Caracterização de novos transcritos humanos localizados no *locus* HPC1 (Hereditary Prostate Cancer) através de identificação *in silico* e validação experimental

Ana Paula Medeiros Silva

Dissertação de Mestrado apresentada à Fundação Antônio Prudente para obtenção do Grau de Mestre em Ciências

Área de concentração: Oncologia

Orientadora: Dra. Anamaria Aranha Camargo

Co-Orientador: Andrew J. G. Simpson

São Paulo

2002

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A. C. Camargo

Silva, Ana Paula Medeiros

Caracterização de novos transcritos humanos localizados no *locus* HPC1 (Hereditary Prostate Cancer) através de identificação *in silico* e validação experimental. / Ana Paula Medeiros Silva - São Paulo, 2002.

75 p.

Dissertação (mestrado) – Fundação Antônio Prudente.

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

Orientadora: Anamaria Aranha Camargo.

Descritores: 1. CÂNCER DE PRÓSTATA HEREDITÁRIO. 2. CROMOSSOMO 1. 3. EXPRESSÃO GÊNICA. 4. PROTEINAS RGS.

"The future belongs to those who believe in the beauty of their dreams".

(Eleanor Roosevelt)

Aos meus pais,
com todo meu amor

Ao Rogério,
com todo meu carinho e admiração

AGRADECIMENTOS

À Dra. Anamaria A. Camargo, por me receber com tanta disponibilidade e carinho. Você é um exemplo de pessoa e de profissional. Muito obrigada pela oportunidade, pela confiança, pela paciência, pela preocupação, por todos os ensinamentos enfim pela chance de poder trabalhar e conviver com a pessoa maravilhosa que você é.

Ao Prof. Dr. Andrew J. G. Simpson pela oportunidade de poder trabalhar em seu grupo de pesquisa e também pela co-orientação que sempre foi acompanhada de excelentes sugestões.

Ao Fabrício, pelos primeiros ensinamentos, por sempre se preocupar em oferecer ajuda, pela paciência, pelo apoio e pela convivência. Você é um grande amigo.

Ao Ricardo Moura, pelas valiosas dicas de trabalho, pela disponibilidade e incentivo constante, pelo convívio diário e também pela amizade que é sempre o mais importante.

À Dirce pelas importantes sugestões, conselhos e pela disponibilidade em ler este trabalho.

À Valéria Paixão e Anna Chris pelo importante e indispensável apoio técnico sempre com muita disponibilidade, amizade e alto astral.

À toda turma do laboratório, Ricardo, Fabricinho, Dani, Anna Chris, Jane, Elis, Fabi, Rapha, Lilian, Zecla e Dirce, pela convivência e pelos bons momentos vividos diariamente no laboratório. Sem vocês, com certeza tudo teria sido mais difícil.

Ao Prof. Dr. Sandro de Souza e toda sua equipe do Laboratório de Bioinformática pela disponibilidade, sugestões e auxílio com o Banco de Dados do Transcriptoma Humano.

Ao Jorge Estefano pela boa vontade e excelente ajuda com a interface gráfica.

À Profa. Dra. Otávia Cabalero e a todo seu grupo, em especial à Adriana Bulgarelli, pela importante ajuda com o Real Time.

Ao João pela disponibiladade e ajuda com o Dot Blot.

A todos os funcionários do ILPC pelo suporte técnico e administrativo e também pela convivência durante este período. Em especial à Eliane, pela boa vontadade e eficiência em resolver os problemas do dia-a-dia.

Aos demais colegas do ILPC pela convivência durante este período.

À Suely Francisco e demais funcionários da biblioteca da Fundação Antônio Prudente pelo auxílio na obtenção e organização do material bibliográfico.

À Ana Maria Kuninari e Márcia Hiratoni pelo excelente trabalho desenvolvido junto a secretaria de pós graduação e pelo carinho com que sempre me receberam.

Ao Prof. Dr. Luiz Fernando Lima Reis pela excelente coordenação do curso de Pós Graduação, pela disponibilidade e também pelos importantes conselhos.

Ao Dr. Humberto Torloni pela fundação e administração do Banco de Tumores que foi importante fonte de material para a realização deste trabalho.

Ao Prof. Dr. Ricardo Renzo Brentani que através da administração do Hospital do Câncer e do Instituto Ludwig, possibilitou o desenvolvimento desse trabalho em um local conceituado e bem estruturado.

Aos meus tios, primos, avós e amigos que sempre torceram por mim.

Ao Walter, Silvana, Juliana, Renata, Wilson, Simone, Paula e Gabriela por serem mais que amigos. Vocês sem dúvida são uma parte muito especial da família. Obrigada pela torcida e incentivo sempre. Agradeço em especial ao Walter, pelo socorro com o computador.

Ao meu irmão que mesmo sem dizer muita coisa, sei que me apoia e torce por mim.

Ao Rogério, pelo amor e amizade, pelo incentivo, por me ensinar a sempre enxergar além, por me apoiar e por me compreender em todos os momentos.

Aos meus pais pelo amor, dedicação, incentivo, confiança e equilíbrio. Vocês são meus grandes exemplos.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pela bolsa de estudo concedida.

RESUMO

Silva APM. **Caracterização de novos transcritos humanos localizados no *locus* HPC1 (Hereditary Prostate Cancer) através de identificação *in silico* e validação experimental**. São Paulo; 2002. [Dissertação de Mestrado - Fundação Antonio Prudente].

A identificação de todos os genes humanos é crucial para entendermos diversos aspectos biológicos e evolutivos da nossa espécie. Em especial, a identificação de genes relacionados a doenças genéticas é fundamental para o desenvolvimento de diagnósticos moleculares e terapias alternativas associadas a essas doenças. Este projeto teve como objetivo a identificação e caracterização de genes humanos no *locus* associado ao câncer de próstata hereditário (HPC1) localizado na região 1q25. Para a identificação dos genes foi utilizada uma estratégia *in silico* baseada no alinhamento entre seqüências genômicas e de cDNA disponíveis nos bancos de dados públicos. Dezessete novos genes foram identificados nessa região (7 clones de cDNA completos e 10 ESTs), a grande maioria dos quais não havia sido anotada na seqüência genômica humana. Oito dos dezessete genes identificados são expressos em próstata e representam, portanto, novos candidatos ao câncer de próstata hereditário. A estratégia adotada nesse trabalho é bastante direta se comparada a estratégias tradicionais e se mostrou extremamente eficiente, permitindo um aumento de 60% no número de genes identificados nessa região. A expressão de dois dos genes identificados nesse trabalho (RGSL1 e RGSL2 que codificam proteínas da família RGS envolvidas na sinalização celular) foi testada por Real-Time PCR em tecido normal e linhagens tumorais de próstata. A expressão foi detectada apenas no tecido normal, sugerindo um possível envolvimento desses genes na doença que será futuramente estudado. * Os resultados do presente trabalho estão contidos no artigo científico, em anexo, submetido em junho de 2002 para a revista GENE.

ABSTRACT

Silva APM. **Caracterização de novos transcritos humanos localizados no *locus* HPC1 (Hereditary Prostate Cancer) através de identificação *in silico* e validação experimental** [Caracterization of new human transcripts within the HPC1 (Hereditary Prostate Cancer) *locus* through *in silico* identification and experimental validation]. São Paulo; 2002. [Dissertação de Mestrado - Fundação Antonio Prudente].

The identification of all human genes is crucial for the understanding of many aspects of the human biology and evolution. Particularly, the identification of genes related to genetic diseases is fundamental to the development of molecular diagnostic tools and alternative therapies for these diseases. The objective of this project was to identify and characterize additional human genes located at the Hereditary Prostate Cancer Locus (HPC1) on chromosome 1q25. An *in silico* strategy, based on the alignment between publicly available genomic and cDNA sequences was used for the identification of new genes located at 1q25. We were able to identify 17 novel genes (7 full-length cDNAa and 10 ESTs), the majority of which were not annotated in the human genome sequence. Eight out of the seventeen new genes are expressed in prostate tissue and represent alternative disease gene candidates for the HPC1 locus. The strategy adopted in this work is straightforward in comparison to other traditional strategies and has proven extremely efficient, increasing the total number of genes identified in this region by approximately 60%. Real-Time PCR was used to test the expression of two identified genes (RGSL1 and RGSL2 coding for proteins from the RGS family involved in cell signaling) in normal prostate tissue and prostate tumor cell lines. Expression was only detected in normal tissue suggesting a possible role for these genes in prostate cancer that deserves further studies.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

µg - Micrograma

µl - Microlitro

µM – Micromolar

ATCC – American Type Culture Collection

BACs – Cromossomo artificial de bactéria (do inglês *Bacterial Artificial Chromosomes*)

Blast - *Basic Local Alignment Search Tool*

cDNA - DNA complementar

CGAP – Projeto Anatomia Genômica do Câncer (do inglês *Cancer Genome Anatomy Project*)

dNTPs - deoxinucleotídeos

DEPC - Di-etil pirocarbonato

DHGP – *German Human Genome Project*

DMSO – Dimetilsulfóxido

DNA - Ácido desoxirribonucléico

DNase - Desoxirribonuclease

DO - Densidade óptica

DTT - Ditiotreitol

EDTA – Ácido etilenodiaminotetracetico disódio

ESTs – Etiquetas de seqüências expressas (do inglês *Expressed Sequence Tags*)

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

GAPDH - Gliceraldeído-3 fosfato desidrogenase

GAPS – GTPases Activating Proteins

GDP – Guanosina difosfato

GTP – Guanosina trifosfato

HCGP – Projeto Genoma Humano do Câncer

HGP – Projeto Genoma Humano

HPC1 – *Hereditary Prostate Cancer*

HTGS – Seqüências genômicas em larga escala (*High-Throughput Genome Sequences*)

ICPCG – *International Consortium on Prostate Cancer Genetics*

ILPC – Instituto Ludwig de Pesquisas sobre o Câncer

InterPro Scan – *Integrated search in PROSITE*

IPTG - iso – propil-β-D-tiogalactopiranosídeo

Kb - kilobases

MGC – *Mammalian Gene Collection*

mRNA - RNA mensageiro

Mb - megabases

NCBI – *National Center for Biotechnology Information*

NIH – *National Institutes of Health*

nm – nanômetros

ORESTES – *Open Reading Frame ESTs*

ORF – *Open Reading Frame*

pb - pares de bases

PCR - Reação em cadeia da polimerase (do inglês *polimerase chain reaction*)

PGH – Projeto Genoma Humano

pH – Potencial hidrogeniônico

Poli-A - cauda poli (A) do RNA mensageiro

PROSITE – Banco de dados de famílias e domínios protéicos

RACE – *Rapid Amplification of cDNA ends*

RGS – Reguladores da sinalização da proteína G (do inglês *Regulators of G Protein Signaling*)

RNA - Ácido ribonucléico

RNAse - Ribonuclease

rpm – rotações por minuto

RT-PCR - Reação em cadeia da polimerase precedida da reação de transcrição reversa (do inglês Reverse transcriptase – polymerase chain reaction)

SDS – Dodecilsulfato de sódio

SNPs – Polimorfismos de nucleotideos únicos (do inglês *Single Nucleotide Polymorphisms*)

TFI – *Transcript Finishing Initiative*

TIGR – *The Institute for Genomic Research*

UCSP map – *University of California Santa Cruz*

UV – Ultravioleta

X-GAL – 5-Bromo-4-chloro-3-indolyl-ß-D-Galactopyranoside

YACs – Cromossomos artificiais de levedura (do inglês *Yeast Artificial Chromosomes*)

ÍNDICE

Introdução

1 INTRODUÇÃO

1.1 O GENOMA HUMANO E A IDENTIFICAÇÃO DE SEUS GENES

O termo genoma foi cunhado há três quartos de século por Hans Winkler, um botânico Alemão, para designar o conjunto cromossômico haplóide dos eucariontes. O genoma humano é composto por 23 pares de cromossomos, contendo aproximadamente 3 bilhões de nucleotídeos e ao contrário dos genomas bacterianos e de eucariotos inferiores, possui uma organização complexa. Estima-se que as seqüências transcritas correspondam a apenas 3% do genoma sendo a maior parte do mesmo composta por seqüências repetitivas de função ainda desconhecida. Os genes humanos também apresentam estrutura complexa e descontínua, sendo as regiões codificantes (exons) interrompidas por regiões não codificantes (íntrons). Essas características tornam o seqüenciamento e a identificação de genes no genoma humano uma tarefa difícil.

O número exato de genes que compõem o genoma humano ainda não foi determinado, mas as estimativas estão entre 35.000 e 120.000 (ROEST et al., 2000; EWING e GREEN 2000; LIANG et al., 2000). Segundo o Unigene (Build#127), existem 22.253 genes humanos representados por seqüências completas de cDNA. O número reduzido de transcritos humanos com seqüências completas em relação ao total esperado se deve principalmente ao fato de que até alguns anos atrás, a grande maioria dos genes humanos era caracterizada de forma isolada e em pequena escala por grupos de pesquisadores interessados em responder perguntas específicas e grande parte deles buscava compreender os mecanismos moleculares associados à doenças genéticas. No entanto, na década de 90, com o desenvolvimento das técnicas de seqüenciamento e da bioinformática, esse posicionamento mudou e o anseio por desvendar todo o código genético humano vem reunindo cientistas de várias partes do mundo.

1.1.1 Identificação de genes humanos em pequena escala

Duas abordagens são comumente utilizadas para a identificação de genes ligados a doenças genéticas: a clonagem funcional e a clonagem posicional. O que determina a utilização de uma ou de outra técnica é o tipo e a quantidade de informações disponíveis a respeito do gene alvo, bem como a disponibilidade de mapas genéticos e físicos (BALLABIO et al., 1998).

1.1.1.1 Clonagem Funcional

A clonagem funcional é utilizada para a identificação de genes que codificam proteínas suposta ou sabidamente associadas aos mecanismos patofisiológicos em estudo (BALLABIO et al., 1998). A partir da seqüência de aminoácidos da proteína, uma vasta gama de técnicas (produção de anticorpos específicos e construção de iniciadores degenerados para o rastreamento de biblliotecas de cDNA) podem ser utilizadas, de forma a identificar a seqüência de nucleotídeos correspondente e determinar seu mapeamento (STRACHAN e READ 1997). Através da clonagem funcional foi possível, por exemplo, determinar a seqüência do gene da fenilalanina-hidroxilase em humanos (LEDLEY et al., 1985) e fazer o rastreamento de uma biblioteca de expressão para identificar o gene tirosinase em camundongos (YAMAMOTO et al., 1987).

1.1.1.2 Clonagem Posicional

Na estratégia posicional, percorre-se o caminho inverso daquele utilizado na clonagem funcional e, por essa razão, a técnica é também conhecida como genética reversa. Em outras palavras, a identificação do gene em questão costuma ser feita sem que haja conhecimento prévio a respeito da função do seu produto protéico, baseando-se apenas na localização de um *locus* mutante no genoma (BALLABIO et al., 1998; BISHOP 1996). Em seguida, genes candidatos isolados a partir do *locus*

mutante são testados para a presença de mutações em portadores da doença em questão.

A primeira etapa no processo de clonagem posicional é estabelecer a posição do gene alvo em um determinado intervalo cromossômico através da identificação de marcadores genéticos polimórficos, de localização cromossômica conhecida, que segreguem com a doença em famílias afetadas (PATTERSON e COOPER 1996). A estratégia clássica utilizada para a identificação desses marcadores, denominada estudo de ligação, baseia-se no fato de que dois *loci* localizados fisicamente próximos em um mesmo cromossomo têm maior probabilidade de serem herdados conjuntamente. A probabilidade de ocorrer recombinação entre dois *loci*, de forma que esses passem a segregar independentemente, é chamada fração de recombinação (EASTON 1996). Quanto maior a distância entre dois *loci* em um mesmo cromossomo maior a freqüência de recombinação. Assim sendo é possível verificar se existe ligação no padrão de herança dos marcadores e da doença genética em famílias afetadas e, desta forma, determinar a localização do *locus* ligado à doença.

Uma vez estabelecida a localização do *locus* cromossômico que abriga o gene alvo, os passos seguintes referem-se à elaboração de um mapa físico de clones genômicos cobrindo a região crítica e à identificação de genes candidatos nessa região. A identificação de genes em uma determinada região do DNA genômico pode ser alcançada através de abordagens diferentes e complementares. Algumas destas abordagens tais como a identificação de ilhas CpG, determinação de regiões conservadas entre diferentes espécies ("zooblots") e hibridação em "Northern blots", são utilizadas para detectar a presença de um gene na região genômica sem realmente isolar o dado gene (BALLABIO et al., 1998). Outras abordagens, tais como "exon trapping", "cDNA selection", rastreamento de bibliotecas de cDNA e seqüenciamento parcial de clones genômicos, permitem o isolamento e uma caracterização parcial da seqüência do gene em estudo, além de fornecer informações a respeito do seu produto protéico (BALLABIO et al., 1998). Alguns genes associados a importantes doenças como a Distrofia Muscular do tipo Duchene, a fibrose cística (CF), a neurofibromatose 1, a doença de Huntington (HD) entre outros foram isolados através da abordagem de clonagem posicional (STRACHAN e READ 1997). A clonagem posicional ainda é

uma abordagem não muito simples e dispendiosa. No entanto, como veremos a seguir, essas limitações vem sendo superadas com a disponibilidade de um número cada vez maior de seqüências de clones genômicos e de cDNA.

1.1.2 Identificação de genes humanos em larga escala

Existem basicamente duas estratégias para a identificação de genes humanos em larga escala: o seqüenciamento de clones de DNA genômico seguido de predição *in silico* e o seqüenciamento de clones de cDNA.

1.1.2.1 Identificação de genes através do seqüenciamento do DNA genômico

Nos últimos anos, com a iniciativa do seqüenciamento do genoma humano e a introdução de ferramentas computacionais, a identificação dos genes humanos está sendo feita em larga escala através de predição *in silico*. O seqüenciamento do genoma humano foi conduzido exclusivamente no meio acadêmico até 1998, quando um grupo do setor privado, representado pela empresa "Celera Genomics" iniciou um projeto paralelo utilizando uma estratégia de seqüenciamento alternativa.

A estratégia adotada pelo consórcio público pode ser dividida em duas etapas: (1) a construção de um mapa físico, composto por clones genômicos de alto peso molecular, que se sobrepõem e cobrem todo o genoma e (2) o seqüenciamento de um conjunto mínimo desses clones genômicos, orientados pelo mapeamento da etapa anterior, através de subclonagem e construção de mini-bibliotecas de shotgun (Figura 1). A seqüência genômica é então reconstituída através de montagem computacional dos fragmentos gerados a partir dos clones genômicos, levando em consideração a sobreposição entre os mesmos e a localização cromossômica dos clones genômicos (AACH et al., 2001; LANDER et al., 2001).

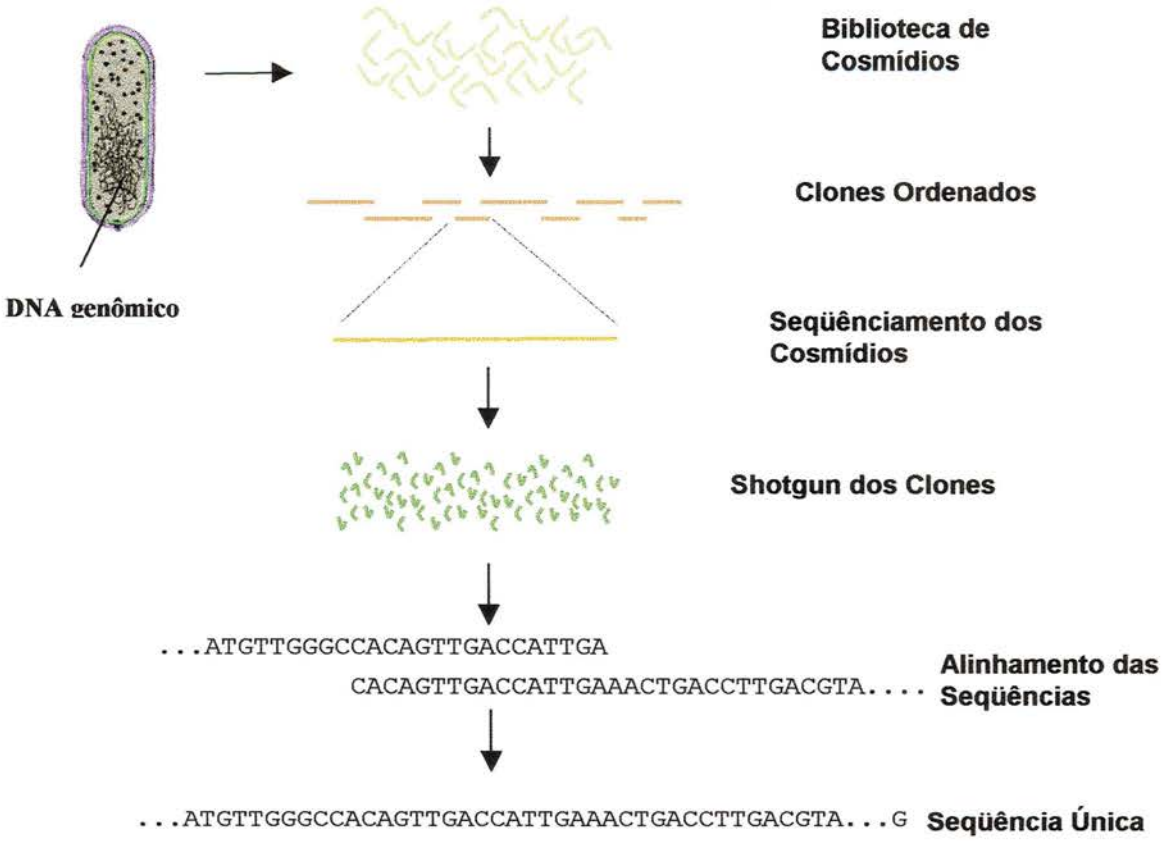
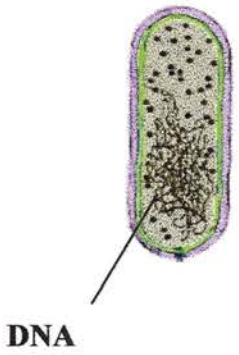


Figura 1 - **Estratégia de seqüenciamento hierárquico adotada pelo consórcio público:** Representação esquemática da estratégia adotada pelo consórcio público para o seqüenciamento do Genoma Humano.

Já o método de "whole genome shotgun", utilizado pela empresa "Celera Genomics", foi inicialmente utilizado no seqüenciamento de genomas bacterianos de menor complexidade. A estratégia consiste em quebrar a molécula de DNA em pequenos fragmentos (2-4 Kb) e em seguida seqüenciar um grande número desses fragmentos, o suficiente para cobrir o genoma em aproximadamente 10 vezes. A seqüência genômica consenso é então montada com base nas sobreposições das seqüências individuais de cada um dos fragmentos através do auxílio de algoritmos computacionais (Figura 2) (VENTER et al., 2001).



