., Paquola, A.C., Pandolfi, J.R., Campos Pardini, M.I., Passetti, F., Quackenbush, J., Schnabel, B., Sogayar, M.C., Souza, J.E., Valentini, S.R., Zaiats, A.C., Amaral, E.J., Arnaldi, L.A., de Araujo, A.G., de Bessa, S.A., Bicknell, D.C., Ribeiro de Camaro, M.E., Carraro, D.M., Carrer, H., Carvalho, A.F., Colin, C., Costa, F., Curcio, C., Guerreiro da Silva, I.D., Pereira da Silva, N., Dellamano, M., El-Dorry, H., Espreafico, E.M., Scattone Ferreira, A.J., Ayres Ferreira, C., Fortes, M.A., Gama, A.H., Giannella-Neto, D., Giannella, M.L., Giorgi, R.R., Goldman, G.H., Goldman, M.H., Hackel, C., Ho, P.L., Kimura, E.M., Kowalski, L.P., Krieger, J.E., Leite, L.C., Lopes, A., Luna, A.M., Mackay, A., Mari, S.K., Marques, A.A., Martins, W.K., Montagnini, A., Mourao Neto, M., Nascimento, A.L., Neville, A.M., Nobrega, M.P., O'Hare, M.J., Otsuka, A.Y., Ruas de Melo, A.I., Paco-Larson, M.L., Guimaraes Pereira, G., Pereira da Silva, N., Pesquero, J.B., Pessoa, J.G., Rahal, P., Rainho, C.A., Rodrigues, V., Rogatto, S.R., Romano, C.M., Romeiro, J.G., Rossi, B.M., Rusticci, M., Guerra de Sa, R., Sant' Anna, S.C., Sarmazo, M.L., Silva, T.C., Soares, F.A., Sonati, M. de F., de Freitas Sousa, J., Queiroz, D., Valente, V., Vettore, A.L., Villanova, F.E., Zago, M.A., Zalcberg, H. and the **Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium; Human Cancer Genome Project Sequencing Consortium** (2003). The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. USA 100*: 13418-13423.

Burge, C. and **Karlin, S.** (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol. 268*: 78-94.

Burset, M. and **Guigó, R.** (1996). Evaluation of gene structure prediction programs. *Genomics 34*: 353-367.

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., Andrade, L.E., Carrer, H., El-Dorry, H.F., Espreafico, E.M., Habr-Gama, A., Giannella-Neto, D., Goldman, G.H., Gruber, A., Hackel, C., Kimura, E.T., Maciel, R.M., Marie, S.K., Martins, E.A., Nobrega, M.P., Paco-Larson, M.L., Pardini, M.I., Pereira, G.G., Pesquero, J.B., Rodrigues, V., Rogatto, S.R., da Silva, I.D., Sogayar, M.C., Sonati, M.F., Tajara, E.H., Valentini, S.R., Alberto, F.L., Amaral, M.E., Aneas, I., Arnaldi, L.A., de Assis, A.M., Bengtson, M.H., Bergamo, N.A., Bombonato, V., de Camargo, M.E., Canevari, R.A., Carraro, D.M., Cerutti, J.M., Correa, M.L., Correa, R.F., Costa, M.C., Curcio, C., Hokama, P.O., Ferreira, A.J., Furuzawa, G.K., Gushiken, T., Ho, P.L., Kimura, E., Krieger, J.E., Leite, L.C., Majumder, P., Marins, M., Marques, E.R., Melo, A.S., Barbosa de Melo, M., Mestriner, C.A., Miracca, E.C., Miranda, D.C., Nascimento, A.L., Nobrega, F.G., Ojopi, E.P., Pandolfi, J.R., Pessoa, L.G., Prevedel, A.C., Rahal, P., Rainho, C.A., Reis, E.M., Ribeiro, M.L., da Ros, N., de Sa, R.G., Sales, M.M., Sant'anna, S.C., dos Santos, M.L., da Silva, A.M., da Silva, N.P., Silva Jr., W.A., da Silveira, R.A., Sousa, J.F., Stecconi, D., Tsukumo, F., Valente, V., Soares, F., Moreira, E.S., Nunes, D.N., Correa, R.G., Zalcberg, H., Carvalho, A.F., Reis, L.F., Brentani, R.R., Simpson, A.J., de Souza, S.J. and **Melo, M.B.** (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. USA 98*: 12103-12108.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and **Hayashizaki, Y.** (2000). Normalization and subtraction of cap-trapper-selected cDNA molecules to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res. 10*: 1617-1630.

Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M. and **Hayashizaki, Y.** (2001). Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics 77*: 79-90.

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

de Souza, S.J., Camargo, A.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A., Andrade, L.E., Carrer, H., El-Dorry, H.F., Espreafico, E.M., Habr-Gama, A., Giannella-Neto, D., Goldman, G.H., Gruber, A., Hackel, C., Kimura, E.T., Maciel, R.M., Marie, S.K., Martins, E.A., Nóbrega, M.P., Paco-Larson, M.L., Pardini, M.I., Pereira, G.G., Pesquero, J.B., Rodrigues, V., Rogatto, S.R., da Silva, I.D., Sogayar, M.C., de Fatima Sonati, M., Tajara, E.H., Valentini, S.R., Acencio, M., Alberto, F.L., Amaral, M.E., Aneas, I., Bengtson, M.H., Carraro, D.M., Carvalho, A.F., Carvalho, L.H., Cerutti, J.M., Correa, M.L., Costa, M.C., Curcio, C., Gushiken, T., Ho, P.L., Kimura, E., Leite,

L.C., Maia, G., Majumder, P., Marins, M., Matsukuma, A., Melo, A.S., Mestriner, C.A., Miracca, E.C., Miranda, D.C., Nascimento, A.N., Nobrega, F.G., Ojopi, E.P., Pandolfi, J.R., Pessoa, L.G., Rahal, P., Rainho, C.A., da Ros, N., de Sa, R.G., Sales, M.M., da Silva, N.P., Silva, T.C., da Silva Jr., W., Simao, D.F., Sousa, J.F., Stecconi, D., Tsukumo, F., Valente, V., Zalcbeg, H., Brentani, R.R., Reis, F.L., Dias-Neto, E. and **Simpson, A.J.** (2000). Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA 97*: 12690-12693.

Dias-Neto, E., Correa, R.G., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva Jr., W., Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., Carvalho, A.F., Matsukuma, A., Baia, G.S., Simpson, D.H., Brunstein, A., de Oliveira, P.S., Bucher, P., Jongeneel, C.V., O'Hare, M.J., Soares, F., Brentani, R.R., Reis, L.F., de Souza, S.J. and **Simpson, A.J.** (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA 97*: 3491-3496.

Guigó, R., Knudsen, S., Drake, N. and **Smith, T.** (1992). Prediction of gene structure. *J. Mol. Biol. 226*: 141-157.

Henke, W., Herdel, K., Jung, K., Schoor, D. and **Lorning, S.A.** (1997). Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res. 25*: 3957-3958.

Hickox, D.M., Gibbs, G., Morrison, J.R., Sebire, K., Edgar, K., Keah, H.H., Alter, K., Loveland, K.L., Hearn, M.T.W., de Kretser, D.M. and **O'Bryan, M.K.** (2002). Identification of a novel testis-specific member of the phosphatidylethanolamine binding protein family, pebp-2. *Biol. Reprod. 67*: 917-927.

Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B. and **Auffray, C.** (1995). The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res. 5*: 272-304.

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

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., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Szustakowki, J., de Jong, P., Catanese, J.J.,

Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. and the **International Human Genome Sequencing Consortium** (2001). Initial sequencing and analysis of the human genome. *Nature 409*: 860-921.

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. *Nature 25*: 239-240.

Makalowski, W., Zhang, J. and **Boguski, M.S.** (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res. 6*: 846-857.

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

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.

Rogic, S., Mackworth, A.K. and **Ouellette, F.B.** (2001). Evaluation of gene-finding programs on mammalian sequences. *Genome Res. 11*: 817-832.

Sakharkar, M.K. and **Kangueane, P.** (2004). Genome SEGE: A database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics 5*: 67.

Sakharkar, M.K., Kangueane, P., Petrov, D.A., Kolaskar, A.S. and **Subbiah, S.** (2002). SEGE: A database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics 18*: 1266-1267.

Sambrook, J., Fritsch, E. and **Maniatis, T.** (1989). *Molecular Cloning*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

Shabalina, S.A., Ogurtsov, A.Y., Rogozin, I.B., Koonin, E.V. and **Lipman, D.J.** (2004). Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res. 32*: 1774-1782.

Shendure, J. and **Church, G.M.** (2002). Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol. 3*: 1-14.

Sogayar, M.C., Camargo, A.A., Bettoni, F., Carraro, D.M., Pires, L.C., Parmigiani, R.B., Ferreira, E.N., de Sa Moreira, E., do Rosario D de O Latorre, M., Simpson, A.J., Cruz, L.O., Degaki, T.L., Festa, F., Massirer, K.B., Sogayar, M.C., Filho, F.C., Camargo, L.P., Cunha, M.A., De Souza, S.J., Faria Jr., M., Giuliatti, S., Kopp, L., de Oliveira, P.S., Paiva, P.B., Pereira, A.A., Pinheiro, D.G., Puga, R.D., S de Souza, J.E., Albuquerque, D.M., Andrade, L.E., Baia, G.S., Briones, M.R., Cavaleiro-Luna, A.M., Cerutti, J.M., Costa, F.F., Costanzi-Strauss, E., Espreafico, E.M., Ferrasi, A.C., Ferro, E.S., Fortes, M.A., Furchi, J.R., Giannella-Neto, D., Goldman, G.H., Goldman, M.H., Gruber, A., Guimaraes, G.S., Hackel, C., Henrique-Silva, F., Kimura, E.T., Leoni, S.G., Macedo, C., Malnic, B., Manzini, B.C.V., Marie, S.K., Martinez-Rossi, N.M., Menossi, M., Miracca, E.C., Nagai, M.A., Nobrega, F.G., Nobrega, M.P., Oba-Shinjo, S.M., Oliveira, M.K., Orabona, G.M., Otsuka, A.Y., Paco-Larson, M.L., Paixao, B.M., Pandolfi, J.R., Pardini, M.I., Passos Bueno, M.R., Passos, G.A., Pesquero, J.B., Pessoa, J.G., Rahal, P., Rainho, C.A., Reis, C.P., Ricca, T.I., Rodrigues, V., Rogatto, S.R., Romano, C.M., Romeiro, J.G., Rossi, A., Sa, R.G., Sales, M.M., Sant'Anna, S.C., Santarosa, P.L., Segato, F., Silva Jr., W.A., Silva, I.D., Silva, N.P., Soares-Costa, A., Sonati, M.F., Strauss, B.E., Tajara, E.H., Valentini, S.R., Villanova, F.E., Ward, L.S., Zanette, D.L. and the **Ludwig-FAPESP Transcript Finishing Initiative** (2004). A transcript finishing initiative for closing gaps in the human transcriptome. *Genome Res. 14*: 1413-1423.

Solovyev, V. (2001). Statistical approaches in eukaryotic gene prediction. In: *Handbook of Statistical Genetics* (Balding, D.J., Bishop, M. and Cannings, C., eds.). John Wiley & Sons Ltd., UK, pp. 83-127.

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

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., Zeeberg, B., Buetow, K.H., Schaefer, C.F., Bhat, N.K., Hopkins, R.F., Jordan, H., Moore, T., Max, S.I., Wang, J., Hsieh, F., Diatchenko, L., Marusina, K., Farmer, A.A., Rubin, G.M., Hong, L., Stapleton, M., Soares, M.B., Bonaldo, M.F., Casavant, T.L., Scheetz, T.E., Brownstein, M.J., Usdin, T.B., Toshiyuki, S., Carninci, P., Prange, C., Raha, S.S., Loquellano, N.A., Peters, G.J., Abramson, R.D., Mullahy, S.J., Bosak, S.A., McEwan, P.J., McKernan, K.J., Malek, J.A., Gunaratne, P.H., Richards, S., Worley, K.C., Hale, S., Garcia, A.M., Gay, L.J., Hulyk, S.W., Villalon, D.K., Muzny, D.M., Sodergren, E.J., Lu, X., Gibbs, R.A., Fahey, J., Helton, E., Ketteman, M., Madan, A., Rodrigues, S., Sanchez, A., Whiting, M., Madan, A., Young, A.C., Shevchenko, Y., Bouffard, G.G., Blakesley, R.W., Touchman, J.W., Green, E.D., Dickson, M.C., Rodriguez, A.C., Grimwood, J., Schmutz, J., Myers, R.M., Butterfield, Y.S., Krzywinski, M.I., Skalska, U., Smailus, D.E., Schnerch, A., Schein, J.E., Jones, S.J., Marra M.A. and the **Mammalian Gene Collection Program Team** (2002). Generation and initial analysis of more than 15,000 full-

length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA 99*: 16899-16903.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science 282*: 2012-2018.

Velculescu, V.E., Zhang, L., Vogelstein, B. and **Kinzler, K.W.** (1995). Serial analysis of gene expression. *Science 270*: 484-487.

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., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and **Zhu, X.** (2001). The Sequence of the Human Genome. *Science 291*: 1304-1351.

Vettore, A.L., da Silva, F.R., Kemper, E.L. and **Arruda, P.** (2001). The libraries that made SUCEST. *Gen. Mol. Biol. 24*: 1-7.

Warrington, J.A., Nair, A., Mahadevappa, M. and **Tsyganskaya, M.** (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics 2*: 143-147.

Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., Lauber, J., Dusterhoft, A., Beyer, A., Kohrer, K., Strack, N., Mewes, H.W., Ottenwalder, B., Obermaier, B., Tampe, J., Heubner, D., Wambutt, R., Korn, B., Klein, M. and **Poustka, A.** (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.

Yao, J., Chiba, T., Sakai, J., Hirose, K., Yamamoto, M., Hada, A., Kuramoto, K., Higuchi, K. and **Mori, M.** (2004). Mouse testis transcriptome revealed using serial analysis of gene expression. *Mamm. Genome 15*: 433-451.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K. and **Rotman, G.** (2003). Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol. 21*: 379-386.

ANEXO 3

A Transcript Finishing Initiative for Closing Gaps in the Human Transcriptome

he Ludwig–FAPESP Transcript Finishing Initiative,[1] Mari Cleide Sogayar,[2] and namaria A. Camargo[2]

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, for 21% of the validated TFUs, a full-length cDNA sequence is not yet available in public databases, and the structure of 69.2% of these TFUs was not correctly predicted by computer programs. The TF strategy provides a significant contribution to the definition of the complete catalog of human genes and transcripts, because 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–CF272733.]

primary objective of the Human Genome Project has been the entification of the complete set of human genes and their de-ved transcripts. A major step towards this goal was achieved at ie beginning of 2001 with the publication of two independent :aft versions of the human genome sequence and the identifi-ition of >30,000 genes (Lander et al. 2001; Venter et al. 2001). owever, it became apparent that extracting exonic sequences irectly from the human genome is not straightforward and that variety of complementary strategies are required for gene iden-fication and characterization.

In this context, microarrays (Penn et al. 2000; Dennis 2001; :hoemaker et al. 2001; Kapranov et al. 2002) and sequence com-arisons with other organisms at an appropriate evolutionary istance (Batzoglou et al. 2000; Roest et al. 2000) constitute pow-ful preliminary approaches to identifying transcribed regions ithin the genome sequence. Nevertheless, transcript sequenc-ıg is necessary both for the final proof of the existence of an xpressed gene and for the precise identification of intron/exon oundaries and alternatively spliced forms (Camargo et al. 2002).

A full-length cDNA sequence, ideally including a transcrip-on initiation site and a polyadenylation site, is the gold stan-ard for transcript definition. Considerable progress has been ıade in the generation of representative full-length cDNA se-uences (Strausberg et al. 1999, 2002; Wiemann et al. 2001; :ikuno et al. 2002; Nakajima et al. 2002), especially following the evelopment of sophisticated protocols for obtaining full-length ranscript molecules and to correct for transcript expression bias Bonaldo et al. 1996, Carninci et al. 2000).

Expressed sequence tags (ESTs) are another major source of ranscript sequence. ESTs either are 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, 1993), but they have also been widely used to build tissue-specific transcript profiles (Bortoluzzi et al. 2000a,b,c; 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 et al. 1998; Jiang and Jacob 1998; Kan et al. 2001), and to study aspects of mRNA structure, such as splicing variants (Hide et al. 2001; Modrek 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 (Garg et al. 1999; Picoult-Newberg et al. 1999; Clifford et al. 2000; Irizarry et al. 2000; Hu et al. 2002).

To date, >5,200,000 human ESTs have been generated from different organs and tissues, deriving mainly from the Merck Gene Index Project (Williamson 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). In addition, 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

A complete list of authors appears at the end of this manuscript.
Corresponding authors.
-MAIL anamaria@compbio.ludwig.org.br; FAX 55-11-3207-7001.
-MAIL mcsoga@iq.usp.br; FAX 55-11-3091-3820.
rticle and publication are at http://www.genome.org/cgi/doi/10.1101/ ır.2111304. Article published online before print in June 2004.

is data set, in conjunction with the draft sequences of the hu-an genome, has been highly informative in the characteriza-ɔn of new human genes (Reymond et al. 2002; Silva et al. 2003). ɛre we describe the utilization of a transcriptome database to ide the generation of novel human transcript sequences on a nome-wide basis. By using the genomic sequence as a scaffold r EST mapping and clustering, we have used RT-PCR to bridge ps between EST clusters that we judged as likely to be derived ɔm the same genes. The resulting novel sequence confirms that e ESTs from different clusters are, in fact, derived from a com-on transcript and defines the intervening region between em.

Because this process is very similar to the finishing phase of nome projects, we called it transcript finishing (TF). This pow-ful, albeit laborious, approach allows the characterization of ɔvel human transcripts and splicing isoforms, which appear to : generally expressed at a low abundance level and/or in a re-ɪicted set of tissues and avoids the necessity of a full-length)NA clone in order to confirm the structure of a gene.

ESULTS

eneration of the Transcriptome Database and EST luster Selection for Experimental Validation

'e have used the publicly available human genome and tran-ript sequences to identify and experimentally validate addi-tional transcribed regions in the human genome. The two data sets were integrated into the transcriptome database by using the BLASTN program to map all transcript sequences onto the as-sembled version of the human genome available from the Na-tional Center for Biotechonology Information (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 graphi-cal interface was also developed (Fig. 1).

We identified 244,148 human transcript clusters, of which 14,598 contained at least one full-length cDNA sequence, and 229,550 clusters that were composed exclusively of partial tran-script sequences. Of the set of 14,598 clusters containing full-length sequences, 13,149 (90%) had at least one corresponding EST, and the remaining 1449 (10%) were composed only of full-length cDNA sequences. These data demonstrate that, despite the fact that >5 million EST sequences are available, they do not fully cover the human transcriptome and that the generation of addi-tional transcribed sequences is still required.

It is noteworthy that clusters composed exclusively by ESTs have a reduced number of sequences (average, 5.9 sequences) derived from fewer different tissues (average, 3.0), compared with clusters containing a full-length cDNA, which have an average of

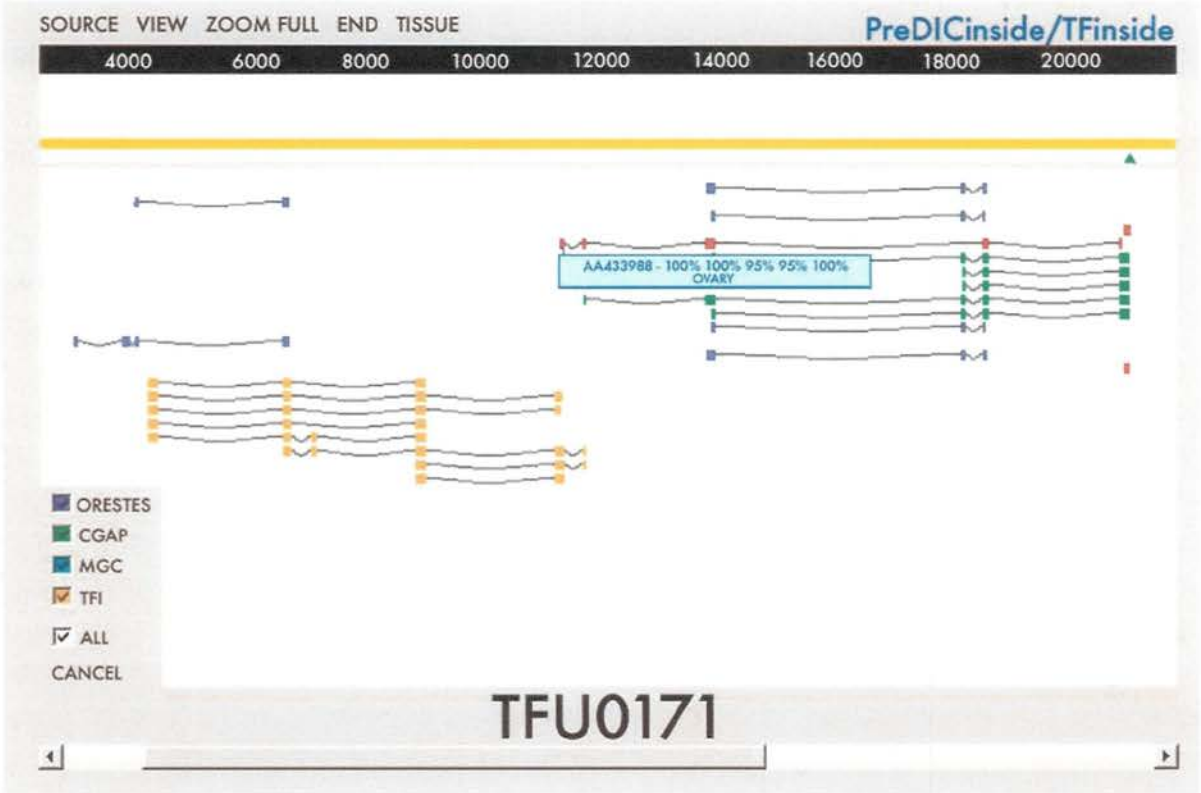


Figure 1 TFI graphical interface. The TFI graphical interface displays a region of the human genome sequence as a yellow line, with a scale in base pairs (bp). Expressed sequence tags (ESTs) that align with the genome sequence are shown in different colors, depending on the project of origin: ORESTES 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 percentage of similarity of each exon with the human genome sequence, and the presence of 3′ tags represented as green triangles.

.5 sequences derived from eight different tissues. Based on ese observations, we conclude that the human transcripts that nain to be defined are expressed at low levels in a restricted set tissues and that their characterization will benefit from a direct proach such as the TF.

Because EST databases contain a significant fraction of artictual and contaminant sequences, we selected, for experimen- validation, pairs of clusters that consist of ESTs that align ncontiguously to the genome, consistent with the presence of splicing structure. We also restricted our validation to pairs of usters that map at a maximum distance of 10 kb from each her, in order to increase the probability that these clusters beng to the same transcript. By using these criteria, a total of 2373 irs of clusters (~2% of the total number of clusters composed of rtial sequences) were initially selected and subjected to manual spection using our graphical interface.

Manual inspection allowed the assessment of similarity and tension of the alignments, as well as the position of the sected pair of clusters relative to the 3'tags. Following this proceure, a subset of 489 pairs of clusters was initially selected for perimental validation. The number of clusters eliminated by anual inspection was very low; therefore, the 489 pairs of clusrs selected for experimental validation can be considered an nbiased sample of the 2373 initially selected clusters. Clusters lected for validation were separated from each other by an avage 2879 bp of intervening genomic sequence and were comosed by an average of 5.92 EST sequences derived from an avage of three distinct tissues. Each pair of EST clusters selected r experimental validation was designated as a single TF unit FU). Information related to the 489 TFUs selected for validation n be accessed at http://200.18.51.201/viewtfi.

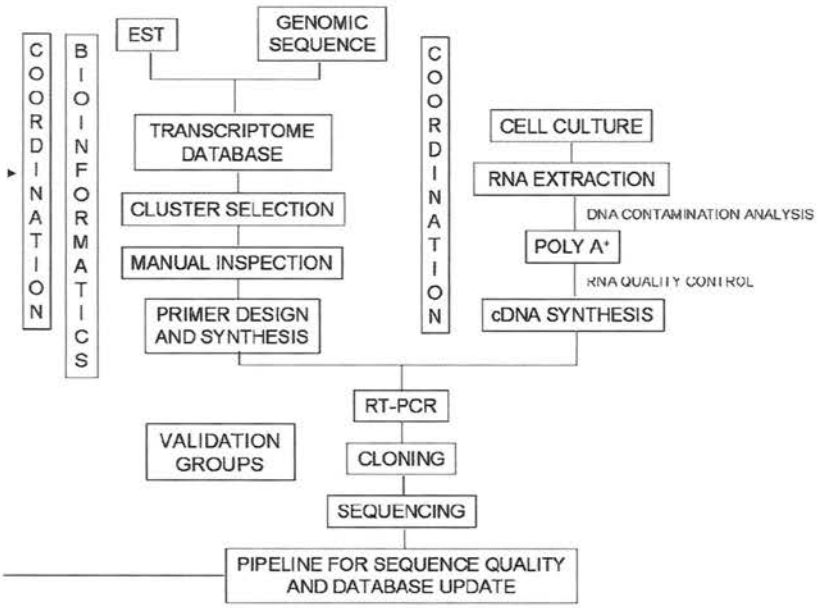
xperimental Validation and the Generation of New ranscribed Sequences

general overview of the computational and experimental valiation strategies is presented in Figure 2. A total of two coordination groups, four 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 because it is generally of a higher quality than are EST sequences. cDNA preparation was also a critical issue, because both the quality and the representation of different tissues directly influence the validation efficiency. As an indicator of genomic DNA contamination, the total RNA preparations were subjected to PCR amplification by using primers within intronic sequences flanking the exon 12 of the *MLH1* gene and found to be negative. The quality of the cDNA product was demonstrated by PCR amplification of sequences located at the 5' extremity of the *NOTCH2* transcript (a long transcript of 11.4 Kb). A total of 22 cDNA preparations, derived from a number of cell lines and representing 18 distinct tissues, were used.

The total of 3019 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 (Fig. 1). We successfully validated 211 of the 489 TFUs that were distributed, yielding an overall validation efficiency of 43.1%.

A single pair of primers was tested for each TFU, and experimental validation was undertaken in a high-throughput single-pass format. Few modifications were adopted when a positive amplification was not achieved (see Methods). To estimate the false-negative amplification rate of the TF strategy, we determined the number of the nonvalidated TFUs for which a full-length cDNA sequence had been made available by other sequencing projects during the course of our project. For 40 of the nonvalidated TFUs, we were able to identify a full-length sequence linking the two EST clusters initially selected for validation. Thus, these cases can be considered to be false-negative amplifications. For 118 of the nonvalidated TFUs, the existence of a full-length sequence matching just one of the two selected clusters allowed us to conclude that the two clusters were in fact part of different transcripts. For these cases, the absence of an RT-PCR product thus reflects true negatives. For the 120 of the remaining nonvalidated TFUs, a conclusive result could not identify any corresponding full-length sequence. Therefore, based on these results, we can estimate that the rate of false-negative amplifications in the TF strategy is ~25% (40/158 nonvalidated TFUs).

In addition, 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 nonvalidated 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



igure 2 General scheme of the TFI strategy. Schematic outline of the strategy used for computational and experimental validation of TFU sequences. Following the development of bioinformatcs tools, the generation of the transcriptome database, and automatic cluster selection, the project asks were divided between the coordination and the validation laboratories.

Table 1. Comparison Between Validated and Nonvalidated TFUs

	Validated TFUs (Std deviation)	Nonvalidated TFUs (Std deviation)	*P* value
Average distance between clusters	2609 (3202)	3105 (2942)	0.008
Average no. of ESTs in each cluster	6.10 (8.91)	5.77 (13.23)	0.010
Average 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

ʼored validation. The presence of ESTs derived from the same sue in both clusters did not influence the likelihood of valida- ɔn according to χ^2 tests.

A total of 59,975 bp of transcribed sequence, organized into 2 exons, were generated, contributing to the definition of the ucture of 211 distinct human transcripts. Each validated TFU d a mean of 281.6 bp and a median of 207 bp of novel se- ience not represented by the original EST clusters and a mean 2.03 and a median of two newly defined exons. The validated U sequences have been submitted to GenBank under the ac- ssion numbers CF272536 to CF272733, which are provided as pplemental Table 1.

onsensus Sequences Generation and Annotation the Validated Human Transcripts

ɔnsensus sequences produced by assembling the sequences de- ʼed from the validation fragment and the sequences from all Ts in both clusters were obtained for 186 of the 211 validated Us. Assembly of a consensus sequence was not possible for TFUs, due mainly to the presence of repetitive sequences d alternative splicing forms. Consensus sequences, with an ʼerage of 1240 bp, can be accessed at (http://200.18.51.201/ ewconsensus/).

Consensus sequences derived from the validated TFUs were igned to the July 2003 version of human genome sequence sembly provided by the University of California, Santa Cruz JCSC), using the BLAT search tool (http://genome.ucsc.edu/cgi- n/hgBlat) to compare the validated consensus sequences with nown genes and gene predictions (Table 2). A significant frac- ɔn (68.8%) of the validated transcripts completely overlapped ith the alignment coordinates of a known gene or full-length uman mRNA submitted to the GenBank during the course of ur project (Fig. 3A), and a smaller fraction (10.2%) represented ctensions (mostly 5′) to partial cDNA sequences deposited in ublic databases (Fig. 3B). However, for 21% of the validated FUs, a full-length cDNA sequence was not available in public atabases as of July 2003. The structure of the majority (69.2%) of ıe validated TFUS without a corresponding full-length cDNA sequence had not been correctly predicted by ab initio gene prediction programs such as Fgenesh++, Geneid, and GenScan. These TFUs can thus be considered as new human transcripts.

The consensus sequences corresponding to new human transcripts were further characterized by BLASTX analysis, and protein domains were predicted by using the Pfam and Prosite databases. Of the 39 consensus sequences representing new human transcripts, 27 (69.2%) contained an ORF of at least 100 amino acids, and eight (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 provided as Supplemental Table 2.

The validated transcripts that completely overlapped with the alignment coordinates of a known gene containing a defined ORF were used to estimate the percentage of consensus sequences that represent complete transcripts. Only a small fraction (9.7%) of the 93 validated TFUs analyzed contained a complete ORF. The low percentage obtained was expected because, in the TF strategy, RT-PCR is used to brigde gaps between partial transcript sequences.

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 transcript w/partial prediction	15	8.0
New transcript w/o prediction	12	6.5
Total	186	100

Identification and Experimental Validation of Alternatively Spliced Isoforms

Several reports have suggested that at least 30% to 35% of human genes undergo alternative splicing (Brett et al. 2000; Modrek et al. 2001); nevertheless, this value is probably underestimated because 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. The possibility of genomic DNA contamination, in those cases in which we have observed the retention of an intron, was excluded due to the presence of processed introns in the same cDNA molecule containing the retained intron. Moreover, the RNAs used for experimental validation of the alternatively spliced forms have been treated with DNase and tested for the absence of intronic sequences, as described in Methods.

We selected six TFUs with alternative exon usage, representing a total of 14 splicing isoforms, for further experimental validation. Each pair of primers used for experimental validation of the alternatively spliced forms was assayed against all 22 cDNA sources, without pooling. Touchdown PCR confirmed 10 (83%) of the putative investigated isoforms. No PCR amplification was achieved for one TFU. Some splicing isoforms were expressed in

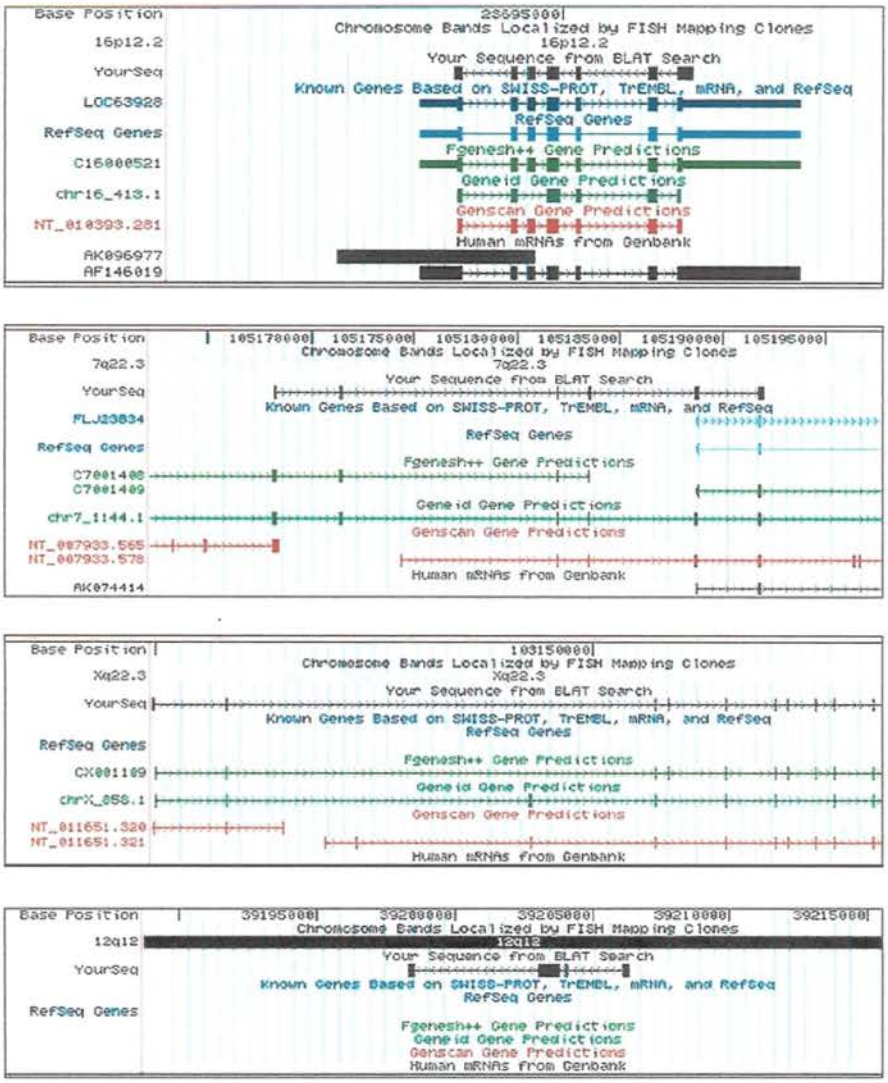


Figure 3 Characterization and annotation of validated TFUs. Alignment of four consensus sequences, derived from the validated TFUs, to the July 2003 version of the UCSC human genome sequence assembly, using the BLAT search tool. (*A*) TFU00023 corresponds to YourSeq (black) completely overlapping with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

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 toward 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, because of 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.

able 3. Alternative Splicing Forms Within Validated TFs

alidated onsensus	Type of alternative splicing	Presence of conserved acceptor and donor sites	No. of alternative isoforms	No. of validated isoforms
FU0118	Exon usage	Yes	2	1
FU0200	Exon usage	Yes	4	4
FU0274	Exon usage	Yes	2	2
FU0351	Exon usage	Yes	2	2
FU1004	Exon usage	Yes	2	1
FU1058	Exon usage	Yes	3	0
FU0155	Exon usage	Yes	2	nd
FU0238	Exon usage	Yes	2	nd
FU0308	Exon usage	Yes	2	nd
FU0003	Intron retention	nd	nd	nd
FU0019	Intron retention	nd	nd	nd
FU0035	Intron retention	nd	nd	nd
FU0052	Intron retention	nd	nd	nd
FU0099	Intron retention	nd	nd	nd
FU0112	Intron retention	nd	nd	nd
FU0125	Intron retention	nd	nd	nd
FU0131	Intron retention	nd	nd	nd
FU0209	Intron retention	nd	nd	nd
FU0285	Intron retention	nd	nd	nd
FU0371	Intron retention	nd	nd	nd
FU0148	Exon skipping	nd	nd	nd
FU1061	Exon skipping	nd	nd	nd

nd = not done.

01; Venter et al. 2001), have uniformly concluded that the uman genome possesses <35,000 genes. This number has been pported by a preliminary analysis of EST coverage of known nes (Ewing and Green 2000) as well as comparative genomics alysis (Roest et al. 2000). Most of these 35,000 genes are al-ady represented by a full-length cDNA sequence in transcript tabases. In UniGene (http://www.ncbi.nlm.nih.gov/entrez/ iery.fcgi?db=unigene), for example, there are currently 28,412 anscript clusters represented by at least one full-length cDNA quence.

Here we have proposed and validated the use of the TF strat-gy for characterization of new human transcripts that are only artially represented by ESTs. Because EST databases contain a gnificant fraction of artifactual and contaminant sequences, we lected pairs of clusters for experimental validation that exhib-ed a clear splicing structure when aligned to the genome. By quiring the occurrence of splicing, the level of contamination the EST databases is significantly reduced, although at the pense of eliminating many genuine 3′ ESTs. The selection cri-ria used in our initial analysis are very restrictive, and the adop-on of less stringent criteria (including clusters without a splic-g structure) will certainly be required to complete the catalog of uman genes using the strategy we described. Given the 2373 itially selected clusters, of which 489 were subjected to experi-ental validation, 1884 pairs of clusters remain to be validated. we assume an overall validation efficiency of 43%, we can stimate that the TF strategy might contribute to the definition f at least 791 additional genes in the human genome.

Several factors may have influenced our validation effi-iency, including experimental limitations related to primer and DNA synthesis, the particular characteristics of human tran-cripts such as low expression level, and the existence of a sig-ificant proportion of sense-antisense transcriptional units on pposite DNA strands of the same genomic locus (Yelin et al 003). A 25% false-negative amplification rate was estimated for the TF strategy and is probably related to the high-throughput single-pass format adopted for the experimental validation. In this context, the use of additional primer pairs and modifications of cycling parameters that would favor the amplification of difficult targets could be added to the process to reduce the negative amplification rate.

We found that validation efficiency was enhanced by implementation of quality controls for cDNA synthesis, the use of polyA+-derived cDNA, 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 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.

For a reasonable fraction (21%) of the validated TFUs, a full-length cDNA sequence was not yet available in public databases. The structure of the majority (69.2%) of these validated TFUs had not been correctly predicted by ab initio gene prediction programs and, consequently, was not annotated in the human genome. In addition, 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 TF strategy provides a convenient and unique means for delineating gene boundaries and new transcribed 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.

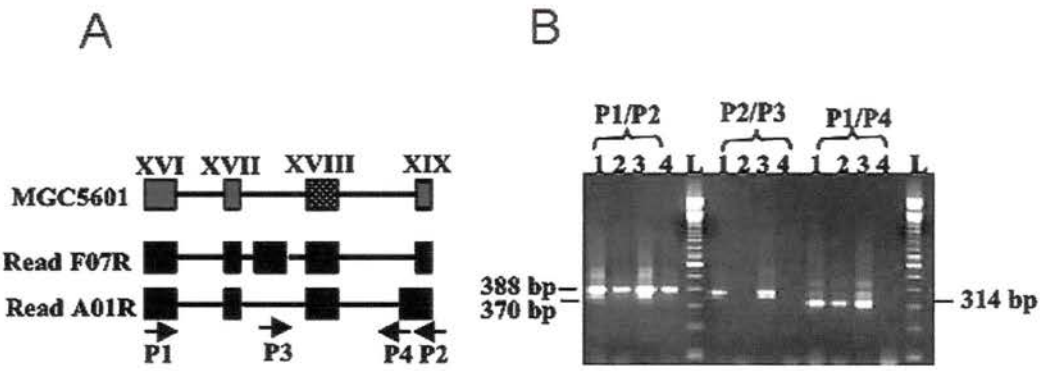
METHODS

Cell Culture

Human cell lines were obtained from the American Type Culture Collection (ATCC) and cultured as recommended (http://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 by using the cesium chloride cushion technique (Chirgwin et al. 1979). Poly A^+ RNA was isolated from 200 µg total RNA with the PolyAttract mRNA isolation kit (Promega), and the total yield of this purification was used for cDNA synthesis. For cDNA synthesis, 100 to 200 µg total RNA or the corresponding purified mRNA were treated with 100 U DNAse I (FPLC-pure, Amersham) and reverse-transcribed by 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



gure 4 Experimental validation of MGC5601 gene alternative splicing isoforms. (*A*) Gene structure for exons XVI-XIX (boxes) of the MGC5601 gene :ated on chromosome 12. Introns are represented by lines. Two alternative exons are shown on TFU reads, and a hypothetical combination of these 'o 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 :signed for validation tests, as indicated in the figure (P1–P4), and each pair of primers were assayed against all 22 cDNA preparations without pooling.) We detected all four of these alternative splicing isoforms in MGC5601. Numbers one through four indicate the tissues from which the cDNA was)tained (1, multiform glioblastoma; 2, glioblastoma; 3, prostate carcinoma; and 4, primary kidney cell culture). The sizes of the bands obtained are :dicated. L indicates 100-bp ladder.

l validation laboratories involved in the project. The quality of ıe cDNA synthesis and the absence of genomic DNA contamiıtion were evaluated for each preparation. Total RNA was subcted to PCR amplification by using primers within intronic seıences flanking exon 12 of the *MLH1* gene (forward, 5′-GGTGTCTCTAGTTCTGG-3′; reverse, 5′-CATTGTTGTAG AGCTCTGC-3′). The quality of the cDNA product was also sted by PCR amplification of sequences located at the 5′ exemity of the *NOTCH2* transcript (a long transcript of 11.4Kb; rward, 5′-ACTGTGGCCAACCAGTTCTC-3′; reverse, 5-′ CTCT ACAGGTGCTCCCTTC-3′).

T-PCR and Sequencing

T-PCR was carried out in 25 µL reaction mixtures containing 1 L cDNA, 10× *Taq* DNA polymerase buffer, 200 µM dNTP, 6 moles of primers, 1.5 mM $MgCl_2$, and 1 U *Taq* DNA polymerase GIBCO BRL). Standard PCR conditions were as follows: 4 min at 4°C (initial denaturation), 40 sec at 94°C, 40 sec at 55°C, and 1 in at 72°C for 35 cycles and a final extension step of 10 min at 2°C. Modifications of the standard protocol included annealing mperature, $MgCl_2$ concentration, addition of PCR enhancers ıch as betaine, and the use of polymerases with hot start activy. PCR products were directly sequenced with the same primers sed for RT-PCR or cloned before sequencing. If more than one agment was obtained for the same TFU using different cDNA ources, all fragments were sequenced. This was also the case if ıultiple bands were obtained in PCR amplifications using a ngle cDNA source. Sequencing different fragments obtained for specific TFU allowed us to characterize a number of alternavely spliced transcripts. Sequencing reactions were carried out y using the DYEnamic ET terminator Cycle Sequencing Kit (Amrsham Pharmacia) and separated by electrophoresis using an BI 377 Prism Sequencer (Applied Biosystems) according to supplier's recommendations.

Transcriptome Database and Graphical Interface

LASTN was used to identify pair-wise similarities between all nown transcript sequences and the draft genome sequence deosited in release 66 (March 2001) of the European Molecular iology Laboratory (EMBL) database. Transcribed sequence data ere extracted from several sources: (1) the human EST section of MBL release 66, (2) human mRNA documented in the human ection of EMBL release 66, (3) ORESTES sequences from the Ludig Institute for Cancer Research (LICR)/FAPESP Human Cancer enome project, and (4) human mRNAs documented in the CBI curated RefSeq database (http://www.ncbi.nlm.nih.gov/efseq). 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 data set 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 by using the PFP software package (Paracel). For each pair of matching transcribed and genomic sequences, local alignments were generated by using Sim4 (Florea et al. 1998), with parameters W = 15, R = 0, A = 4, and 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 30 nt. 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 by 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

nomic interval where the two noncontiguous EST clusters map d where the system searched for primers. A single pair of prim- ; was designed for each TFU, which usually targeted the two ons flanking the putative gap. In a few cases, in which the esence of repetitive sequences or atypical base composition evented the design of primers, adjacent exons were used. The imer3 program (version 0.9) developed by the Whitehead In- tute for Biomedical Research was used for primer design, opting the following parameters: primer size of a minimum of bp, optimal 18 bp and maximum 21 bp; melting temperature a minimum of 55°C, optimal 60°C and maximum 65°C; and clamp set to one. The output of Primer3 was processed in der to filter primers that had alternate annealing sites in the ven genomic sequence. The system uses a Web-based interface at allows submission of files containing information on primer sign, retrieval of primers found, and the modification of de- ılt parameters for primer picking.

quence Analysis and Database Update

quences were subjected to an automated protocol to (1) assess quence quality, (2) trim vector sequences, (3) mask repetitive ements, and (4) remove undesirable sequences such as bacte- ıl, mitochondrial, and fungi sequences. The sequence quality as determined by Phred analysis using a trimCutOff of 0.06171 wing and Green 1998; Ewing et al. 1998). Sequences with <100 ıses were excluded. Mitochondrial, bacterial, and fungi se- ıences were identified by BLAST searches against the GenBank ıtry corresponding to the human mitochondrial complete ge- ome sequence and against a locally developed bacterial and ngal database, respectively. Significant hits were determined by ing an E value of 10^{-5} for searches against mitochondrial ge- ome and an E value of 10^{-30} for searches against bacterial da- bases. Masking of repetitive elements was undertaken by using e RepeatMasker (http://www.repeatmasker.org) under default ırameters. The remaining high-quality sequences were aligned gainst the original genomic clone by using the BLASTN pro- am, and alignment coordinates and scores were loaded into the ySQL database on a daily basis.

onsensus Assembly

ıe reads corresponding to validated TFs were assembled into contig by using the PhredPhrap. The contig sequence was igned with both EST clusters by using the BLASTN program, ıd alignment coordinates were used for consensus generation. Web-based interface was developed to monitor the assembly ıd access the consensus sequences (http://200.18.51.201/ ewconsensus/).

haracterization of Validated Transcripts

haracterization of validated transcripts was pursued by using ıe UCSC Genome Browser (Kent et al. 2002), which is available : http://genome.ucsc.edu. This allowed determination of se- uence overlap between the validated consensus sequences, nown genes, and gene predictions. Consensus sequences de- ved from the validated TFUs were aligned to the July 2003 ver- on of the human genome sequence assembly provided by CSC using the BLAT search tool. The annotation tracks used for mparison to already known genes were known genes, RefSeq enes, and human mRNAs from the GenBank. A validated tran- ript was considered a new gene if its alignment coordinates did ot match the coordinates of any other sequence available ırough the known genes, RefSeq genes, or human mRNA anno- ation tracks. For comparison to gene predictions, the following acks were used: Fgenesh++, Geneid, and GenScan predictions. he prediction of individual exons instead of the full transcript rediction was considered. A validated exon was considered as redicted if it aligned within the coordinates defined by any of he three gene prediction programs (not necessarily sharing bor- ers) and a new validated transcript was considered not predicted all exons were not predicted by the computer programs. The onsensus sequences corresponding to new validated transcripts were further characterized by BLASTX analysis, and protein domains were determined by 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 by 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 by using the UCSC browser. 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 by using Primer3 with default parameters. The presence of alternative isoforms was analyzed by using a cDNA panel composed of 20 different normal and tumor tissues. *GAPD* 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,[3] Dirce Maria Carraro,[3] Lilian C. Pires,[3] Raphael B. Parmigiani,[3] Elisa N. Ferreira,[3] Eloísa de Sá Moreira,[3,32] Maria do Rosário D. de O. Latorre,[4] Andrew J.G. Simpson,[3] and Anamaria A. Camargo[3]

Coordination Group University of São Paulo Chemistry Institute

Luciana Oliveira Cruz,[5] Theri Leica Degaki,[5] Fernanda Festa,[5] Katlin B. Massirer,[5] and Mari C. Sogayar[5]

Bioinformatics Groups

Fernando Camargo Filho,[6] Luiz Paulo Camargo,[6] Marco A.V. Cunha,[7] Sandro J. De Souza,[8] Milton Faria Junior,[6] Silvana Giuliatti,[6] Leonardo Kopp,[9] Paulo S.L. de Oliveira,[9] Paulo B. Paiva,[10] Anderson A. Pereira,[6] Daniel G. Pinheiro,[7] Renato D. Puga,[6] and Jorge Estefano S. de Souza[8]

Validation Groups

Dulcineia M. Albuquerque,[11] Luís E.C. Andrade,[12] Gilson S. Baia,[13] Marcelo R.S. Briones,[14] Ana M.S. Cavaleiro–Luna,[15] Janete