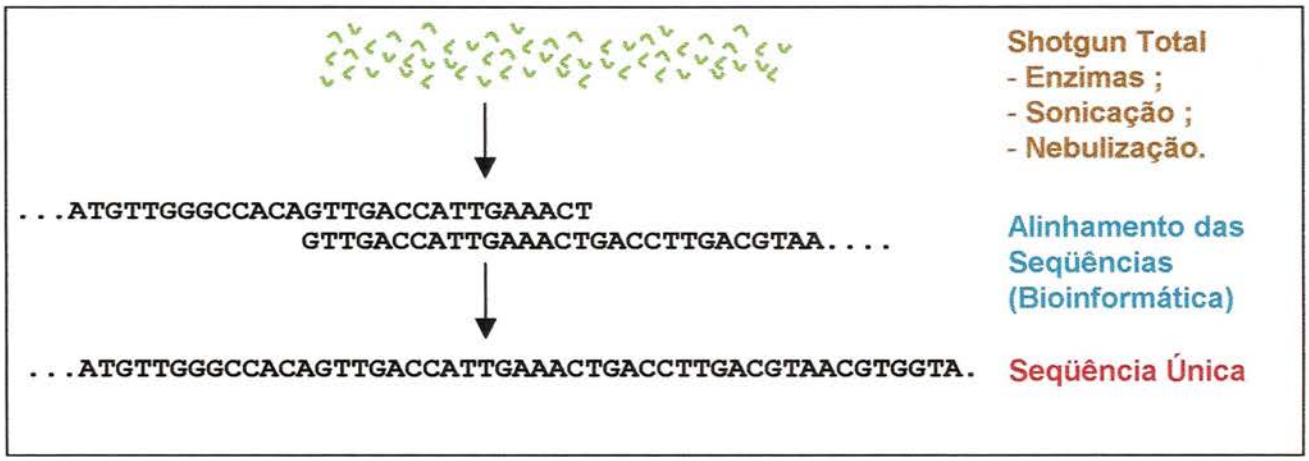


Figura 2 - **Estratégia de "shotgun" total adotada pela empresa Celera Genomics.** Representação esquemática da estratégia adotada pela empresa Celera Genomics para o seqüenciamento do Genoma Humano.

Apesar de terem utilizado estratégias de seqüenciamento diferentes, os dois grupos utilizaram uma abordagem muito semelhante para a identificação de genes a partir da seqüência genômica. A identificação foi feita inicialmente através do uso de programas de predição de genes na análise da seqüência genômica. Os programas de predição são desenhados para reconhecer na seqüência genômica estruturas características de genes humanos como, por exemplo, sinais de início da transcrição, sítios de "splicing" e fases abertas de leitura, entre outras. Em uma segunda etapa, somente as predições que puderam ser confirmadas através de evidência biológica foram mantidas. Como fonte de evidência biológica foram consideradas a presença de similaridade com seqüências humanas de cDNA completas e ESTs (expressed sequence tags) assim como similaridade com genes e proteínas de outros organismos. Um total aproximado de 30.000 genes humanos foi identificado por ambos os grupos utilizando essa estratégia (AACH et al., 2001, LANDER et al., 2001, VENTER et al., 2001).

Apesar de bastante direta, a identificação de genes *in silico* apresenta uma grande margem de erro e estima-se que 90% das estruturas gênicas preditas, por programas de computador, estejam incorretas. Análises preliminares com a seqüência completa do cromossomo 22 revelaram que, apesar de 94% dos transcritos humanos validados e localizados no cromossomo 22 terem sido identificados por programas de predição, apenas 20% desses tiveram sua estrutura corretamente predita e 16% dos exons experimentalmente validados não foram detectados *in silico* (DUNHAM et al., 1999). Em estudo semelhante conduzido por ROGIC et al. (2001), foi verificado que a confiabilidade de quatro programas de predição de genes comumente utilizados diminuía significativamente quando as regiões genômicas a serem estudadas eram muito extensas (GUIGO et al., 2000; ROGIC et al. 2001).

Recentemente, HOGENESCH et al. (2001) com o objetivo de comparar a anotação feita pelos dois grupos de seqüenciamento do genoma humano, realizou um estudo com um conjunto de genes de estrutura já bem caracterizada e possuindo seqüência completa de mRNA catalogada no banco de dados RefSeq (é um banco de dados não redundante que contém um grande número de seqüências completas de mRNA) e verificou que aproximadamente 84% desses transcritos foram identificados

por ambos os grupos. No entanto, quando a mesma comparação foi feita com o conjunto de genes definidos somente a partir da seqüência genômica, 80% dos transcritos foram identificados por apenas um dos dois grupos, evidenciando dessa forma as falhas existentes no processo de predição *in silico* de genes (CAMARGO et al., 2001; HOGENESCH et al., 2001).

O consenso atual é de que o baixo número de genes identificados no genoma humano se deve principalmente, a uma insuficiência de evidências biológicas que suportem as predições computacionais. Em outras palavras, está ficando cada vez mais evidente a necessidade de gerar seqüências de cDNA completas e ESTs ou até mesmo de seqüenciar, paralelamente, genomas de organismos evolutivamente relacionados com a espécie humana, tornando as análises *in silico* mais acuradas.

1.1.2.2 Identificação de genes através do seqüenciamento de clones de cDNA

Como mencionado anteriormente, os genes representam apenas 3% de todo o genoma humano, 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, correspondem à fração do genoma que carrega a informação dos genes. O cDNA obtido através da transcrição reversa pode ser inserido em vetores e replicado em células hospedeiras, dando origem às chamadas bibliotecas de cDNA.

Dois pontos muito importantes devem ser considerados na construção de bibliotecas de cDNA, pois representam potenciais limitações da metodologia em relação à caracterização de genes humanos. O primeiro diz respeito à eficiência da transcrição reversa e é de especial importância para a caracterização de transcritos longos, maiores que 5 kb. A síntese da molécula de cDNA é iniciada a partir da extremidade 3' da molécula de mRNA (cauda poli-A) e prossegue em direção à extremidade 5' da molécula. Se a eficiência da transcrição reversa não for satisfatória,

a porcentagem de moléculas de cDNA representando transcritos completos na biblioteca pode ser prejudicada.

Outro aspecto importante a ser considerado é a redundância da biblioteca de cDNA. Transcritos altamente expressos tendem a estar mais representados em bibliotecas de cDNA, de forma que, a partir de um certo ponto, o seqüenciamento de diferentes clones acrescenta muito pouco em termos de diversidade de genes caracterizados (BONALDO et al., 1996). Várias metodologias foram desenvolvidas para solucionar essa limitação, como por exemplo a normalização e subtração de bibliotecas por BONALDO et al. (1996); no entanto, elas são extremamente trabalhosas e requerem grandes quantidades de mRNA inicial que nem sempre é disponível.

Os clones de cDNA podem ser seqüenciados completa ou parcialmente a partir de suas extremidades 5' e 3' gerando as ESTs. Apesar de partirem do mesmo material (clones representados em bibliotecas de cDNA) as duas metodologias geram dados relativamente distintos e as limitações encontradas na caracterização de genes humanos são distintas como discutiremos a seguir.

1.1.2.2.1 Seqüenciamento completo de clones de cDNA

A estratégia mais direta e mais confiável para a determinação de transcritos humanos é o seqüenciamento completo de clones de cDNA, previamente selecionados por conterem transcritos inteiros. A seleção dos clones representativos que serão completamente seqüenciados é feita após o seqüenciamento parcial das extremidades 5' e 3' e análise computacional de um grande número de clones por biblioteca. Os insertos dos clones selecionados são então subclonados em fragmentos menores compatíveis com o seqüenciamento e posteriormente montados com o auxílio de programas de computador (STRAUSBERG et al., 1999). Vários projetos que visam o seqüenciamento completo de clones de cDNA estão em andamento como por exemplo o MGC (Mammalian Gene Collection) (STRAUSBERG et al., 1999). No entanto, a construção de bibliotecas enriquecidas para mRNAs completos e normalizadas em

relação a abundância dos transcritos é um processo trabalhoso, acarretando na diminuição da escala de produção. Em dois anos de projeto o MGC conseguiu isolar e seqüenciar clones de cDNA correspondentes a pouco mais de 9.000 genes humanos.

1.1.2.2.2 Seqüenciamento parcial de clones de cDNA: EST

A produção de ESTs teve suas origens em 1980, mas somente no início desta década elas começaram a ser produzidas em larga escala (ADAMS et al., 1995; KORENBERG et al., 1995; HILLIER et al., 1996). Até Junho de 2002, mais de 4,464,710 ESTs humanas obtidas a partir de diversos órgãos e tecidos foram depositadas no GeneBank, derivadas principalmente do "Merck Gene Index Project" (ECKMAN et al., 1998; WILLIAMSON 1999), do "Cancer Genome Anatomy Project" (CGAP) (STRAUSBERG et al., 2000) e do Projeto Genoma Humano do Câncer (HCGP). Devido a limitações técnicas da reação de seqüenciamento, as ESTs são geradas a partir das extremidades 5′ (26% das seqüências) ou 3′ (65% das seqüências) das moléculas de cDNA e, desta forma, as regiões centrais dos genes estão sub-representadas em bancos de dados de ESTs dificultando a reconstrução do transcrito completo a partir desses dados. Sabe-se que apenas 14% dos genes humanos já caracterizados estão representados por ESTs em toda a sua extensão (DE SOUZA et al., 2000).

Esta limitação vem sendo superada com o desenvolvimento de uma nova metodologia para a produção de ESTs denominada ORESTES ("Open Reading Frame ESTs") (DIAS NETO et al., 2000; DE SOUZA et al., 2000; CAMARGO et al., 2001). A nova metodologia permite que seqüências derivadas da região central dos transcritos sejam geradas em larga escala. A base desta abordagem é a amplificação por PCR de seqüências expressas utilizando iniciadores aleatórios sob condições de baixa estringência. Outras vantagens oferecidas pela técnica são a pequena quantidade de mRNA requerida para a construção das bibliotecas e a possibilidade de caracterização de transcritos pouco abundantes. Até o momento, mais de 1 milhão de ORESTES foram geradas dentro do Projeto Genoma Humano do Câncer, financiado

pelo Instituto Ludwig e pela FAPESP (BONALUME NETO 1999). Por serem derivados das regiões centrais dos transcritos, os dados de ORESTES complementam os dados de ESTs convencionais e possibilitam teoricamente a montagem de transcritos completos a partir de seqüências parciais com o auxílio de programas de computador.

Os programas de computador como por exemplo CAP3 (HUANG e MADAN 1999) e "TIGR Assembler" (QUACKENBUSH et al., 2000) identificam trechos de identidade e sobreposição entre a seqüência das ESTs, produzindo uma seqüência consenso ("contig") que representa o transcrito. Alguns grupos já utilizam essa estratégia para a construção de bancos de dados representando os diferentes transcritos humanos; entre eles, cabe destacar o Unigene (SCHULER 1997) e o "TIGR Gene Index" (QUACKENBUSH et al., 2000). No entanto, a grande limitação dessa metodologia diz respeito aos artefatos gerados durante o processo de montagem das seqüências. Os artefatos ocorrem principalmente devido a: i) baixa qualidade das seqüências, que dificulta a determinação precisa de sobreposições, ii) presença de formas alternativas de "splicing", que pode dificultar o agrupamento de seqüências derivadas de um mesmo gene, iii) existência de famílias gênicas com grande similaridade em nível de nucleotídeos e cujas seqüências podem ser artificialmente agrupadas em um mesmo consenso (QUACKENBUSH et al., 2000).

1.2 O PROJETO DE CARACTERIZAÇÃO DE TRANSCRITOS HUMANOS

O Projeto de Caracterização de Transcritos Humanos http://www.compbionet.org.br/transcript/ lançado recentemente pela FAPESP e o Instituto Ludwig, tem como objetivo gerar um banco de dados, validado experimentalmente, contendo informações sobre a maioria dos transcritos derivados do genoma humano. A estratégia proposta nesse projeto é complementar as abordagens empregadas por outros grupos, utilizando tanto seqüências de cDNA quanto seqüências genômicas disponíveis em banco de dados públicos. Ao contrário

dos projetos de seqüenciamento de clones de cDNA, o projeto irá gerar a seqüência de fragmentos parciais de cDNA evitando assim todo o trabalho relacionado com a preparação de bibliotecas de alta qualidade. Esses fragmentos parciais serão gerados através de RT-PCR de forma a validar e complementar a estrutura de genes humanos parcialmente representados por ESTs.

A estrutura dos diferentes transcritos foi definida *in silico* através do alinhamento entre seqüências de cDNA e seqüências genômicas e está sendo validada experimentalmente por RT-PCR e seqüenciamento. A definição *in silico* da estrutura dos diferentes transcritos foi feita em duas etapas distintas, no Laboratório de Bioinformática do Instituto Ludwig:

→ mapeamento das seqüências de cDNA na seqüência genômica disponível nos bancos de dados públicos e agrupamento dessas seqüências em diferentes "clusters" com base nos dados do mapeamento.

→ produção de "tags" de 50 nucleotídeos derivadas da extremidade 3'dos diferentes transcritos humanos que depois de mapeadas no genoma servem de âncoras para delimitar o final de transcritos.

O alinhamento entre seqüências transcritas e as seqüências genômicas foi feito com o programa BlastN. As coordenadas referentes aos exons e íntrons definidas através dos alinhamentos foram indexadas tanto em relação à seqüência genômica, quanto ao transcrito. Todas as informações sobre o mapeamento foram armazenadas em um banco de dados MySQL.

O agrupamento das seqüências transcritas em "clusters" foi feito com base nas coordenadas dos exons definidas na seqüência genômica, de forma que dois transcritos foram considerados parte de um mesmo "cluster" quando compartilhavam as coordenadas de pelo menos um exon. Utilizando esta estratégia foram definidos 124,855 "clusters", dos quais 10,538 incluíam seqüências de cDNA completas. Essa estratégia de agrupamento evita parcialmente os artefatos gerados através do alinhamento de seqüências transcritas, geralmente associados à baixa qualidade das seqüências, à presença de extensas famílias gênicas no genoma e formas alternativas de

"splicing". No entanto, artefatos referentes ao mapeamento de pseudogenes e seqüências contaminantes de DNA genômico presentes em bibliotecas de cDNA podem ser introduzidos no banco e devem ser cuidadosamente considerados no processo de validação.

A produção de "tags" derivadas das extremidades 3' dos diferentes transcritos é uma maneira prática e direta para definir a localização espacial dos transcritos no genoma, demarcando com precisão o final de um transcrito. Como "tags", foram escolhidos os últimos 50 nucleotídeos imediatamente adjacentes à cauda poli-A, de forma a garantir a individualidade de cada "tag" em relação aos diferentes transcritos. Após a exclusão das "tags" contendo elementos repetitivos, as demais foram então mapeadas no genoma através de alinhamentos e agrupadas em "clusters" de acordo com o seu mapeamento (ISELI et al., 2002). No total, foram identificadas 100.261 "tags", armazenadas no banco de dados MySQL.

1.3 PREDISPOSIÇÃO GENÉTICA PARA O CÂNCER

O estudo das bases moleculares do desenvolvimento de tumores gerou o conceito de que o câncer é uma doença genética causada pelo acúmulo de mutações em genes que codificam proteínas envolvidas principalmente na regulação da proliferação celular (VOGELSTEIN e KINZLE 1998). Os genes envolvidos na tumorigênese podem ser funcionalmente divididos em duas classes: os proto-oncogenes e os genes supressores de tumor.

Os proto-oncogenes, no geral, estão envolvidos na regulação positiva da proliferação celular e quando alterados são denominados oncogenes. Mutações em oncogenes estão freqüentemente associadas a um ganho de função e, portanto, são ditas dominantes em relação à tumorigênese. Basta a alteração de um único alelo para a manifestação do fenótipo celular alterado. Os genes supressores de tumor, ao contrário, atuam restringindo a proliferação e alterações nesses genes estão geralmente associadas a uma perda de função. Alterações em genes supressores de tumor são, portanto, recessivas em relação à manifestação do fenótipo maligno, havendo necessidade de inativação dos dois alelos (VOGELSTEIN e KINZLE 1998).

A grande maioria dos tumores é do tipo esporádico e não apresenta nenhuma relação familial. Esses tumores surgem como resultado de mutações em oncogenes e genes supressores de tumor que ocorrem nas células somáticas do indivíduo. Entretanto, uma pequena porcentagem dos tumores pode resultar inicialmente de mutações germinativas que passam a ser transmitidas de geração em geração. No geral, as síndromes de câncer familial estão associadas aos genes supressores de tumor. Acredita-se que, pelo fato de mutações nesses genes serem recessivas em relação à manifestação do fenótipo maligno, as mesmas são mais compatíveis com o desenvolvimento do embrião.

Embora sejam raros, os tipos de câncer familial trazem informações genéticas que são relevantes para o estudo e melhor entendimento dos cânceres esporádicos. O gene supressor de tumor Rb foi primeiro identificado através de sua inativação, em famílias portadoras de retinoblastoma e agora se sabe que mutações no gene Rb também ocorrem em diversos tipos de câncer esporádico. Um outro exemplo são as mutações herdadas no gene APC e que estão envolvidas tanto na forma familial quanto na forma esporádica do câncer de cólon (KING 2000).

1.3.1 Câncer de próstata hereditário

O câncer de próstata representa um sério problema de saúde pública no Brasil, em função de suas altas taxas de incidência e mortalidade. Esse tipo de câncer é o segundo mais comum em homens, sendo superado pelo câncer de pele. Segundo as estimativas de incidência e mortalidade por câncer no Brasil do Instituto Nacional de Câncer, deverão ocorrer 25.600 novos casos de câncer de próstata em 2002. Em taxa de mortalidade, o câncer de próstata aparece como o quarto tipo mais mortal, sendo responsável por 7.223 óbitos em 1999. Enquanto a incidência está ligada às características demográficas da população, a alta taxa de mortalidade é causada principalmente pelo retardo do diagnóstico, que favorece a ocorrência de tumores com alta capacidade biológica de invasão local e de disseminação para outros órgãos. Tais tumores são incuráveis quando tratados em fase metastática (www.inca.gov.br).

Embora homens de qualquer idade possam desenvolver câncer de próstata, a doença costuma afetar homens com mais de 50 anos de idade. Na realidade, mais de 70% de todos os casos são diagnosticados em homens com idades acima de 65 anos. Esse tipo de câncer é menos comum na Ásia, África e América do Sul (http://www.cancer.org). Mundialmente, sua incidência está aumentando regularmente em cerca de 3% ao ano. Grandes variações tanto na incidência quanto na mortalidade desse tipo de câncer têm sido observadas entre países e grupos étnicos. A incidência do câncer de próstata está relacionada à idade e não há nenhuma outra evidência que correlacione a doença a fatores ambientais ou relacionados ao estilo de vida dos indivíduos afetados (CUSSENOT e VALERI 2001).

Dados de história familial têm sugerido a existência de um componente genético na etiologia do câncer de próstata. Estudos realizados em irmãos gêmeos também tem proporcionado evidência adicional de que em determinados casos, o câncer de próstata seja de origem hereditária (CUSSENOT e VALERI 2001). No entanto a caracterização desses casos não é trivial devido ao elevado número de casos esporádicos de câncer de próstata que podem ocorrer em uma mesma família. Para o diagnóstico de câncer de próstata hereditário é necessária a presença de pelo menos três membros de primeiro grau afetados pela doença ou dois parentes diagnosticados antes dos 55 anos de idade (CARTER et al., 1993; CUSSENOT e VALERI 2001).

Com o objetivo de identificar genes relacionados ao câncer de próstata hereditário diversos estudos de ligação foram realizados. Esses estudos conduziram à identificação de pelo menos dez *loci* cromossômicos, associados com a predisposição a esse tipo de câncer sendo três no cromossomo 1, (HPC1, PcaP e CAPB) e um *locus* em cada um dos cromossomos X, 2q, 8p, 12p, 15q, 16p, 16q e 20q13 (XU et al., 1998; Suarez et al., 2000; Berry et al., 2000; SOOD et al., 2001; PLAETKE et al., 2002; XU et al., 2002). O primeiro *locus* mapeado, em 1996, foi o HPC1 ("hereditary prostate cancer" 1) localizado na região 1q24-25, através de estudos de ligação (CUSSENOT e VALERI 2001; DE LA CHAPELLE e PELTOMÄKI 1998; ISAACS e COFFEY 1999). Este *locus* foi identificado em um estudo incluindo famílias norte-americanas e suíças com pelo menos três indivíduos afetados em cada uma. No entanto, uma extensão desta análise, utilizando um grande número de famílias de

mesma origem, mostrou que um subgrupo de famílias diagnosticadas em idade precoce e com pelo menos cinco casos eram os responsáveis pela evidência de ligação (CUSSENOT e VALERI 2001).

Muitas outras análises de ligação foram realizadas com o objetivo de confirmar esta localização, no entanto os resultados sempre se mostraram controversos (CUSSENOT e VALERI 2001). A mais forte confirmação para esta localização foi obtida quando famílias de Utah foram examinadas (CUSSENOT e VALERI 2001). Utah apresenta uma taxa de casos de câncer de próstata aumentada em 1.2 vezes quando comparada ao restante dos Estados Unidos (EELES e CANNON-ALBRIGHT 1996). Essas famílias apresentaram um "LOD score" de 2.43 e outra vez o diagnóstico em idade precoce foi o que proporcionou a forte evidência da ocorrência de ligação no *locus* HPC1. Recentemente, o ICPCG ("International Consortium on Prostate Cancer Genetics") realizou uma grande análise de ligação em 772 famílias de diversas origens e encontraram que 6% dessas apresentavam ligação para o *locus* 1q25 (CUSSENOT e VALERI 2001).

1.3.2 A região 1q25 e os genes candidatos

Vários genes associados a doenças humanas têm sido mapeados no intervalo cromossômico 1q24-q31. Dentre eles, destaca-se o *locus* HPC1, associado a uma forma de câncer de próstata hereditário (SMITH et al., 1996), mapeado através de estudo de ligação. Outros exemplos são a síndrome de pericardite, artropatia e campodactilia (BAHABRI et al., 1998) e uma das síndromes de hiperparatiroidismo (HOBBS et al., 1999). A caracterização dos transcritos codificados nessa região representa uma etapa fundamental para a identificação dos genes associados às doenças citadas acima.

Recentemente, CARPTEN et al. (2000) realizaram o mapeamento físico da região 1q24-31. Dois tipos de "contigs" foram construídos: o primeiro deles, constituído apenas por fragmentos genômicos clonados em cromossomos artificiais de levedura (YACs), cobre uma região de 20 Mb em 1q24-q31; o segundo, composto por

um conjunto de clones sob a forma de cromossomos artificiais de bactéria (BACs), abrange um segmento de 6 Mb em 1q25 e está sendo seqüenciado pelo "Sanger Genome Sequencing Center", como parte do Projeto Genoma Humano (CARPTEN et al., 2000).

Esse mesmo grupo deu início a identificação de genes e ESTs já mapeados na região 1q24-31 e a identificação de novas seqüências transcritas a partir dos clones genômicos que compõem o "contig" de 6 Mb em 1q25 (CARPTEN et al., 2000). Inicialmente, foi selecionado um grupo de genes e ESTs da região de interesse através da consulta a um banco de dados de mapeamento experimental de ESTs através da utilização de painéis de híbridos (DELOUKAS et al., 1998). Além disso, vários outros transcritos dessa região foram identificados através de "exon trapping", "cDNA selection" e seqüenciamento amostral a partir de alguns BACs da região 1q25. É interessante notar que, segundo os resultados de Blast, todos estes transcritos já apresentavam pelo menos uma EST correspondente nos bancos de dados públicos.

As abordagens citadas acima permitiram a identificação e o mapeamento preciso de 18 genes já conhecidos e 31 ESTs no "contig" de 6 Mb da região 1q25 e de outras 11 seqüências transcritas no "contig" de YACs que cobre o intervalo 1q24-q31. No entanto, a determinação da seqüência de nucleotídeos completa destes cDNAs, assim como a identificação de novos transcritos constituem uma etapa fundamental para que estes dados possam ser utilizados pelos grupos de pesquisa que estudam as doenças mapeadas nesta região do cromossomo 1 (CARPTEN et al., 2000).

Em 12 de janeiro de 2001, o grupo de CARPTEN et al., (2000), relatou a caracterização completa de 13 ESTs previamente identificadas pelo grupo e do gene RGS8 humano homólogo ao gene RGS8 de rato. O trabalho também apresenta dados relacionados ao padrão de expressão dos transcritos em tecidos diversos, a organização genômica e a homologia desses transcritos com genes já conhecidos. Dos 13 novos transcritos, 10 estão expressos em próstata e representam prováveis candidatos a gene supressor de tumor relacionado ao câncer de próstata hereditário (SOOD et al., 2001).

Recentemente, CARPTEN et al. (2002), relatou que mutações no gene RNASEL segregam independentemente em 2 de 8 famílias selecionadas para o *locus*

HPC1, sugerindo o envolvimento desse gene no câncer de próstata hereditário. O gene RNASEL (2',5' oligoisoadenylate synthetase dependent) codifica uma ribonuclease que atua regulando a proliferação celular e a apoptose e é apontado como um candidato a gene supressor de tumor (WANG et al., 2002; RÖKMAN et al., 2002). Entretanto, alelos inativos do gene RNASEL estão presentes em uma baixa freqüência na população geral. E em uma das famílias que apresenta a mutação, apenas 4 dos 6 indivíduos são portadores da referida mutação. Por essas razões, a identificação de outras mutações funcionalmente significativas ainda é necessária para a comprovação do envolvimento do gene RNASEL no câncer de próstata hereditário. Enquanto isso, a existência de outros genes na região 1q25 relacionados a doença não pode ser descartada (CARPTEN et al., 2002; RÖKMAN et al., 2002).

Em um estudo semelhante, Rökman e colaboradores fizeram uma triagem em diversos indivíduos portadores e não portadores de câncer de próstata hereditário, a fim de avaliar a presença de mutação germinativa no gene da RNASEL. A mutação truncada, E265X, foi encontrada em 5 (4.3%) dos 116 pacientes oriundos de famílias afetadas pelo HPC. A freqüência da mutação foi bastante significativa quando comparada a freqüência de 1.8% encontrada no grupo controle. A maior freqüência de ocorrência da mutação E265X (9.5%) foi encontrada em pacientes de famílias com quatro ou mais indivíduos afetados. No entanto, a segregação foi detectada apenas em uma família com três indivíduos afetados. A média de idade de início da doença para portadores da mutação foi diminuída em 11 anos em relação aos indivíduos não portadores da mesma família. Além disso, das quatro variantes "missense" encontradas, a variante R462Q mostrou associação com HPC. Nenhuma das variantes mostrou qualquer diferença entre o grupo controle, o grupo dos indivíduos com hiperplasia benigna prostática ou com câncer de próstata não caracterizado. Embora as mutações no gene da RNASEL não expliquem a segregação da doença nas famílias com HPC, as formas variantes estão presentes em famílias com HPC que possuem mais do que 2 indivíduos afetados podendo estar associadas com a idade de início da doença (RÖKMAN et al., 2002).

1.4 SINALIZAÇÃO CELULAR E RGS

1.4.1 A sinalização celular e o câncer

A sobrevivência nos organismos multicelulares depende de uma elaborada rede de comunicação intercelular que coordena o crescimento, a diferenciação e o metabolismo das células em diversos tecidos e órgãos. Um importante mecanismo através do qual as células podem se comunicar é através de moléculas de sinalização extracelular. A resposta de uma célula a uma molécula de sinalização depende de sua ligação a receptores específicos que podem estar localizados na superfície da célula alvo, no núcleo ou no citosol. A ligação da molécula de sinalização ao receptor promove uma mudança na conformação do mesmo, a qual, por sua vez, é responsável por desencadear uma seqüência de reações envolvendo enzimas efetoras e mensageiros secundários que transduzem e amplificam o sinal até o núcleo da célula e promovem uma mudança na função celular através de alterações no padrão de expressão gênica (LODISH et al., 1995; KING 2000).

O funcionamento da maquinaria de sinalização extracelular se encontra freqüentemente alterado nas células tumorais, tornando-as independentes da sinalização para a execução de suas atividades, principalmente no que se refere à proliferação celular e à diferenciação. Essa independência está associada a um aumento na sensibilidade aos fatores de crescimento e diminuição da resposta aos sinais inibitórios de células adjacentes e da matriz extracelular. O oncogene *ras*, por exemplo, envolvido na transdução de sinal é um dos mais freqüentemente alterados em tumores (THOMPSON et al., 1999). As alterações em *ras*, podem ser mutações envolvendo o próprio gene ou proteínas que regulam sua função. Um outro exemplo de oncogene associado a sinalização celular é o receptor *neu*. Esse tipo de receptor de superfície celular apresenta um domínio citoplasmático com atividade tirosina-kinase que transmite o sinal de crescimento dando início à cascata de sinalização através da fosforilação dos resíduos de tirosina. Uma única alteração, em um aminoácido (valina→glutamina) presente no domínio transmembrana do receptor codificado pelo

gene *neu* torna-o constitutivamente ativo mesmo na ausência do ligante (LODISH et al., 1995).

1.4.2 Receptores acoplados à proteína G

Os receptores acoplados à proteína G estão localizados na superfície celular e interagem com ligantes solúveis em água como, por exemplo, a epinephrina, a serotonina e o glucagon. Embora esses receptores se liguem a diferentes ligantes e possam mediar respostas celulares bastante distintas, eles apresentam uma estrutura protéica bastante similar composta por sete alças de 22-24 resíduos hidrofóbicos que formam sete α-hélices transmembrânicas. A alça localizada entre as α-hélices 5 e 6 e o segmento C-terminal é responsável pela interação física com a proteína G.

As proteínas G são compostas por três subunidades designadas α, β e γ que estão complexadas quando a proteína está inativa e atuam como um "interruptor" molecular, o qual está no estado inativado quando ligado ao GDP, e ativado quando ligado ao GTP. A interação ligante-receptor faz com que a subunidade α libere o GDP e se ligue ao GTP, tornando-a ativa. Neste momento a subunidade α se dissocia das demais subunidades e se liga a uma enzima efetora (por exemplo a adenilato ciclase) inibindo ou estimulando a formação de um mensageiro secundário. A ativação da proteína G é rápida, porque o GTP ligado a subunidade α é hidrolizado a GDP em segundos, promovendo a reassociação das subunidades α, β e γ e conseqüentemente inativando a enzima (LODISH et al., 1995; ROSS e WILKIE 2000). Dessa forma, a sinalização mediada pela proteína G é intrinsecamente cinética. A amplitude do sinal é determinada pelo balanço das taxas de troca GDP/GTP (ativação) e das taxas de hidrólise de GTP (inativação).

1.4.3 Inativação da proteína G e as RGSs

A inativação das proteínas G pode ser estimulada por proteínas que aceleram a hidrólise do GTP e são comumente denominadas proteínas ativadoras de GTPases (GTPases Activating Proteins – GAPs). As GAPs foram inicialmente identificadas em estudos genéticos em leveduras, *C. elegans* e *Aspergillus nidulans* que tinham como objetivo identificar proteínas capazes de inibir a sinalização das proteínas G. Várias GAPs foram identificadas e dentre elas um grupo denominado RGS (Regulator of G-Protein Signaling) que apresentava grande similaridade em uma região da proteína composta por 120 amino-acidos e denominada domínio RGS. Hoje sabe-se que o domínio RGS é capaz de se ligar a subunidade α da proteína G acelerando a hidrólise de GTP. As proteínas RGS foram posteriormente caracterizadas em eucariotos superiores e no caso de mamíferos cerca de 20 proteínas RGS já foram identificadas (KOELLE 1997).

As proteínas que compõem a família RGS podem ser divididas em dois grupos: um grupo composto basicamente por proteínas que apresentam apenas o domínio RGS e outro que apresenta, além do domínio RGS, outros domínios capazes de interagir com outras proteínas. As RGSs atuam inibindo a cascata de sinalização desencadeada pela proteína G e estão envolvidas na modulação de uma variedade de funções celulares tais como: proliferação, diferenciação, e resposta a neurotransmissores (DE VRIES e FARQUHAR 1999).

Evidências da atuação das RGS no processo de desenvolvimento tumoral começam a aparecer na literatura, despertando nosso interesse por essa família de proteínas. Recentemente, um novo membro da família RGS, a RGS14, foi caracterizado como sendo um gene induzido por p53 em resposta ao "stress" genotóxico. Foi também demonstrado que a superexpressão da proteína RGS14 inibe ambas subunidades da proteína G acoplada a um receptor de fator de crescimento regulando a sinalização mitogênica desencadeada por enzimas da família MAP-kinases (BUCKBINDER et al., 1997).

Justificativa

2 JUSTIFICATIVA

A identificação de todos os genes é o principal objetivo do Projeto Genoma Humano. No entanto, devido à complexidade da organização do genoma humano, várias abordagens complementares vêm sendo utilizadas por diferentes grupos. Certamente, uma das áreas que mais se beneficiará com esses resultados é a medicina (BRODER e VENTER 2000). Atualmente, com todos os dados de seqüência disponíveis, é possível acessar os genes localizados em uma determinada região de uma maneira muito mais direta e rápida.

Grande avanço já foi obtido com o mapeamento experimental em larga escala de seqüências transcritas, mas com a disponibilidade de um número cada vez maior de seqüências genômicas e transcritas (ESTs e cDNAs), esta metodologia tende a se tornar obsoleta. Com as ferramentas computacionais disponíveis, já é possível localizar todas as seqüências transcritas no DNA genômico de maneira rápida e eficiente. Esta é a proposta do Projeto de Caracterização de Transcritos Humanos Ludwig / FAPESP, que visa construir e validar um banco de dados de todos os transcritos derivados do genoma humano. Desta forma, através de uma simples consulta a este banco de dados, será possível identificar de uma forma muito precisa seqüências transcritas localizadas em uma região genômica de interesse, como por exemplo a região 1q25.

O primeiro objetivo deste projeto é utilizar o banco de dados do Projeto de Caracterização de Transcritos Humanos para identificar os transcritos localizados na região 1q25. Esta região foi escolhida por estar associada a uma forma hereditária de câncer de próstata e a outras doenças genéticas, além de ter sido extensivamente caracterizada por outras metodologias. Isso nos permitirá avaliar a confiabilidade da estratégia utilizada para a construção do banco de dados, assim como identificar novos transcritos ainda não caracterizados por outras metodologias e localizados nessa região. O segundo objetivo deste projeto é a validação experimental e produção de seqüências completas destes transcritos, que constitui o primeiro passo para a identificação e caracterização de genes candidatos para as diversas doenças genéticas mapedas em 1q25.

Objetivos

3 OBJETIVOS

- Identificar os transcritos localizados na região1q25 de acordo com o banco de dados do Transcriptoma Humano.
- Comparar os resultados obtidos a partir das buscas no banco de dados do Transcriptoma Humano com os dados de transcritos mapeados na mesma região por outras estratégias e descritos na literatura.
- Verificar através de RT-PCR a expressão em próstata dos transcritos mapeados na região.
- Gerar a seqüência completa de prováveis candidatos a gene supressor de tumor mapeados na região 1q25.
- Gerar a anotação completa dos transcritos identificados na região 1q25 a fim de levantar genes candidatos ligados ao *locus* HPC1 envolvido no câncer de próstata hereditário.

Materiais & Métodos

4 MATERIAS E MÉTODOS

4.1 ANÁLISE DO BANCO DE DADOS DO PROJETO TRANSCRIPTOMA HUMANO

4.1.1 Identificação de clones genômicos da região 1q25

Os clones genômicos foram identificados através de buscas em bancos de dados do Projeto Genoma Humano (HGP) utilizando o programa BlastN. Como "query" foram utilizadas seqüências de marcadores genéticos, D1S2818 (Z54047) e D1S1642 (G09777), para a região 1q25 descritas no trabalho de CARPTEN et al. (2000). A disposição física dos clones genômicos que cobrem a região 1q25 foi determinada através da ferramenta Map Viewer (versão 02) disponível via web no Centro Nacional de Informação em Biotecnologia (NCBI) (http://www.ncbi.nlm.nih.gov/genome/guide/).

4.1.2 Buscas no banco de dados do Projeto Transcriptoma Humano

A seqüência genômica correspondente a cada um dos clones selecionados foi obtida a partir do banco de dados EMBL (European Molecular Biology Laboratory) versão 66. Os dados correspondentes às seqüências transcritas, foram obtidos a partir de diversas fontes: (1) seção de ESTs humanas do banco de dados EMBL versão 66; (2) seqüências de mRNA humano presentes no EMBL versão 66; (3) seqüências de ORESTES provenientes do Projeto Genoma Humano do câncer LICR/FAPESP (CAMARGO et al., 2001; DIAS NETO et al., 2000) e (4) seqüências de mRNA humano presentes no banco de dados RefSeq disponibilizado pelo NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). O Blast (disponibilizado pelo NCBI) foi utilizado para identificar similaridade entre todas as seqüências transcritas e os dados genômicos.

As seqüências transcritas foram filtradas quanto a presença de contaminação e os elementos repetitivos foram mascarados usando o "software" PFP (Paracel, Pasadena, CA). Para cada par correspondente de seqüências transcritas e genômicas, foi feito um alinhamento local utilizando o programa Sim4 (FLOREA et al., 1998), com parâmetros W=15 R=0 A=4 P=1 a fim de que as bordas entre íntrons e éxons, assim como a direção do transcrito em relação à seqüência genômica fossem definidas. O dado obtido a partir do Sim4 foi filtrado para eliminar todos os alinhamentos que não continham pelo menos uma região correspondente com pelo menos 95% de identidade acima de 30 nucleotídeos. Para a extração das "tags" foram utilizados os dados dos cromatogramas, evitando problemas decorrentes da qualidade de seqüência e a constante remoção da cauda de poli-A durante a submissão das seqüências de ESTs aos bancos de dados.

As coordenadas dos alinhamentos e informações relacionadas foram armazenadas em um banco de dados relacional MySQL.

Nesse tipo de banco, os dados ficam armazenados em tabelas de fácil correlação. Dessa forma é possível identificar quais ESTs e seqüências completas de mRNAs estão localizadas em um determinado clone gênomico, qual a posição exata do alinhamento, e qual a posição dessas seqüências em relação a uma 3′ "tag".

Os dados armazenados no banco de dados relacional, foram utilizados para criar "clusters" de seqüências transcritas baseadas nas posições de cada um dentro dos clones genômicos individuais. Os membros de cada "cluster" foram determinados através das coordenadas dos éxons em relação a seqüência genômica. Se as coordenadas de pelo menos um éxon fossem comuns a dois transcritos, então os mesmos eram considerados parte de um mesmo "cluster".

O número de acesso do GeneBank dos clones genômicos localizados na região 1q25 serviu para iniciar todas as buscas no banco de dados. Foram realizadas dois tipos de buscas com os seguintes objetivos: selecionar clones completos de mRNA (Full-length) mapeados pela primeira vez na região 1q25 e que portanto não haviam sido descritos pelo grupo do NIH e selecionar seqüências parciais de transcritos (ESTs) também mapeadas pela primeira vez na região. As buscas foram realizadas

através de "selects" específicos para cada uma das informações desejadas. Os selects utilizados estão abaixo relacionados.

- Selects utilizados para a obtenção de dados relacionados aos clones genômicos da região 1q25:
  1. Número de seqüências mapeadas em um determinado clone genômico – *select distinct gene_id from sequences where genomic_id like "AC019075%";*

  2. Número de ESTs mapeadas em um determinado clone genômico - *select distinct gene_id from sequences where genomic_id like "AC019075%" and full_length="E";*

  3. Número de "Clusters" de sequências mapeados em um determinado clone genômico - *select count(cluster) from clusters where genomic_id like "AC019075%";*

  4. Número de Clusters contendo Full-Lenght - *select count(cluster) from clusters where genomic_id like "AC019075%" and full_length="Y";*

  5. Número de Clusters contendo ESTs - *select count(cluster) from clusters where genomic_id like "AC019075%" and full_length="n";*

  6. Número de Clusters contendo ESTs com Splicing - *select count(cluster) from clusters where genomic_id like "AC019075%" and full_length="n" and splicing="y";*

  7. Número Total de ESTs com splicing – *select distinct gene_id from sequences where genomic_id like "AC019075%" and full_length="E" and splicing="Y";*

4.1.3 Inspeção manual das seqüências de cDNAs completas e ESTs Selecionadas

Os dados extraídos do banco do Projeto Transcriptoma Humano foram manualmente revistos para evitar possíveis artefatos gerados durante o processo de alinhamento. Para tanto, os alinhamentos entre as seqüências expressas identificadas em nosso banco de dados e a seqüência do genoma humano foram repetidos manualmente. Durante esse processo foram eliminadas seqüências expressas que apresentaram melhor alinhamento com clones genômicos localizados fora do intervalo 1q25 assim como ESTs que quando alinhadas contra a seqüência genômica não apresentaram estrutura de "splicing".

4.2 VALIDAÇÃO EXPERIMENTAL DOS TRANSCRITOS

A expressão dos transcritos selecionados foi testada através de RT-PCR. Nas reações foram utilizados cDNAs de tecido normal de próstata além de cDNAs correspondentes aos tecidos de origem das diferentes seqüências expressas selecionadas e linhagens tumorais de próstata. Dezenove outros tecidos normais humanos (testículo, pulmão, intestino, coração, útero, medula óssea, placenta, cólon, cérebro fetal, fígado, fígado fetal, timo, glândula salivar, cordão espinhal, rim, baço, músculo esquelético, traquéia e glândula adrenal) foram comprados da Clontech (Palo Alto, CA) e utilizados para a síntese de cDNA. Os produtos da RT-PCR foram então clonados e seqüenciados para confirmar a especificidade da amplificação.

4.2.1 Linhagens Celulares

As linhagens tumorais de próstata DU145 e PC3 foram obtidas através da ATCC (American Type Culture Collection) e cultivadas em meio apropriado até obtenção de massa suficiente ($3X10^6$ células) para a obtenção de RNA.

4.2.2 Extração de RNA total e controle de contaminação com DNA Genômico

A extração de RNA foi feita a partir das linhagens celulares acima descritas através da técnica de Sedimentação em Cloreto de Césio. As células foram homogeneizadas em 9 ml de solução de lise (4M de isotiocianato de guanidina / 25mM citrato – pH 7.0 / 0.1M β-mercapto). O lisado foi, em seguida, aplicado sobre 4 ml de um gradiente de cloreto de césio (5.7M CsCl e 25mM de NaAc) e centrifugado a 29.000 rpm em ultracentrífuga Beckman utilizando o rotor SW40Ti durante 17 horas a 20°C. Após a centrifugação, o RNA formou um precipitado no fundo do tubo que foi solubilizado em 50 a 200µl de água biodestilada tratada com DEPC.

A quantificação do RNA foi feita através da leitura em espectrofotômetro com comprimento de onda equivalente a 260nm, considerando-se que 1DO260nm equivale a 40µg/ml de RNA. A relação entre as leituras realizadas a 260 e 280nm foi utilizada como parâmetro na estimativa do grau de contaminação do RNA por proteínas. Para que seja considerado puro, o valor desta relação deve ser superior a 2.

Para se ter a garantia de que o RNA extraído não estava contaminado com DNA genômico, aplicamos o teste de hMLH1. Esta reação de PCR utiliza iniciadores desenhados nos introns 12 e 13 do gene MLH1 (Forward - 5'TGG TGT CTC TAG TTC TGG3' e Reverse - 5'CAT TGT TGT AGT AGC TCT GC3') e 200ng de RNA total como molde.

A reação teve 35 ciclos com uma temperatura de "annealing" de 55°C e um tempo de extensão de 45 segundos. Caso o RNA estivesse contaminado com DNA seria amplificado um fragmento de aproximadamente 250pb. Nas amostras em que foi detectada contaminação foi aplicado, então, o tratamento com DNAse e, logo após, as amostras foram novamente submetidas ao teste de hMLH1.

4.2.3 Tratamento do RNA total com DNAse I

O RNA total, quando necessário, foi tratado com DNAse livre de RNAse (1U/µl, Promega), utilizando 10 unidades da enzima para tratar 50µg de RNA por 60 minutos a 37°C. A reação foi interrompida pela adição de uma solução de neutralização (2,5 µl EDTA 0,5M, 12,5µl NaOAC 3M, 1,25µl SDS 20% e água q.s.p. 25µl), e o RNA foi extraído com fenol/clorofórmio, e precipitado a -20°C por 18 horas após adição de NaCl 250mM e etanol 100% com um volume igual a duas vezes o volume inicial. Após centrifugação, o RNA foi solubilizado em 30,0µl de água tratada com DEPC. A quantificação foi feita através da leitura de uma alíquota da amostra no espectrofotômetro a 260 e 280 nm conforme descrito anteriormente.

4.2.4 RT-PCR

A síntese da primeira fita de cDNA foi realizada segundo especificações do fornecedor da enzima SuperScript II (GibcoBRL). Um micrograma de RNA total livre de DNA genômico foi misturado com 500ng de oligo (dT)12-18 e o volume foi acertado para 12µl. Esta mistura foi aquecida a 70°C por 10 minutos e rapidamente resfriada em gelo. Em seguida a reação foi processada em presença de 1X tampão de obtenção da primeira fita, 10mM de DTT, 500uM de dNTPs e 200 unidades de transcriptase reversa (SuperScript II, GibcoBRL). A reação foi processada a 42°C durante 50 minutos. A reação foi inativada através do aquecimento a 70°C durante 15 minutos.