[3]Ludwig Institute for Cancer Research, São Paulo, SP, 01509–010, Brazil
[4]Department of Epidemiology, School of Public Health, University of São Paulo, SP, 01246–904, Brazil
[5]Instituto de Química, Universidade de São Paulo, São Paulo, SP, 05513–970, Brazil
[6]Dep. de Eng. Química e de Informática, Bioinformática, Universidade de Ribeirão Preto, Ribeirão Preto, SP, 14096–380, Brazil
[7]Centro de Terapia Celular, Hemocentro e Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 14051–140, Brazil
[8]Laboratório de Biologia Computacional, Instituto Ludwig, São Paulo, SP, 01509–010, Brazil
[9]Laboratório de Genética e Cardiologia Molecular, Instituto do Coração, Universidade de São Paulo, São Paulo, SP, 05403–000, Brazil
[10]Bioinformatics Laboratory, Health Informatics Department, Federal University of São Paulo, São Paulo, SP, 04039–032, Brazil
[11]Departamento de Clínica Médica, Hemocentro, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, SP, 13083–970, Brazil
[12]Rheumatology Division, Federal University of São Paulo, São Paulo, SP, 04113–001, Brazil
[13]Departamento de Histologia e Embriologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, 05508–900, Brazil
[14]Department of Microbiology, Immunology and Parasitology, Federal University of São Paulo, São Paulo, SP, 04023–062, Brazil
[15]Laboratory for Cellular and Molecular Endocrinology, School of Medicine, University of São Paulo, São Paulo, SP, 01246–903, Brazil

. Cerutti,[16] Fernando F. Costa,[11] Eugenia Costanzi-Strauss,[17] ilza M. Espreafico,[18] Adriana C. Ferrasi,[19] Emer S. Ferro,[13] aria A.H.Z. Fortes,[15] Joelma R.F. Furchi,[20] Daniel Giannella-to,[15] Gustavo H. Goldman,[21] Maria H.S. Goldman,[22] Arthur uber,[23] Gustavo S. Guimarães,[16] Christine Hackel,[24] Flavio nrique-Silva,[20] Edna T. Kimura,[13] Suzana G. Leoni,[11] Cláudia acedo,[25] Bettina Malnic,[26] Carina V. Manzini B.,[26] Suely K.N. arie,[27] Nilce M. Martinez-Rossi,[25] Marcelo Menossi,[28,29] Elisa-te C. Miracca,[30] Maria A. Nagai,[30] Francisco G. Nobrega,[31] Ma-ia P. Nobrega,[31] Sueli M. Oba-Shinjo,[27] Márika K. Oliveira,[18] uilherme M. Orabona,[32] Audrey Y. Otsuka,[33] Maria L. Paço-rson,[18] Beatriz M.C. Paixão,[7] Jose R.C. Pandolfi,[34] Maria I.M.C. rdini,[19] Maria R. Passos Bueno,[32] Geraldo A.S. Passos,[35] Joao B. squero,[36] Juliana G. Pessoa,[36] Paula Rahal,[37] Cláudia A. inho,[38] Caroline P. Reis,[28] Tatiana I. Ricca,[14] Vanderlei Rod-gues,[39] Silvia R. Rogatto,[38] Camila M. Romano,[23] Janaína G. meiro,[37] Antonio Rossi,[39] Renata G. Sá,[39] Magaly M. Sales,[19] mone C. Sant'Anna,[24] Patrícia L. Santarosa,[40] Fernando Segato,[25] Wilson A. Silva Jr.,[7,25] Ismael D.C.G. Silva,[33] Neusa P. Silva,[12] Andrea Soares-Costa,[20] Maria F. Sonati,[41] Bryan E. Strauss,[42] Eloiza H. Tajara,[37] Sandro R. Valentini,[34] Fabiola E. Villanova,[33] Laura S. Ward,[40] and Dalila L. Zanette[7]

ACKNOWLEDGMENTS

We dedicate this work to Dr. Ricardo R. Brentani (Director of the Ludwig Institute-São Paulo Branch and of the A.C. Camargo Hospital) and Dr. José Fernando Perez (Scientific Director of the São Paulo Research Foundation-FAPESP) for unconditional support and constant incentive to the Brazilian Genome Initiative. We thank Fernanda G. Barbuzano, Mário H. Bengtson, Ana P. Bogossian, 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).

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

REFERENCES

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., 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 2375 human brain genes. *Nature* **355:** 632–634.

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., and 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.

———. 2000b. A novel resource for the study of genes expressed in the adult human retina. *Invest. 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., Samaia, H.P., Dias-Neto, E., Simao, D.F., Migotto, I.A.,

Laboratório de Endocrinologia Molecular, Disciplina de Endocrinologia, Departamento de Medicina, Universidade Federal de São ulo, São Paulo, 04039–002, Brazil
Laboratório de Transferência Gênica, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, 05508–900, Bra-

Departamento de Biologia Celular e Molecular e Bioagentes Patogenicos, Faculdade de Medicina de Ribeirão Preto, Universidade de o Paulo, Ribeirão Preto, SP, 14049–900, Brazil
Laboratório de Biologia Molecular, Hemocentro, Faculdade de Medicina, Universidade Estadual Paulista, Botucatu, SP, 18618–970, azil
Departamento de Genética e Evolução, Universidade Federal de o Carlos, São Carlos, SP, 13565–905, Brazil
Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 14040–903, Brazil
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 14040–901, Brazil
Depto. de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, SP, 05508–000, Brazil
Departamento de Genética Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, SP, 13081–970, azil
Departamento de Genética, Faculdade de Medicina de Ribeirão eto, Universidade de São Paulo, Ribeirão Preto, SP, 14040–900, azil
Departamento de Bioquímica, Instituto de Química, Universidade e São Paulo, São Paulo, SP, 05599–970, Brazil
Departamento de Neurologia, Faculdade de Medicina, Universidade de São Paulo, São Paulo, SP, 01246–903, Brazil
Laboratório de Genoma Funcional, Centro de Biologia Molecular e ngenharia Genética, Universidade Estadual de Campinas, Campinas, SP, 13083–970, Brazil
Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, 13084–971, Brazil
Departamento de Radiologia, Disciplina de Oncologia, Faculdade e Medicina, Universidade de São Paulo, São Paulo, SP, 01246–903, azil
Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do araíba, São José dos Campos, SP, 12244–000, Brazil
Departamento de Biologia, Centro de Estudos do Genoma Humano, Instituto de Biociências, Universidade de São Paulo, São aulo, SP, 05508–900, Brazil
Molecular Gynecology Laboratory, Gynecology Department, Federal University of São Paulo, São Paulo, SP, 04039–001, Brazil
Department of Biological Sciences, School of Pharmacy, São Paulo tate University, Araraquara, SP, 14801–902, Brazil
Disciplina de Genética, Faculdade de Odontologia, Universidade e São Paulo, Ribeirão Preto, SP, 14040–900, Brazil
Departamento de Biofísica, Universidade Federal de São Paulo, São aulo, SP, 04023–062, Brazil
Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São Jose do Rio Preto, SP 5054–000, Brazil
Departamento de Genética, Instituto de Biociências, Universidade stadual Paulista, Botucatu, SP, 18618–000, Brazil
Departamento de Bioquímica e Imunologia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, P, 14049–900, Brazil
Laboratório de Genética Molecular do Câncer, Departamento de Clínica Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, SP, 13083–970, Brazil
Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, SP, 13083–970, Brazil
Setor de Vetores Virais, Laboratório de Cardiologia Molecular, Instituto do Coração, Faculdade de Medicina, Universidade de São Paulo, São Paulo, SP, 05403–000, Brazil

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.* **98:** 12103–12108.
margo, 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.
rninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., and 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:** 1617–1630.
irgwin, 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.
ark, 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.
ifford, 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.
nnis, C. 2001. Tiled arrays for gene hunting. *Nat. Rev. Genet.* **2:** 161.
as, N.E., Garcia, C.R., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva Jr., W., 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.* **97:** 3491–3496.
nham, 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.
ring, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred, II: Error probabilities. *Genome Res.* **8:** 186–194.
——. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.
ring, 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.
orea, 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.
arg, 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.
autheret, 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.
attori, 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.
ide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C., and Kelso, J.F. 2001. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.* **11:** 1848–1853.
u, 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.
udson, 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.
uminiecki, L. and Bicknell, R. 2000. In silico cloning of novel endothelial-specific genes. *Genome Res.* **10:** 1796–1806.
izarry, 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.
eli, 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.
ang, J. and Jacob, H.J. 1998. EbEST: An automated tool using expressed sequence tags to delineate gene structure. *Genome Res.* **8:** 268–275.
an, 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.
an, Z., States, D., and Gish, W. 2002. Selecting for functional alternative splices in ESTs. *Genome Res.* **12:** 1837–1845.
apranov, 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.* **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., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286:** 455–457.
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.* **99:** 16899–16903.
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.
emann, 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.
lliamson, A.R. 1999. The Merck Gene Index project. *Drug Discov. Today* **4:** 115–122.
e, 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.
i, 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.
lin, 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.
, 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.

WEB SITE REFERENCES

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene; Unigene home page.
http://www.atcc.org; American Type Culture Collection home page.
http://www.ncbi.nlm.nih.gov/refseq; RefSeq home page.
http://www.repeatmasker.org; RepeatMasker program.
http://genome.ucsc.edu/cgi-bin/hgBlat; University of California Santa Cruz, Genome Browser.
http://200.18.51.201/viewtfi; Information related to TFUs selected for experimental validation.
http://200.18.51.201/transcript/Participants.html; Full list of participant groups of The Ludwig-FAPESP Transcript Finishing Initiative.
http://200.18.51.201/viewconsensus; Access to consensus sequences generated for validated TFUs.

Received October 23, 2003; accepted in revised form March 12, 2004.

ANEXO 4

Nineteen Additional Unpredicted Transcripts from Human Chromosome 21

Alexandre Reymond,[1,*] Anamaria A. Camargo,[2,*] Samuel Deutsch,[1,*] Brian J. Stevenson,[3,4,*] Raphael B. Parmigiani,[2] Catherine Ucla,[1] Fabiana Bettoni,[2] Colette Rossier,[1] Robert Lyle,[1] Michel Guipponi,[1] Sandro de Souza,[2] Christian Iseli,[3,4] C. Victor Jongeneel,[3,4] Philipp Bucher,[3,4,5] Andrew J. G. Simpson,[2] and Stylianos E. Antonarakis[1,†]

[1]*Division of Medical Genetics, University of Geneva Medical School, Geneva, Switzerland*
[2]*Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente, 109, 01509-010, São Paulo, SP, Brazil*
[3]*Swiss Institute of Bioinformatics (SIB), Epalinges, Switzerland*
[4]*Office of Information Technology, Ludwig Institute for Cancer Research, Epalinges, Switzerland*
[5]*Swiss Institute for Experimental Cancer Research (ISREC), Epalinges, Switzerland*
*****[AUTHORS: PROVIDE POSTAL CODES FOR ALL]******

[*]*These authors contributed equally to this work.*

[†]*To whom correspondence and reprint requests should be addressed. Fax: 0041227025706. E-mail: Stylianos.Antonarakis@medecine.unige.ch.*

The identification of all human chromosome 21 (HC21) genes is a necessary step in understanding the molecular pathogenesis of trisomy 21 (Down syndrome). The first analysis of the sequence of 21q included 127 previously characterized genes and predicted an additional 98 novel anonymous genes. Recently we evaluated the quality of this annotation by characterizing a set of HC21 open reading frames (C21orfs) identified by mapping spliced expressed sequence tags (ESTs) and predicted genes (PREDs), identified only *in silico*. This study underscored the limitations of *in silico*-only gene prediction, as many PREDs were incorrectly predicted. To refine the HC21 annotation, we have developed a reliable algorithm to extract and stringently map sequences that contain bona fide 3′ transcript ends to the genome. We then created a specific 21q Acedb [AUTHORS: DEFINE] that incorporates new ESTs as well as features such as CpG islands, repeats, and gene predictions. Using these tools we identified 27 new putative genes. To validate these, we sequenced previously cloned cDNAs and carried out RT-PCR, 5′- and 3′-RACE procedures, and comparative mapping. These approaches substantiated 19 new transcripts, thus increasing the HC21 gene count by 9.5%. These transcripts were likely not previously identified because they are small and encode small proteins. We also identified four transcriptional units that are spliced but contain no obvious open reading frame. The HC21 data presented here further emphasize that current gene prediction algorithms miss a substantial number of transcripts that nevertheless can be identified using a combination of experimental approaches and multiple refined algorithms.

Key Words: human chromosome 21, transcription map, genomic sequences annotation, gene prediction, Down syndrome, trisomy 21, expression pattern, Acedb

INTRODUCTION

Trisomy 21 causes Down syndrome (DS), which is the most common genetic cause of mental retardation and affects approximately 1 in 700 live births [1]. The understanding of the molecular pathogenesis of DS necessitates the identification of all human chromosome 21 (HC21) genes and assessment of their contribution in the different phenotypes of the syndrome [2]. This annotation should also simplify the identification of the genes responsible for monogenic diseases (for example, holoprosencephaly and Usher syndrome 1E [3,4]), complex common phenotypes (for example, bipolar disease and familial combined hyperlipidemia [5,6]), and malignancies mapping to this chromosome (for example, childhood leukemias, transient neonatal leukemia, and squamous non-

0888-7543/02 $35.00

doi:10.1006/geno.2002.6781, available online at http://www.idealibrary.com on IDEAL

ll lung carcinoma [7–9]). In addition, this work may ;ht on epidemiological studies that suggest that DS ıals are "protected" from certain cancers because their :e is lower than that of the general population [10].

sequence of HC21, sequences from eukaryotic s, and sequences of expressed sequence tags (ESTs) :ilitated the cataloging of human genes [11–19]. The nnotation of the complete sequence of the long arm 1 revealed 127 genes and 98 predicted transcripts ;) [16]. We previously evaluated the quality of the experimentally for a set of C21orfs defined by match-:ed ESTs and PREDs defined solely by computer pre- [20]. We showed that HC21 open reading frames) gene models have a stronger predictive value than :D models. This study also underscored the limita- *in silico*-only gene prediction, as many PREDs were :tly predicted [20]. Similar results were obtained in an ; of the HC21 region syntenic to mouse chromosome n the uromodulin-like gene to KIAA0179), where 6, PRED47, PRED48, PRED49, and PRED50 were o be incorrectly predicted [21].

versely, we expect that yet-undescribed HC21 genes identified, as the gene prediction strategy for chromo- and the rest of the genome is likely to have been biased genes with small ORFs, and/or large 5′ and/or 3′ UTRs 9]. Furthermore, additional ESTs and genomic :es of other species have been deposited in the data- :er the initial annotation of chromosome 21 facilitating :ch for as yet unidentified genes. Indeed, we and oth- e already identified six new HC21 genes/gene models 'RF sizes not exceeding 690 bp (*C21orf67*, *C21orf69*, '0, PRED67, PRED68, and PRED69) [20,22].

'urther refine the HC21 sequence annotation, we have yzed the entire chromosome sequence incorporating ESTs [23–25] (http://www.genoscope.cns.fr/, 'mgc.nci.nih.gov/, http://www.china-zj.com/eng-jingquqiye/rd/genome.htm, and http://www.nedo. :nglish/index.html) and creating a reliable algorithm ıgently extract and map to the genome bona fide 3′ ·ipt ends from EST traces. The combined data are zed as an Acedb **[AUTHORS: DEFINE]** database, 21ace, freely available to the scientific community. The graphical display allows an integrated view of the data.

have identified 19 novel genes not previously predicted chromosome 21 mapping and sequencing consortium ıus increasing the gene content of this human chromo- ɔy almost 10% (163 known genes, 36 predicted genes, novel genes). We also identified four additional HC21 :iptional units that are spliced but contain no obvious eading frame. Laboratory experiments (EST sequencing, R, and RACE) and human–mouse comparative map- onfirmed the existence of these newly identified tran- and transcriptional units. The medical implications of :esults for the study of DS phenotypes and for the gene ıt of the entire human genome are discussed.

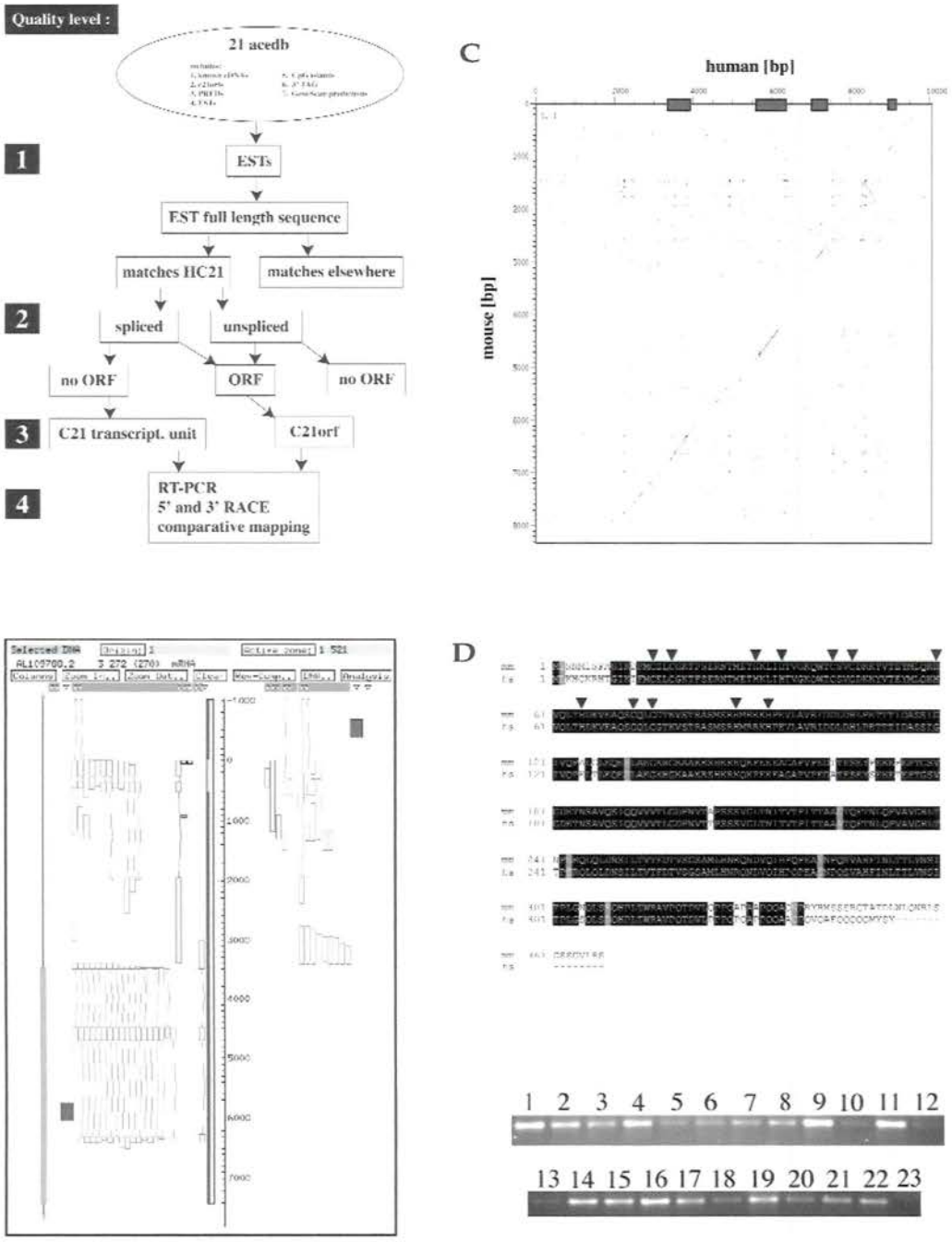
RESULTS

Selection and Evaluation of Candidate Genes

To refine the HC21 sequence annotation, we re-analyzed the entire chromosome sequence incorporating new ESTs and created a reliable algorithm to stringently extract and map 3′ polyadenylation tags from EST traces. The combined data are organized as an Acedb database, called 21ace. (A copy of the 21ace database can be downloaded via anonymous ftp from ftp.licr.org/pub. Xace (Unix) and WinAce (PC) programs are available from http://www.acedb.org/Software/Downloads/.) The 21ace graphical display allowed us to rapidly exclude genes and gene models already described [16,20,22] and their corresponding ESTs. Remaining ESTs and/or EST clusters were analyzed for the presence of a possible open reading frame following the construction of a contig for each cluster to minimize the effect of sequencing errors. The spliced ESTs and/or EST clusters with a possible open reading frame and at least one of the following were considered as candidate genes (quality level 1; Fig. 1A): a 3′ tag; the presence of a CpG island; or a GeneScan prediction. Unspliced ESTs and/or EST clusters with a possible open reading frame were selected only if at least two of the above criteria were fulfilled. An example of such a candidate gene is shown in Fig. 1B, as visualized in 21ace. These candidate genes were subsequently shown to be bona fide genes and were termed *C21orf81* and *C21orf83*. We are aware of the fact that this conservative approach probably excluded some candidates that may represent new HC21 transcripts.

To validate selected transcripts, we followed the strategy outlined in Fig. 1A. We first sequenced the full length of the cDNAs from which the corresponding ESTs were derived. We retained the sequences that matched to HC21 (quality level 2 gene candidates), while the remainder was discarded. Two subcategories were then created: spliced and unspliced cDNAs. cDNAs colinear with the genomic sequence and with open reading frame ≥ 300 bp and spliced cDNAs with open reading frame ≥ 100 bp were retained and elevated to quality level 3 gene candidates. To confirm level 3 candidates as bona fide transcripts, we carried out RT-PCR on a panel of 22 normalized cDNA pools. We found evidence of expression for all of the level 3 candidates, except *C21orf87*, in at least one tested tissue allowing re-classification to quality level 4 gene candidates ("confirmed genes"). Most of these transcripts were only detected following nested PCR reactions. Some transcripts are expressed in almost all tissues analyzed (for example, *C21orf83*, *C21orf93*, and *C21orf101*), whereas others (*C21orf86*) showed a more restricted expression pattern (Table 2). We completed some sequences by 5′- or 3′-RACE as required, providing additional evidence that the quality level 4 candidate genes are genuine. The identification of 5′-RACE *C21orf87* amplimers in lung and placenta provided the evidence that allowed this transcript to be upgraded to quality level 4.