Na PCR, foram utilizados 2µl da reação de cDNA para um volume final igual a 50µl. A PCR foi realizada em presença de tampão 1X, 1,5mM de $MgCl_2$, 0,2mM de dNTPs, 0,2uM de cada iniciador (sense e antisense específicos para cada transcrito) e 0,025 U/µl de Taq DNA polimerase. As reações ocorreram a repetidos ciclos de 94°C/30seg., 55°C/45seg., 72°C/1min. precedidos por um ciclo inicial de 94°C/4min. e seguidos por um ciclo final de extensão de 72°C/7min.. As tabelas abaixo contêm a seqüência de cada um dos pares de primers utilizados para a validação experimental

das ESTs candidatas e das seqüências completas de mRNA respectivamente. Para a validação do padrão de expressão das seqüências completas de mRNA e para algumas das ESTs foi adotada a estratégia de Nested RT-PCR. Para tanto, 1,0 µl do produto da primeira PCR foi utilizado como molde juntamente com iniciadores internos em um volume final igual a 25µl. A segunda PCR foi realizada em presença de tampão 1X, 1,5mM de $MgCl_2$, 1,25mM de dNTPs, 0,2 uM de cada iniciador (sense e antisense específicos para cada transcrito) e 0,025 U/µl de Taq DNA polimerase. As reações ocorreram a repetidos ciclos de 94°C/30seg., 65°C/45seg., 72°C/1min. precedidos por um ciclo inicial de 94°C/4minutos e seguidos por um ciclo final de extensão de 72°C/6minutos.

cDNA completo	Primers
Full-length 01 DKFZp43E169	5'GGATTTCTGGATCCTTGCTG3'(F) 5'TGGGTGTAGCAAACGCTG3'(R)
Full-length 01 DKFZp43E169	Nested: 5'AGGGCTCCAGGGTGGTAAC3'(F) 5'CCACTTTCTGCATGTGGC3'(R)
Full-length 02 FLJ13112	5'TGCTTGCATGCCTGAATATC3'(F) 5'TTCACCGTGTTGCCCAG3'(R)
Full-length 02 FLJ13112	Nested: 5'GAGCGAGCACAGGAAAGG3'(F) 5'TGCAAGTGAGAAAACAGAGGC3'(R)
Full-length 03 YQ80H05	5'GGAGAACAATCACATGAAG3'(F) 5'ATTCTGCCTGGCTAACGC3'(R)
Full-length 03 YQ80H05	Nested: 5'CCATAGATATAATGAGCCCAG3'(F) 5'TTGGATAACCTTTTGTGTTCC3'(R)
Full-length 04 FLJ10580	5'CCCTCACAGGATGAGTTTGAG3'(F) 5'GCTTCGCAACAAAGCCG3'(R)
Full-length 04 FLJ10580	Nested: 5'TGGTTTCCAATTTGTTGGC3'(F) 5'GGACCGTGTGTGACTCTGG3'(R)
Full-length 05 OCLM	5'TTGCCTGAGGTCACTCTGC3'(F) 5'TGCATTGCAGGTTTCAGG3'(R)
Full-length 05 OCLM	Nested: 5'CCCAATTGCCAGTAAAAAC3'(F) 5'TTGCAGGGTATGGGGATG3'(R)
Full-length 06 FLJ12315	5'AGGGTCTCACTCCCTGTGC3'(F) 5'TGGCTGTTGTTATGGGGTG3'(R)
Full-length 06 FLJ12315	Nested: 5'ATACAGGCGTGTGCCGTC3'(F) 5'GAAGCCATTGGCATGAGC3'(R)
Full-length 07 FLJ13301	5'CATGTAATGCAGGGGTTGAG3'(F) 5'AGGCCCTCCAAACCAGAG3'(R)
Full-length 07 FLJ13301	Nested: 5'ACTGCAACCTCCGCCTC3'(F) 5'GCCAAGGCAGGTGGATAAC3'(R)

Candidato	Primers
EST 01 AA757244	5'TGGAGGCTCTCAAAATGAGG 3'(F) 5' GTGCAGACCAAGAGAATGGAG3'(R)
EST 01 AA757244	Nested : 5'GCAGGTGTCAGTCACTCTTGC3'(F) 5'TCAGCCAAGACTCGTGGTATC3'(R)
EST02 BE566336	5'AACAGCTCCGGTCTACG3'(F) 5'ATGTACAGCCGATGACGTT3'(R)
EST 02 BE566336	Nested : 5'TCCCAGCGTGAGAGATGC 3'(F) 5'TTGGCTCCAGACACATGC 3'(R)
EST 03 AI187857	5'AACCTCCACAGGATCAAAGG3'(F) 5'ACGCACCAATCAGCATCC3'(R)
EST 04 AI018001	5'GCTCACCACTAGGAAGAGGG 3'(F) 5'TTGCGTGGCATGAGATTG 3'(R)
EST 05 AW205088	5'AAAGATGTTCCCGCCCTG 3'(F) 5'ATCAGCCACACGGCAGAG 3'(R)
EST 05 AW205088	Nested : 5'CTTTCCAGCGTGGTCCAG3'(F) 5'AGCAAAACTCTGCCCAGG3'(R)
EST 06 AI201365	5'TGACCAGCTCAACTTCATGC 3'(F) 5'TCCAGTGGGCAGCAAAAG 3'(R)
EST 06 AI201365	Nested : 5'GAATTTTTGTCGTGATGCAAG3'(F) 5'CGGTATATAGACTTTGCAGG3'(R)
EST 07 BE906931	5'TGTCCGCATAAAAACCTG 3'(F) 5'TCCAGTGGGCAGCAAAAG 3'(R)
EST 08 AI243690	5'GACTGACCAAAAGAGCCGAG 3'(F) 5'CCCAAAACTCCATATCCCAG 3'(R)
EST 08 AI243690	Nested: 5'GACCAGCTGCCTTCATCA3'(F) 5'TTGCGATGCCACCATAAC3'(R)
EST 09 AA460184	5'AAGGACTCATGGTGACGACC 3'(F) 5'CTTCCAGGACCATACGGTTT 3'(R)
EST 10 AI025599	5'GCAAATTTCAGGCATGACTT 3'(F) 5'CCCTGATCCCAAGTGTGAAG 3'(R)
EST 10 AI025599	Nested: 5'AACGTTTCAAGTGTCTCGCC3'(F) 5'TAGCGGAAAGGCAACAGAAT3'(R)

4.2.5 Preparação de bactérias competentes

Para a preparação de bactérias competentes foram utilizadas células JM 109 e o protocolo descrito por INOUE et al. (1990). Uma alíquota de estoque destas células foi plaqueada em ágar-LB (Peptona de caseína 1%, extrato de levedura 0,5%, NaCl 0,5% e Agar 1,5%) e incubada por 16 horas a 37°C. Em seguida, foram inoculadas dez colônias isoladas em 260,0 ml de meio LB (Peptona de caseína 1%, extrato de levedura 0,5%, NaCl 0,5%) contido em um frasco de 2,0 litros. O meio foi então incubado entre 18-22°C sob agitação vigorosa (200rpm) até atingir a DO a 600nm igual 0,6. Ao atingir esta DO, o frasco foi imediatamente colocado em gelo por 10 minutos e depois distribuído em seis alíquotas de 42ml em tubos de centrífuga de 50,0ml. Os tubos foram centrifugados a 3.500rpm (centrífuga Sorvall RC5B rotor GSA) por 15 minutos a 4°C. Cada precipitado foi ressuspendido em 12,0ml de TB (Transformation Buffer: 10mM Pipes, 55mM $CaCl_2$, 250mM KCl) gelado e foi reunido o conteúdo final de três tubos resultando apenas dois tubos com 36,0ml cada um. Os tubos foram incubados no gelo por 10 minutos e centrifugados novamente como descrito acima. Cada precipitado foi ressuspendido em 9,3ml de TB gelado e reunidos em um tubo resultando em um volume total de 18,6ml. Em seguida foi adicionado, vagarosamente, 1,4ml de DMSO (dimetilsulfóxido) atingindo uma concentração final de 7%. A suspensão de células foi incubada em gelo por 10 minutos, e distribuída em tubos de congelamento de células e imediatamente imersas em nitrogênio líquido, onde foram mantidas até sua utilização (INOUE et al., 1990).

4.2.6 Clonagem e transformação bacteriana

Os fragmentos de validação dos diferentes transcritos foram clonados utilizando o TA Cloning Kit (Invitrogen) segundo especificações do fornecedor. A transformação bacteriana foi realizada utilizando 5µl da reação de ligação e 100µl de bactérias competentes. A bactéria foi descongelada e mantida em gelo por 20 minutos. A mistura de produto de ligação e bactéria foi homogeneizada cuidadosamente e

incubada por 30 minutos em gelo. Após esse período foi feito o choque térmico incubando por 30 segundos a 42°C e trinta segundos em gelo. Em seguida foi adicionado 1,0ml de meio LB sem antibiótico e a cultura foi incubada por 60 minutos em banho-maria a 37°C. Esta cultura foi centrifugada (3.000 rpm durante 10 min. em centrífuga eppendorf 5415C rotor F-45-18-11) e ao sedimento bacteriano foram adicionados 100µl de meio LB sem antibiótico. Esta suspensão foi plaqueada sobre meio LB-ágar contendo 50µg/ml de ampicilina, 20µg/ml de IPTG (isopropiltio-β-D-galactosídeo) e 4µg/ml de X-gal (5-bromo-4-cloro-3indolil-β-D-galactosídeo) e incubada a 37°C por 18 horas. As colônias de bactérias transformadas foram coletadas da placa, inoculadas em 5ml de meio LB contendo 50µg/ml de ampicilina, incubadas a 37°C por 12 a 16 horas. O DNA plasmidial foi purificado desta cultura utilizando o kit Wizard Plus Miniprep DNA Purification System (Promega) segundo especificações do fornecedor.

4.2.7 Seqüenciamento automático e submissão dos dados

De 300 a 500 nanogramas dos plasmídeos obtidos na clonagem contendo os insertos de interesse foram seqüenciados utilizando o kit "DYEnamicTM ET Terminator Cycle Sequencing" (Amersham Life Science) disponível para o seqüenciamento automático no seqüenciador ABI377 (Applied Biosystems- Perkin Elmer) segundo especificações dos fornecedores. Os iniciadores utilizados foram M13-40 5'CGC CAG GGT TTT CCC AGT CAC GAC3' e M13 Rev 5'TTT CAC ACA GGA AAC AGC TAT GAC3'. As seqüências foram enviadas ao laboratório de Bioinformática do Instituto Ludwig via web para serem incorporadas no banco de dados do Projeto Transcriptoma Humano.

4.3 PRODUÇÃO DE SEQÜÊNCIAS DE cDNA COMPLETAS CORRESPONDENTES AS ESTs CANDIDATAS

Duas abordagens foram utilizadas para a produção da seqüência completa dos transcritos correspondentes as ESTs que apresentaram expressão em próstata: seqüenciamento completo dos clones de cDNA correspondentes as ESTs e extensão da seqüência transcrita através de RACE.(Rapid Amplification of cDNA Ends)

4.3.1 Seqüenciamento de insertos de clones de cDNA

A seleção de clones de cDNA correspondentes as ESTs candidatas foi feita junto ao DHGP (German Human Genome Project) (http://www.dhgp.de/). Para tanto as seqüências das ESTs foram utilizadas como "query" em buscas por BalstN contra um banco de dados de clones de cDNA disponibilizados pelo consórcio IMAGE. Os clones foram solicitados ao IMAGE e a identidade dos mesmos foi confirmada através de seqüenciamento das extremidades dos clones. Em seguida a seqüência completa de cada um desses clones foi submetida ao GenBank e receberam a seguinte nomeação: AF508902, AF508903, AF508904, AF508905, AF508906, AF508907, AF508908, AF508909, AF508910 e AF508911. A Tabela abaixo mostra a relação de clones do IMAGE correspondentes às ESTs candidatas selecionadas. Os iniciadores utilizados foram T7 5'TAA TAC GAC TCA CTA TAG GG3'; SP6 5'TTT AGG TGA CAC TAT AG3'; T3 5'ATT AAC CCT CAC TAA AGG GA3'.

Candidato	Clones do IMAGE
Candidato 01	IMAGp998A223330Q2
Candidato 02	IMAGp958G08365Q2
Candidato 03	IMAGp998D064417Q2
Candidato 04	IMAGp998N074130Q2
Candidato 05	IMAGp998C217408Q2
Candidato 06	IMAGp998O054461Q2
Candidato 07	IMAGp998O089705Q2
Candidato 08	IMAGp998F114536Q2
Candidato 09	IMAGp998F071962Q2
Candidato 10	IMAGp998G244172Q2

4.3.2 RACE (Rapid Amplification of cDNA ends)

Outra possibilidade para obter a seqüência completa dos transcritos , é a utilização de RACE. Para tanto, utilizamos o Marathon™ Amplification cDNA Kit (Clontech® nº: K1802-1), que contém bibliotecas de cDNA fita dupla já ligados a adaptadores em ambas extremidades. Primeiramente foi feita uma PCR com iniciadores específicos para os adaptadores e para a porção interna do transcrito de interesse. Os iniciadores internos foram direcionados para a porção 5' ou 3'. A reação primária foi realizada utilizando 2ng de cDNA e 200nM de cada iniciador externo, 200nM de dNTPs, 1X cDNA PCR reaction buffer e 1X Advantage cDNA Polymerase Mix (Clontech) para um volume final de reação de 50µl. Todas as reações foram realizadas em uma Thermocycler 9700 (Perkin-Elmer) utilizando as seguintes condições para a ciclagem: denaturação inicial a 94°C durante 2 minutos; 5 ciclos de

30 segundos a 94°C e 2 minutos a 68°C; 25 ciclos de 30 segundos a 94°C, 30 segundos a 60°C e 2 minutos a 72°C; seguida por uma extensão final a 72°C durante 10 minutos.

Para aumetar a especificidade das reações a estratégia de Nested-PCR sugerida pelo fornecedor foi adotada. Os produtos obtidos a partir da primeira reação foram utilizados como molde para a segunda reação de RACE, agora utilizando o iniciador adaptador AP2 e o primer "nested" específico para o gene em questão. Os produtos obtidos foram analisados em gel de agarose 1%, purificados utilizando o QIAquick Gel Extraction Kit (Qiagen) e clonados utilizando o TOPO TA Cloning Kit (Invitrogen) de acordo com as especificações do fabricante. Os clones foram seqüenciados como descrito acima, o que permitiu a extensão da seqüência dos transcritos. Os transcritos que tiveram suas seqüências completamente geradas através da estratégia de RACE são representados pelas ESTs 01 e 05. Na Tabela abaixo temos representado as seqüências dos iniciadores que foram construídos para as reações.

Candidato 01 – GSP1 (Gene Specific Primer)	5'GGG CCA CCT CCT CCT CCG GAT GAC AC 3'
Candidato 01 – GSPN1 (Gene Specific Primer Nested)	5'GCT TCT TTC CAT CAC CCT TTA GTC TGA ATG G 3'
Candidato 05 – GSP05	5'CAT GGG GGG TTT GCC GGT GCC AGG TCT G 3'
Candidato 05 – GSPN05	5'CTT TCT GTT GAT CAG GGA TTA CGG GAA TG 3'

4.3.3 Real Time PCR

Um PCR quantitativo foi realizado com o aparelho LigthCycler™ (Roche Diagnostics, Mannheim, Germany) em um volume total de reação de 10µl por capilar. Cada 10µl de reação contém 0.75U de Platinum Taq DNA Polymerase (Gibco BRL, Life Technologies), 1µl de buffer 10X, 5mM $MgCl_2$, 0.2mM de dNTPs (Promega), 0.4 µM de cada iniciador, 5% DMSO e 0.1 µl de SYBR® Green I (Sigma) (diluído

1:100) e 1µl de cDNA preparado como descrito previamente. Cada experimento foi feito em duplicata e repetido duas vezes.

As condições de amplificação utilizadas pelo aparelho são: 2 minutos de denaturação inicial a 95°C, seguida de 45 ciclos de denaturação a 94°C durante 15 segundos, anelamento dos iniciadores a 62°C durante 20 segundos e extensão a 72°C durante 30 segundos. A detecção do produto fluorescente foi realizada na última etapa de cada ciclo a uma temperatura variando entre 3-4°C abaixo da TM (melting temperature) do produto. Para confirmar a especificidade da amplificação, os produtos da PCR foram submetidos a análise da curva de "melting". Ao final de cada reação de PCR através do resfriamento das amostras `a 65°C e então aumentando a temperatura a 95°C a 0.1°C/s, com a contínua aquisição da fluorescência. Nas reações foram utilizados cDNA de testículo e próstata normal, e cDNA das linhagens tumorais de próstata DU145 e PC3. Os genes que tiveram seus níveis de expressão avaliados foram os genes RGSL1 (EST1) e o RGSL2 (FL1).

Os genes de referência utilizados nas reações para corrigir variações nas quantidades iniciais de RNA foram o GAPDH e a β-actina. As seqüências dos primers utilizados nas amplificações tanto dos genes constitutivos quanto dos genes de interesse estão relacionadas na tabela abaixo.

Gene	Seqüência
GAPDH (FW)	5'CTG CAC CAC CAA CTG CTT A3'
GAPDH (REV)	5'CAT GAC GGC AGG TCA GGT C3'
β-actina (FW)	5'CAC TGT GTT GGC GTA CAG GT3'
β-actina (REV)	5'TCA TCA CCA TTG GCA ATG AG3'
RGSL1 (FW)	5'TGG AGG CTC TCA AAA TGA GG3'
RGSL1 (REV)	5'GTG CAG ACC AAG AGA ATG GAG3'
RGSL2 (FW)	5'GGA TTT CTG GAT CCT TGC TG3'
RGSL2 (REV)	5'TGG GTG TAG CAA ACG CTG3'

Valores quantitativos foram obtidos no ponto durante a ciclagem no qual a amplificação do produto da PCR foi detectada, ao invés da quantidade do produto de PCR acumulado depois de um número fixo de ciclos. O parâmetro CP (crossing point) é definido como uma fração do número de ciclos na qual a fluorescência gerada ultrapasse um padrão fixo. O dado de quantificação foi analisado pelo "software" de análise do aparelho como descrito previamente (MORRISON et al., 1998). A amplificação dos genes de interesse foi quantificada e então dividida pelos valores obtidos a partir dos genes de referência (GAPDH e β-actina) para obter um valor normalizado de cada amostra para a variabilidade na quantidade e integridade do RNA (XIE et al., 2001). Como controle positivo para a expressão dos genes RGSL1 e RGSL2 utilizamos cDNA de testículo e como controle negativo utilizamos capilares contendo apenas o mix necessário para a reação sem nenhum cDNA.

Resultados e Discussão

5 RESULTADOS E DISCUSSÃO

5.1 LEVANTAMENTO DOS MARCADORES MOLECULARES E DOS CLONES GENÔMICOS QUE COBREM A REGIÃO 1q25

Os primeiros dados obtidos foram relacionados aos dois marcadores, já identificados por CARPTEN et al. (2000) que flanqueiam o intervalo 1q25, delimitando um intervalo de 6 Mb. Esses marcadores de microsatélite são denominados D1S2818 (Z54047) e D1S1642 (GO9777) e foram identificados através de estudo de ligação (CARPTEN et al., 2000; SOOD et al., 2001). Após a obtenção das seqüências de cada um dos marcadores, através de uma consulta feita ao GenBank, os mesmos foram alinhados contra os dados de seqüência do genoma humano, através da utilização do programa BlastN, a fim de que os clones genômicos, onde as seqüências dos marcadores estivessem contidas, fossem identificados. O banco de dados utilizado para a obtenção dessas informações foi o NR e o HTGS. Os clones genômicos AL139344 e AL391065 foram identificados através dessas buscas delimitando a região em questão (Figuras 3 e 4). De posse da informação a respeito da identidade do primeiro e do último clone genômico nos quais os marcadores estão localizados, o passo seguinte foi obter os demais clones genômicos que cobrem o intervalo compreendido entre os dois marcadores. Essa informação foi obtida através da utilização do Map Viewer (versão 02) uma ferramenta disponibilizada pelo NCBI que nos permite a visualização do mapa físico de uma dada região genômica (Figura 5 e 6).

Os clones genômicos que compõem a montagem do genoma humano disponibilizada pelo Consórcio Público estão divididos em duas categorias principais de acordo com a qualidade da seqüência e estágio da montagem: clones "draft" e "finished". Uma seqüência "draft" têm a seqüência do inserto coberta pelo menos 3 a 4 vezes. Pelo fato dessas seqüências não serem de alta qualidade, a ordem e a orientação dos clones podem muitas vezes não estarem corretas. Por outro lado, as seqüências

"finished", possuem os insertos continuamente seqüenciados (em geral não existem "gaps" na seqüência) com um alto padrão de qualidade. Nesses casos a taxa de erro no seqüenciamento é de 0,01%.

```
>gi|16972777|emb|AL139344.23|AL139344 Human DNA sequence from clone RP1-223H12 on
chromosome 1q25.1-25.3,
            complete sequence [Homo sapiens]
          Length = 152058

 Score =  595 bits (300), Expect = e-167
 Identities = 332/343 (96%), Gaps = 10/343 (2%)
 Strand = Plus / Plus


Query: 1     aattagccagttgtggtggcaggtacctataatctcagatactcaggaggctgaagcaag 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 26468 aattagccagttgtggtggcaggtacctataatctcagatactcaggaggctgaagcaag 26527


Query: 61    agaatcacttgaacctgggaggcagaagttgcagtgaacagagatggtgccactgcactc 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 26528 agaatcacttgaacctgggaggcagaagttgcagtgaacagagatggtgccactgcactc 26587


Query: 121   aagggcaatagagtgggactcagtctcaaacacacac----------acacacacacaca 170
              ||||||||||||||||||||||||||||||||||          ||||||||||||
Sbjct: 26588 aagggcaatagagtgggactcagtctcaaacacacacacacacacacacacacacacaca 26647


Query: 171   cacacacacataataataataaaatggggatagtaatagtaacactacctcatagagtta 230
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 26648 cacacacacataataataataaaatggggatagtaatagtaacactacctcatagagtta 26707


Query: 231   ttgttcaattaaatgagtcaacacaacttagcatgcaacaactacaatgcagggaaaaaa 290
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 26708 ttgttcaattaaatgagtcaacacaacttagcatgcaacaactacaatgcagggaaaaaa 26767


Query: 291   aacaacagacctttttttttgggtgaggggcaggggganaggatc 333
              |||||||||||||||||||||||||||||||||| ||||||
Sbjct: 26768 aacaacagacctttttttttgggtgaggggcaggggacaggatc 26810
```

Figura 3 - **Identificação do primeiro clone genômico da região 1q25:** Alinhamento entre a seqüência do marcador de microsatélite D1S2818 e a seqüência genômica AL139344.

```
>gi|14018283|emb|AL391065.6|AL391065 Human DNA sequence from clone RP11-108M21 on
chromosome 1, complete
            sequence [Homo sapiens]
          Length = 170639

 Score =  517 bits (261), Expect = e-144
 Identities = 287/295 (97%), Gaps = 8/295 (2%)
 Strand = Plus / Minus


Query: 1     cccactaagtatttttgaagtgtatacttgcagtgtagctgacacctctctaaaagaagt 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54739 cccactaagtatttttgaagtgtatacttgcagtgtagctgacacctctctaaaagaagt 54680


Query: 61    ttagctcattgaataagatttttaataatttttagtacagtatgttaaatatttagtaagt 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54679 ttagctcattgaataagatttttaataatttttagtacagtatgttaaatatttagtaagt 54620


Query: 121   aaatatatatacacatatatatgtagatagataaatagatagatagatagatagataga- 179
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54619 aaatatatatacacatatatatgtagatagataaatagatagatagatagatagatagat 54560


Query: 180   -------tagatagaacaaatagacaaattatgctatagtaattaaaacatagtcatttt 232
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54559 agatagatagatagaacaaatagacaaattatgctatagtaattaaaacatagtcatttt 54500


Query: 233   ttatttgtcctaacttatccctaggtacttcagatacgggcattggaagaagcaa 287
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54499 ttatttgtcctaacttatccctaggtacttcagatacgggcattggaagaagcaa 54445
```

Figura 4 - **Identificação do último clone genômico da região 1q25:** Alinhamento entre a seqüência do marcador de microsatélite D1S1642 e a seqüência genômica AL391065.