Some quality level 2 candidates that map to HC21, are spliced, and show specific expression patterns have no obvi-



G. 1. *C21orf83* is a candidate gene upgraded to a bona fide gene. (A) Schematic outline of the strategy used to confirm the candidate genes. The ality level is shown on the left. (B) The 21ace output for the *C21orf83* region (light blue, known gene cDNA; dark blue, cDNA in database; red, Ts; green, Orestes ESTs; brown, Genescan prediction; salmon, Celera Genomics prediction; black bar, 3′ tags; framed burgundy box, repeat). (C) *1orf83* human (x axis) and mouse (y axis) comparative mapping with the dotter program. Boxes define the positions of the human *C21orf83* exons.) Alignment of the human (bottom) and mouse (top) C21orf83 putative peptides. Conserved and similar amino acids are boxed. Key residues of e three C_2H_2-type zinc fingers are indicated with arrowheads. (E) C21orf83 expression pattern (1, testis; 2, lung; 3, prostate; 4, small intestine; 5, east; 6, brain; 7, heart; 8, uterus; 9, bone marrow; 10, placenta; 11, colon; 12, fetal brain; 13, liver; 14, fetal liver; 15, thymus; 16, salivary gland; 17, inal cord; 18, kidney; 19, spleen; 20, skeletal muscle; 21, trachea; 22, adrenal gland; 23, no DNA).

doi:10.1006/geno.2002.6781, available online at http://www.idealibrary.com on IDEAL

Gene[a]	*Hs* accession[b]	5′ STOP in frame[c]	*Hs* isoforms[d]	*Hs* size [a.a.][e]	*Mm* accession[b]	*Mm* size [a.a.][e]	Identity [%][f]	Dotter hit[g]	Pfam/Inter-Pro/Prosite[h]	genomic seq.[i]	from[j]	to[j]	exons [n][k]
C21orf65	AF426256 AF321193	+		91				yes		AP001730	323951		4
									AP001731		19010		
C21orf81	AF426257	+		89				no		AL163204	338454	312634	6
C21orf82	AF426258	+		64				yes		AP001719	111744	113842	2
C21orf83[l]	AY063456 AF426259	+[m]	+	353/298	AY063457	368	88%	yes	PF00096 PS00028 zinc finger	AP002955 or gap and AP001743	41759[q]	28441[q]	8
	AF426260 AL109788	+	+	178									
C21orf84	AF426261	+		77				in seq. gap[n]		AP001751	210752	204218	4
C21orf85	AF426262	+		96				yes		AP001759	259994	260284	2
C21orf86	AF426264	+		165				no		AL163301	332374	331877	2
C21orf87	AF426265	+		145				no		AL163279	158702	158265	1
C21orf88	AF426266-267	+	+	145/64				no		AL163280	115261	100600	3
C21orf89	AF426268	+		33				no		AL163301	296287	296388	3
C21orf90	AF426269-270	+	+	65/37				no		AP001754	233415	234330	3
C21orf93	AF427488	+		139				no		AL163302 AL163301	3247[o]	1856[o]	3
C21orf94	AF427489	+		62				no		AL163246	56932	57120	2
C21orf95	AY061853 AF401639	+[m]		154	AY061854	165	84%	yes PS50099	IPR000694	AP001696	302124	197685	3
C21orf99	AF427490	+		68				in seq. gap[n]		AL163202	79429	86095	4
C21orf100	AY063458 AY063459	+		55				no		AL163247	234589	234756	2
C21orf101	AY061855 AB049942	-		125	AY061856	125	86%	yes	IPR000529 PS01048	AP001719	4650	73549	3
C21orf102	AY061857 AB058646	+[m]		257	AY061858	257	80%	yes	IPR000372 IPR001611	AP001754	172266	173039	1
MCM3APAS[l]	AF426263 AK001370	-		≥123				no		AP001759	248088	267331	4
	+		122							267377	267745		
D21S2088E	AY063451							not done		AP001687	164890	141161	5
D21S2089E	AY063452 AY063453		+					yes		AP001671	83252	68149	3
D21S2090E	AY063454							no		AL163252	88742	96857	2
D21S2091E	AY063455							no		AL163252	210808	208177	3

[a]Nomenclature committee/GDB approved gene symbol.
[b]*Homo sapiens* and *Mus musculus* GenBank accession numbers.
[c]Presence (+) or absence (-) of a 5′ STOP codon in-frame.
[d]The (+) sign specifies the identification of multiple *H. sapiens* isoforms.
[e]Size of the encoded peptides in number of amino acid residues.
[f]Percentage of identity at the amino acid level between *H. sapiens* and *M. musculus* encoded peptide.
[g]Dotter program comparative mapping was found (yes) or was absent (no) between human and mouse.
[h]Identified domains.
[i]*H. sapiens* genomic sequence GenBank accession number.
[j]C21orfs open reading frame and transcription units start and end positions.
[k]Number of exons.
[l]Potentially bicistronic gene.
[m]No 5′ STOP codon in-frame in *H. sapiens*, but a 5′ STOP codon in frame was identified in *M. musculus*.
[n]Murine systemic genomic sequence probably in sequencing gap.
[o]Base pair numbering from AL163302.
[p]Murine EST. **[AUTHORS: WHERE IS THIS ON THE TABLE?]**
[q]Base pair numbering from AP002955.

s open reading frame (ORF lessthan100 bp). We termed ıse transcripts "chromosome 21 transcription units" (Table *D21S2088E* to *D21S2091E*). Similar to the C21orfs, the ɔression profiles of these units were investigated by nested R (Table 2).

Interspecies comparative mapping and sequencing have ɐn useful to confirm and/or identify genes. The availability the mouse genome sequence (http://www.celera.com/) ɔwed screening of the mouse regions syntenic to HC21 (on ɔuse chromosomes 16, 17, and 10) for the presence of ɹuences homologous to the newly identified genes. Regions mologous to *C21orf65*, *C21orf83*, *C21orf95*, and *C21orf101* ɹre readily identified by the BLASTn option at Celera ɲomics. These sequences and related mouse ESTs allowed to reconstruct the cDNA sequences of the mouse orthologs the *C21orf83*, *C21orf95*, and *C21orf101* genes (Table 1, acc. s. AY06345, AY061854, and AY061856, and Fig. 1D). We reaɲed that the algorithm used in the basic local alignment ɹrch tool (BLAST) sequence comparison might not recognize ɔrt homologies hidden in long DNA sequences. Thus, we ɔ undertook comparative mapping by aligning the correɔnding human and mouse genomic sequences, whereby man genomic sequences between two known genes (cenɹmeric and telomeric "anchors") and containing one of the wly identified HC21 genes were compared with the correɔnding mouse genomic sequence delimited by the ortholoɹus "anchors" using a dot-plot program [26]. This approach owed us to identify sequences similar to *C21orf65*, *C21orf82*, *1orf83*, *C21orf85*, *C21orf95*, and *C21orf101* in the mouse synɹic region, demonstrating that these genes were conserved ɔm rodents to primates (Table 1). The D21S2089E transcripɲn unit has also been partially conserved through mammalian olution (Table 1). An example of the different steps leading the identification of a bona fide gene is shown in Fig. 1. iefly, *C21orf83* was identified as a candidate gene using the ace (Fig. 1B); the comparative mapping showed that this gene ɹd the corresponding peptides are conserved in the mouse igs. 1C and 1D) and RT-PCR, showed that its expression is ɔiquitous (Fig. 1E).

The reliability of our approach to identify new HC21 ɹnes is illustrated further by the fact that during the prepation of this manuscript, 3 of 19 of our quality level 4 tranɹipts were reported in the sequence databases by other vestigators (Table 1). *C21orf65* was identified by the Construction of Transcripts Map from Chromosome q22.2: Down Syndrome Critical Region" effort (GenBank c. no. AF321193) and named *DSCR8* (Down syndrome critɹl region gene 8). Likewise, *C21orf95* and *C21orf101* were ɔned (GenBank acc. nos. AF401639 and AB049942; unpubɹhed data) and called *CYYR1* (cysteine and tyrosine-rich ɹotein 1) and *MRPS6* (mRNA for mitochondrial ribosomal ɹotein S6), respectively [27] (Table 1). We also serendipiɹusly identified a contig of mouse ESTs mapping to a region mouse chromosome 10 syntenic to HC 21. The putative ɔen reading frame present on the resulting consensus sequence (GenBank acc. no. AY061858) is conserved across species allowing reconstruction of the cDNA sequence of the human gene (*C21orf102*, GenBank acc. no. AY061857).

Characteristics of the Novel HC21 Genes

A description of the novel HC21 genes and their RNA/protein products is presented in Table 1. These transcripts were designated *C21orf65*, *C21orf81* to *C21orf90*, *C21orf93* to *C21orf95*, *C21orf99* to *C21orf102* (for human chromosome 21 open reading frame), and *MCM3APAS* (MCM3-associated protein antisense). Most of the 19 putative peptides show no significant similarity to any characterized proteins, with the following four exceptions: the C21orf83 protein contains three zinc fingers of the C_2H_2 type (Pfam domain PF00096; Fig. 1D); the C21orf102 protein is a member of the ribosomal S6 family (InterPro domain IPR000529); the C21orf95 protein harbors a proline-rich domain (Prosite signature PS50099) at its carboxy terminus; and the C21orf102 protein contains a set of leucine-rich stretches (InterPro domains IPR000372 and IPR001611). *C21orf83* and *MCM3APAS* each contain two equally likely open reading frames and hence are potentially bicistronic, encoding putative peptides of 353 and 178, and ≥ 123 and 122 residues, respectively. The newly identified genes encode short transcripts (mean = 1220 bp, n = 19) divided into one to eight exons (mean = 3.2 exons, n = 19, Refseq/SwissProt/TrEMBL data set has a mean of 8.8 exons [18]). They encode putative proteins of 33–353 residues, with a mean size of 121 residues (n = 19). They are expressed at a relatively low level as they are represented by an average of 11.5 ESTs in GenBank (n = 18, SD = 11.7, mouse ESTs for *C21orf102*), if we exclude *C21orf101* (represented by 101 ESTs). In contrast the previously described HC21 genes are represented by an average of 75 ESTs (n = 144, SD = 45.6).

Discussion

HC21 has been extensively annotated in the course of the search for candidate genes involved in DS and the many monogenic disorders that have been linked to this chromosome [16,20,22,28–36]. Nevertheless, we suspected that the high-throughput gene identification strategy pursued for chromosome 22, chromosome 21, and later for the rest of the genome may be biased against poorly expressed genes and/or genes with small ORFs [14,16,18,19]. To improve gene annotation in HC21 and to evaluate the sensitivity of currently used methods for gene prediction, we first developed an HC21 Acedb database, 21ace, containing all available ESTs that map to this chromosome, as well as all known transcripts, predicted transcripts, and pseudogenes described so far [16,20–22]. In addition we incorporated a series of tools to facilitate gene discovery such as CpG islands, Genescan predictions, repeats, and pfam analysis. We developed a novel algorithm (3′ tag mapping) to identify potential 3′ ends of transcripts as defined by two "biological signatures" of

doi:10.1006/geno.2002.6781, available online at http://www.idealibrary.com on IDEAL

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

[a]The white area specifies untested tissues.
[b]5′-RACE data.

nes: poly(A) stretches and poly(A) signals, in the context of derived ESTs. This tool allowed us to visualize on the nome experimentally verified polyadenylation sites, and ıs to precisely map the 3′ ends of genes.

:w Genes on HC21

e re-analysis of the HC21 coupled with experimental doc- ıentation identified 19 new transcripts and four HC21 tran- ·iption units. The C21orfs are novel genes encoding puta- e proteins, whereas C21 transcription units represent liced transcripts with no open reading frame. We were able to identify mouse orthologs of some of the novel 1orf genes during our comparative mapping. This could be e to some of the orthologs being in sequencing gaps; some these genes encoding noncoding RNAs; or some of these nes not being present in rodents, a finding which has been ·eviously reported for *C21orf21* and *C21orf22* [21].

ıplication for the Rest of the Genome

precise and complete analysis of constituent genes and cod- ; regions has proven difficult in complex eukaryotic nomes using only computational analysis. While earlier imates of gene number based on EST clustering predicted tween 45,000 and 140,000 genes [14,37,38], the initial quence of the human genome predicted around 35,000 nes [18,19,39,40]. A more recent comparison of the genes edicted by the International Human Genome Sequencing nsortium (Ensembl) and Celera Genomics has shown that th predicted sets were of comparable quality. However, e two gene sets were largely non-overlapping, allowing the thors to conclude that the methods used for gene predic- n by either group are individually incomplete [41]. nilarly, prediction methods used to analyze the *Drosophila* nscriptome have shown to be too conservative, as a novel notation approach has increased the number of predicted nscripts by 7.7% [11,42]. In the absence of more sensitive mputational approaches, gene identification will depend the alignment of finished genomic sequence with quences from experimentally validated transcripts.

The first annotation of the sequence of 21q confirmed 127 nes, and predicted an additional 98 anonymous gene mod- ; based on exon prediction programs and/or on matching liced ESTs [16]. So far the number of confirmed genes has creased to 163, while there remain 63 predicted genes, yield- g a total of 226 genes [16,20–22,28]. This number contains e entire set of PREDs and C21orfs from the chromosome 21 apping and Sequencing Consortium minus a small number gene models, which were upgraded to confirmed genes or ere eliminated as false-positive predictions [16,20–22,28]. If e subtract from the above total the 27 PRED gene models lely predicted *in silico* (subcategory 4.2 in [16]), which have) related spliced ESTs and/or are not conserved in the ouse, the total number is only 199 genes. The identification 19 novel HC21 genes described here brings this total to 8, an increase of 9.5%. These 19 new genes were not pre- cted by the Chromosome 21 Mapping and Sequencing Consortium or by the Ensembl annotation effort, or described in subsequent analyses (first selection step in our novel gene selection procedure) [16,19,20,22,28]. Celera Genomics predicted only three of these novel genes, *C21orf81*, *C21orf83*, and *C21orf102*. For example, *C21orf83*, encoding a member of the zinc finger family (which partially maps in a sequencing gap that was filled by Celera Genomics), was not detected by the public consortium [19]. The HC21 data we present suggest that the current gene predictions contain many false negatives and numerous false positives as previously shown by us and others [20,22]. As the newly identified genes reported here encode putative proteins with a mean size of 121 residues, clearly below the 575-residue average length of the HC21 described genes or the 469-residue average length of the Refseq/SwissProt/TrEMBL data set [18], our data also suggest that the prediction algorithms are biased against genes with small ORFs. Despite our careful validation of the HC21 annotations our actual total of 218 genes (this manuscript and [20]) remains lower than the 449 gene models described by Celera Genomics for the same chromosome [19]. The large proportion of pseudogenes in their prediction, as well as in the Ensembl predictions, might explain this difference (A.R., S.D., and S.E.A., unpublished data). Our data suggest therefore that Hogenesch *et al.* largely overestimated the total number of human genes by using these databases [41]. If the number of HC21 genes ranges between 218 and 250, we could predict that the total number of human genes ranges between 21,500 and 24,500 by extrapolating from the proportion of unbiased mapping of 305 of 30,181 human ESTs to this chromosome [43].

We have identified 19 novel transcripts on HC21, although this chromosome was relatively well annotated (for example, many pseudogenes were correctly annotated) [16,28,44]. This shows that the identification of the complete human transcript map cannot rely on gene prediction alone. The definitive gene annotation of the human genome will require a gene-by-gene approach coupled with experimental verification. This endeavor would greatly benefit of gene annotation algorithms showing unidirectional error, thus generating only false positives or only false negatives.

Materials and Methods

The HC21 Acedb database. The genomic sequence data were extracted from the EMBL database release 66. Transcribed sequence data were extracted from several sources: human EST section of EMBL release 66; human mRNA documented in the human section of EMBL release 66; ORESTES sequences from the LICR/FAPESP Human Cancer Genome project; human mRNAs documented in the NCBI curated RefSeq database; and published HC21 genes, C21orfs, and HC21 PREDs. The majority of ORESTES sequences are also found in the EMBL EST databank, but we classified them differently in the mapping output to indicate that they are not derived from oligo-dT primed cDNA. The transcript sequences were filtered for contaminants, and repetitive elements were masked out using the PFP software package (Paracel, Pasadena, CA). We used Megablast (from the NCBI BLAST package [45]) to identify pairwise similarities between all transcribed sequences and the genomic data. For each pair of matching transcribed and genomic sequences, local alignments were gener-

g sim4 [46], with parameters W = 15, R = 0, A = 4, P = 1. We filtered output to eliminate all alignments that did not contain at least one atching with at least 95% identity over 30 nucleotides. The 3′ tags erated from about 2,400,000 sequence trace files, by extracting the 50 es that lay directly before the longest poly(A) or poly(T) in the trace ract_seq from the Staden package. A minimal length of 10 A's (T's) rced. The Repsim algorithm was used to eliminate tags containing p of repetitive sequence. CpG islands were located with a CpG island-rofile (available on request) and the program pfsearch from the pftools ftp://ftp.isrec.isb-sib.ch/pub/sib-isrec/pftools/), implementing the match search method for generalized profiles [47]. The in-built graph-ays of Acedb were used to create an integrated view of the HC21 data ple parsing scripts were used to convert the filtered sim4 output, 3′ nation, RepeatMasker output (A. F. Smit, Institute for Systems Biology, hed data), CpG island predictions, GenScan predictions [49], and sequence data into ace format, suitable for uploading to Acedb. g a method (determining position and display color) controlled each rrangement within Acedb, with minor modifications to the sequence lel. The original data are visible within Acedb via the LongText func-

oning. The cDNA clones used in this study were obtained from Genetics (http://www.resgen.com) unless otherwise indicated, and uenced directly using an ABI377 or an ABI PRISM3100 (Applied ns). The ORF sequences were determined using the following tem-*21orf65*, IMAGE clone 781289; *C21orf81*, IMAGE clones 2304590, and 5289033 and subsequent 5′-RACE products derived from lung, l heart cDNAs; *C21orf82*, IMAGE clones 429071 and 2097381; *C21orf83*, lones 814590, 1853490, 1467262, and 3477187 and subsequent RACE from testis cDNA; *C21orf84*, IMAGE clones 2723484 and 2723574 and nt RACE products from placenta cDNA; *C21orf85*, IMAGE clone *C21orf86*, IMAGE clone 1756203; *C21orf87*, IMAGE clones 705099 and *C21orf88*, IMAGE clones 526903 and 2097151; *C21orf89*, IMAGE clone *C21orf90*, IMAGE clone 430058; *C21orf93*, IMAGE clones 1756203 and *C21orf94*, IMAGE clones 2464987 and 1170079; *C21orf95*, IMAGE 50623, 4106483, 162308, and 2338862; *C21orf99*, IMAGE clones 1461135 444; *C21orf100*, IMAGE clone 3289153; *C21orf101*, IMAGE clones 1744284, and 4764462; *C21orf102*, mouse H3100H09 clone from the Institute of Aging (NIA) 15K Mouse cDNA Clone Set gsun.grc.nia.nih.gov/cDNA/15k.html); D21S2088E (GDB:11500011), lones 307666 and 773409; D21S2089E (GDB:11500012), IMAGE clones 2910016, and 1755433; D21S2090E (GDB:11500013), IMAGE clone D21S2091E (GDB:11500020), IMAGE clone 1468536.

Ve carried out both 5′- and 3′-RACE on various poly(A)$^+$ RNAs using thon cDNA Amplification Kit (Clontech). Double-strand cDNA syn-d adaptor ligations to the synthesized cDNAs were carried out accord-e manufacturer's instructions. PCR products were cloned using the A kit (Invitrogen) and sequenced. The sequence and amplification con-sed are available on request.

structure. The intron–exon junctions of the newly identified tran-ere determined by comparison of the genomic sequence to the cDNA using the est_genome software, available through the UK Human Mapping Project (http://www.hgmp.mrc.ac.uk).

on pattern studies. We tested the expression of the new HC21 tran-RT-PCR. Total RNA derived from 22 different normal human tissues ng, prostate, small intestine, breast, brain, heart, uterus, bone marrow, colon, fetal brain, liver, fetal liver, thymus, salivary gland, spinal cord, spleen, skeletal muscle, trachea, and adrenal gland; Clontech, Palo) was used for cDNA synthesis. The quality of total RNA was tested using MLH1 primers located at intronic sequences flanking exon 12 , 5′-TGGTGTCTCTAGTTCTGG-3′, and reverse, 5′-CATTGTTGTAG-TGC-3′), as an indicator of possible genomic DNA contamination. transcription was carried out using the Superscript First Strand s Kit according to the manufacturer's recommendations (Life gies). A nested PCR approach was adopted because the expression he new transcripts was expected to be low. Primers for RT-PCR were l from the sequence of distinct exons so that the possible amplification nic DNA could be distinguished from cDNA amplification. RT-PCR and nested PCR conditions, primer sequences, and the expected size of PCR fragments are available on request.

ACKNOWLEDGMENTS

We thank A. Bairoch, L. Cimasoni, M. Dermitzakis, A. Estreicher, M. Friedli, O. Menzel, L. Rougemont, and M. Wattenhofer for suggestions and/or critical reading of the manuscript; and M.-P. Papasavvas and N. Scamuffa for core assistance ***[AUTHORS: PLEASE PROVIDE FULL FIRST NAMES AND AFFILIATIONS FOR ALL].*** *We gratefully acknowledge the support of the Genomics Program of the Ludwig Institute for Cancer Research. This work was supported by grants from the Jérôme Lejeune Foundation to R.L., A.R., and M.G.; from the Swiss FNRS 31.57149.99, the Swiss FNRS NPR38, and the European Union/OFES and ChildCare foundation to S.E.A.; from the Ludwig Institute for Cancer Research and from the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) to A.A.C., S.D.S., and A.J.G.S.*

RECEIVED FOR PUBLICATION FEBRUARY 20;
ACCEPTED MARCH 27, 2002.

REFERENCES

1. Epstein, C. J. (1995). Down syndrome (trisomy 21). In *The Metabolic and Molecular Bases of Inherited Disease* (C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle, Eds.), pp. 749–794. McGraw Hill, New York.
2. Delabar, J. M., *et al.* (1993). Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.* **1:** 114–124.
3. Chaib, H., *et al.* (1997). A newly identified locus for Usher syndrome type I, USH1E, maps to chromosome 21q21. *Hum. Mol. Genet.* **6:** 27–31.
4. Muenke, M., *et al.* (1995). Physical mapping of the holoprosencephaly critical region in 21q22.3, exclusion of SIM2 as a candidate gene for holoprosencephaly, and mapping of SIM2 to a region of chromosome 21 important for Down syndrome. *Am. J. Hum. Genet.* **57:** 1074–1079.
5. Pajukanta, P., *et al.* (1999). Genomewide scan for familial combined hyperlipidemia genes in finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels. *Am. J. Hum. Genet.* **64:** 1453–1463.
6. Straub, R. E., *et al.* (1994). A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. *Nat. Genet.* **8:** 291–296.
7. Groet, J., *et al.* (2000). Narrowing of the region of allelic loss in 21q11-21 in squamous non-small cell lung carcinoma and cloning of a novel ubiquitin-specific protease gene from the deleted segment. *Genes Chromosomes Cancer* **27:** 153–161.
8. Iselius, L., Jacobs, P., and Morton, N. (1990). Leukaemia and transient leukaemia in Down syndrome. *Hum. Genet.* **85:** 477–485.
9. Zipursky, A., Brown, E. J., Christensen, H., and Doyle, J. (1999). Transient myeloproliferative disorder (transient leukemia) and hematologic manifestations of Down syndrome. *Clin. Lab. Med.* **19:** 157–167.
10. Hasle, H., Clemmensen, I. H., and Mikkelsen, M. (2000). Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet* **355:** 165–169.
11. Adams, M. D., *et al.* (2000). The genome sequence of Drosophila melanogaster. *Science* **287:** 2185–2195.
12. Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST—database for "expressed sequence tags". *Nat. Genet.* **4:** 332–333.
13. The C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282:** 2012–2018.
14. Dunham, I., *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.
15. Goffeau, A., *et al.* (1996). Life with 6000 genes. *Science* **274:** 563–567.
16. Hattori, M., *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.
17. Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.
18. Lander, E. S., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
19. Venter, J. C., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304–1351.
20. Reymond, A., *et al.* (2001). From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. *Genomics* **78:** 46–54.
21. Davisson, M. T., *et al.* (2001). Evolutionary breakpoints on human chromosome 21. *Genomics* **78:** 99–106.
22. Pletcher, M. T., Wiltshire, T., Cabin, D. E., Villanueva, M., and Reeves, R. H. (2001). Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics* **74:** 45–54.
23. Camargo, A. A., *et al.* (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. USA* **98:** 12103–12108.
24. Dias Neto, E., *et al.* (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **97:** 3491–3496.
25. Kawai, J., *et al.* (2001). Functional annotation of a full-length mouse cDNA collection.

lature **409:** 685–690.
'onnhammer, E. L., and Durbin, R. (1995). A dot-matrix program with dynamic threshld control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–10.
'uzuki, T., *et al.* (2001). Proteomic analysis of the mammalian mitochondrial ribosome. dentification of protein components in the 28 S small subunit. *J. Biol. Chem.* **276:** 3181–33195.
\ntonarakis, S. E. (2001). Chromosome 21: from sequence to applications. *Curr. Opin. Genet. Dev.* **11:** 241–246.
Chen, H., *et al.* (1996). Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. *Genome Res.* **6:** 747–760.
The Finnish-German APECED Consortium. (1997). An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy. *Nat. Genet.* **17:** 399–403.
Lieman-Hurwitz, J., Dafni, N., Lavie, V., and Groner, Y. (1982). Human cytoplasmic superoxide dismutase cDNA clone: a probe for studying the molecular biology of Down syndrome. *Proc. Natl. Acad. Sci. USA* **79:** 2808–2811.
Nagamine, K., *et al.* (1997). Positional cloning of the APECED gene. *Nat. Genet.* **17:** 393–398.
Pennacchio, L. A., *et al.* (1996). Mutations in the gene encoding cystatin B in progressive myoclonus epilepsy (EPM1). *Science* **271:** 1731–1734.
Scott, H. S., *et al.* (2001). Insertion of β-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. *Nat. Genet.* **27:** 59–63.
Sertie, A. L., *et al.* (2000). Collagen XVIII, containing an endogenous inhibitor of angiogenesis and tumor growth, plays a critical role in the maintenance of retinal structure and in neural tube closure (Knobloch syndrome). *Hum. Mol. Genet.* **9:** 2051–2058.
36. Song, W. J., *et al.* (1999). Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat. Genet.* **23:** 166–175.
37. Fields, C., Adams, M. D., White, O., and Venter, J. C. (1994). How many genes in the human genome? *Nat. Genet.* **7:** 345–346.
38. Liang, F., *et al.* (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25:** 239–240.
39. Ewing, B., and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.
40. Roest Crollius, H., *et al.* (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25:** 235–238.
41. Hogenesch, J. B., *et al.* (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.
42. Gopal, S., *et al.* (2001). Homology-based annotation yields 1,042 new candidate genes in the Drosophila melanogaster genome. *Nat. Genet.* **27:** 337–340.
43. Deloukas, P., *et al.* (1998). A physical map of 30,000 human genes. *Science* **282:** 744–746.
44. Gardiner, K., and Davisson, M. T. (2000). The sequence of human chromosome 21 and implications for research into Down syndrome. *Genome Biol.* **1:** 1–9.
45. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.
46. 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.
47. Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20:** 3–23.
48. Kelley, S. (2000). Getting started with Acedb. *Brief Bioinform.* **1:** 131–137.
49. Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Physiol Genomics 21: 000–000, 2005.
First published March 22, 2005; doi:10.1152/physiolgenomics.00237.2004.

ıtification of human exons overexpressed in tumors through the use of ome and expressed sequence data

Natanja Kirschbaum-Slager, Raphael Bessa Parmigiani, Anamaria Aranha Camargo, and Sandro José de Souza
Ludwig Institute for Cancer Research, São Paulo Branch, Sao Paulo, Brazil

Submitted 12 October 2004; accepted in final form 15 March 2005

ger-Kirschbaum, Natanja, Raphael Bessa Parmigiani, Ana- Aranha Camargo, and Sandro José de Souza. Identification ıan exons overexpressed in tumors through the use of genome pressed sequence data. *Physiol Genomics* 21: 000–000, 2005. ublished March 22, 2005; doi:10.1152/physiolgenomics.00237. —Alternative splicing is one of the major sources of the large iptional diversity found in human cells. Splicing variants have hown to be associated with features like spreading and progres- ı several human tumors. Therefore, such variants may be of mportance as both diagnostic and therapeutic tools. Here, by a set of criteria regarding the expression pattern of splicing ts and statistical analyses, we were able to screen the genome ons overexpressed in tumors of specific tissues. However, as in analyses attempting to identify tumor-associated variants, our candidates was seriously inflated with cases of genes differen- expressed in tumors. To exclude these cases and increase the ility of finding bona fide regulated splicing variants, we per- l a serial analysis of gene expression (SAGE), excluding those that were shown to be upregulated in tumors. This allowed us dict the overexpression of single exons in specific tumors. Our group of candidates includes 1,386 exons belonging to 638 Experimental validation of a few candidates in normal tissue, cell lines, and patient samples suggests that most of these lates are indeed tumor-associated exons. Further functional ication of our candidate genes shows that our final list is slightly d with cancer-related genes.

ıtive splicing; tumor; transcriptome; serial analysis of gene ssion

NATIVE SPLICING is one of the main sources of the vari- y found in the human transcriptome (3). There are four ent types of alternative splicing: exon skipping/usage, ıative usage of a donor site, alternative usage of an otor site, and intron retention (20). Several bioinformatics ses have indicated that at least one-half of all human undergo alternative splicing (7, 11, 17, 22, 23). In ily 80% of these cases, alternative splicing invokes ges in the coding region (CDS) of genes, resulting in ural changes of the respective protein product (14, 23). e biological impact of alternative splicing is perceptible, xample, in *Drosophila*, in which sex determination is ered by alternative splicing of a master gene (25). Fur- ıore, ~15% of all human genetic diseases are believed to used by mutations in the splicing acceptor/donor sites, rating changes in the splicing pattern of one or more genes, which implies that alternative splicing also plays an important role in pathogenicity (19).

An apparent link between certain cancer types and alternative splicing is being investigated (for a review, see Caballero et al., Ref. 8). Several splicing variants from different genes, including *cd44, wt1, cd79b, bin1*, and *Syk*, have been shown to be associated with different aspects of tumorigenesis (1, 10, 13, 26, 32).

The increasing amount of cDNA libraries constructed from a diversity of both tumor and normal tissues and cell lines allows several types of computational analyses. This, together with the release of the final sequence of the human genome (16, 31), permits genome-wide analyses of alternative splicing and the search for tumor-associated splicing variants. Several groups have performed such analyses and have reported the differential expression of splicing variants in tumors (15, 33–35). None of these studies, however, systematically verified the expression pattern of the prototype variant of the same candidate gene (33, 15). Hence, it cannot be ruled out that the variants selected by their analyses as being tumor specific are variants of genes that are generally overexpressed in tumors. Furthermore, none of those studies has investigated the expression of splicing variants within tumors of one specific tissue.

Here, by using strict selection and statistical criteria, we were able to screen the genome for exons overexpressed in tumors. Tumor-associated exons are those that appear preferentially in splicing isoforms found to be overexpressed in tumors. Such exons could be of major diagnostic value, allowing the early detection of tumors based on their specific expression. New epitopes encoded by tumor-associated exons may be targeted by antibodies as well. Eventually, this should permit drug design, as the protein encoded by a spliced variant may be a therapeutic target. Here, we show by experimental, statistical, and literature validation that our set of candidates is enriched with bona fide tumor-associated splicing variants.

MATERIALS AND METHODS

Alignment and clustering of all cDNA sequences: cDNA mapping and clustering. All human cDNAs available in dbEST (July 2002, Ref. 4) and mRNA sequences from known human genes from UniGene release 153 (29) were aligned to the masked human genome sequence [build 29, obtained from the National Center for Biotechnology Information (NCBI)] by use of pp-Blast (27), an implementation of MEGABLAST (37) for a parallel cluster. The parameters used in MEGABLAST were: −f T −J F −F F −W 24. The MEGABLAST output was parsed, and a MySQL database was loaded with the mapping information. Spurious hits were excluded from the mapping database by use of an additional set of alignment criteria. These include a minimum degree of identity for a cDNA/genome alignment set to 93% over at least 45% of the total expressed sequence tag (EST) length or 55% of the total length of the full-insert sequence.

AQ: 2

AQ: 18

cle published online before print. See web site for date of publication /physiolgenomics.physiology.org).
lress for reprint requests and other correspondence: S. J. de Souza, g Institute for Cancer Research, São Paulo Branch, Rua Prof. Antonio nte 109, 4 andar, São Paulo, 01509-010, SP, Brazil (e-mail: o@compbio.ludwig.org.br).

more, for sequences mapping to more than one location on the e, a score associated with a higher identity over a longer ent was assigned. Clustering of cDNA sequences was based on enomic coordinates as described by Sakabe et al. (28). Briefly, sequences shared at least partially the same gene structure, they oined into the same cluster. If no exon/intron boundary was l, a sequence had to have at least a 100-bp overlap with another ce at the genome level to be added to the respective cluster.

struction of the binary matrices. All sequences were repre- as binary matrices, and each expressed exon was represented ne) and each skipped exon by 0 (zero). Variants were defined an exon when they included two flanking exons next to an one (represented as 10+1, meaning that at least one exon is d between two flanking exons).

atistics. After a screening for variants that included exons at the osition of an exon skipping in another variant of the same , a *Z*-statistic was calculated for each exon. This way, the ility of tumor association of the exon to a specific tissue, based numbers of ESTs confirming the variant in either tumor or l tissue, was evaluated (33)

$$Z = (p_t - p_n)/\sqrt{p(1-p)(1/n_n + 1/n_t)}$$

given exon, p_t and p_n are the expression frequencies of the exon or and normal tissues, respectively, in a specific tissue (the no. or or normal ESTs containing the specific exon ÷ total no. of or normal ESTs from all libraries). To minimize sampling bias ll libraries, we only took into account libraries that had at least e of the smallest library in which a transcript containing the c exon was found in the specific tissue. The p is the geometric e frequency of the exon in tumor and normal libraries, and n_n are the numbers of ESTs in the normal and tumor libraries, tively, taken into account for each specific exon in each tissue. h tissue, *Z*-values having a $P \leq 0.05$ were considered signifi- It should therefore be noted that the statistically significant lates still have a probability $P < 0.05$ of being a false-positive late.)

ial analysis of gene expression tag assignment. A virtual serial is of gene expression (SAGE) tag is a prediction of the 10-bp ıce downstream of the 3′-most *Nla*III site of the transcript that theoretically be produced by a SAGE experiment (5). One entative full-insert mRNA was selected from those candidate rs that included at least one full-insert mRNA showing at least a poly A signal and/or a poly A tail. This full insert was then ed a virtual SAGE tag (5). The tag was assigned only to the st *NlaIII* site of the transcript. This tag was used to query all libraries of the same tissue in which we characterized the ve overexpressed exon. The frequency of each tag was counted ıor and normal libraries of the same tissue.

ain a *Z*-statistic was calculated

$$Z = (p_t - p_n)/\sqrt{p(1-p)(1/n_n + 1/n_t)}$$

given gene, p_t and p_n are the expression frequencies of the ic 3′-most SAGE tag in tumor and normal libraries, respec- , in a specific tissue (the no. of tumor or normal tags ÷ total no. nor or normal tags from all libraries in that tissue). The p is the etric average frequency of the tag in tumor and normal libraries, $_n$ and n_t are the numbers of tags in the normal and tumor libraries into account for each specific exon. *Z*-values having a $P \leq 0.05$ considered significant (It should therefore be noted that the ically significant candidates still have a probability $P < 0.05$ of a false-positive candidate.)

perimental validation. Total RNA derived from four different al human tissues (lung, prostate, breast, brain, colon) was pur- d from Clontech Laboratories and used for cDNA synthesis.

ıman tumor cell lines were obtained from the American Type re Collection (ATCC) and maintained in appropriated medium as recommended by this organization (**http://www.atcc.org**). The following human tumor cell lines were used: A172 and T98G (glioblastoma), DU145 and PC3 (prostate), MCF-7 and MDA-MB- (breast), H1155 and H358 (lung), and SW480 (colon).

Patient samples were obtained from the Hospital A. C. Camargo tumor collection and prepared by manual dissection. All patient samples were collected after explicit informed consent, and the study was approved by the Institutional Ethics Committee. AQ: 7

Total RNA was extracted from tumor cell lines and tumor/normal patient samples by a conventional CsCl-guanidine thiocyanate gradient method (9), and RNA integrity was analyzed using agarose gels. Genomic DNA contamination of the total RNA was tested with PCR, using hMLH1 primers located at intronic sequences flanking exon 12 (forward, 5′-TGG TGT CTC TAG TTC TGG-3′; reverse, 5′-CAT TGT TGT AGT AGC TCT GC-3′).

Reverse transcription was carried out using the Superscript First Strand Synthesis Kit, according to the manufacturer's instructions (Invitrogen). RT-PCR reactions were carried out in a 25-μl reaction mixture containing 1 μl of cDNA, 1× *Taq* DNA polymerase buffer, 0.1 mM dNTPs, 6 pmol of each primer (for sequences of primers, see Supplemental Material; available at the *Physiological Genomics* web site),[1] 1 mM $MgCl_2$, and 1 U *Taq* DNA polymerase (Invitrogen). Standard PCR conditions were as follows: 4 min at 94°C (initial denaturation), 35 cycles of 45 s at 94°C, 45 s at 58°C, and 1 min at 72°C, with a final extension step of 10 min at 72°C. RT-PCR products were analyzed on 8% silver-stained polyacrylamide gels and on 2% ethidium bromide-agarose gels. Sequencing reactions were carried out using DYEnamic (ET Terminator Cycle Sequencing Kit, Amersham Pharmacia) and an ABI 377 prism sequencer (Perkin Elmer), according to the supplier's recommendations. AQ: 8 Fn1

RESULTS

Transcriptome database. The database used in this work contains data obtained from alignments of all cDNA sequences to the human genome sequence (12, 28). In addition to the representation of all data concerning the alignment and clustering of the sequences, the database also contains binary matrices that were constructed for each transcript (28). In such a matrix, a transcribed exon is represented by a one (1) and an absent exon is represented by a zero (0). This approach facilitates the analysis of exon skipping/exon usage throughout the genome and the comparison of the different transcripts and exons with each other.

Our database contains 3,475,514 expressed sequences from 7,167 cDNA libraries from different tissues (see Table 1), of which 52,903 represent full-insert sequences (completely sequenced cDNA clones). Four thousand, two hundred and forty-nine (4,249) of these libraries were constructed from tumor samples and tumor cell lines, generating 1,427,390 sequences, while the remaining 2,918 libraries were constructed from normal samples, generating 2,048,124 sequences. We will refer to libraries constructed from either tumor samples or tumor cell lines as tumor libraries. Our analysis was performed on both normalized and nonnormalized libraries. T1 AQ: 9

Database validation. Our clustering strategy (28) generated 318,272 cDNA clusters, 21,306 containing at least one full-insert mRNA. Of all clusters containing at least one full-insert mRNA, 52% undergo exon skipping (12), which is in agree-

[1]The Supplemental Material for this article (Supplemental Tables S1–S6, Supplemental Figs. S1–S4, Supplemental File S1) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00237.2004/DC1**.

1. *No. of tumor and normal libraries, ESTs, and groups in the exon-skipping database*

	Libraries	Tissue Groups	ESTs
aries	7,167	60	3,475,514
libraries	4,249	42	1,427,390
libraries	2,918	53	2,048,124

types of libraries include cell lines and patient tissue. EST, expressed ce tag.

with the splicing rate reported in the literature (11, 17, Considering all clusters in the database, we found that 3 present at least one exon-skipping variant. Our analysis erformed on this latter set of clusters.

e suitability of our exon-skipping database for the current sis was manually evaluated through the analysis of 61 described in the literature to have at least 2 splicing ms. Compared with the literature, 62% of those genes l) were shown to have the same or a larger number of ts in the exon-skipping database than the number of ts published (see Supplemental Table S1). It should be that, although examples from the literature include all of alternative splicing, our database considers only exon ng/usage. This validation step confirmed that the exon- ng database sufficiently covers the repertoire of splicing ts represented in the sequence databases.

nor-specific exons. We screened our database for poten- mor-associated exons, which appear in isoforms found to clusively expressed in tumor samples and tumor cell Our clustering strategy and matrix representation al- l us to screen our database for exons that were not ssed in transcripts from any normal tissue but were ssed in transcripts from tumor libraries. For each gene, at one transcript showing exon skipping was chosen to sent the cluster; the exon-skipping event should be con- d by at least two cDNA sequences from different librar- We screened for variants that would show the expression exon at the exact position of the skipped exon in this type transcript. Exons fitting into this category had to be ed by at least two other exons; they should be represented quences derived from tumor libraries only. We increased tringency of our analysis by only selecting those exon events that were confirmed by at least two cDNA nces derived from different tumor libraries. Because of stringent criteria, we were able to identify only 11 r-specific exons from 11 different genes (see Supplemen- aterial). The variants skipping these exons were expressed veral normal tissues.

mor-associated exons expressed in specific tissues. On the of the low number of candidates identified by our first pach, we decided to investigate whether a given exon l be tumor specific when its expression pattern was ana- within one specific tissue. The search criteria for our dates and the number of clusters filtered in each step are narized in Fig. 1. A certain variant was defined as a date when it was associated with tumor samples and cell within one tissue only, although it could appear in normal les of other tissues. For this purpose, all libraries were ed into tissue groups according to their annotations. in each of the 60 selected groups, libraries were subdi-

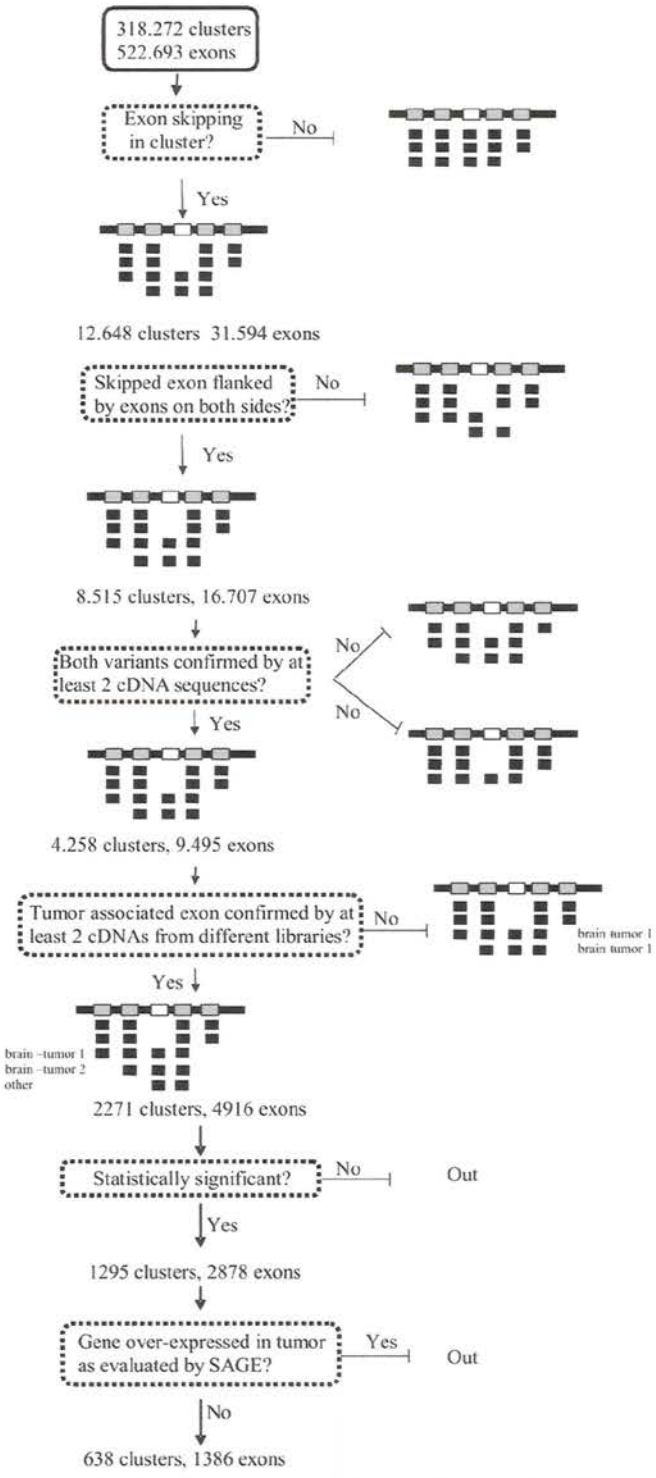


Fig. 1. Flow chart describing the approach used here to identify tumor-associated exons. Thick black lines with white boxes represent clusters; the black boxes under the clusters represent the exons present in the cDNA sequences that align to the cluster. The nos. of genes and exons obtained after each step of the screening are listed.

2. *No. of candidate exons after original screening 'a, after statistical filter, and after SAGE filter*

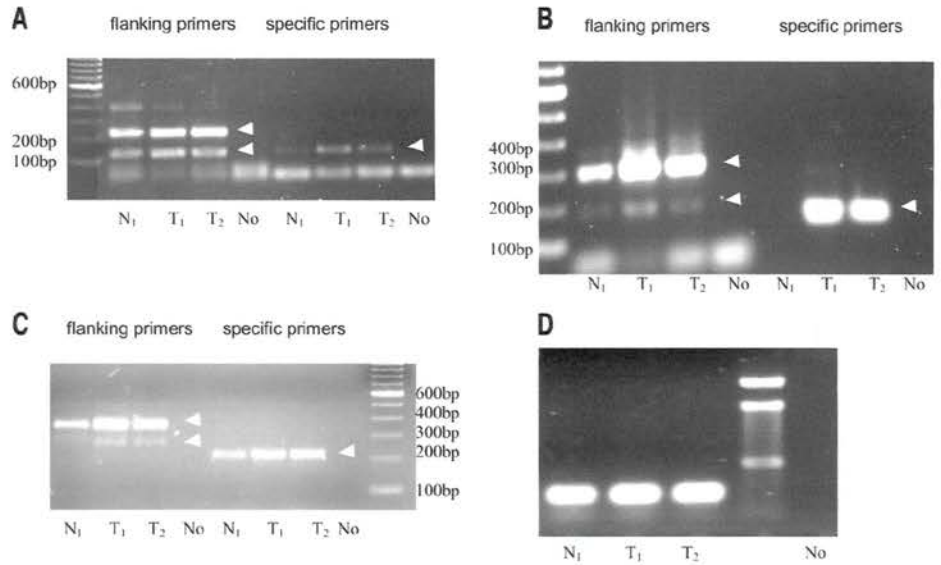
	After Initial Criteria	After Statistical Filter	After SAGE Filter
associated exons in all tissues	4,916	2,878	1,386
associated exons in brain	269	233	172
associated exons in breast	461	272	183
associated exons in prostate	203	192	138
associated exons in lung	239	193	156
associated exons in colon	847	266	235

E, serial analysis of gene expression.

into those derived from either tumor or normal tissue Supplemental Table S2). Only those 37 groups that in- d both tumor and normal libraries were used. This tissue- ic analysis increased the number of candidates to 2,271 , including 4,916 tumor-associated exons within different types (a list of all candidate genes is available; see emental Table S3). Of these genes, 2,108 contained at ne full-insert mRNA (containing 4,647 candidate exons). *tistical filter for the tumor-associated variants.* Tumor iation of each of the candidate exons was tested for its ical significance. A *Z*-score was calculated for each date exon (see MATERIALS AND METHODS) per tissue (33). statistical approach takes into account the total number of for each tissue group in either normal or tumor libraries IATERIALS AND METHODS). Of the total number of candidate (4,916), 2,878 (59%) were shown to be significantly iated with tumors ($P < 0.05$). Of all candidates poten- associated with brain tumors (269 candidate exons), 233 (87%) exons presented a significant *Z*-score ($P < 0.05$). For prostate, lung, breast, and colon, 192 of 203 (95%), 193 of 239 (81%), 272 of 461 (59%), and 266 of 847 (31%) candidate exons, respectively, were shown to be significantly associated with tumors within the respective tissue (Table 2, *column 3*, and Supplemental Table S4). T2

Experimental validation. Seven candidates were randomly selected to be screened for expression of their putative tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Five candidates were selected from brain, one from breast, and one from prostate, all of them passing the statistical test ($P < 0.05$) in the respective tissue. Three primers were designed for each candidate: two on the exons flanking the candidate tumor-associated exon (flanking primers), and one on the exon itself (specific primer). The products from the reaction using one flanking and one specific primer showed overexpression of the candidate variant in the respective tumor cell line (Fig. 2, data for 3 candidates). However, when using the two flanking primers, we observed that the variant skipping the exon was also overexpressed in the tumor cell lines (Fig. 2). This raised the possibility that our analysis was inflated with genes overexpressed in tumors instead of tumor-associated variants. F2

SAGE analysis. On the basis of the above observations, we implemented an additional filter selecting those candidate exons that did not belong to genes overexpressed in tumors. A virtual SAGE analysis was performed to verify whether the candidate genes were overexpressed in tumors from the respective tissue (5). We computationally assigned a virtual SAGE tag to one full-insert transcript of each gene (see MATERIALS AND METHODS) and statistically verified the tumor-to-normal ratio for