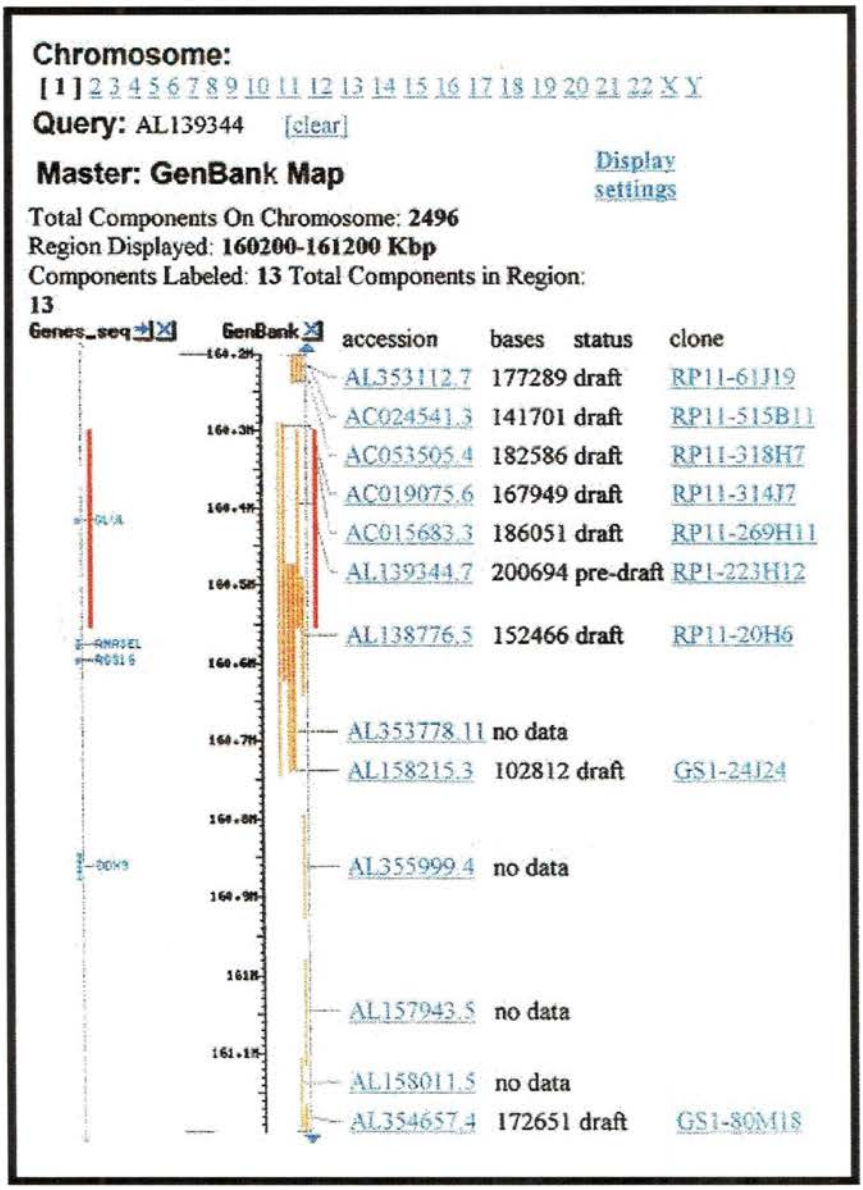


Figura 5 - **Mapa físico da região 1q25 disponibilizado pelo NCBI, destacando o primeiro clone genômico do intervalo.** Representação gráfica, disponibilizada pelo MapViewer, do mapa físico da região 1q25 e dos clones genômicos que cobrem o intervalo. O clone AC019075 é o clone no qual o marcador de microsatélite D1S2818 foi localizado e portanto esse foi considerado o primeiro clone genômico da região 1q25 em nosso projeto.

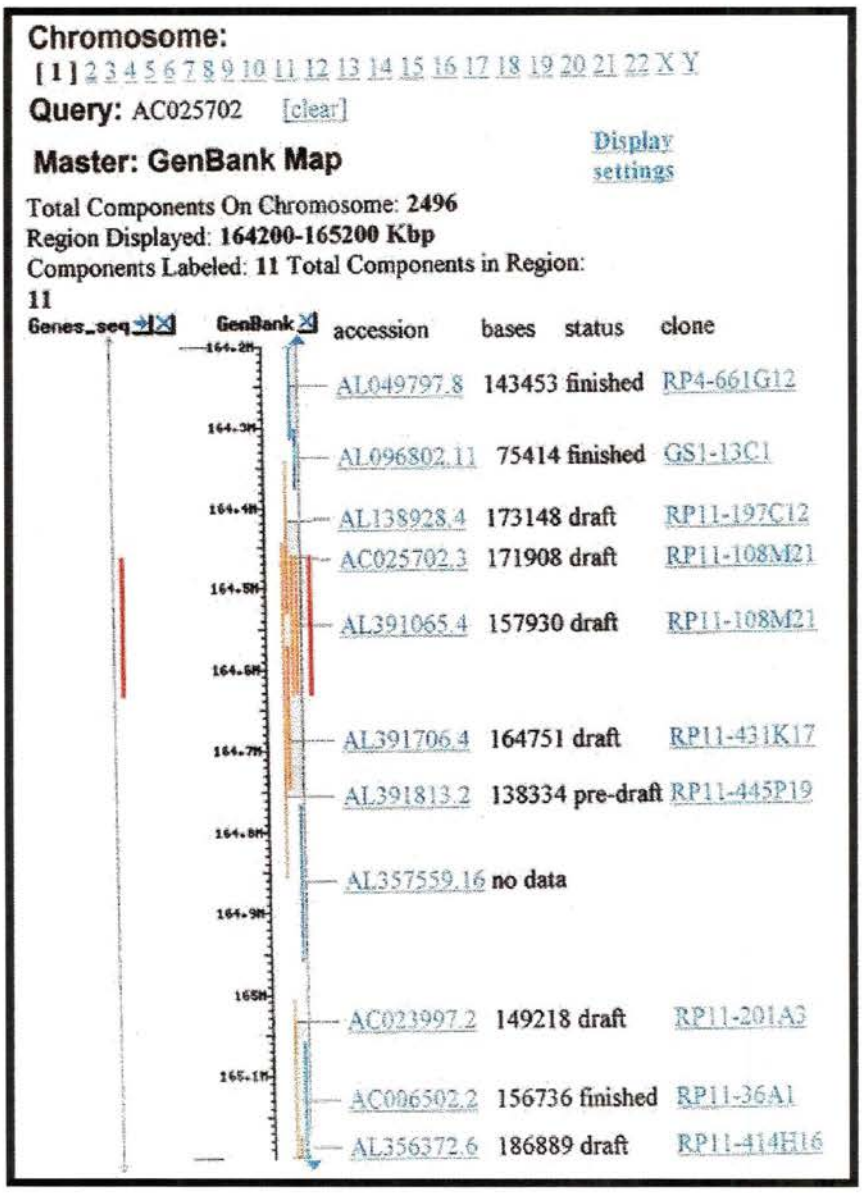


Figura 6 - **Mapa físico da região 1q25 disponibilizado pelo NCBI, destacando o último clone genômico do intervalo.** Representação gráfica, disponibilizada pelo MapViewer, do mapa físico da região 1q25 e dos clones genômicos que cobrem o intervalo. O clone AL391065 é o clone no qual o marcador de microsatélite D1S1642 foi localizado e portanto esse foi considerado o último clone genômico da região 1q25 em nosso projeto.

Um total de 44 clones genômicos cobrindo o intervalo entre os marcadores D1S2818 e D1S1642 foram selecionados através do MapViewer (Tabela1). A seleção dos clones foi feita de forma a evitar redundância e sobreposição entre os clones genômicos escolhidos. No entanto, o intervalo apresenta 24 clones (54,5%) que estão em estado "draft" e nesses casos todos os clones mapeados nessa região foram considerados em nossa análise a fim de que nenhum "gap" pudesse ser gerado, levando a perda de informações. De acordo com o MapViewer, o tamanho estimado da região seqüenciada é de aproximadamente 5.3Mb, existindo 3 "gaps" que juntos possuem o tamanho estimado de 0,14Mb. Considerando a qualidade da montagem e da seqüência, 55,12% da seqüência disponibilizada pode ser considerada de alta qualidade ("finished").

O "status" da montagem é de extrema importância para nossa estratégia pois a qualidade da seqüência e a existência de "gaps" interferem diretamente nos alinhamentos realizados. Ao trabalharmos com uma região ainda não completamente seqüenciada devemos considerar a hipótese da existência de transcritos adicionais na região que eventualmente estejam localizados em "gaps" ou que não tenham sido identificados pois a qualidade da seqüência genômica disponibiliazada não permite o alinhamento com as seqüências expressas dentro dos parâmetros estabelecidos.

Com o objetivo de confirmar a presença dos demais marcadores de microsatélites identificados por CARPTEN et al. (2000) nos clones genômicos selecionados, 14 outros marcadores foram selecionados e suas seqüências foram alinhadas contra a seqüência genômica humana (Tabela 2). O resultado desses alinhamentos, nos permitiu confirmar a presença dos diferentes marcadores nos clones genômicos selecionados através do Map Viewer e dessa forma minimizar o impacto de eventuais erros presentes na montagem do genoma humano em nossa análise.

Tabela 1 - **Clones genômicos localizados na região 1q25:** Dados obtidos para cada clone genômico apartir do Banco de Dados.

Acc. No.	Tamanho	Status	No. contigs	No. Seqs	No. Clusters	No. Clusters FL	No. Clusters EST	No. ESTs	No clusters EST SPL	No. de ESTs com splicing
01. AC019075	167949	Draft	08	42	15	03	12	41	02	04
02. AC015683	186051	Draft	09	232	24	07	17	228	03	63
03. AL139344	200694	Pre-draft	11	936	22	07	15	929	01	259
04. AL138776	152466	Draft	01	43	06	00	06	43	03	06
05. AL353778	No data	No data	01	206	14	04	10	200	01	59
06. AL158215	102812	Draft	06	246	19	05	14	240	02	63
07. AL355999	No data	No data	09	409	42	06	36	403	03	259
08. AL157943	No data	No data	05	115	21	05	16	112	01	37
09. AL158011	No data	No data	05	97	08	01	07	96	02	10
10. AL354657	172651	Draft	04	07	04	00	04	07	01	01
11. AL445228	No data	No data	05	07	07	00	07	07	01	01
12. AL078645	206089	Finished	01	78	30	02	28	76	05	29
13. AL358152	No data	No data	04	211	24	02	22	208	03	20
14. AC074116	No data	No data	13	492	34	14	20	479	01	323
15. AL391824	67390	Draft	01	98	08	01	07	94	03	19
16. AL121996	No data	No data	11	25	16	05	11	23	01	02
17. AC027182	140952	Draft	02	24	02	00	02	24	00	15
18. AC027690	197650	Draft	03	275	32	03	29	271	01	26
19. AL356273	142459	Draft	05	445	28	08	20	430	00	191
20. AL096819	136188	Finished	01	401	23	03	20	392	00	20
21. AL136086	103347	Finished	01	248	22	01	21	244	01	17
22. AL109956	No data	No data	01	33	13	01	12	32	00	05
23. AL109865	201823	Finished	01	166	27	03	24	160	03	26
24. AL136133	85565	Finished	01	31	08	02	06	28	02	10
25. AL078644	130177	Finished	01	304	22	01	21	294	02	192
26. AL117347	61698	Finished	00	00	00	00	00	00	00	00
27. AL133383	204158	Finished	01	27	04	00	04	27	04	19
28. AC023275	180248	Finished	01	11	08	00	08	11	00	00
29. AL135796	157029	Finished	01	05	04	00	04	05	00	00
30. AL133515	47170	Finished	01	94	07	01	06	90	02	15
31. AL118512	96783	Finished	01	10	04	00	04	10	00	00
32. AL135797	No data	No data	01	03	02	00	02	03	02	03
33. AL133553	190655	Finished	01	297	17	01	16	294	05	206
34. AL162722	171179	Draft	14	171	30	00	30	156	00	66
35. AL096803	129723	Finished	01	11	07	00	07	11	00	00
36. AL049198	41015	Finished	01	01	01	00	01	01	01	01
37. AL033533	84412	Finished	01	77	05	01	04	75	00	11
38. AL022241	101046	Finished	00	00	00	00	00	00	00	00
39. AL022147	114638	Finished	01	20	12	02	10	17	00	04
40. AL049797	143453	Finished	01	34	09	01	08	32	00	13
41. AL096802	75414	Finished	01	02	02	00	02	02	00	00
42. AL138928	173148	Draft	04	04	04	00	04	04	00	00
43. AC025702	171908	Draft	02	03	03	00	03	03	00	00
44. AL391065	157930	Draft	02	03	03	00	03	03	00	00
44	5,3 Mb	-	146	6110	620	93	527	5965	59	1995

Tabela 2 - **Marcadores de microsatélites da região 1q25 identificados por Carpten et al (2000):** Outros marcadores de microsatélites existentes na região 1q25.

D1S2818	AL139344
D1S290E	AL139344
D1S3425	AL159743
D1S2472	AL157943
D1S158	AL133383
D1S2127	AL136086
D1S240	AL356273
D1S254	AL135796
D1S444	AL135796
D1S191	AC023275
D1S2165	AL391824
D1S1599	AL162722
D1S202	AL049797
D1S1642	AC025702

O número de acesso de cada um dos clones genômicos identificados através do MapViewer foi utilizado como dado de entrada para as buscas no Banco de Dados do Transcriptoma Humano. Esse é um banco de dados relacional em linguagem MySQL, no qual os dados são armazenados em forma de tabelas de fácil correlação. As buscas foram feitas através de "selects" específicos, listados nos Materiais e Métodos, para cada uma das informações necessárias.

Um total de 6.107 seqüências expressas foi mapeado nos clones genômicos selecionados as quais foram agrupadas em 617 "clusters". Cabe ressaltar que existe uma certa redundância nesses dados uma vez que em regiões "draft" desse intervalo mais de um clone genômico foi selecionado. Essa redundância, no entanto, foi eliminada posteriormente durante o processo de seleção de candidatos. Desses "clusters", 93 são compostos por pelo menos uma seqüência completa de mRNA e 524 "clusters" são compostos apenas por seqüências de ESTs. Apenas 59 dos 524 clusters compostos por ESTs apresentaram estrutura de "splicing" quando alinhados com a seqüência genômica e foram selecionados para análise posterior (Tabela 1).

5.2 IDENTIFICAÇÃO DE SEQÜÊNCIAS COMPLETAS DE CDNA DA REGIÃO 1q25

Com base nas informações acima descritas, a etapa seguinte foi avaliar quantos, dos genes já identificados e anotados, tanto pelo Projeto Genoma quanto pelo grupo de Carpten, poderiam ser identificados em nosso banco de dados e principalmente identificar novos transcritos que apesar de já possuírem a seqüência completa ainda não haviam sido mapeados nessa região. Dessa forma, foram selecionados através de consulta feita ao Banco de Dados do Transcriptoma Humano e após eliminação de redundância, 42 seqüências completas de cDNA. A etapa seguinte foi obter informações a respeito de cada um dos 42 transcritos identificados utilizando para tanto o número de acesso do GenBank das seqüências selecionadas. O Banco de Dados de seqüências expressas do Unigene foi utilizado para a obtenção de informações relacionadas a cada uma das seqüências de mRNAs completas identificadas tais como: o nome do gene, o número do "cluster" do Unigene, uma breve descrição a respeito do gene, o tamanho de cada Unigene cluster (quantas ESTs representam aquele transcrito) e também dados de expressão, obtidos a partir das bibliotecas a partir das quais as ESTs foram geradas.

O Unigene é um Banco de Dados que separa todas as seqüências expressas depositadas no GenBank e depois as agrupa, com base na similaridade das seqüências, em conjuntos não redundantes denominados UniGene "clusters". Dessa forma, cada Unigene "cluster" representa um único transcrito e contém informações a respeito do tecido no qual um dado gene é expresso e da localização cromossômica dentre outras. De acordo com os dados de expressão do Unigene, 16 dos 30 cDNA "full-length" selecionados em nosso banco de dados estão expressos em próstata.

Houve ainda a necessidade de fazer uma inspeção manual do alinhamento de cada uma das 42 seqüências de cDNA completas, para avaliar a ocorrência de um melhor alinhamento do gene em outra região cromossômica. Após inspeção manual, verificamos que 14 seqüências alinhavam melhor em outras regiões cromossômicas devido a presença de seqüências retrovirais no clone de cDNA. Das 28 seqüências

remanescentes, 21 seqüências já haviam sido mapeadas na região 1q25 e por essa razão haviam previamente sido consideradas prováveis candidatos a genes supressores de tumor envolvidos no câncer de próstata hereditário. Os 07 clones de cDNA "full-length" por nós mapeados pela primeira vez na região 1q25 estão listados na Tabela 3 e Figura 7. Somente três dessas seqüências apresentam uma clara fase aberta de leitura (open reading frame – ORF). Essas seqüências são representadas pelas seqüências "full-length" (FL) 1, 4 e 5 das quais duas apresentam uma evidente estrutura de "splicing", FL1 e FL4, quando alinhadas com a seqüência genômica. Três dessas seqüências candidatas foram recentemente anotadas pelo Projeto Genoma Humano.

O candidato FL1 (AL136902) representa uma proteína hipotética com uma fase aberta de leitura composta por 540 aminoácidos. A análise da seqüência da proteína predita usando InterProScan revelou a presença de um domínio RGS (Regulador de Sinalização da Proteína G), altamente conservado, composto por 120 aminoácidos. Esse dado está representado na Figura 8. O candidato FL4 (AK001442) corresponde a uma proteína de função desconhecida e o candidato FL5 (AF142063) corresponde ao gene da oculomedina que é expresso nas células trabeculares da retina quando as mesmas sofrem algum tipo de "stress" mecânico (SATO et al., 1999). As outras seqüências de cDNA "full-length" não possuem uma clara estrutura de "splicing" nem uma fase aberta de leitura e com exceção do candidato FL7 também não possuem um típico sinal de poliadenilação, sugerindo que esses clones de cDNA possam ser artefatos. O baixo número de ESTs associadas a cada um dos clones de cDNA "full-length" também sugere um padrão de expressão bastante restrito e suporta a hipótese de que cada um desses clones possa representar artefatos (Tabela 3). Entretanto, essa hipótese foi descartada uma vez que todos os candidatos "full-length", exceto o candidato FL4, mostraram-se positivamente expressos em cDNA de próstata (que foi extensivamente testado quanto a presença de contaminação por DNA genômico) através de detecção por RT-PCR. Dessa forma, esses candidatos "full-length" representam potenciais candidatos a gene supressor de tumor envolvido no câncer de próstata hereditário.

Tabela 3 - **cDNAs completos mapeados na região 1q25 através do banco de dados do Transcriptoma Humano:** Características e padrão de expressão em próstata referentes às seqüências de cDNA completas mapeadas na região 1q25.