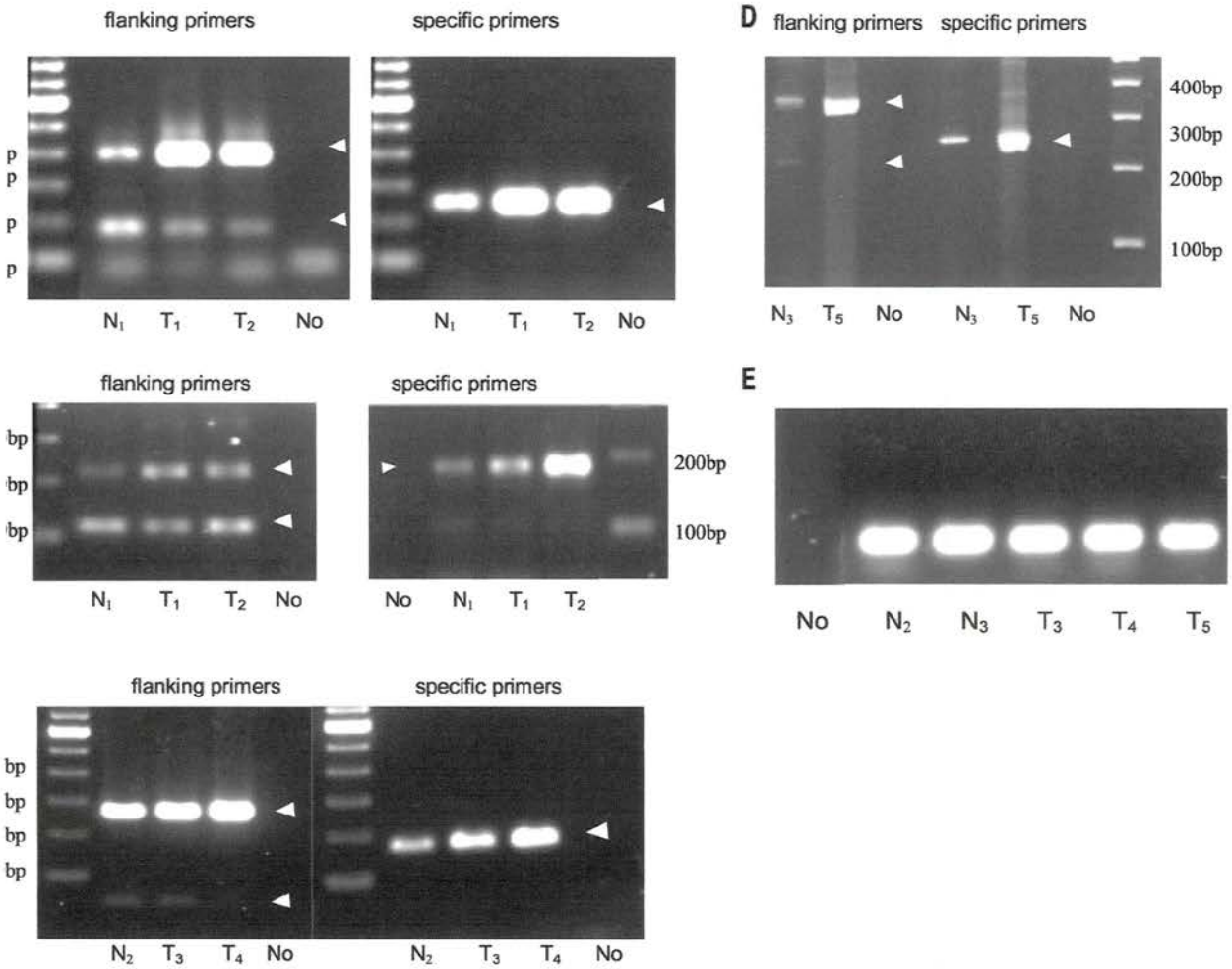
Experimental validation of 3 brain tumor-associated candidates after the statistical filter. Candidates that passed the statistical filter were randomly chosen creened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers were designed for ndidate: 2 on the exons flanking the candidate tumor-associated exon and 1 annealing to the tumor-associated exon itself. We performed 2 sets of reactions, g the flanking primers, which should amplify both variants, and 1 using 1 flanking and the specific primer, which should amplify only the variant expressing ndidate exon. N_1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, A172 glioblastoma cell line; No, "no DNA" control. Either flanking s or specific primers were used for the amplification of variants of the following genes: *THC211630* (AJ010070; *A*), *CDK-2* (NM_052827; *B*), and *calponin* 2 (AK057960; *C*). The products from the 2nd reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant respective tumor cell line. However, the variant skipping the specific exon was also overexpressed in the tumor cell lines. Amplification of *GAPDH* as ive control is shown for all samples in *D*.

spective tag. All candidate genes showing a statistically cantly higher tag count in tumors were excluded from the candidates (see MATERIALS AND METHODS).

er this additional filter, 638 candidate genes, including exons, remained in our list of candidates (Table 2, *n* 4, and Supplemental Table S5). The distribution of ripts with more than one candidate exon is shown in emental Fig. S1. The final list contained 172 candidate containing potential tumor-associated exons in brain, breast, 138 in prostate, 156 in lung, and 235 in colon. selected a few candidates for experimental validation. RNA extracted from tumor cell lines and normal tissues, served that, of the 10 candidates with conclusive results, didate genes showed that the variant containing the late exon was overexpressed in either brain, lung, or colon tumor cell lines, whereas the exon-skipping prototype was not (Fig. 3). The other six candidates showed a pattern similar to those in Fig. 2: both the exon-skipping variant and the variant including the selected tumor-associated exon were overexpressed in tumor tissue (results not shown). F3 AQ: 13

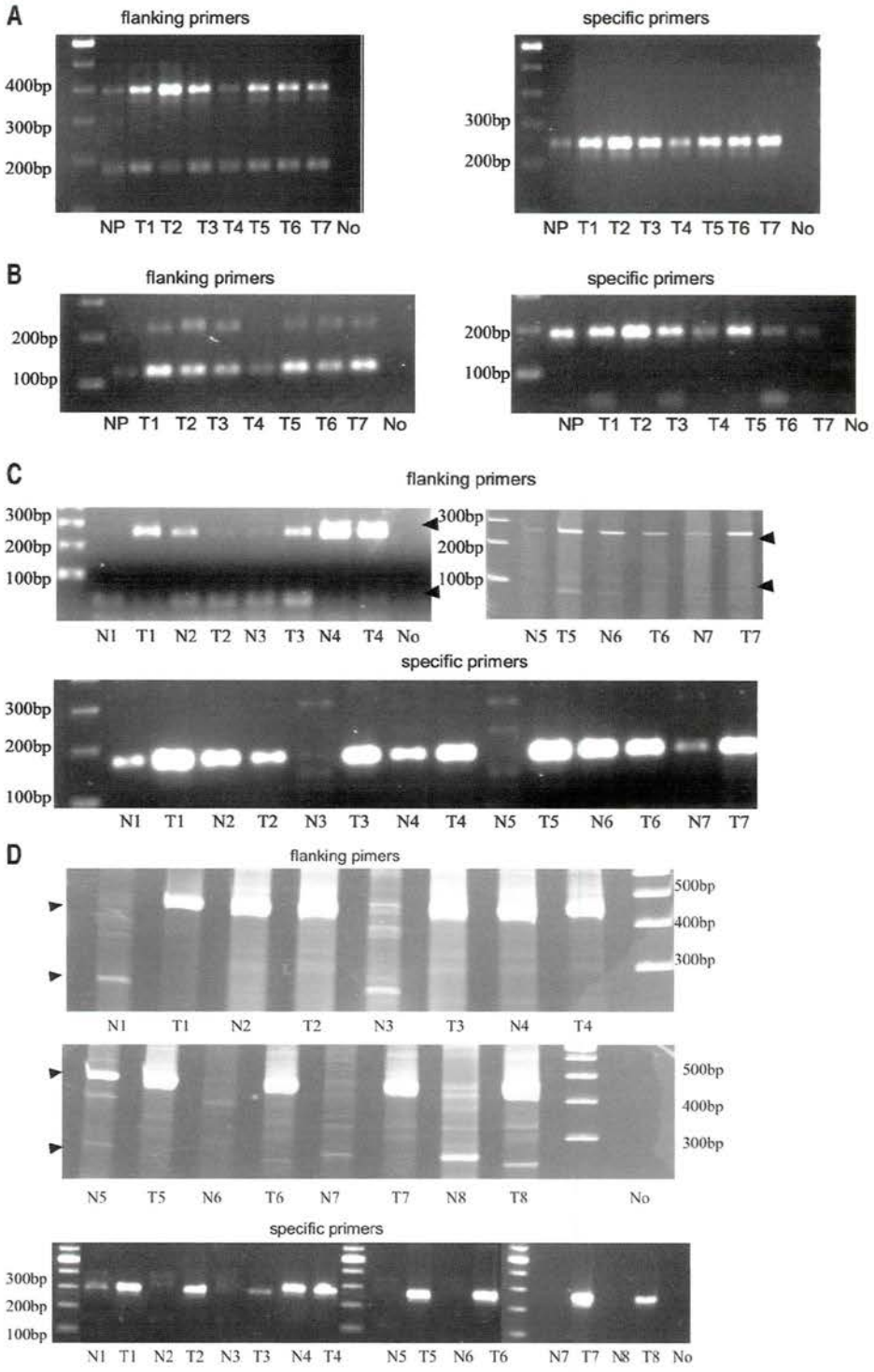
Three of the four positively validated cases in the cell lines were also validated in patient samples (Fig. 4). Interestingly, when testing two of the six cases that were negative in cell lines, both were positively validated in some of the patient samples (Fig. 5). F4 F5

Five of the six exons validated in patient samples are located inside the coding region of their respective gene. Of those five, the length of four exons is not a multiple of three and can therefore be expected to cause a change in the reading frame of its gene.



Experimental validation after the serial analysis of gene expression (SAGE) filter. As in Fig. 2, candidates that passed both the statistical filter and the filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. primers were designed for each candidate: 2 on the exons flanking the candidate tumor-associated exon and 1 on the exon itself. Candidate exons onded to the following genes: "*Delta Tubulin*" (BC000258; *A*), Zinc finger protein 585A (AK074345)-T1 (*B*), *RNA terminal phosphate cyclase-like 1* 1025; *C*), and *karyopherin* (*importin*) *beta 1* (NM_002265; *D*). N1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, A172 glioblastoma e; N_2, normal lung tissue; T_3, H1155 lung tumor cell line; T_4, H358 lung tumor cell line; N_3, normal colon tissue; T_5, SW480 colon tumor cell line; No, A control. The products from the second reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant in the tive tumor cell line. However, for the primers flanking the candidate exon, it is shown that only the variant expressing the candidate exon was overexpressed tumor cell lines. *A–C*: agarose gels. *D*: a silver-stained polyacrylamide gel (the variant skipping the candidate exon for this gene was only visualized this ue to its very low expression level). Amplification of *GAPDH* as a positive control is shown for all samples in *E*.

Experimental validation in patient amples. The 4 candidate genes that sitively validated in tumor cell lines rther validated by RT-PCR on cDNA tient tumor samples using the same and conditions as before. For brain we used 7 tissue samples from glio- a patients and compared those with rcially obtained normal pool brain Clontech Laboratories). For lung and we compared RNA from paired nor- or patient samples. For 3 of the 4 tes, we obtained the same expression as in cell lines in at least 2 patient . *A*: *Delta Tubulin* (BC000258) was d in brain samples. The products e reaction using the primers anneal- he exon itself showed overexpression andidate variant in 6 of the 7 patient . The primers flanking the candidate ow that only the variant expressing didate exon was overexpressed in atients. NB, pool of normal whole ssue; T_1–T_7, 7 different brain tumor samples; No, no DNA control. *B*: *iger protein 585A* (AK074345) was d in brain samples. The products e reaction using the primers anneal- he exon itself showed overexpression andidate variant in 4 of the 7 patient . However, the primers flanking the te exon showed that both the variant ing the candidate exon and the pro- were overexpressed in these patients. ol of normal whole brain tissue; T_1– ifferent brain tumor patient samples; DNA control. *C*: *RNA terminal phos- cyclase-like 1* (BC001025) was vali- n paired lung samples. The products e reaction using the primers anneal- he exon itself showed overexpression andidate variant in 5 of the 7 patient . The primers flanking the candidate owed that only the variant expressing didate exon was overexpressed in at of these patients (*patients 1* and *7*). and T_1–T_7, different paired normal/ lung samples, respectively; No, no control. *D*: *karyopherin* (*importin*) (NM_002265) was validated in paired amples. The products from the reac- ing the primers annealing to the exon howed overexpression of the candi- riant in 7 of 8 patient samples. The flanking the candidate exon showed ly the variant expressing the candi- on was overexpressed in at least 5 of atients (*patients 1*, *3*, *5*, *7*, and *8*). and T_1–T_8, different paired normal/ colon samples, respectively; No, no ontrol. Amplification of *GAPDH* as a e control is shown in Supplemental .



nctional classification of the final candidate list. To ana- he functional characteristics of the final list of genes, we ared our candidates to a list of 1,127 cancer-related (CR) (15a). This list was a manually curated compilation on queries of various public databases using the words er" and "tumor" (for more details, see Brentani et al., Ref. 15a). The CR genes in the list constitute 5.3% of the known genes of our transcriptome database. When analyzing our 638 candidates, we found an overlap of 60 candidates (9.4%) in the CR list. Among these genes, we found *Syk* and *bin1*, which are known to have tumor-associated variants (13, 33) (for a whole list, see Supplemental Table S6). Thus there is an excess of

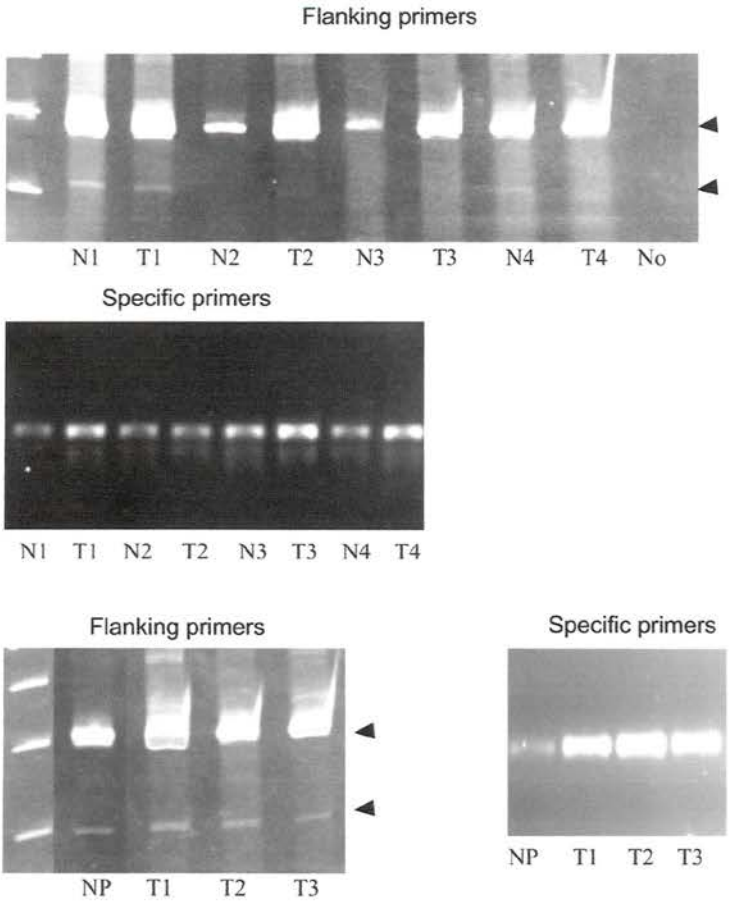


Fig. 5. We evaluated the expression pattern of 2 more candidates that were originally negative in the validation using tumor cell lines. *A*: the gene NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49-kDa (NADH-coenzyme Q reductase) (BC001456), was validated in paired lung samples. N_1–N_4 and T_1-T_4, same paired lung normal/tumor patient samples, respectively, as in Fig. 4*C*; No, no DNA control. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 of 4 paired samples. The primers flanking the candidate exon show that the variant expressing the candidate exon was overexpressed in 3 tumor samples, whereas the prototype was not overexpressed in tumors. *B*: validation of the gene "proteasome (prosome, macropain) 26S subunit, non-ATPase, 10 (NM_002814)" in prostate normal pool and 3 patient tumor samples. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 patient tumor samples. The primers flanking the candidate exon show that only the variant including the candidate exon is overexpressed in patient tumor samples. NP, normal pooled prostate cDNA; T_1–T_3, different prostate tumor patient samples; No, no DNA control.

r-related genes in our final list of candidates (chi- = 7.97, 1 degree of freedom, $P = 0.005$). Comparing results to a simulation of 200 randomly chosen sets of lusters out of all UniGene clusters, we found that none of sets presented >60 CR genes ($P < 0.005$).

Gene Ontology (GO) terms of the final list of 638 dates were obtained with the GOTM program (36). For of the 638 genes, a GO term could be assigned (see emental Fig. S3 for an overview of the distribution of the rms in the different GO categories and Supplemental Fig. r all levels of the GO tree). The program GOstat (2) was o analyze whether any GO category was overrepresented final list of candidates relative to the representation of all gy terms in the ontology database (see Supplemental File n each category, the lowest P value resulting in biolog- meaningful GO terms was used. In the category "bio- l process," using a stringent P value cutoff of 10^{-5}, we the GOs "intracellular protein transport" and "cell h and maintenance" to be significantly overrepresented. category "molecular process," the GOs "actin binding," ptor activity," "cytoskeletal protein binding," and "ATP ng" were significantly overrepresented (cutoff P value of). Finally, in the category "cellular components," the GO xisome" was significantly overrepresented (P value cut- 0.001).

erature validation. In one reported study (33) of tumor- iated splicing variants, experimental validation was performed in 76 genes chosen either by statistical criteria or by knowledge of their tumor association. All 76 candidates were experimentally validated (M. P. Lee, personal communication). To validate our candidates once more, we verified whether we could find any of our candidates in the published list of experimentally validated genes (Table 3). Thirteen of the T3 validated candidates of this reported work were found in our initial set of candidates before the SAGE analysis. In our final list of candidates, we could only find three of their candidates.

Table 3. *Overlap of our candidates with experimentally validated candidates from Wang et al. (33)*

Gene	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
NME1	+	−
CDC25C	+	−
DVL1	+	−
ERCC1	+	−
GSS	+	−
GTF3C1	+	+
IRAK1	+	−
NKTR	+	−
POLB	+	+
RAD51	+	−
SHC1	+	−
ST5	+	+
TNFRSF1A	+	−

ume comparison with a different study (35) indicated an pping of 21 candidates of 89 published genes (Table 4). stingly, we observed that 11 of the candidate genes in this eport (35) presented an overexpression in tumors as ted by SAGE. Taken together, these comparisons high- he importance of filtering off genes generally overex- d in tumors to increase the likelihood of finding bona fide -associated splicing variants.

SSION

characterization of splicing variants associated with s is critical for the development of new diagnostic and eutic strategies for the treatment of cancer. Few attempts been made to search the human genome for tumor- ated splicing isoforms (35, 33, 15). An interesting aspect se studies is the fact that none of them take into consid- n whether the differential expression was specific to the tive splicing variant or common to all transcripts from ene. Although some of these reports provided statistical ents corroborating the association between the splicing m and tumors, most of them lack experimental valida- Wang et al. (33) showed experimental validation for the *RABIA* (Fig. 2 in Wang et al., Ref. 33). There, it is le to see that, in some samples, the prototype variant is verexpressed in tumors. In our attempt to define exons pressed in tumors, we faced the same problem.

clustering of all human cDNAs onto the human genome nce allowed us to focus our strategy on determining the ssion pattern of all exons in the human genome. We were o seek for exons that were exclusively represented by ript sequences derived from tumor cDNA libraries. A computational analysis revealed that only 11 exons were to be expressed in tumors with no expression at all in l tissues. This motivated us to search for exons expressed n tumor tissues or tumor cell lines within a specific tissue. this strategy, we found 4,916 tumor-associated exons.

4. *Overlap of our candidates with those of Xu and 35), having log score >3*

e	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
	+	+
	+	−
l	+	−
	+	−
	+	+
)MB	+	+
F	+	−
K	+	−
	+	+
	+	+
	+	+
1257	+	+
4L2	+	−
l	+	+
l	+	+
	+	−
1	+	−
	+	−
	+	−
	+	+
	+	−

The expression pattern of all variants was defined based on the annotation provided with the cDNA libraries publicly available. This information, however, is not always precise and clear. Because further tissue characterization of a transcript is dependent on this information, efforts are currently being made to verify the real character of all publicly available libraries (18).

To verify whether the presence of data from normalized libraries would compromise any analyses based on the relative expression levels of transcripts, we tested whether our set of candidates was inflated with sequences derived from normalized libraries. We found a number proportional to the total number of normalized data in our initial set (30% of candidates, 30% of normalized data in the initial set). This excludes the possibility that our set of candidates is enriched with artifacts due to data from normalized libraries.

To refine our analysis of tumor-associated exons, the candidates were further screened by a statistical analysis of each exon and, for a few cases, by RT-PCR validation in tumor cell lines. Roughly 41% of our candidate exons were excluded by the statistical filter.

Nowadays, experimental analysis by RT-PCR is one of the most specific ways of verifying the expression pattern of mRNA transcripts. However, while amplifying two different variants by the use of two flanking primers, competition in transcript amplification may not reflect correctly the intrinsic difference in the expression level of the two transcripts. When using a specific primer for one exon only, however, the primers may be of such specificity that they may amplify a maximum of the transcript regardless of its expression level within the cell. More sensitive methods like real-time PCR or single molecule profiling may be used to better quantify splicing variants (30, 38).

Experimental validation also showed that the whole gene, not only the candidate exon, was overexpressed in tumor cell lines. We found support for this when we performed a SAGE analysis for all of our candidates. For this approach, we assumed that the 3′-most SAGE tag is representative of the most abundant transcript of the candidate genes. We also assumed that the prototype is more abundantly expressed than the candidate variant and that the SAGE tag count is therefore an indication of the prototype expression pattern. We found that ~52% of our candidates obtained after the statistical filter represented genes overexpressed in tumors. All those cases were excluded, and a new list of candidates was produced containing 1,386 exons. Validation with those candidates showed a success rate of 40% (4/10) when tumor cell lines were used. When we used a panel of patient samples, the success rate was much higher, ~85% (5/6). This probably occurs because of both the limitations of RT-PCR and the still-limited number of SAGE libraries available today. Furthermore, the heterogeneity of tumor samples and cell line cultures provides another variable to the whole system. Only a large-scale validation scheme will allow the definitive test of our bioinformatics pipeline. However, here we show that the combination of our computational analysis with experimental validation is successful in screening for real cancer-associated exons at a success rate that would not be achieved using either a computational analysis or experimental validation alone.

Our final list of candidate genes is enriched with cancer-related genes ($P = 0.005$). As stated before, it is likely that

ts associated with cancer are found in genes that are l to cancer. On the other hand, this does not mean that ts from genes not involved in cancer would not have a ɔnal impact on tumorigenesis (35). An ontology analysis uggested that genes involved in intracellular protein ort and cell growth and maintenance are overrepresented final list of candidates. One could expect that any change expression level of splicing variants of genes that are ɔted to cell cycle and maintenance might influence cell ɔrmation. In the category cellular components, the GO some was significantly overrepresented. Interestingly, ɔf the peroxisome proliferator-activated receptor (PPAR) es, for example, has been shown to be involved in the ɜenesis of several tumors. PPARα induces hepatocarci-; PPARγ has an anti-proliferation, pro-apoptotic effect therefore thought to have an anti-carcinogenic effect; PARβ/δ is involved in the control of cell proliferation ɔoptosis (21). Further investigation on the impact of the ɛpression of the variants of any genes involved in the pathways might give some insight on the possible ɔn of specific splicing variants.

our knowledge, this is the first report that attempts to specifically for exons overexpressed in tumors while ling genes that are generally overexpressed in the same s. Such exons may provide valuable information for investigations on the regulation of tumor-associated ative splicing. Finally, further experimental analysis will te the extent to which our candidate exons are of poten-agnostic and/or therapeutic value.

OWLEDGMENTS

thank Dr. Ricardo R. Brentani, Dr. Helena P. Brentani, Maria D. ɔvski, Noboru J. Sakabe, and Pedro A. F. Galante for helpful discus-nd/or careful reading of the manuscript.

TS

Kirschbaum-Slager is supported by PhD fellowships from CAPES (to and FAPESP (from 01/04). R. B. Parmigiani is supported by a PhD hip from CAPES. This project was supported by a CEPID Grant from P.

RENCES

udry D, Hamelin M, Cabanis MO, Fournet JC, Tournade MF, rnacki S, Junien C, and Jeanpierre C. WT1 splicing alterations in ilms' tumors. *Clin Cancer Res* 6: 3957–3965, 2000.

issbarth T and Speed TP. GOstat: find statistically overrepresented ne Ontologies within a group of genes. *Bioinformatics* 20: 1464–1465, 04.

ack DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu v Biochem* 72: 291–336, 2003.

guski MS, Lowe TMJ, and Tolstoshev CM. dbEST—database for xpressed sequence tags." *Nat Genet* 4: 332–333, 1993.

on K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak Morin PJ, Buetow KH, Strausberg RL, de Souza SJ, and Riggins J. An anatomy of normal and malignant gene expression. *Proc Natl ad Sci USA* 99: 11287–11292, 2002.