"Full-length"	**Número de acesso**	**Unigene "cluster"**	**Presença de ORF**	**"Splicing"**	**Sinal de Poly A**	**Anotação**	**Expressão em próstata**
1	AL136902	HS.307080 (size 1)	Sim	Sim	Sim	Domínio RGS	Sim
2	AK023174	HS.333467 (size 2)	Não	Não	Não	-	Sim
3	AF075081	-	Não	Não	Não	-	Sim
4	AK001442	HS.140402 (size 3)	Sim	Sim	Sim	Proteína hipotética	Não
5	AF142063	HS.283759 (size 1)	Sim	Não	Não	OCLM oculomedina	Sim
6	AK022377	-	Não	Não	Não	-	Sim
7	AK023363	HS.300827 (size 1)	Não	Não	Sim	-	Sim

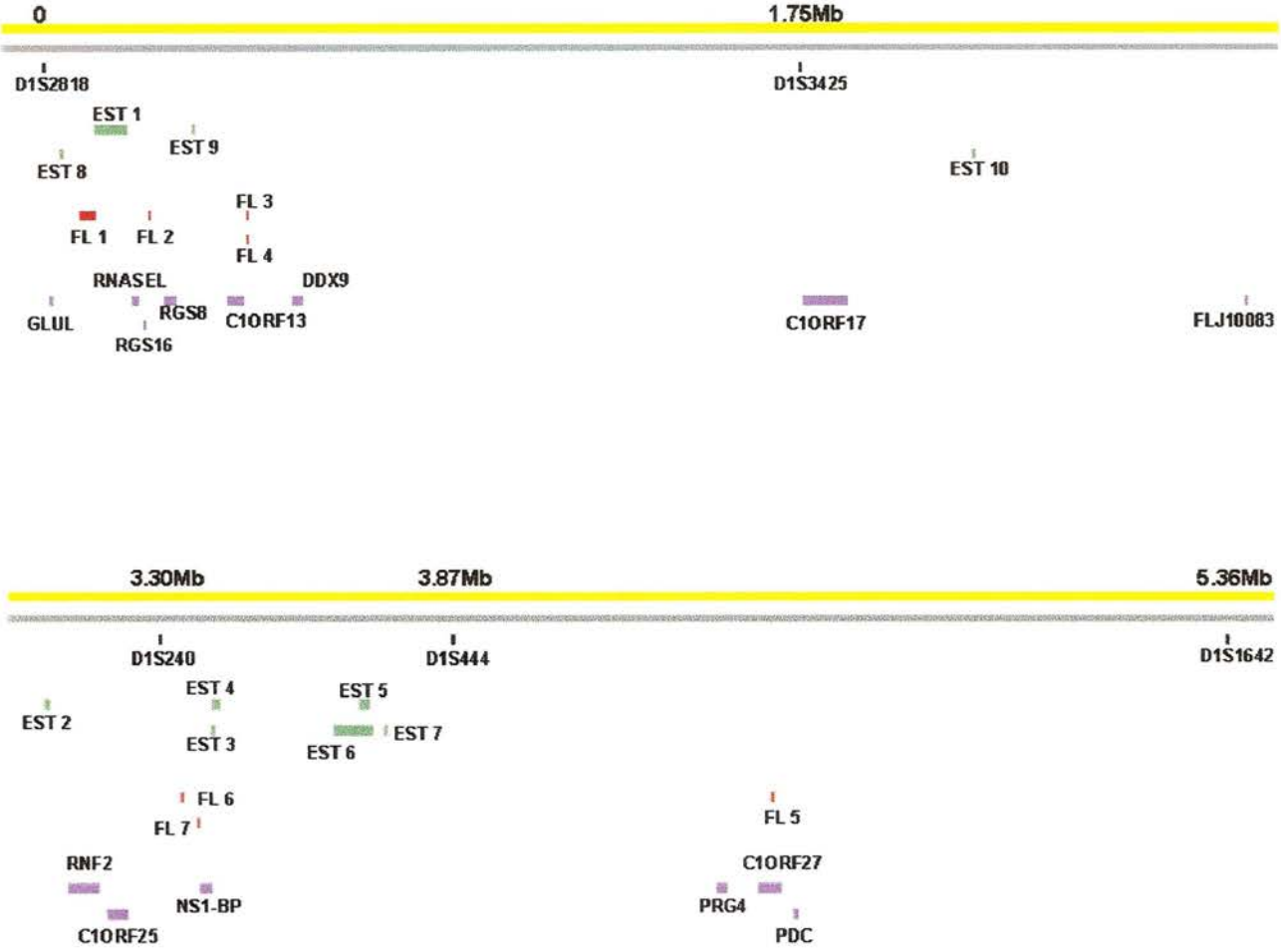


Figura 7 - **Representação gráfica dos transcritos mapeados na região 1q25:** Alguns dos marcadores genéticos da região, D1S2818, D1S3425, D1S240, D1S444, D1S1642 estão representados em preto; alguns dos transcritos já mapeados na região são fornecidos como pontos de referência e estão representados em roxo. As ESTs e os "Full-length" identificados no presente trabalho estão representados em verde e vermelho respectivamente. A escala é dada em Megabases (Mb).

```
catggcaacatgagcagtgctgagataattggttctacaaatcttataattctgctagag
         M  S  S  A  E  I  I  G  S  T  N  L  I  I  L  L  E
gatgaagtctttgccgatttttcaacacatttctttccctcccggtttttggtcagaca
 D  E  V  F  A  D  F  F  N  T  F  L  S  L  P  V  F  G  Q  T
ccattttatactgttgaaaattcacagtggagcttgtggccagaaatacattgtaacttg
 P  F  Y  T  V  E  N  S  Q  W  S  L  W  P  E  I  P  C  N  L
attgccaaatacaaagggttattgacctggttggaaaaatgccgattacctttcttctgt
 I  A  K  Y  K  G  L  L  T  W  L  E  K  C  R  L  P  F  F  C
aaaacaaacttgtgtttccattacattctctgtcaggagttcatcagtttcattaagtcc
 K  T  N  L  C  F  H  Y  I  L  C  Q  E  F  I  S  F  I  K  S
ccagaaggagccaagatgatgagatggaaaaaggcagaccagtggctactccagaaatgc
 P  E  G  A  K  M  M  R  W  K  K  A  D  Q  W  L  L  Q  K  C
attggcggggtcagagggatgtggcgcttctattcctacctcacaggcagtgcaggtgaa
 I  G  G  V  R  G  M  W  R  F  Y  S  Y  L  T  G  S  A  G  E
gaattggtggatttctggatccttgctgagaacatcctgagcatagatgagatggacctg
 E  L  V  D  F  W  I  L  A  E  N  I  L  S  I  D  E  M  D  L
gaagtgagagactactacctgtccctcctcctcatgctgagggccactcatctgcaggag
 E  V  R  D  Y  Y  L  S  L  L  L  M  L  R  A  T  H  L  Q  E
ggctccagggtggtaaccctctgtaacatgaacatcaagtccctcctgaacctctccatc
 G  S  R  V  V  T  L  C  N  M  N  I  K  S  L  L  N  L  S  I
tggcatcccaaccaatcaaccactaggagggagatcctgagccacatgcagaaagtggct
 W  H  P  N  Q  S  T  T  R  R  E  I  L  S  H  M  Q  K  V  A
ctgttcaaactccagagctattggcttcccaacttttacacccacaccaagatgaccatg
 L  F  K  L  Q  S  Y  W  L  P  N  F  Y  T  H  T  K  M  T  M
gccaaggaggaagcatgccatggtctgatgcaagagtacgagactcgcttatacagcgtt
 A  K  E  E  A  C  H  G  L  M  Q  E  Y  E  T  R  L  Y  S  V
tgctacacccacataggagggctccctctgaacatgagcatcaagaagtgccaccacttt
 C  Y  T  H  I  G  G  L  P  L  N  M  S  I  K  K  C  H  H  F
cagaaacggtactcaagcaggaaagccaagaggaagatgtggcaattggtagatcctgac
 Q  K  R  Y  S  S  R  K  A  K  R  K  M  W  Q  L  V  D  P  D
tcttggtctctggaaatggatctcaagccagatgctattggtatgcccctacaggagaca
 S  W  S  L  E  M  D  L  K  P  D  A  I  G  M  P  L  Q  E  T
tgtcctcaagagaaggtggttatacaaatgccttccctgaaaatggcttcttcaaaggaa
 C  P  Q  E  K  V  V  I  Q  M  P  S  L  K  M  A  S  S  K  E
acaagaatcagttccctggaaaaggatatgcattatgcaaaaatatccagcatggagaat
 T  R  I  S  S  L  E  K  D  M  H  Y  A  K  I  S  S  M  E  N
aaagccaagagccacctccacatggaagccccctttgagacaaaggtctctacccacctg
 K  A  K  S  H  L  H  M  E  A  P  F  E  T  K  V  S  T  H  L
aggactgtcatccccattgtcaatcactcctccaagatgacaattcagaaggccatcaag
 R  T  V  I  P  I  V  N  H  S  S  K  M  T  I  Q  K  A  I  K
caaagcttctccttaggatacatccacttggccttgtgtgctgatgcctgtgcagggaac
 Q  S  F  S  L  G  Y  I  H  L  A  L  C  A  D  A  C  A  G  N
cctttccgggaccacctgaagaagctgaatttgaaagtggagatccaacttcttgacctc
 P  F  R  D  H  L  K  K  L  N  L  K  V  E  I  Q  L  L  D  L
tggcaggacttgcagcatttcctcagtgtccttctgaataacaaaaagaatgggaatgca
 W  Q  D  L  Q  H  F  L  S  V  L  L  N  N  K  K  N  G  N  A
atctttcgtcacttgctgggtgacagaatctgcgagctctacctgaatgagcagattggt
 I  F  R  H  L  L  G  D  R  I  C  E  L  Y  L  N  E  Q  I  G
ccgtgcttaccactcaaatcccaaaccattcagggcctgaaggaactattgccctctggg
 P  C  L  P  L  K  S  Q  T  I  Q  G  L  K  E  L  L  P  S  G
gatgtgatcccctggattcccaaagcccagaaggagatttgcaagatgctcagtccctgg
 D  V  I  P  W  I  P  K  A  Q  K  E  I  C  K  M  L  S  P  W
tatgatgagtttctagatgaagaggactactggtttctccttttacggtaggaaggact
 Y  D  E  F  L  D  E  E  D  Y  W  F  L  L  F  T  V  G  R  T
ttgggttaggaaggaatcatgaggatgagggaagaagaaagagtaattactgttttaaaa
 L  G  -
gggttatgtgttaaagtaaatgaaattgttatttttcctagagtcaaccaaagatcagca
tggtccctgttgttctaaagctaaacctctcaaggaaaaggactcagtgcataagatgac
tttggtgaaaccccgtctctactaaaaatacaaaaaattagccgggcgtagtggcgggcg
cctgtagtcccagctacttgggaggctgaggcaggagaatggtgtgaacccgggaggcgg
agcttgcagtgagccgagatcccgccactgcacgccagcctgggcgacagagcgagactc
cgtctcaaaaaaaaaaaaaaaaaaaaaaaaag
```

Figura 8 - **Seqüência completa do gene RGSL2:** Seqüência completa do transcrito correspondente ao FL1 e a fase aberta de leitura predita, codificando uma proteína de 540 aminoácidos. O domínio RGS é destacado em vermelho.

Com base na comparação entre as seqüências completas de cDNA identificadas a partir do nosso banco de dados e as seqüências anotadas tanto pelo projeto Genoma Humano, quanto pelo grupo de CARPTEN et al. (2000) pudemos observar uma divergência nos dados. Um total de 12 seqüências de cDNAs completas foram identificadas exclusivamente pelo grupo de CARPTEN et al. (2000). Com o objetivo de verificar o porque esses genes não haviam sido identificados através do nosso banco de dados, os mesmos foram inspecionados manualmente. Após inspeção manual, a divergência dos dados, pôde ser explicada pelo fato de que:

08 genes, LAMC1; LAMC2; KIAA0479; KIAA0250; NCF2; C1ORF28; C1ORF15 e B3GALT2, não mapearam nos clones genômicos da região 1q25 levantados através do Map Viewer;

02 genes, GE36 e o C1ORF13, tiveram suas seqüências depositadas no GenBank após nosso banco de dados ter sido atualizado em março de 2001;

02 genes, RNASEL e a PDC, estão identificados como ESTs em nosso banco de dados.

O número de genes mapeados na região 1q25 por CARPTEN et al. (2000) e não identificados através de nossa abordagem é significativo, sugerindo a existência de incongruências entre o mapa físico elaborado pelo grupo e o mapa físico disponibilizado pelo consórcio público. No entanto, é prematuro afirmar qual dos mapas está correto, uma vez que a resposta definitiva provavelmente será alcançada em alguns anos quando a região estiver completamente seqüenciada e montada.

5.3 IDENTIFICAÇÃO DE SEQÜÊNCIAS PARCIAIS (ESTS) DA REGIÃO 1q25

A segunda etapa do trabalho foi então isolar outros transcritos ainda não completamente caracterizados e mapeados na região 1q25. Sabe-se que a identidade entre seqüência genômica e uma seqüência transcrita é um indicador de que a seqüência genômica em questão representa um gene. Entretanto, os bancos de dados de seqüências transcritas, em especial, os bancos de dados de ESTs contêm uma fração significativa de seqüências contaminantes ou que representam artefatos derivados de DNA intrônico que sem dúvida, restringem sua utilidade. Para superar essa limitação, essa etapa do trabalho, foi baseada na seleção de ESTs seguindo os seguintes critérios: estar presente nos clones genômicos pertencentes a região 1q25 e apresentar estrutura de "splicing" quando alinhadas contra a seqüência genômica. Considerar a ocorrência de "splicing", representa um critério bastante estringente que reduz drasticamente o nível de contaminação genômica nos bancos de dados de ESTs embora acabe eliminando muitas ESTs verdadeiras derivadas da porção 3'dos transcritos. Na figura 9, podemos observar o alinhamento entre a EST 1 e o clone genômico correspondente, pertencente ao intervalo 1q25, e a presença da estrutura de "splicing".

Um número total de 59 "clusters" de ESTs com "splicing" foi encontrado em nosso Banco de Dados. Após eliminação manual das redundâncias existentes, devido à sobreposição dos clones genômicos, esse número de "clusters" foi reduzido a 32 candidatos. Os alinhamentos entre as ESTs selecionadas e os clones genômicos foram inspecionados manualmente, através de BlastN, e o número de ESTs candidatas novamente sofreu uma redução. Nesse caso, a inspeção manual dos alinhamentos teve por objetivo: i) excluir a possibilidade de que a seqüência selecionada alinhasse melhor em outra região cromossômica que não a região 1q25, devido a presença de regiões repetitivas, ii) identificar ESTs que correspondessem a clones de cDNA "full-length" que tivessem sido submetidos ao GenBank após a atualização de nosso banco de dados em Março de 2001 e iii) confirmar a presença de estrutura de "splicing"

representada não só pela descontinuidade do alinhamento entre a EST e a seqüência genômica mas também pela identificação de sítios de "splincing" conservados.

Dessa forma, nosso conjunto de candidatos, sofreu nova redução uma vez que das 32 seqüências selecionadas em nossa análise inicial 4 seqüências foram eliminadas por representarem seqüências repetitivas não detectadas por nosso procedimento de mascarar elementos repetitivos (essas seqüências alinhavam em diversas regiões genômicas), 4 seqüências correspondiam a genes "full-length" que haviam sido submetidos ao GenBank após a atualização de nosso banco de dados e 14 seqüências não apresentavam sítios de "splicing" conservados. Dessa forma, nosso conjunto de ESTs candidatas passou a ser composto por 10 ESTs que estão listadas na Tabela 4 e Figura 7. Cabe ressaltar que os programas de predição de genes não conseguiram identificar essas regiões transcritas e por essa razão nenhuma dessas seqüências está anotada no UCSC map (Golden Path).

Tabela 4 - **ESTs mapeadas na região 1q25 através do banco de dados do Transcriptoma Humano:** Características e padrão de expressão em próstata referentes às ESTs candidatas mapeadas na região 1q25.

EST	Número de acesso	Unigene "cluster"	Sinal de Poly A	IMAGE clone	Expressão em próstata
1	AA757244	Hs.121200 (size 1)	Sim	IMAGp998A223330Q2	Sim
2	BE566336	-	Não	IMAGp958G08365Q2	Não
3	AI187857	-	Não	IMAGp998D064417Q2	Pseudogene
4	AI018001	Hs.131220 (size 5)	Sim	IMAGp998N074130Q2	Pseudogene
5	AW205088	Hs.116070 (size 19)	Sim	IMAGp998C217408Q2	Sim
6	AI201365	Hs.344374 (size 2)	Não	IMAGp998O054461Q2	Não
7	BE906931	-	Não	IMAGp998O089705Q2	Não
8	AI243690	-	Não	IMAGp998F114536Q2	Não
9	AA460184	-	Não	IMAGp998F071962Q2	Não
10	AI025599	-	Não	IMAGp998G244172Q2	Não

```
>gi|12666205|emb|AL138776.10|AL138776  Human DNA sequence from clone RP11-20H6 on
chromosome 1q25.1-31.1,
            complete sequence [Homo sapiens]
          Length = 100549

 Score = 87.7 bits (44), Expect = 8e-15
 Identities = 44/44 (100%)
 Strand = Plus / Plus


Query: 1     aatgggttgacagcccagatatggcaggaccagactgcaatggGT 44
              ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 62765 aatgggttgacagcccagatatggcaggaccagactgcaatggg 62808

Score =  634 bits (320), Expect = e-179
 Identities = 320/320 (100%)
 Strand = Plus / Plus


Query: 43  AGgaatctcatacagcccattcagactaaagggtgatggaaagaagcagaggaaagcagaa 102
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 66863 ggaatctcatacagcccattcagactaaagggtgatggaaagaagcagaggaaagcagaa 66922


Query: 103   acagcaagagtgactgagacctgctgtcatctatatagaagcctttctgatacatgtcag 162
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 66923 acagcaagagtgactgagacctgctgtcatctatatagaagcctttctgatacatgtcag 66982


Query: 163   agagtgacatgaagtctcccctctttatacacacgtatgaaaatacatctatccctcatt 222
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 66983 agagtgacatgaagtctcccctctttatacacacgtatgaaaatacatctatccctcatt 67042


Query: 223   ttgagagcctccaactgatccaagtgtcatccggaggaggaggtggcccgtatgtgcagg 282
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 67043 ttgagagcctccaactgatccaagtgtcatccggaggaggaggtggcccgtatgtgcagg 67102


Query: 283   acctcagggtgccagccctcacccatacacctccagggctgcaggtcattagccagaata 342
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 67103 acctcagggtgccagccctcacccatacacctccagggctgcaggtcattagccagaata 67162


Query: 343   aaggatgtccacattcacaa 362
              ||||||||||||||||||
Sbjct: 67163 aaggatgtccacattcacaa 67182
```

Figura 9 - **Representação do alinhamento de uma das ESTs candidatas, destacando a estrutura de "splicing":** Alinhamento entre a EST 1 (AA757244) e o clone genômico AL138776. Em vermelho estão representados os sítios doador e aceptor de "splicing".

Os clones IMAGE correspondentes a cada uma das ESTs selecionadas foram comprados e os insertos de cDNA foram completamente seqüenciados. As análises das seqüências correspondentes as ESTs 3 e 4 revelaram uma alta similaridade ao nível de aminoácidos com uma proteína associada a microtúbulo (MAP1A/1B acc. No. NM_022818). As seqüências, entretanto, continham diversos códons de terminação e inserções de elementos repetitivos. Com base nessas características, nós concluímos que essas duas ESTs correspondem a pseudogenes não processados mapeados na região 1q25. Essa observação foi confirmada através da anotação da seqüência do BAC correspondente (AL078644) feita pelo Consórcio de Seqüenciamento do Genoma Humano.

Uma vez que nosso tecido de interesse é próstata, a expressão de cada um dos novos transcritos foi testada através de RT-PCR. Das 10 ESTs selecionadas, 2 (EST1 e EST5) mostraram-se positivamente expressas em próstata. Uma vez que essas duas ESTs podem representar novos candidatos a genes supressores de tumor para o câncer de próstata hereditário, suas seqüências completas foram geradas. Para tanto utilizamos a abordagem de RACE. As seqüências geradas foram analisadas quanto a presença de fase aberta de leitura (ORF) e também quanto a similaridade com genes já identificados ou domínios protéicos. Além disso, o padrão de expressão desses dois candidatos foi avaliado através de RT-PCR utilizando um painel de cDNA composto por 20 diferentes tecidos normais.

A EST1 corresponde a um transcrito de 2.060 pares de bases (AF510428) com uma alta similaridade (94%) em nível de nucleotídeos com uma proteína hipotética de *Macaca fascicularis* (AB071118). O gene está organizado em 15 éxons (Tabela 5) distribuídos ao longo de 67.182 pares de bases da seqüência genômica. A análise da seqüência gerada revelou uma fase aberta de leitura que codifica para uma proteína composta por 500 aminoácidos. A seqüência completa do transcrito pode ser vista na Figura 10. A análise da seqüência da proteína predita utilizando o InterProScan (http://www.ebi.ac.uk/interpro/scan.html) revelou a presença de um domínio RGS (Regulador de Sinalização da proteína G) altamente conservado composto por 120 aminoácidos. O transcrito gerado correspondente a EST1, foi nomeado RGSL1 após consulta ao HUGO (Genome Nomenclature Committee). Além disso os dados

referentes ao padrão de expressão, obtidos através de RT-PCR, revelaram que esse transcrito é altamente expresso em testículo e é expresso em níveis mais baixos em placenta, traquéia, medula óssea, pulmão e próstata. Os dados de expressão podem ser vistos na Figura 11A.

A EST5 corresponde a um transcrito de 612 pares de bases (AF510427). A análise da seqüência do transcrito não revelou uma fase aberta de leitura. O gene está organizado em 4 éxons (Tabela 6) distribuídos ao longo de 70.087 pares de bases da seqüência genômica. Os dados de expressão revelaram que o transcrito é expresso em testículo, placenta, traquéia, medula óssea e próstata. Os dados de expressão podem ser vistos na Figura 11B.

```
agatgctcagtccctggtatgatgagtttctagatgaagaggactactggtttctcctttt
tacgtaactaaaatgtcctttgaggagctttgctacaagaacccaaagatggccatacag
            M  S  F  E  E  L  C  Y  K  N  P  K  M  A  I  Q
aagatcagtgatgactacaaaatatactgtgagaaagcacctaaaatagattttaaaatg
 K  I  S  D  D  Y  K  I  Y  C  E  K  A  P  K  I  D  F  K  M
gaaattatcaaggaaacaaagacagtttcacgctcaaataggaaaatgtccttgctcaaa
 E  I  I  K  E  T  K  T  V  S  R  S  N  R  K  M  S  L  L  K
agaactcttgtaaggaagccatcaatgagacccagaaacttgacagaagtcctcttaaat
 R  T  L  V  R  K  P  S  M  R  P  R  N  L  T  E  V  L  L  N
acacagcacttggagttcttcagggagttcctcaaggaacggaaggctaaaatcccattg
 T  Q  H  L  E  F  F  R  E  F  L  K  E  R  K  A  K  I  P  L
caatttctcacagctgtacagaagatcagtatagagaccaatgaaaagatttgcaagtct
 Q  F  L  T  A  V  Q  K  I  S  I  E  T  N  E  K  I  C  K  S
ctcatagaaaatgtaatcaagactttcttccaaggccaactctctcctgaagaaatgctg
 L  I  E  N  V  I  K  T  F  F  Q  G  Q  L  S  P  E  E  M  L
cagtgtgatgcccctattatcaaagaaatcgcttccatgcgtcatgtcaccacaagcaca
 Q  C  D  A  P  I  I  K  E  I  A  S  M  R  H  V  T  T  S  T
ctgttaacgctccagggacatgttatgaaatctatagaagaaaagtggtttaaagattat
 L  L  T  L  Q  G  H  V  M  K  S  I  E  E  K  W  F  K  D  Y
caagacctgttcccacctcaccatcaggaggtggaagtgcaaagtgaagtacaaatttcg
 Q  D  L  F  P  P  H  H  Q  E  V  E  V  Q  S  E  V  Q  I  S
tctaggaagccctcaaagatagtgtcaacttacctacaggaatcccagaagaaaggctgg
 S  R  K  P  S  K  I  V  S  T  Y  L  Q  E  S  Q  K  K  G  W
atgagaatgatcagctttatcaggagtttttgcaagtaccgcagatttatgttgaatcct
 M  R  M  I  S  F  I  R  S  F  C  K  Y  R  R  F  M  L  N  P
agtaagcgccaggaatttgaagactatcttcaccaggaaatgcaaaatagcaaggaaaat
 S  K  R  Q  E  F  E  D  Y  L  H  Q  E  M  Q  N  S  K  E  N
ttcaccacagcacacaacacatcgggacgttcagctccaccctccacaaatgtccggagt
 F  T  T  A  H  N  T  S  G  R  S  A  P  P  S  T  N  V  R  S
gcagaccaagagaatggagaaataacccttgtaaagcgtcgtatatttggccacaggatt
 A  D  Q  E  N  G  E  I  T  L  V  K  R  R  I  F  G  H  R  I
atcactgtcaactttgcgatcaatgatctatatttcttttctgaaatggagaaatttaat
 I  T  V  N  F  A  I  N  D  L  Y  F  F  S  E  M  E  K  F  N
gatctggtcagttcagcccacatgctgcaggtcaaccgggcatataatgagaatgatgtg
 D  L  V  S  S  A  H  M  L  Q  V  N  R  A  Y  N  E  N  D  V
atcctaatgcggtccaaaatgaacattatccaaaaactcttcctgaattctgacatccct
 I  L  M  R  S  K  M  N  I  I  Q  K  L  F  L  N  S  D  I  P
ccaaagctgagggtgaatgtccctgagttccagaaggatgccatccttgctgccatcaca
 P  K  L  R  V  N  V  P  E  F  Q  K  D  A  I  L  A  A  I  T
gagggctacctagatcggagcgtcttccatggggctatcatgtctgtcttccccgttgtt
 E  G  Y  L  D  R  S  V  F  H  G  A  I  M  S  V  F  P  V  V
atgtacttctggaaaaggttttgtttctggaaggcaacccgctcttacttacagtatagg
 M  Y  F  W  K  R  F  C  F  W  K  A  T  R  S  Y  L  Q  Y  R
gggaagaagttcaaggacagaaaaagccctcctaaatctacggacaagtatcctttctcg
 G  K  K  F  K  D  R  K  S  P  P  K  S  T  D  K  Y  P  F  S
agtggaggagacaatgccatcttaaggttcaccttgctcagaggtattgagtggttgcag
 S  G  G  D  N  A  I  L  R  F  T  L  L  R  G  I  E  W  L  Q
cctcaacgggaagcaataagttcagttcaaaattcttcatcaagcaaacttactcagcca
 P  Q  R  E  A  I  S  S  V  Q  N  S  S  S  S  K  L  T  Q  P
agactcgtggtatctgccatgcagctgcatcccgtccagggacaaaaattatcctacatc
 R  L  V  V  S  A  M  Q  L  H  P  V  Q  G  Q  K  L  S  Y  I
aaaaaagagaagtaatcaagcgagacccccagcagagataaatcatctcttagaggcctc
 K  K  E  K  -
CtaacactgacagaaacttttctggtcaaaaatgaaaggtcctgGtaccatcttccccag
AaacgatcaagaagggcaatgggttgacagcccagatatggcagGaccagactgcaatgg
GaatctcatacagcccattcagactaaagggtgatggaaagaagCagaggaaagcagaaa
CagcaagagtgactgagacctgctgtcatctatatagaagccttTctgatacatgtcaga
GagtgacatgaagtctcccctctttatacacacgtatgaaaataCatctatccctcattt
TgagagcctccaactgatccaagtgtcatccggaggaggaggtgGcccgtatgtgcagga
Cctcagggtgccagccctcacccatacacctccagggctgcaggtcattagccagaataa
aggatgtccacattcacaa
```

Figura 10 - **Seqüência completa do gene RGSL1:** Seqüência completa do transcrito correspondente a EST1 e a fase aberta de leitura predita, codificando uma proteína de 500 aminoácidos. O domínio RGS é destacado em vermelho.