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

iballero OL, de Souza SJ, Brentani RR, and Simpson AJG. Alter-tive spliced transcripts as cancer markers. *Dis Markers* 17: 67–75, 2001.

hirgwin JM, Przybyla AE, MacDonald RJ, and Rutter WJ. Isolation biologically active ribonucleic acid from sources enriched in ribonu-ease. *Biochemistry* 18: 5294–5299, 1979.

ragg MS, Chan HTC, Fox MD, Tutt A, Smith A, Oscier DG, amblin TJ, and Glennie MJ. The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. *Blood* 100: 3068–3076, 2002.

11. **Croft L, Schandorff S, Clark F, Burrage K, Arctander P, and Mattick JS.** ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet* 24: 340–341, 2000.
12. **Galante PA, Sakabe NJ, Kirschbaum-Slager N, and de Souza SJ.** Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10: 757–765, 2004.
13. **Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, and Prendergast GC.** Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. *Proc Natl Acad Sci USA* 96: 9689–9694, 1999.
14. **Hide WA, Babenko VN, van Heusden PA, Seoighe C, and Kelso JF.** The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res* 11: 1848–1853, 2001.
15. **Hui L, Zhang X, Wu X, Lin Z, Wang Q, Li Y, and Hu G.** Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* 23: 3013–3023, 2004.

15a. **Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium; Human Cancer Genome Project Sequencing Consortium.** The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc Natl Acad Sci USA* 100: 13418–13423, 2003. AQ: 17

16. **International Human Genome Sequencing Consortium.** Initial sequencing and analysis of the human genome. *Nature* 409: 860–921, 2001.
17. **Kan Z, States D, and Gish W.** Selecting for functional alternative splices in ESTs. *Genome Res* 12: 1837–1845, 2002.
18. **Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, and Hide W.** eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 13: 1222–1230, 2003.
19. **Krawczak M, Reiss J, and Cooper DN.** The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 90: 41–54, 1992.
20. **McKeown M.** Alternative mRNA splicing. *Annu Rev Cell Biol* 8: 133–155, 1992.
21. **Michalik L, Desvergne B, and Wahli W.** Peroxisome-proliferator-activated receptors and cancers: complex stories. *Nat Rev Cancer* 4: 61–70, 2004.
22. **Mironov AA, Fickett JW, and Gelfand MS.** Frequent alternative splicing of human genes. *Genome Res* 9: 1288–1293, 1999.
23. **Modrek B, Resch A, Grasso C, and Lee C.** Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 29: 2850–2859, 2001.
24. **Modrek B and Lee C.** A genomic view of alternative splicing. *Nat Genet* 30: 13–19, 2002.
25. **Nagoshi RN, McKeown M, Burtis KC, Belote JM, and Baker BS.** The control of alternative splicing at genes regulating sexual differentiation in *D. melanogaster*. *Cell* 53: 229–236, 1988.
26. **Naor D, Sionov RV, and Ish-Shalom D.** CD44: structure, function, and association with the malignant process. *Adv Cancer Res* 71: 241–319, 1997.
27. **Osorio EC, de Souza JE, Zaiats AC, de Oliveira PS, and de Souza SJ.** pp-Blast: a "pseudo-parallel" Blast. *Braz J Med Biol Res* 36: 463–464, 2003.
28. **Sakabe NJ, de Souza JES, Galante PAF, de Oliveira PSL, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, Brentani RR, Strausberg RL, Simpson AJG, and de Souza SJ.** ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *CR Biol* 326: 979–985, 2003.
29. **Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR Jr, Ide SE, Iorio KR, Lee WY, Seki N,**

gase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, adusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, ela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James R, Bentley D, Deloukas P, Lander ES, and Hudson TJ. A gene map the human genome. *Science* 274: 540–546, 1996.

ndenbroucke II, Vandesompele J, Paepe AD, and Messiaen L. antification of splice variants using real-time PCR. *Nucleic Acids Res* E68, 2001.

nter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, ith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng l, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, bor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKu- k VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, nkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, rea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, inert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, nazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, angelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, iman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, ik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, ao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, des R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, ao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, umhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, sam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, venport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, stin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, ine L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, atts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, mblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, ong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, and Zhu X. The sequence of the human genome. *Science* 291: 1304–1351, 2001.

32. **Wang L, Duke L, Zhang PS, Arlinghaus RB, Symmans WF, Sahin A, Mendez R, and Dai JL.** Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Res* 63: 4724–4730, 2003.
33. **Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, and Lee MP.** Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res* 63: 655–657, 2003.
34. **Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A, and Mintz L.** Computational analysis of alternative splicing using EST tissue information. *Genomics* 80: 326–330, 2002.
35. **Xu Q and Lee C.** Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res* 31: 5635–5643, 2003.
36. **Zhang B, Schmoyer D, Kirov S, and Snoddy J.** GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5: 16, 2004.
37. **Zhang Z, Schwartz S, Wagner L, and Miller W.** A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7: 203–214, 2000.
38. **Zhu J, Shendure J, Mitra RD, and Church GM.** Single molecule profiling of alternative pre-mRNA splicing. *Science* 301: 836–838, 2003.

ANEXO 6

Capítulo 1 - O Genoma Humano e o Câncer

Raphael Bessa Parmigiani
Anamaria Aranha Camargo

Introdução

O aparecimento de um tumor está associado com a ocorrência de alterações genéticas que se acumulam progressivamente no material genético (DNA) de uma célula normal. Dessa forma, o câncer pode ser considerado uma doença genética e a identificação e caracterização dos genes alterados é fundamental para a compreensão das bases moleculares da doença. A identificação desses genes consiste também no primeiro passo para o desenvolvimento de métodos de diagnóstico mais sensíveis e de terapias alternativas mais específicas e eficazes.

Até pouco tempo a identificação de genes relacionados ao câncer dependia exclusivamente da detecção freqüente e constante de uma determinada alteração cromossômica em tipos específicos de tumores e do posterior mapeamento de genes candidatos na região cromossômica freqüentemente alterada. No entanto, esta estratégia é bastante trabalhosa e apresenta algumas limitações, já que nem sempre alterações cromossômicas são detectadas nos tumores.

Esse cenário foi recentemente alterado devido à disponibilidade da seqüência completa do genoma humano e de um número eqüilavente de seqüências que correspondem a genes expressos em diferentes órgãos e tecidos. Em conjunto, os dados de sequenciamento permitiram a identificação de mais de 35 mil genes em nosso genoma. Essa informação certamente irá acelerar a busca por novos genes relacionados ao câncer e, em um futuro não muito distante, será utilizada para o combate mais efetivo da doença.

O câncer e os nossos genes

A integridade de um determinado tecido, assim como sua função, é conferida por um equilíbrio estabelecido entre proliferação e morte celular. Este equilíbrio é mantido através de um complexo sistema de sinalização intra e extracelular. Quando este equilíbrio é perdido, as células passam a se proliferar de forma anômala, formando uma massa de células desordenadas que constituem um tumor primário. No processo de progressão tumoral, algumas células do tumor primário perdem a capacidade de adesão, invadem a membrana basal do tecido de origem através da produção de enzimas proteolíticas, atravessam a parede dos vasos sanguíneos, caem na circulação e formam áreas de proliferação em outros tecidos denominadas metástases.

A perda do controle da proliferação e a aquisição de características associadas com a progressão tumoral são conseqüências de alterações que ocorrem no conteúdo genético das células normais. A célula alterada, por adquirir uma maior capacidade de proliferação em relação às células vizinhas, sofre uma expansão clonal, transmitindo geneticamente a alteração para todas as células que, a partir dela, se originam. Devido à complexidade e à existência de vias alternativas no controle da proliferação celular, é necessária a ocorrência de alterações adicionais e sucessivas em diferentes genes para que haja a formação de um tumor. Cada nova alteração é acompanhada de uma nova onda de expansão clonal e, ao final desse processo, surge uma população celular com grande potencial de crescimento e invasão, um tumor malígno.

Assim sendo, o câncer pode ser considerado uma doença genética complexa que resulta de alterações simultâneas em genes geralmente relacionados a proliferação, diferenciação e morte celular . Nos últimos anos, um grande avanço na identificação de genes relacionados a formação e progressão tumoral foi alcançado e atualmente mais de 200 genes dessa categoria já estão caracterizados. Como veremos a seguir, a disponibilidade da seqüência completa do genoma humano e de um número eqüilavente de seqüências que correspondem a genes expressos em diferentes órgãos e tecidos certamente irá acelerar o processo de identificação de genes relacionados ao câncer. No entanto, é difícil prever exatamente o impacto que a disponibilidade dessas seqüências terá, a curto prazo, no combate efetivo à doença.

Apesar disso, com base em alguns exemplos que já se encontram relatados na literatura e que serão discutidos no final desse capítulo, é possível identificar algumas aplicações imediatas. Dentre essas aplicações, destacam-se: (I) a avaliação da susceptibilidade individual a determinados tipos de câncer; (II) o desenvolvimento de métodos de diagnóstico molecular mais precisos e precoces; (III) a determinação da resposta a um determinado tratamento; (IV) e o desenvolvimento de terapias alternativas e novos fármacos.

O sequenciamento completo do genoma humano

O termo genoma é designado para representar o conjunto de genes e seqüências regulatórias de um dado organismo (Figura 1.1). Os genes, por sua vez, carregam as informações genéticas que determinam todas as características de um organismo e a sua existência foi inicialmente inferida nos experimentos realizados por Gregor Mendel em 1865. A natureza química dos genes, entretanto, só foi revelada na década de 40 e 50 através de experimentos que demonstraram que os genes estavam contidos na molécula de DNA. Alguns anos depois, em 1953, James Watson e Francis Crick determinaram a estrutura física do DNA e o modelo proposto da dupla fita foi fundamental para a compreensão do mecanismo de transmissão e execução da informação genética.

Com o desenvolvimento das técnicas de manipulação do DNA (técnicas do DNA recombinante) e, em especial, da técnica de seqüenciamento, tornou-se possível isolar e determinar a seqüência dos genes. No entanto, no início da década de 80, isolar e caracterizar o conjunto completo de genes de um organismo parecia um objetivo muito distante. A forma manual e limitada com que os dados de seqüenciamento eram gerados e analisados era o fator limitante. Essa situação foi revertida com o desenvolvimento dos primeiros seqüenciadores semi-automáticos de DNA e com o aparecimento de uma nova área da biologia voltada para a análise computacional dos dados de sequenciamento: a bioinformática. Nascia assim a era da Genômica.

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. O genoma humano haplóide é dividido em 23 moléculas lineares de DNA, os

cromossomos, sendo a mais curta com 55Mb 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. Devido à sua complexidade, a determinação da seqüência correta desses 3,2 bilhões de nucleotídeos e a identificação dos genes contidos nessa seqüência 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.

O seqüenciamento do genoma humano foi primeira e exclusivamente 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 e bastante controversa, pois, até então, tal estratégia só havia sido utilizada no sequenciamento de genomas bacterianos, infinitamente menos complexos que o genoma humano (Figura 1.2). A proposta da Celera Genomics era terminar o sequenciamento do genoma humano em um menor tempo e com um menor custo e, para tanto, se associou com a Applied Biosystems, uma das maiores empresas fabricantes de seqüenciadores semi-automáticos de DNA. Após inúmeros debates e ataques públicos entre os dois grupos concorrentes, o anúncio do término do sequenciamento do genoma humano foi feito, em conjunto pelos dois grupos, no dia 26 de junho de 2000 em sessão solene na Casa Branca em Washington.

A determinação da sequência dos 3,2 bilhões de nucleotídeos que compõem o genoma humano (10 e 15) certamente representou um grande marco na ciência, no entanto, apesar do enorme entusiasmo gerado, um grande percurso ainda precisa ser percorrido para que possamos conhecer o significado da seqüência obtida e converter essa informação em aplicações práticas. Uma das primeiras questões que surgem é: como identificar os genes a partir de uma seqüência monótona de nucleotídeos A, T, C e G?

Para a identificação de genes a partir da seqüência genômica são utilizados programas computacionais de predição gência, capazes de reconhecer, na seqüência bruta de nucleotídeos, seqüências sinais características dos genes humanos. No

entanto essas sequencias são bastante degeneradas e inespecíficas e em vista disso os programas de predição gênica não são cem por cento eficazes. Além disso, os genes humanos correspondem a apenas 3% do genoma humano e 50% das seqüências de nucleotídeos correspondem a seqüências de DNA repetitivas de função ainda desconhecida. Em consequência disso, estima-se que 90% das estruturas gênicas preditas por programas de computador estejam incorretas e que um número ainda não estimado de genes não seja detectado por esses programas.

Utilizando programas computacionais ambos os grupos de sequenciamento do genoma humano conseguiram identificar um número aproximado de 35.000 genes. Mas como identificar dentre esses genes aqueles relacionados com a formação de tumores? Uma possibilidade seria procurar entre os novos genes identificados, genes que apresentassem similaridade a nível de nucleotídeos ou aminoácidos com genes já reconhecidamente envolvidos com câncer. A maioria dos genes já associados à essa doença fazem parte de famílias gênicas extensas e é possível, a exemplo do que ocorre para o gene supressor de tumor p53, que outros membros da mesma família estejam igualmente associados ao câncer. Outra possibilidade seria a identificação de genes localizados em regiões específicas do genoma as quais são freqüentemente afetadas por rearranjos cromossômicos em tumores. Contudo, em um estudo preliminar realizado por Futreal e colaboradores (6), essas abordagens não foram efetivas na identificação de novos genes relacionados ao câncer. Uma possível explicação para este fato seria a existência de genes humanos ainda não identificados a partir da seqüência genômica.

Nesse sentido, desde o anúncio do sequenciamento completo do genoma humano, o número de relatos na literatura que apontam a existência de um número superior de genes no genoma humano tem crescido. Dados recentes suportam a existência de pelo menos 60 mil genes no genoma humano e sugerem que a identificação de todos os genes humanos só será possível através da utilização de abordagens complementares, como por exemplo o sequenciamento de moléculas de cDNA (2).

O sequenciamento em larga escala de moléculas de cDNA

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), 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 clonagem e replicado em células bacterianas, dando origem às chamadas bibliotecas de cDNA. Os clones que compõem uma biblioteca de cDNA representam os genes expressos em um determinado tecido do qual o mRNA foi extraído e utilizado para a construção da biblioteca. Os clones de cDNA podem ser, então, seqüenciados de duas formas: seqüenciamento completo do inserto clonado ou seqüenciamento parcial das extremidades desses clones gerando as ESTs (do inglês Expressed Sequence Tags).

A produção de seqüências completas de cDNA e ESTs vem sendo realizada por diversos grupos simultaneamente ao seqüenciamento do genoma humano. Mais especificamente, a produção de ESTs teve suas origens em 1980, mas somente no início da década de 90 elas começaram a ser produzidas em larga escala. As ESTs foram inicialmente exploradas na identificação de novos genes, no entanto, atualmente, estão sendo intensamente utilizadas na construção de perfis de transcritos tecido-específicos, na construção de mapas físicos, na comparação de genomas de diferentes organismos, no estudo qualitativo da expressão de mRNA e, ainda, na identificação de genes a partir da seqüência genômica.

Um dos projetos pioneiros na produção de seqüências de ESTs, iniciado em 1997, foi o "Cancer Genome Anatomy Project" (CGAP) (13). O projeto foi financiado pelo Instituto Nacional de Câncer do Estados Unidos e teve como objetivo gerar um catálogo de genes expressos em uma grande variedade de tecidos normais e tumorais. Mais de 2 milhões de seqüências foram geradas e, uma vez terminada a fase de sequenciamento em larga escala, o enfoque foi voltado para a integração das seqüências produzidas com informações disponibilizadas por outros projetos como, por exemplo, o projeto de mapeamento de anormalidades

cromossômicas em tumores ou mesmo o projeto de sequenciamento completo do genoma humano.

Em 1999, a FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo) e o Instituto Ludwig de Pesquisa sobre o Câncer lançaram, também, o Projeto Genoma Humano do Câncer (HCGP) com o objetivo de gerar 1 milhão de ESTs utilizando uma nova metodologia denominada ORESTES ("Open Reading Frame ESTs") (1 e 4). 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. 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.

As informações geradas pelos projetos CGAP e HCGP são complementares e correspondem a mais de 50% das seqüências de ESTs depositadas em bancos de dados públicos. Recentemente as informações geradas pelos dois grupos foram integradas em um banco de dados internacional de expressão gênica em câncer com o objetivo de acelerar a descoberta de novos genes relacionados com a doença (14). Para facilitar a análise e utilização dos dados, uma série de ferramentas computacionais foi desenvolvida e disponibilizada na página do CGAP (**http://cgap.nci.nhi.gov**). Essas ferramentas permitem a análise individual de cada banco de dados, mas ao mesmo tempo é capaz de integrar as informações geradas por ambos os projetos permitindo que os pesquisadores analisem *in silico* o padrão de expressão gênica de diferentes tumores e identifiquem genes que, quando alterados, estejam associados com a formação de tumores.

O impacto dos dados de sequenciamento disponibilizados pelos dois projetos na pesquisa sobre o câncer já começam a aparecer na literatura. Através da análise dos dados disponibilizados pelo CGAP, por exemplo, uma EST correspondente à subunidade catalítica da telomerase humana foi identificada devido a sua similaridade com seqüências de telomerases de outros organismos e serviu de base para a caracterização funcional desse gene em humanos (12). A enzima telomerase cataliza a adição de repetições teloméricas na extremidade dos cromossomos

humanos, garantindo dessa forma que a cada ciclo de divisão celular o material genético da célula seja replicado de maneira estável e fiel. A atividade da enzima telomerase é detectada em células com alto índice de proliferação como, por exemplo, as células tumorais. Em vista disso, esta enzima tem sido considerada um alvo potencial para o desenvolvimento de drogas. Da mesma forma, a análise dos dados de ESTs tem permitido a identificação de genes preferencialmente expressos em tumores de próstata, pâncreas, mama, cérebro, intestino e ovário. A identificação de genes preferencialmente expressos em tumores é uma das possíveis contribuições para o desenvolvimento de novas ferramentas diagnósticas e terapias alternativas como veremos a seguir.

Avaliação da susceptibilidade a determinados tipos de câncer: síndromes hereditárias e SNPs

As alterações genéticas relacionadas ao câncer podem ocorrer tanto em células da linhagem germinativa quanto em células somáticas do indivíduo adulto. O tipo celular no qual ocorrem as alteracões é de primordial importância, já que determina se estas alterações serão transmitidas para gerações futuras ou se seus efeitos serão limitados ao tecido em que ocorreu a alteração. Alterações que ocorrem na linhagem germinativas são propagadas a todas as células do indivíduo e são transmitidas hereditariamente para os seus descendentes. Já as alterações somáticas estão restritas às células originadas a partir da célula alterada, ou seja, são encontradas apenas no tecido em questão, não sendo transmitidas geneticamente para seus descendentes. Alterações somáticas estão associadas com o aparecimento de tumores esporádicos ao passo que alterações germinativas são responsáveis pelos casos de câncer com caráter hereditário.

Os tumores hereditários correspondem a uma pequena fração dos casos de câncer (5%), no geral se manifestam em idade precoce e acometem vários indivíduos de uma mesma família. Alguns exemplo de síndromes de câncer hereditário já estão bem caracterizados e os genes que quando alterados são capazes de levar a formação dos tumores já foram identificados. Nestes casos a busca de alterações que afetam esses genes em familiares diretos de portadores da síndrome permite um diagnóstico pré-sintomático da doença e conseqüentemente um acompanhamento clínico mais

rigoroso o qual, em última instância, resulta no diagnóstico precoce do câncer e em uma melhor evolução da doença.

Um bom exemplo são os genes envolvidos no câncer de mama hereditário, *BRCA-1* e *BRCA-2,* para os quais diferentes mutações já foram identificadas. Uma vez detectada uma paciente portadora de mutação germinativa em um desses dois genes, todos os parentes diretos podem ser testados para a presença da mutação identificada e determinar, desta forma, o risco individual de desenvolver tumor de mama. Esse teste permite identificar parentes que herdaram a mutação e que portanto apresentam um risco elevado de desenvolver tumor de mama e que certamente irão se beneficiar de um acompanhamento clínico mais rigoroso com mamografias e ultrasonografias mais freqüentes ou mesmo de terapias endócrinas preventivas. Da mesma forma, o teste pode identificar familiares que não herdaram a mutação e que, por apresentarem o mesmo risco de desenvolver tumores de mama que a população em geral, não necessitariam de um acompanhamento rigoroso e muitas vezes de custo elevado.

Outro exemplo deste tipo de associação foi feita entre a presença de alterações no gene *VHL* e pacientes portadores da síndrome de Von Hippel-Lindau. De maneira geral, os portadores desta síndrome desenvolvem precocemente diferentes tipos de tumor maligno, em especial hemangioblastoma cerebelar, angioma de retina e carcinoma renal. Somam-se ainda os tumores benignos, também muito freqüentes, dos quais destacam-se angiomas, cistos e adenomas. O gene *VHL* é um gene supressor de tumor e por isso, alterações que resultem na formação de uma proteína não funcional podem levar a um descontrole do ciclo celular e propiciar assim o desenvolvimento de uma neoplasia (11). A exemplo das alterações em *BRCA1* e *BRCA2,* a importância da identificação destas mutações está na possibilidade do desenvolvimento de testes genéticos em que se possa avaliar a presença das mesmas em parentes diretos de portadores da síndrome. É evidente que a confirmação de mutações em um determinado indivíduo não evitará o desenvolvimento do tumor, mas cuidados poderão ser tomados no sentido de diminuir a exposição do indivíduo a fatores de risco para determinados tipos de câncer, além de se identificar precocemente lesões sub-clínicas. Assim, estes

pacientes ganham um considerável aumento na sobrevida e no tempo livre de doença, bem como uma melhora da qualidade de vida.

Além das alterações hereditárias, existe outro tipo de alteração genética, denominada SNP (*Single Nucleotide Polymorphism*), que pode determinar a susceptibilidade de um indivíduo ao desenvolvimento de tumores. Os SNPs são variações de um único nucleotídeo que ocorrem entre as sequências de DNA de dois indivíduos. São exatamente estas variações que determinam as diferenças fenotípicas entre os indivíduos, ou seja, a individualidade de cada um. O que diferencia um SNP de uma mutação é a frequência com que os mesmos ocorrem na população, sendo considerado um SNP toda variação de um único nucleotídeo que ocorre em mais de 1 % da população. Outra diferença é o caráter delério da alteração. Mutações geralmente estão relacionadas a um fenótipo grave, uma disfunção evidente. Já os SNPs não necessariamente estão associados a um fenótipo característico e no geral se manifestam através de um susceptibilidade discreta que pode ser agravada por fatores externos.

Em um estudo recente, por exemplo, foi comprovado que a presença de um determinado SNP no gene da endostatina está relacionado a uma maior susceptibilidade ao desenvolvimento de câncer de próstata (8). Esta proteína está envolvida na angiogênese (formação de novos vasos a partir de um endotélio pré-existente), funcionando como um inibidor da proliferação e migração de células endoteliais. Considerando-se que a angiogênese é um passo fundamental para a progressão tumoral e formação de metástases, alterações na função desta proteína poderiam estar associadas ao desenvolvimento de tumores. Neste estudo foi demonstrado que a alteração de um nucleotídeo, que posteriormente alterava o aminoácido codificado por ele, é muito mais freqüente em pacientes com câncer de próstata do que em indivíduos sadios. Embora estes pacientes produzam a mesma quantidade de proteína encontrada nos indivíduos sadios, acredita-se que esta mudança de um único aminoácido leva a uma perda significativa de sua função e conseqüentemente a um maior risco de desenvolvimento do tumor.

A identificação de uma lista de SNPs em todo o genoma humano, especialmente aqueles localizados nos genes ou em regiões regulatórias da expressão dos mesmos, pode ajudar a esclarecer porque determinadas pessoas apresentam uma

maior susceptibilidade a determinado tipo de câncer. Da mesma maneira como para as mutações germinativas, testes que avaliem a presença de determinados SNPs associados ao desenvolvimento de tumores podem ajudar na prevenção da doença, principalmente nos casos em que o surgimento da mesma está estreitamente relacionado à exposição a fatores de risco como, por exemplo, o fumo e a bebida alcoólica.