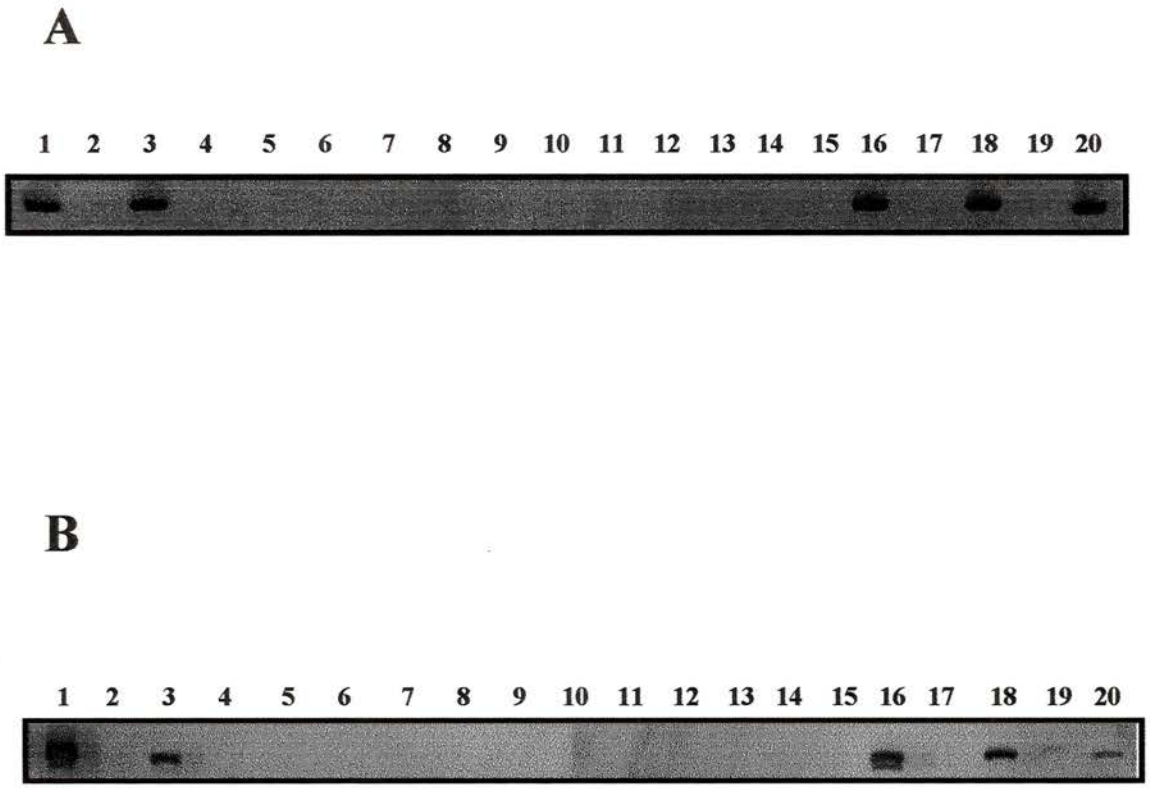


Figura 11 - **Padrão de expressão do gene RGSL1 e EST 5 mapeados na região 1q25:** Amostras de cDNA obtidas a partir de 20 tecidos humanos normais diferentes foram utilizadas nas reações de RT-PCR nested com iniciadores específicos para a EST1 (A) e para a EST5 (B) mapeadas na região 1q25. Cada canaleta corresponde a: 1. traquéia; 2. cérebro fetal; 3. medula óssea; 4. glândula adrenal; 5. músculo esquelético; 6. baço; 7. fígado fetal; 8. fígado; 9. glândula salivar; 10. timo; 11. rim; 12. cordão espinhal; 13. útero; 14. coração; 15. cólon; 16. placenta; 17. intestino; 18. testículo; 19. pulmão e 20. próstata.

Tabela 5 - **Intron/Exon Boundaries do gene RGSL1:** As letras maiúsculas e minúsculas correspondem as seqüências dos exons e introns respectivamente. Os sítios doador e aceptor de "splicing" estão representados em negrito.

Exon	Sítio aceptor	Sítio doador	Tamanho do exon (bp)
1	----------	CCTTTTTACG**gt**aggaagga	65 bp
2	ttcacagacaTAACTAAAAT	CCTAAAATAGgcaagtgttt	105 bp
3	tttttttt**ag**ATTTTAAAAT	TGAGACCCAG**gt**gagatata	106 bp
4	tgttttcc**ag**AAACTTGACA	CTCTCTCCTG**gt**atgcatcc	194 bp
5	cattttct**ag**AAGAAATGCT	AAGAAAAGTG**gt**gagtatac	118 bp
6	cttctgac**ag**GTTTAAAGAT	GGAATCCCAG**gt**tagtgaag	121 bp
7	ctttcccc**ag**AAGAAAGGCT	AGCAAGGAAA**gt**aagtcatt	130 bp
8	tctctctc**ag**ATTTCACCAC	AAATGGAGAA**gt**aagttttc	175 bp
9	tctttttt**ag**ATTTAATGAT	AAAGCTGAGG**gt**aagaaaga	139 bp
10	ctgtgggt**ag**GTGAATGTCC	TCTGGAAAAG**gt**aactgttt	125 bp
11	Cttcccac**ag**GTTTTGTTTC	TCGAGTGGAG**gt**aagctcca	110 bp
12	ccttcccc**ag**GAGACAATGC	GTTCAAAATT**gt**gagttgga	87 bp
13	ttctcttt**ag**CTTCATCAAG	CCGTCCAGGG**gt**aggcatga	67 bp
14	ctcccttc**ag**ACAAAAATTA	ACTGCAATGG**gt**aagacttg	199 bp
15	tgtttttt**ag**GAATCTCATA	----------	319 bp

Tabela 6 - **Intron/Exon Boundaries d o gene correspondente à EST 5:** As letras maiúsculas e minúsculas correspondem as seqüências dos exons e introns respectivamente. Os sítios doador e aceptor de "splicing" estão representados em negrito.

Exon	Sítio aceptor	Sítio doador	Tamanho do exon (bp)
1	----------	GGACTATAAG**gt**aggagaaa	72 bp
2	ttctatctac**ag**GCACAATG	CTGCAGAGAA**gt**gagtcaaa	64 bp
3	tccttcctttt**ag**TGTAAATA	AGCATGAATG**gt**cagtacac	161 bp
4	ctgggttcac**ag**CTCAACCA	----------	285 bp

5.4 GENES RGS MAPEADOS NA REGIÃO 1q25

Até o momento, como resultado de nosso trabalho, nós conseguimos identificar e caracterizar duas proteínas RGS distintas (a partir da EST1 e do "Full-length"1) que correspondem, segundo a nomenclatura HUGO, a RGSL1 e RGSL2 respectivamente mapeadas no intervalo 1q25. Ambas são expressas em próstata e por essa razão podem ser consideradas possíveis candidatos a gene supressor de tumor para o câncer de próstata hereditário. As proteínas RGS são caracterizadas pela presença de um domínio constituído de 120 aminoácidos e possuem função GTPase. Essas proteínas regulam negativamente a sinalização das proteínas G heterotriméricas. Elas estão envolvidas na modulação de uma variedade de funções celulares incluindo proliferação, diferenciação, resposta a neurotransmissores, desenvolvimento embrionário e além disso são potenciais genes supressores de tumor. Recentemente, um novo membro da família RGS, a RGS14, foi caracterizada como sendo um gene alvo induzido por p53 em resposta ao "stress" genotóxico. A super expressão de RGS14 inibe a proteína G acoplada a um receptor de fator de crescimento mediando a ativação do padrão de sinalização MAP-kinase.

Para que nós pudéssemos começar a explorar o possível envolvimento das duas proteínas RGSL mapeadas no intervalo 1q25, nós realizamos experimentos utilizando o LightCycler-Based Real-Time RT-PCR para medir os níveis de expressão de cada um dos transcritos em cDNA de próstata normal e também em duas linhagens tumorais de próstata (DU145 e PC3). A utilização de linhagens tumorais obtidas a partir de tumores de próstata hereditários não é ideal, mas pode ser justificada pelo fato de genes envolvidos em tumores hereditários estarem freqüentemente envolvidos em tumores esporádicos correspondentes.

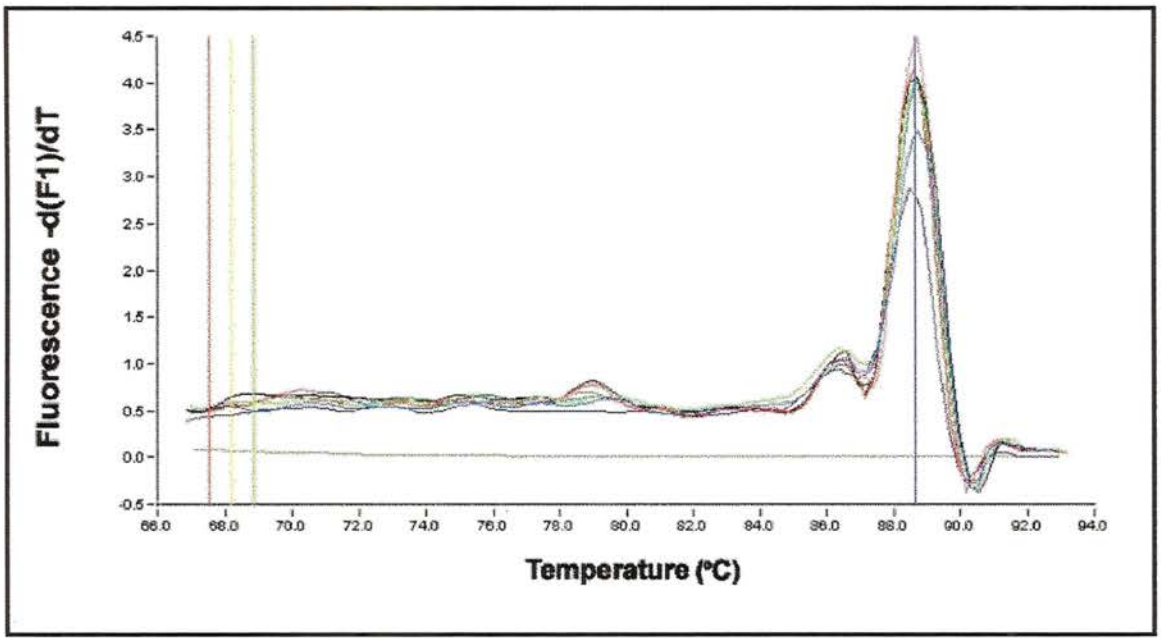
Os níveis de expressão foram determinados como uma taxa entre os genes RGSL1 e RGSL2 os genes de referência GAPDH e β-actina para corrigir variações nas quantidades iniciais de RNA. A expressão de ambos os transcritos foi facilmente detectada em próstata normal. Entretanto, a expressão dos transcritos não pode ser detectada nas duas linhagens tumorais testadas (DU145 e PC3) o que sugere que esses dois genes possam desempenhar algum papel no câncer de próstata. Os resultados

obtidos a partir dos experimentos de Real-Time PCR podem ser visualizados na Tabela 7 e Figura 12. Atualmente estamos testando o nível de expressão dos genes RGS em amostras de próstata de pacientes do Hospital A. C. Camargo.

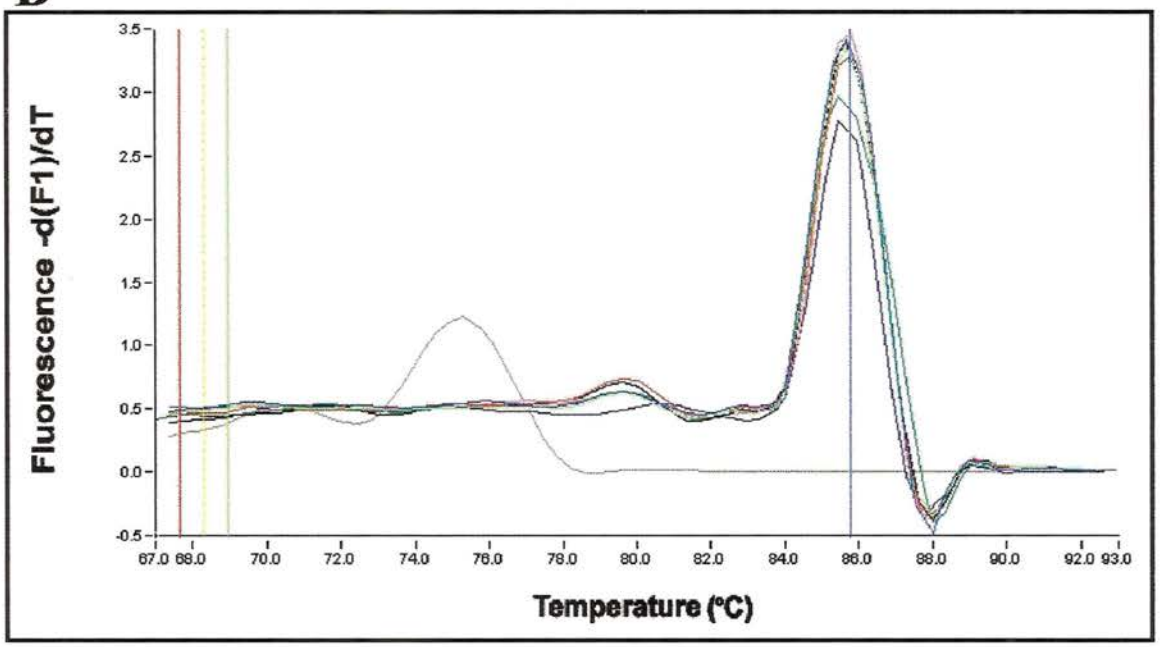
Tabela 7 - **Dados de expressão diferencial dos genes RGSL1 e RGSL2 em tecido normal e linhagens tumorais de próstata.**

Gene	Amostras	CP médio	Anneling/Aquisição
GAPDH	Normal	19,10	62°C / 85°C
	DU-145	14,99	
	PC3	14,51	
	Testículo	26,60	
	No	-	
β-actina	Normal	14,88	62°C / 82°C
	DU-145	12,10	
	PC3	11,94	
	Testículo	22,94	
	No	-	
RGSL2	Normal	29,91	62°C / 82°C
	DU-145	-	
	PC3	-	
	Testículo	35,41	
	No	-	
RGSL1	Normal	38.23	65°C / 80°C
	DU-145	-	
	PC3	-	
	Testículo	39.84	
	No	-	

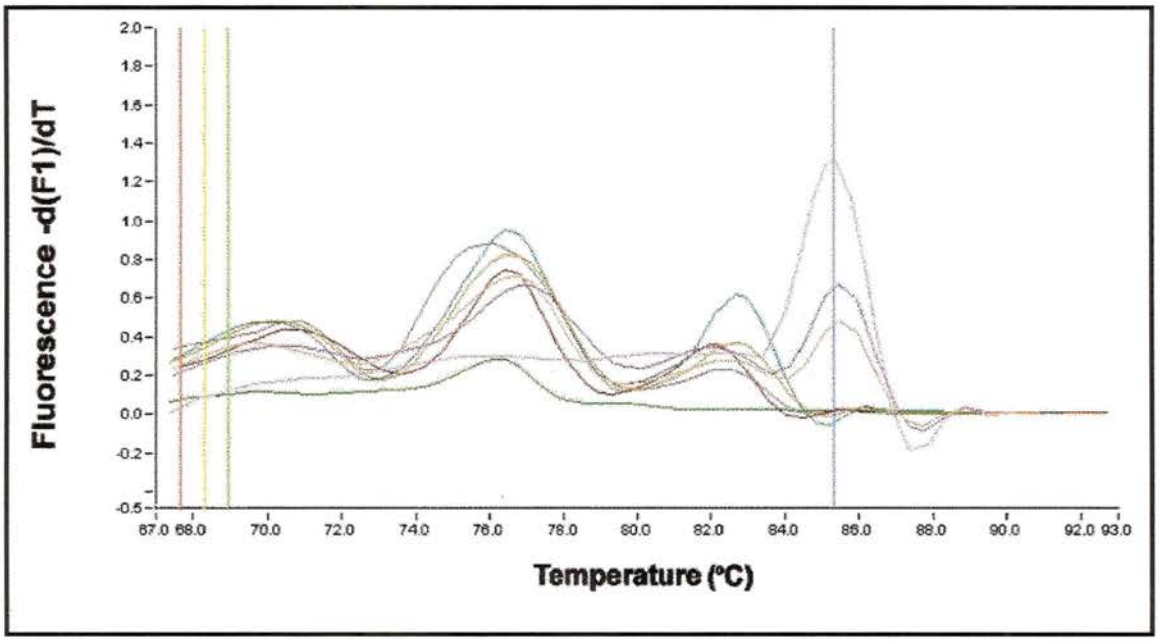
A


Fluorescence -d(F1)/dT
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
66.0
68.0
70.0
72.0
74.0
76.0
78.0
80.0
82.0
84.0
86.0
88.0
90.0
92.0
94.0
Temperature (°C)

B


Fluorescence -d(F1)/dT
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
67.0
68.0
70.0
72.0
74.0
76.0
78.0
80.0
82.0
84.0
86.0
88.0
90.0
92.0
93.0
Temperature (°C)

C



D

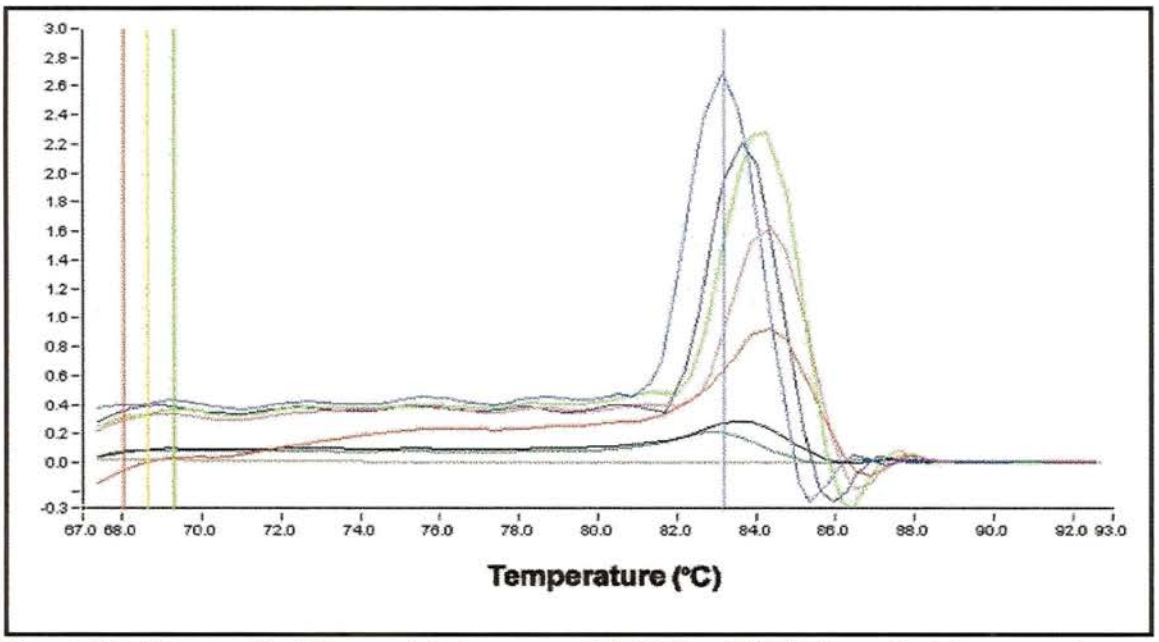


Figura 12: **Expressão diferencial dos genes RGSL1 e RGSL2 em tecido normal e linhagens tumorais de próstata:** (A) Perfil da curva de "melting" para o gene constitutivo GAPDH; (B) Perfil da curva de "melting" para o gene constitutivo β-actina; (C) Perfil da curva de "melting" para o gene RGSL2; (D) Perfil da curva de "melting" para o gene RGSL1.