Desenvolvimento de diagnóstico moleculares e a evolução da doença

Tão importante quanto se diagnosticar um tumor é saber qual o tipo histológico do tumor que foi identificado. Isso porque a histologia de diferentes tumores pode ser semelhante, mas suas alterações moleculares e comportamento clínico bem diferentes. Convencionalmente a análise e classificação histopatológica do tumor é feita através de técnicas de microscopia óptica e imunohistoquímica. Diferentes ensaios são necessários para tal classificação e o conhecimento sobre a perda da expressão ou super-expressão de um determinado gene (e conseqüentemente de sua proteína) pode ser decisivo para se avaliar a evolução clínica de um paciente ou mesmo a conduta terapêutica a ser tomada (como veremos no item seguinte). Entretanto, existem muitos casos nos quais os marcadores tumorais disponíveis são insuficientes para distinguir tumores histologicamente semelhantes e que possuem evolução e resposta ao tratamento muito diferentes.

Além de se classificar um tumor corretamente, outra dificuldade existente se refere à avaliação e conduta das lesões precursoras de tumor. Muitas vezes as mesmas são consideradas inofensivas e facilmente tratáveis por ainda não apresentarem grandes alterações morfológicas. Entretanto, estas lesões podem apresentar algumas das alterações moleculares importantes, características de uma neoplasia verdadeira. Assim, mais uma vez estudos moleculares poderiam avaliar a presença de tais alterações e assim auxiliar o médico na decisão a ser tomada no tratamento deste tipo de lesão. Em ambos os casos citados, a busca por novos marcadores tumorais é imprescindível.

Considerando-se que o câncer é uma doença genética causada por alterações genéticas que modificam o perfil de expressão gênica da célula, espera-se que genes com expressão diferencial entre uma célula normal e tumoral possam ser

identificados, os quais poderão ser utilizados como assinaturas moleculares. Após o sequenciamento do genoma humano, o desenvolvimento de técnicas eficientes de clonagem e a identificação de uma grande quantidade de novos genes, criaram-se as condições necessárias para o desenvolvimento de ferramentas que avaliem o nível de expressão gênico em larga escala. Desta maneira, a identificação de novos marcadores moleculares para o câncer tem se tornado menos trabalhosa, principalmente com o uso de uma poderosa técnica, denominada "microarray".

Esta técnica, também conhecida como chip de DNA, permite a análise do padrão de expressão de milhares de genes simultaneamente, gerando assim um perfil de expressão gênica do tecido analisado. A avaliação do perfil de expressão pode ser feita no sentido de se comparar, por exemplo, os genes expressos em um tecido normal e em um tumor, e assim encontrar diferenças na expressão de alguns genes entre as duas amostras. Neste caso, estes genes diferencialmente expressos seriam bons candidatos a marcadores moleculares. Mas além destes marcadores que diferenciam amostras normais de tumorais, a comparação entre amostras em diferentes estadios da doença, desde uma lesão precursora, um carcinoma *in situ*, até um carcinoma invasivo e metastático, poderia levar também à identificação de marcadores relacionados, à progressão tumoral e formação de metástases. Todos estes marcadores serão de grande utilidade na conduta a ser tomada, principalmente nos casos em que tratamentos mais agressivos podem ser evitados.

Entretanto, como mencionado anteriormente, existem situações mais complexas em que se pretende diferenciar dois tumores histologicamente muito semelhantes, mas com evolução clínica e resposta ao tratamento bem diferentes. Nestes casos, marcadores moleculares isolados muitas vezes não são eficientes para permitir tal distinção entre as amostras, necessitando assim de uma análise molecular mais abrangente. Novamente a técnica de "microarray" poderia ser aplicada, uma vez que, ao compararmos o perfil de expressão gênica de diferentes amostras, seria possível classificá-las em sub-grupos que possuíssem perfis de expressão semelhantes. Assim, a diferenciação poderia ser feita baseando-se no comportamento de um grupo de genes e não apenas em um marcador isoladamente.

A idéia de que o perfil de expressão gerado pelo "microarray" pode refletir as características e o comportamento de um determinado tipo de câncer foi inicialmente

comprovado em um estudo no qual foram comparados os perfis de expressão de leucemia mielóide aguda (LMA) e leucemia linfocítica aguda (LLA) (7). Neste caso os perfis mostraram-se muito informativos, possibilitando uma clara distinção entre os dois tipos de leucemia. Embora o diagnóstico histopatológico destes tipos de leucemia não seja muito difícil, o estudo serviu para demonstrar a eficiência das análises feitas com o perfil de expressão gência de diferentes amostras, as quais podem ser muito úteis no diagnóstico diferencial em outros casos em que a análise histopatológica seja insuficiente para a classificação do tumor e avaliação da evolução da doença.

A farmacogenômica e a determinação de resposta ao tratamento

A farmacogenômica pode ser definida como o uso de marcadores biológicos (DNA, RNA e proteínas) para a predição da eficácia de uma determinada droga e para a determinação da probabilidade de ocorrência de efeitos colaterais e adversos. Devido à baixa eficácia (30%) da maioria dos agentes terapêuticos utilizados no tratamento do câncer, a farmacogenômica se torna essencial na prática oncológica, pois permite a seleção prévia de pacientes que irão de beneficiar do tratamento.

Nesse sentido, a farmacogenômica tem sido aplicada na oncologia basicamente de duas formas: I) na identificação de SNPs presentes em genes que codificam proteínas relacionadas com o metabolismo e/ou o transporte de drogas e que correlacionam com uma melhor ou pior resposta ao tratamento, e II) na determinação de perfis de expressão gênica de tumores capazes de indicar uma pior ou melhor resposta ao tratamento (5). Uma vez estabelecidos, esses marcadores de resposta ao tratamento poderão servir como guias para a otimização de protocolos terapêuticos para cada indivíduo.

Um exemplo marcante de variabilidade genética relacionada a resposta ao tratamento do câncer é o caso da tiopurina S-metiltransferase. SNPs presentes no genes que codifica essas enzima são capazes de determinar a resposta a droga mercaptopurina comumente utilizada no tratamento de leucemias infantis. Pacientes com uma determinada alteração nesse gene são classificados com maus metabolizadores da droga e possuem um grande risco de apresentar reações adversas quando tratados com as doses convencionais do medicamento (9).

Variações nos perfis de expressão gênica também têm sido amplamente utilizadas como marcadores de resposta ao tratamento do câncer. Como mencionado anteriormente tais perfis podem ser determinados através da análise simultânea do padrão de expressão de milhares de genes em experimentos de "microarray". Esses experimentos partem da premissa de que os tumores apresentam alterações no padrão de expressão gênica, as quais podem ser utilizadas como assinaturas que refletem suas característica biológicas. Assim, comparando o perfil de expressão de tumores de pacientes que respondem melhor ou pior ao tratamento, é possível identificar genes relacionados com a resposta a droga utilizada.

Em um estudo recente utilizando linhagens celulares tumorais com diferentes sensibilidades a uma nova droga desenvolvida para o tratamento de câncer, foi possível identificar genes marcadores de reposta ao tratamento. O estudo avaliou, através de experimentos de "microarray", genes que tinham sua expressão aumentada nas linhagens celulares e que após o tratamento apresentaram uma redução significativa nos níveis de expressão (16). Seis genes com esse padrão de expressão foram isolados e de forma interessante, dois deles fazem parte da via de regulação do gene alvo inativado pela droga. A expressão desses seis genes pode ser utilizada como fator preditivo de uma boa resposta ao tratamento e dessa forma ser utilizada para a seleção de pacientes que se beneficiariam do uso da nova droga. Em análises posteriores, utilizando tumores de pacientes com câncer, 30% das amostras testadas apresentaram a expressão dos seis genes isolados nos experimentos com as linhagens celulares, havendo uma correlação direta entre tumores com expressão positiva e boa resposta ao tratamento.

Desenvolvimento de novas terapias

O desenvolvimento de agentes terapêuticos mais eficazes é fundamental para o progresso da quimioterapia contra o câncer. Nesse contexto, a farmacogenômica também pode ser útil e avanços significativos na descoberta de novos agentes foram obtidos recentemente através da inibição de moléculas alvos cuja relação direta com a formação de tumores já estava bem estabelecida.

A identificação de mais de 35.000 genes através do sequenciamento completo do genoma humano aumentou consideravelmente o número de genes potenciais para

o desenvolvimento de novas drogas. Merecem especial atenção novos genes que pertencem a famílias gênicas de alvos já estabelecidos para o tratamento de câncer como por exemplo, genes que codificam para receptores associados a proteína G, quinases e proteases os quais podem ser bloqueados por drogas específicas.

Um bom exemplo de como os estudos das alterações moleculares de tumores levou ao desenvolvimento de uma droga com poucos efeitos colaterais e de grande efeito terapêutico é o **Gleevec** (3). Esta droga, também conhecida como Imatinib, vem sendo muito utilizada no tratamento da leucemia mielóide crônica (CML). Esta neoplasia possui como característica marcante a translocação recíproca de parte dos braços longos dos cromossomos 9 e 22, resultando assim, na formação do cromossomo Philadelphia. Como resultado dessa translocação, há a formação de um gene quimérico, o *BCR-ABL*, constituído pela justaposição do proto-oncogene *C-ABL*, oriundo do cromossomo 9, com sequências do gene *BCR* (*Breakpoint Cluster Region*), do cromossomo 22. A proteína quimérica resultante possui uma forte atividade tirosina quinase e é encontrada em 95% dos pacientes com CML.

Esta proteína leva à ativação de muitas vias de sinalização intracelular, causando assim, alterações nas propriedades proliferativas, adesivas e de sobrevivência das células tumorais. Por isso, o gene *BCR-ABL* tem sido considerado um oncogene típico de leucemias. Além disso, como esta atividade tirosina quinase é essencial para o processo de transformação, a proteína BCR-ABL tornou-se um alvo importante para tratamento deste tipo de neoplasia. Na tentativa de se obter um inibidor, foi desenvolvido o Gleevec, cuja grande vantagem é sua alta especificidade pela quinase ABL, uma vez que reduz os efeitos colaterias normalmente encontrados no tratamento do câncer (Figura 1.3). Além de sua especificidade por um alvo presente apenas nas células alteradas, o Gleevec tem ampla aplicação, uma vez que seu alvo está presente em 95% dos pacientes com CML.

Além do Gleevec, outras drogas já estão sendo desenvolvidas com o objetivo de se bloquear a ação de alvos recém identificados (17). Outros exemplos são:

1. **Gefitinib**: inibidor do receptor do fator de crescimento epidermal utilizado no tratamento de câncer de cabeça e pescoço, de tumores de próstata hormônio refratários e também de tumores de pulmão de células não pequenas.

2. **Herceptina**: anticorpo monoclonal utilizado no tratamento de tumores de mama com alta expressão do oncogene *ERB-B2*.
3. **17AAG**: inibidor da chaperonina HSP90, comumente superexpressa em diferentes tipos de tumor, que está completando a fase de estudos clínicos com resultados muito promissores.

Por outro lado, o sequenciamento completo do genoma humano e a produção de um catálogo completo dos nossos genes também podem auxiliar no entendimento dos efeitos colaterais de algumas drogas. Muitos desses efeitos colaterais estão relacionados com a atuação inespecífica dessas drogas em proteínas similares a proteína alvo para a qual a droga foi desenhada. Com o catálogo completo de genes humanos disponíveis, novas drogas podem ser desenvolvidas com o conhecimento prévio de outros genes similares que poderão ser afetados, minimizando dessa forma seus efeitos colaterais.

Conclusões

No dia 26 de junho de 2000 um grande marco cientifico e tecnológico foi alcançado com o término do sequenciamento do genoma humano. A informação disponibilizada por esse projeto, em conjunto com seqüências de clones de cDNA geradas por diferentes grupos de pesquisa, estão sendo utilizadas para se construir um catálogo completo de genes humanos. Este catálogo inicialmente composto por aproximadamente 35 mil genes ainda não está completo e alguns anos serão necessários para completá-lo através da utilização de abordagens complementares para a identificação gênica.

Apesar de ainda incompleto, o catálogo disponibilizado irá acelerar indiscutivelmente a descoberta de genes que quando alterados geneticamente levam a formação de tumores. Como vimos neste capítulo, resultados promissores nessa direção já estão descritos na literatura e informações genéticas a respeito de genes específicos, ou de um conjunto deles, já estão sendo utilizadas no desenvolvimento de métodos de diagnóstico molecular mais precisos, na determinação da resposta a um determinado tratamento e no desenvolvimento de terapias alternativas e novos fármacos. O próximo grande desafio será a aplicação dessas informações na prática

clínica. Para tanto adaptações metodológicas terão que ser feitas, visando a diminuição de custos e a automatização dos protocolos experimentais utilizados. Também será fundamental que a linguagem muitas vezes complicada dos nossos genes, seja traduzida e transmitida a médicos e pacientes. Esperamos que este capítulo tenha contribuído de alguma forma no processo de tradução e que a leitura do mesmo sirva como um veículo efetivo de transmissão dessas informações.

Referências Bibliográficas

1. Camargo AA, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci U S A**. 98(21):12103-8, 2001.
2. Camargo AA, de Souza SJ, Brentani RR, Simpson AJ. Human gene discovery through experimental definition of transcribed regions of the human genome. **Curr Opin Chem Biol**. 6(1):13-6, 2002.
3. Capdeville R, Silberman S. Imatinib: A targeted clinical drug development. **Semin Hematol.**;40(2):15-20, 2003.
4. Dias Neto E, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci U S A**. 97(7):3491-6, 2000.
5. Dracopoli NC. Pharmacogenomic applications in clinical drug development. **Cancer Chemother Pharmacol**. 52 Suppl 1:57-60, 2003.
6. Futreal PA, Kasprzyk A, Birney E, Mullikin JC, Wooster R, Stratton MR. Cancer and genomics. **Nature**. 409(6822):850-2, 2001.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**. 286(5439):531-7, 1999.
8. Iughetti P, Suzuki O, Godoi PH, Alves VA, Sertie AL, Zorick T, Soares F, Camargo A, Moreira ES, di Loreto C, Moreira-Filho CA, Simpson A, Oliva G, Passos-Bueno MR. A polymorphism in endostatin, an angiogenesis inhibitor, predisposes for the development of prostatic adenocarcinoma. **Cancer Res**. 61(20):7375-8, 2001.
9. Krynetski EY, Evans WE. Pharmacogenetics as a molecular basis for individualized drug therapy: the thiopurine S-methyltransferase paradigm. **Pharm Res**. 16(3):342-9, 1999.
10. Lander ES, et al. Initial sequencing and analysis of the human genome. **Nature**. 409(6822):860-921, 2001.
11. Lonser RR, Glenn GM, Walther M, Chew EY, Libutti SK, Linehan WM, Oldfield EH. Von Hippel-Lindau disease. **Lancet** 361(9374):2059-67, 2003.

12. Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH, Lingner J, Harley CB, Cech TR. Telomerase catalytic subunit homologs from fission yeast and human. **Science**. 277 (5328):955-9, 1997.
13. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. The cancer genome anatomy project: building an annotated gene index. **Trends Genet**. 16(3):103-6, 2000.
14. Strausberg RL, Camargo AA, Riggins GJ, Schaefer CF, de Souza SJ, Grouse LH, Lal A, Buetow KH, Boon K, Greenhut SF, Simpson AJ. An international database and integrated analysis tools for the study of cancer gene expression. **Pharmacogenomics J**.;2(3):156-64, 2002.
15. Venter JC, et al. The sequence of the human genome. **Science**. 291(5507):1304-51, 2001.
16. Yamori T. Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. **Cancer Chemother Pharmacol**. 52(1):74-9, 2003.
17. Workman P. The opportunities and challenges of personalized genome-based molecular therapies for cancer: targets, technologies, and molecular chaperones. **Cancer Chemother Pharmacol**. 52(1):45-56, 2003.

Glossário

cDNA - DNA complementar: Molécula de DNA sintetizada *in vitro* a partir de uma molécula de RNA mensageiro.

CGAP - Cancer Genome Anatomy Project: um dos projetos pioneiros na produção de seqüências de ESTs.

DNA – ácido desoxiribonucléico. Molécula biológica que carrega as informações genéticas de um organismo vivo.

EST - Expressed Sequence Tag: Sequência parcial de uma molécula de cDNA.

Genoma: conjunto de genes e seqüências regulatórias de um dado organismo.

Genes Supressores de Tumor: categoria de genes capazes de inibir a proliferação celular e, consequentemente, a formação de tumores.

Gleevec: Medicamento comumente utilizado no tratamento da leucemia mielóide crônica (*CML*).

HCGP - Projeto Genoma Humano do Câncer: Iniciativa brasileira para a produção de seqüências expressas de diferentes tipos de tumor.

LLA: leucemia linfocítica aguda

LMA: leucemia mielóide aguda.

Microarray: Técnica que permite a avaliação da expressão de vários genes, simultaneamente.

Nucleotídeo: Unidade formadora dos ácidos nucléicos, DNA e RNA.

ORESTES - Open Reading Frame ESTs: Seqüências derivadas das porções centrais dos transcritos, desenvolvida no Brasil.

RNA – ácido ribonucléico. Molécula biológica sintetizada a partir do DNA e que serve de molde para a síntese protéica.

SNP - Single Nucleotide Polymorphism: Variação de um único nucleotídeo que ocorre entre sequências de DNA de dois indivíduos, com uma freqüência maior que 1% na população.

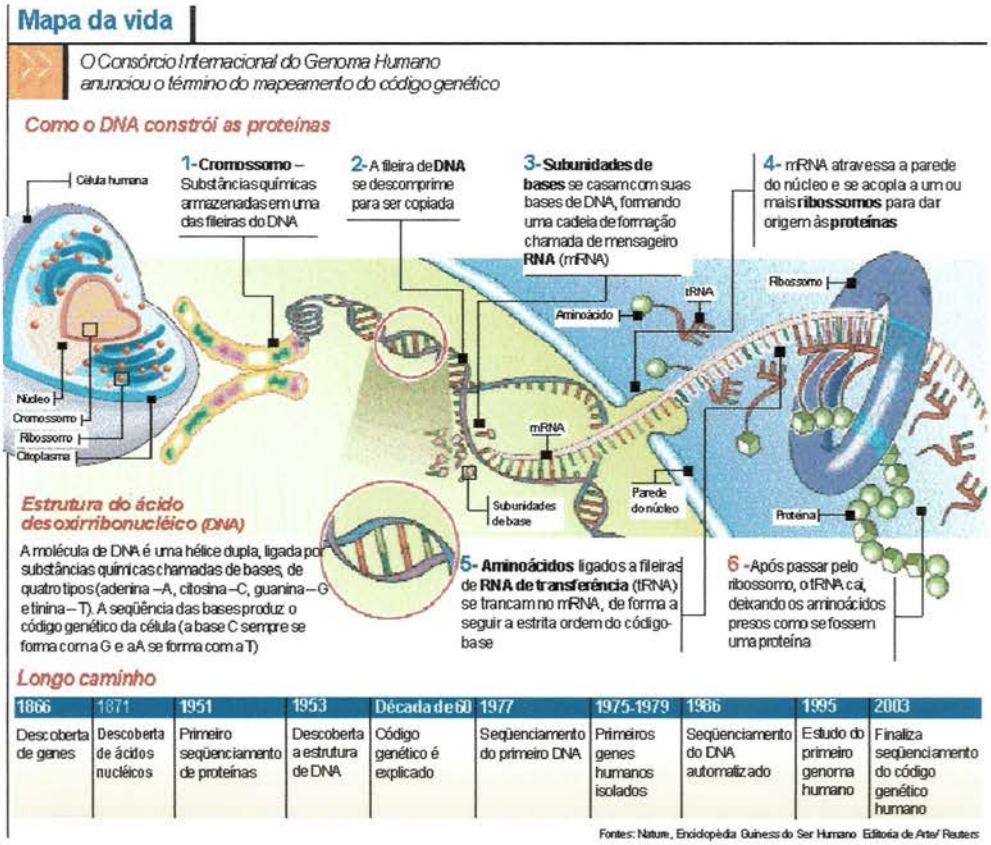


Figura 1.1: Esquema representando o fluxo da informação genética: DNA-RNA-Proteína e as principais descobertas que levaram ao nascimento da Genômica.

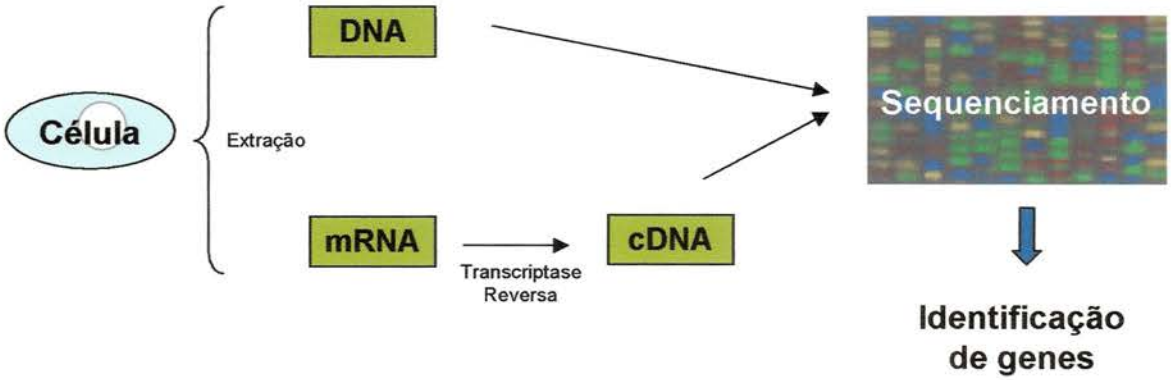


Figura 1.2: Estratégias de sequenciamento do material genético humano. Os genes humanos podem ser identificados através da utilização de duas abordagens complementares: o sequenciamento da molécula de DNA e o sequenciamento de moléculas de cDNA. As moléculas de cDNA são sintetizadas in vitro a partir da molécula de mRNA, através da utilização da enzima Transcriptase Reversa.

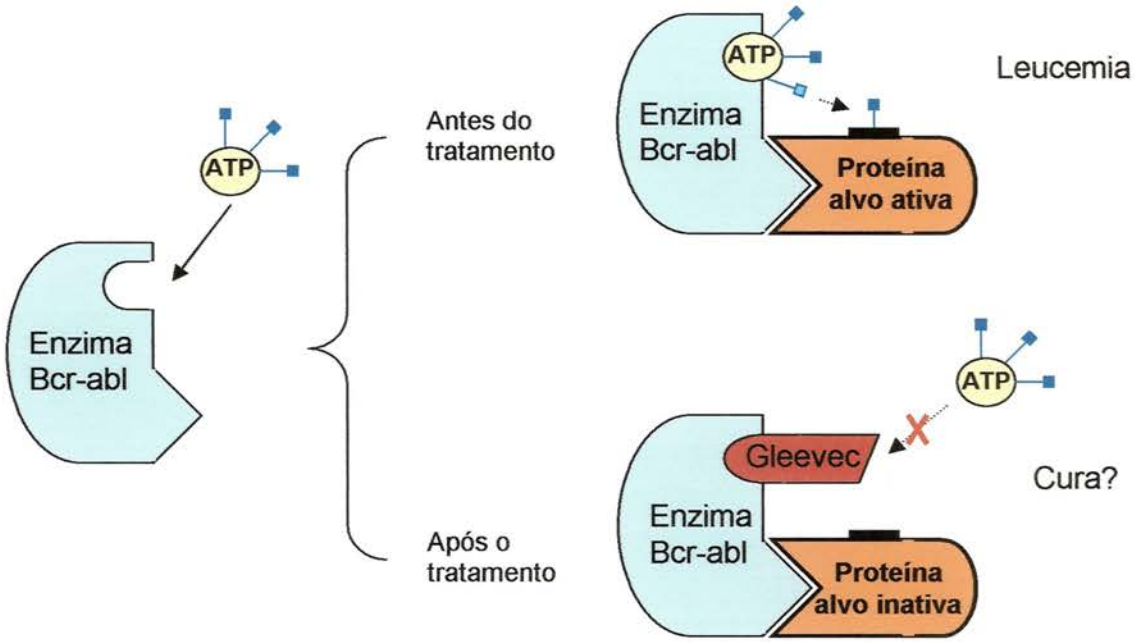


Figura 1.3: Mecanismo de atuação do Gleevec. A protéina resultante da fusão dos genes bcr-abl possui um sítio de ligação a uma molécula de ATP. Essa ligação é necessária para a atividade biológica da proteína que, quando ativada, altera a capacidade proliferativa da célula através da ativação de outras proteínas alvo. O Gleevec se liga especificamente ao sítio de ligação do ATP impedindo a ativação da proteína bcr-abl e consequentemente a ativação das proteínas alvo.