5.5 CONSIDERAÇÕES FINAIS

Uma etapa fundamental para a análise de uma região candidata a abrigar genes relacionados a determinadas doenças genéticas é gerar um mapa da região em questão contendo todos os transcritos ali mapeados. Com a disponibilidade da seqüência genômica humana e com o desenvolvimento de abordagens computacionais, a busca pela identificação de genes em regiões candidatas foi acelerada. Entretanto, os programas de computador desenvolvidos para fazer predição gênica representam apenas uma solução parcial para a identificação de genes no genoma. Isso se deve principalmente às altas taxas de falso negativo e falso positivo obtidas na identificação "in silico" de genes. Assim, a melhor alternativa parece ser fazer uma cuidadosa seleção de seqüências transcritas, identificadas *in silico*, seguida de validação experimental.

Utilizando esse tipo de abordagem, nosso grupo recentemente identificou 19 novos genes no cromossomo 21 que não haviam sido preditos pelos programas de computador e por essa razão permaneciam sem nenhum tipo de anotação. Em contrapartida, Reymond *et al.* mostrou que a maior parte dos transcritos do cromossomo 21 definidos somente por programas de predição gênica não correspondem a genes verdadeiros, confirmando assim, a importância da validação experimental.

Embora seja possível que outros genes dentro do intervalo 1q25 ainda permaneçam não identificados, nós conseguimos identificar 17 genes adicionais dentro desse intervalo, o que aumentou o número total de genes descritos nessa região em aproximadamente 60% (Figura 1). Oito dos transcritos identificados representam novos candidatos a gene supressor de tumor para o câncer de próstata hereditário uma vez que mostraram um padrão de expressão positivo em próstata. É importante lembrar que se houverem genes ainda não representados por ESTs, estes não poderão ser identificados neste trabalho. Daí a importância de se utilizar outras abordagens tais como buscas de similaridade com outros organismos ou hibridizações em microarray.

A abordagem desenvolvida nesse trabalho é simples quando comparada a outras mais dispendiosas como por exemplo "cDNA selection" e "exon trapping".

Além disso, essa metodologia permite a identificação de transcritos expressos em baixos níveis e representados por poucas ESTs. Em oposição a predição de genes feita por programas de computador, essa abordagem se mostra mais eficiente uma vez que não requer uma seqüência genômica de alta qualidade, consegue detectar genes com pequenas fases abertas de leitura e até mesmo RNAs não codificantes. Dessa forma, nós concluimos que a abordagem utilizada é funcional, representando uma estratégia alternativa para a identificação de genes que podem estar envolvidos em diversas doenças de interesse.

Referências Bibliográficas

6 REFERÊNCIAS BIBLIOGRÁFICAS

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

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

Bahabri SA, Suwairi WM, Laxer RM, et al. The camptodactyly-arthropathy-coxa vara-pericarditis syndrome: clinical features and genetic mapping to human chromosome 1. **Arthritis Rheum** 1998; 41:730-5.

Ballabio A, Brown S, Fisher E. Strategies for gene discovery in mammalian systems. In: Birren B, Green ED, Klapholz S, Myers RM, Roskams J, editors. **Genome analysis: detecting genes**. New York; Cold Spring Harbor Laboratory Press; 1998. p.1-4.

Bishop DT. Genetic predisposition to cancer: an introduction. In: Eeles RA, Ponder BAJ, Easton DF, Horwich A, editors. **Genetic predisposition to cancer.** Oxford: Chapman & Hall Medical; 1996. p.3-15.

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

Bonalume Neto R. Brazilian scientists team up for cancer genome project. **Nature** 1999; 398:450.

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

Buckbinder L, Velasco-Miguel S, Chen Y, et al. The p53 tumor suppressor targets a novel regulator of G protein signaling. **Proc Natl Acad Sci USA** 1997; 94:7868-72.

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

Câncer de próstata: epidemiologia. Disponível em: http://www.inca.gov.br/cancer/prostata/. Acesso em: 15 ago. 2002.

Carpten JD, Makalowska I, Robbins CM, et al. A 6-Mb high-resolution physical and transcription map encompassing the hereditary prostate cancer 1 (HPC1) region. **Genomics** 2000; 64:1-14.

Carpten J, Nupponen N, Isaacs S, et al. Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. **Nat Genet** 2002; 30:181-4.

Carter BS, Bova GS, Beaty TH, Steinberg GD, Childs B, Isaacs WB, Walsh PC. Hereditary prostate cancer: epidemiologic and clinical features. **J Urol** 1993; 150:797-802.

Cussenot O, Valeri A. Heterogeneity in genetic susceptibility to prostate cancer. **Eur J Intern Med** 2001; 12:11-16.

De La Chapelle A, Peltomaki P. The genetics of hereditary common cancers. **Curr Opin Genet Dev** 1998; 8:298-303.

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

De Vries L, Gist Farquhar M. RGS proteins: more than just GAPs for heterotrimeric G proteins. **Trends Cell Biol** 1999; 9:138-44.

Deloukas P, Schuler GD, Gyapay G, et al. A physical map of 30,000 human genes. **Science** 1998; 282:744-6.

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

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

Easton DF. From families to chromosomes: genetic linkage, and other methods for finding cancer-predisposition genes. In: Eeles RA, Ponder BAJ, Easton DF, Horwich A, editors. **Genetic predisposition to cancer.** Oxford: Chapman & Hall Medical; 1996. p.16-39.

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

Eeles RA, Cannon-Albright L. Familial prostate cancer and its management. In: Eeles RA, Ponder BAJ, Easton DF, Horwich A, editors. **Genetic predisposition to cancer.** Oxford: Chapman & Hall Medical; 1996. p.320-30.

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

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

German Human Genome Project. http://www.dhgp.de/>. Acesso em: 20 jun. 2001.

Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. **Genome Res** 2000; 10:1631-42.

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

How Many Men Get Prostate Cancer? Disponível em: http://www.cancer.org/eprise/main/docroot/CRI/content/. Acesso em: 15 ago. 2002.

Hobbs MR, Pole AR, Pidwirny GN, et al. Hyperparathyroidism-jaw tumor syndrome: the HRPT2 locus is within a 0.7-cM region on chromosome 1q. **Am J Hum Genet** 1999; 64:518-25.

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

Homo sapiens genome view. Disponível em: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search>. Acesso em: 8 mar. 2001.

Huang X, Madan A. CAP3: A DNA sequence assembly program. **Genome Res** 1999; 9:868-77.

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

InterProScan Sequence Search. <http://www.ebi.ac.uk/interpro/scan.html>. Acesso em: 12 jun. 2002.

Isaacs WB, Coffey DS. Molecular genetics of prostate cancer. In: Vogelzang NJ, Shipley WU, Scardino PT, Coffey DS, editors. **Comprehensive textbook of genitourinary oncology**. Philadelphia: Lippincott Willians & Wilkins; 1999. p.545-52.

Iseli C, Stevenson BJ, de Souza SJ, et al. Long-Range Heterogeneity at the 3' Ends of Human mRNAs. **Genome Res** 2002; 12:1068-74.

King R. **Cancer biology**. 2nd ed. London: Prentice Hall; 2000. p.174-212: Responding to the environment: growth regulation and signal transduction.

Koelle MR. A new family of G-protein regulators - the RGS proteins. **Curr Opin Cell Biol** 1997; 9:143-7.

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

Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome.
Nature 2001; 15:860-921.

Ledley FD, DiLella AG, Kwok SC, Woo SL. Homology between phenylalanine and tyrosine hydroxylases reveals common structural and functional domains. **Biochemistry** 1985; 24:3389-94.

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

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. **Molecular cell biology**. 3th ed. New York: W. H. Freeman; 1995. p.853-924: Cell-to-cell signaling: hormones and receptors.

Morrison TB, Weis JJ, Wittwer CT. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. **Biotechniques** 1998; 24:954-8, 960, 962.

NCBI Reference Sequences. Disponível em: <http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>. Acesso em: 10 maio 2001.

Patterson H, Cooper C. From chromosomes to genes: how to isolate cancer-predisposition genes. In: Eeles RA, Ponder BAJ, Easton DF, Horwich A, editors. **Genetic predisposition to cancer.** Oxford: Chapman & Hall Medical; 1996. p.40-53.

Plaetke R, Thompson I, Sarosdy M, Harris JM, Troyer D, Arar NH. Genetic fieldwork for hereditary prostate cancer studies. **Urol Oncol** 2002; 7:19-27.

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

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

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

Rökman A, Ikonen T, Seppala EH, et al. Germline alterations of the RNASEL gene, a candidate HPC1 gene at 1q25, in patients and families with prostate cancer. **Am J Hum Genet** 2002; 70:1299-304.

Ross EM, Wilkie TM. GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins. **Annu Rev Biochem** 2000; 69:795-827.

Sato Y, Matsuo T, Ohtsuki H. A novel gene (oculomedin) induced by mechanical stretching in human trabecular cells of the eye. **Biochem Biophys Res Commun**. 1999; 259:349-51.

Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. **J Mol Med** 1997; 75:694-8.

Smith JR, Freije D, Carpten JD, Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. **Science** 1996; 274:1371-4.

Sood R, Bonner TI, Makalowska I, et al.Cloning and characterization of 13 novel transcripts and the human RGS8 gene from the 1q25 region encompassing the hereditary prostate cancer (HPC1) locus. **Genomics** 2001; 73:211-22.

Strachan T, Read AP. **Human molecular genetics**. London: BIOS Scientific; 1997. p.367-98: Human genetic diseases: Identifying human diseases genes

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

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

Thompson TC, Timme TL, Bangma CH, et al. Molecular biology of prostate cancer. In: Vogelzang NJ, Shipley WU, Scardino PT, Coffey DS, editors. **Comprehensive textbook of genitourinary oncology**. Philadelphia: Lippincott Willians & Wilkins; 1999. p.553-64.

Transcript finishing initiative. Disponível em: http://www.compbionet.org.br/transcript/. Acesso em: 10 jan. 2002.

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

Vogelstein B, Kinzler KW. The genetic basis of human cancer. New York: McGraw-Hill, 1998.

Wang L, McDonnell SK, Elkins DA, et al. Analysis of the RNASEL gene in familial and sporadic prostate cancer. **Am J Hum Genet** 2002; 71:116-23.

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

Xie D, Nakachi K, Wang H, Elashoff R, Koeffler HP. Elevated levels of connective tissue growth factor, WISP-1, and CYR61 in primary breast cancers associated with more advanced features. **Cancer Res** 2001; 61:8917-23.

Xu J, Meyers D, Freije D, et al. Evidence for a prostate cancer susceptibility locus on the X chromosome. **Nat Genet**. 1998 Oct;20(2):175-9.

Xu J, Zheng SL, Komiya A, et al. Germline mutations and sequence variants of the macrophage scavenger receptor 1 gene are associated with prostate cancer risk. **Nat Genet**. 2002 Oct;32(2):321-5.

Yamamoto H, Ito K, Ishiguro S, Takeuchi T. Gene controlling a differentiation step in the quail melanocyte. **Dev Genet** 1987; 8:179-85.

ANEXOS

ANEXO 1

QUESTÕES ÉTICAS

Todo o material biológico a ser utilizado na validação experimental do projeto em questão, será proveniente de linhagens celulares e RNA fornecido pelo Projeto Genoma Humano do Câncer / FAPESP / Ludwig.

ANEXO 2

Artigo científico contendo os resultados mostrados no presente trabalho, submetido para a revista GENE em junho de 2002.

Identification of 17 Additional Expressed Genes within the Hereditary Prostate Cancer Locus (HPC1) at 1q25.

Ana Paula M. Silva[1], Anna Christina M. Salim[1], Adriana Bulgarelli[1], Jorge Estefano S. de Souza[1], Otavia L. Caballero[1], Christian Iseli[2], Brian J. Stevenson[2], C. Victor Jongeneel[2,3], Sandro J. de Souza[1], Andrew J.G.Simpson[1] and Anamaria A. Camargo[1]*.

1 Ludwig Institute for Cancer Research – São Paulo, Brazil
2 Swiss Institute of Bioinformatics (SIB), Epalinges, Switzerland
3 Office of Information Technology, Ludwig Institute for Cancer Research-Epalinges, Switzerland.

*corresponding author
Rua Antonio Prudente 109, 4th floor.
01509-010 São Paulo SP Brazil
Phone 55 11 3388-3249 ext 248.
Fax 55 11 3207-7001
e-mail: anamaria@compbio.ludwig.org.br

Abstract

We have utilized the human genome sequence and publicly available transcript sequences (ESTs and full-length cDNA clones) to identify additional expressed sequences located at the Hereditary Prostate Cancer Locus (HPC1) on chromosome 1q25. All transcripts already described for the 1q25 locus were identified and we were able to define 17 additional expressed sequences (7 full-length cDNA clones and 10 ESTs) within this region. The majority of these newly identified transcripts were not annotated in the human genome sequence. Eight out of the 17 additional expressed sequences identified are expressed in prostate tissue and represent alternative disease gene candidates for the HPC1 locus. Expression of two identified candidates corresponding to RGS family members was detected by Real-Time PCR in normal prostate tissue but could not be detected in prostate tumor cell lines, suggesting they might have a role in prostate cancer.

Key words: Hereditary Prostate Cancer (HPC1), 1q25 region, EST, Transcription map, Genomic sequences annotation, Gene identification, RGS protein.

Introduction

The identification of disease genes is crucial for the development of molecular diagnosis and alternative therapies for inherited diseases. For the most part, this has been achieved through positional cloning, a powerful but laborious and time-consuming strategy. The availability of the physical map and draft sequence of the human genome [1, 2] now permits the acceleration of this approach via the "in silico" identification of candidate genes. At least 30 disease genes have already been positionally cloned with the aid of publicly available genome sequences [1].

However, gene identification within the human genome using gene-prediction computer programs is not straightforward [3-6] and requires confirmation from transcript data. Since there is now a massive accumulation of publicly available human transcript sequences both in the form of expressed sequence tags (ESTs) as well as full-length cDNA clones, the most efficient approach to gene identification is through the direct use of these data [7]. Identity with an authentic transcript sequence is the most reliable indicator that a genomic sequence forms part of a gene. Using this criterion, we have recently identified a number of transcripts that were not annotated within the finished sequence of chromosomes 22 and 21 by other methodologies [8, 9].

Here we have utilized the human genome sequence and publicly available transcript sequences (ESTs and full-length cDNA clones) to identify additional expressed sequences located at the Hereditary Prostate Cancer Locus (HPC1) on chromosome 1q25 [10]. Several candidate genes have already been identified in this region by experimental approaches [11, 12]. In particular, Carpten et al. recently reported that mutations in the RNASEL gene segregate independently in two out of 8

HPC1-linked families suggesting its involvement in hereditary prostate cancer [13]. However inactive RNASEL alleles are present at low frequency in the general population and the identification of other functionally significant mutations will be necessary to confirm this gene as the prostate-cancer susceptibility gene [13, 14].

Based on exhaustive pair-wise comparison between transcript sequences and the publicly available human genome sequence, we were able to identify all transcripts already described for the 1q25 locus and define 17 expressed sequences (7 full-length cDNA clones and 10 ESTs) that were not previously know to lie within this region. Gene-prediction computer programs alone did not detect the majority of these newly identified transcripts that were hence not annotated in the human genome sequence. Eight of the 17 additional expressed sequences identified (6 full-length cDNA clones and 2 ESTs) are expressed in normal prostate tissue and thus represent alternatives candidates for the HPC1 locus. Expression of two identified candidates corresponding to putative RGS family members was detected by Real-Time RT-PCR in normal prostate tissue but could not be detected in prostate tumor cell lines.

Results and Discussion

Transcriptome database and the identification of transcribed sequences mapped to the 1q25 region.

The HPC1 locus has been genetically mapped to 1q25 region in an interval flanked by the genetic markers D1S2818 and D1S1642 [12]. The physical position of these markers in the human genome draft was determined using the NCBI MapViewer tool (www.ncbi.nlm.nih.gov/genome/guide/human/). Genomic clones AL139344 and AL391065 were identified as containing the sequences related to D1S2818 and D1S1642 respectively that flank the HPC1 locus at 1q25 [12]. A total

of 44 genomic clones were then selected to cover the intervening 5.3Mb (Table 1). The slightly smaller size of this region than reported elsewhere [12] is due to the unfinished sequencing status of the interval and the presence of 3 sequencing gaps in the assembly corresponding approximately to 0.14MB. Of the 44 genomic clones selected, 24 (54.5%) were still in the form of unfinished sequence at the time of analysis. However, this did not present a problem for the transcript-guided gene identification we have used, but greatly complicates strategies based on computer predictions that require high quality sequence.

The accession number of the selected genomic clones was used as entry in our Transcriptome Database (see materials and methods). We have identified 308 clusters of transcribed sequences that map within the defined genomic region. Of these, 42 contained at least one sequence corresponding to a full-length cDNA clone (FL) and 266 were composed only of partial transcript sequences (ESTs).

Full-length cDNA clones (FL) mapped to 1q25.

We manually examined the 42 mapped FL sequences to investigate whether they were better aligned with other genomic regions. This was found to be the case in 14 instances, most of which were due to the presence of shared retroviral sequences within the full-length cDNA sequence. Of the 28 remaining sequences, 21 have already been mapped to the 1q25 region and were previously considered as putative candidate genes for HPC1 [11-12]. The seven newly identified full-length cDNA clones that we have mapped to HPC1 are listed on Table 2 and Fig. 1. Only three of these contain an obvious open reading frame (FL1, FL4 and FL5) of which two have clear splicing structure (FL1 and FL4) when aligned to the genome. These three candidates have been recently annotated by the human genome project.

Candidate FL1 (AL136902) corresponds to a hypothetical protein with a 540aa open reading frame. Similarity searches against a protein domain database revealed the presence of a 120aa RGS domain (Regulator of G Protein Signaling - IPR000342). FL4 corresponds to a protein of unknown function and candidate FL5 corresponds to the oculomedin gene that is expressed in trabecular cells of the retina when subjected to cyclic mechanical stretching [15]. The remaining full-length cDNA sequences do not have a clear ORF or splicing structure, and with exception of FL7 also do not contain a typical polyadenylation signal, suggesting that they might represent cDNA cloning artifacts. The low number of ESTs associated with each FL also suggests a very restricted expression pattern, and supports the hypothesis that they could be cloning artifacts (Table 2). This was ruled out, since all these FL cDNA sequences with the exception of FL4, were detected by RT-PCR in the prostate cDNA extensively tested for the presence of DNA contamination. Thus, they represent potential HPC1 candidate disease genes.

Partial transcript sequences (ESTs) mapped to 1q25.

Identity with a transcript sequence is an indicator that a genomic sequence represents a gene. However, transcript sequence databases, particularly ESTs databases, contain a significant fraction of artifactual and contaminant sequences derived from intronic or intergenic DNA that restricts their utility [16, 17]. To overcome this limitation, we opted to work with clusters composed of ESTs that align non-contiguously to the genome, suggesting the presence of a splicing structure. By requiring the occurrence of splicing, the level of contamination in the EST databases is drastically reduced although at the expense of eliminating many genuine 3' ESTs. Indeed, of 266 clusters composed only of ESTs that aligned with HPC1 locus, as few

as 32 showed a splicing structure, which were consequently selected for further analysis.

The alignments between the candidate ESTs and genomic clones were manually inspected to i) exclude the possibility that the selected sequence aligns to another genomic region due to the presence of unknown repeats, ii) identify ESTs that correspond to full-length genes that were submitted to GenBank after the construction of our database iii) confirm the presence of a splicing structure through the identification of conserved splicing sites.

Of the 32 EST clusters initially selected in our analysis, four were found to correspond to repetitive sequences not detected by our masking procedure and which aligned to several distinct genomic regions, four were found to correspond to full-length genes that had been submitted to GenBank after the construction of our database and 14 did not present conserved splicing sites. The remaining 10 ESTs clusters are listed in Table 3 and Fig. 1. Gene prediction computer programs did not predict any of these transcribed regions. As a consequence none are annotated as putative genes on the UCSC map (Golden Path).

IMAGE clones corresponding to the selected ESTs were purchased and the cDNA inserts were completely sequenced. The sequences corresponding to ESTs 3 and 4 revealed a high similarity at the amino acid level with a microtubule associated protein (MAP1A/1B acc.no. NM_022818). The sequences, however, contained several stop codons and insertions of repetitive elements. We conclude that these ESTs correspond to an unprocessed pseudogene mapped to 1q25. This observation was confirmed in the annotation of the corresponding BAC sequence (AL078644) provided by the Human Genome Sequencing Consortium.

Since our tissue of interest is prostate, the expression of the new transcripts was tested by RT-PCR. Of the ten selected ESTs, two (EST1, EST5) were shown to be expressed in prostate and were selected for full-length sequence characterization since they might represent new candidate disease genes for HPC1. Both 5' and 3' RACE were used for transcript extension. The sequences were searched for the presence of ORFs and similarity to any known genes or protein domains. In addition, the expression profiles of the additional transcripts were further explored by RT-PCR using a cDNA panel derived from 20 different normal tissues.

EST1 corresponds to a 2,060 bp transcript (AF510428) with a high similarity (94%) at the nucleotide level with a hypothetical protein from *Macaca fascicularis* (AB071118). The gene is organized into 15 exons distributed over 67,182 bp of genomic sequence. Analysis of the sequence revealed a single putative open reading frame coding for a protein of 500 amino acids (Fig. 2). Analysis of the predicted protein sequence using InterPro Scan (http://www.ebi.ac.uk/interpro/scan.html) revealed the presence of a highly conserved RGS domain (Regulator of G protein signaling) of 120aa. The transcript corresponding to EST1 was named RGSL1 after consulting the HUGO Genome Nomenclature Committee and is expressed at high levels in testis and at lower levels in normal placenta, trachea, bone marrow, lung and prostate where transcripts could only be detected by nested- RT-PCR reactions (Fig. 3A). EST5 corresponds to a 612 bp transcript (AF510427) with no detectable ORF. The gene is organized into 4 exons distributed over an extension of 70,087bp and is expressed in testis, placenta, trachea, bone marrow and prostate (Fig. 3B).

RGS genes mapped to 1q25.

We have, as a result of the work described here, identified and characterized two distinct RGS like proteins (EST1 and FL1) that correspond to RGSL1 and RGSL2 respectively at 1q25 interval and that were named according to HUGO Genome Nomenclature Committee. Both are expressed in the prostate and thus can be considered as candidate genes for HPC1. RGS proteins are characterized by the presence of a 120 amino acid residue domain and function as GTPase-activating proteins (GAP) that negatively regulate heterotrimeric G-proteins responsible for the rapid turnoff of G protein-coupled receptor signaling pathways [18,19]. They are involved in modulating a variety of cellular functions, including proliferation, differentiation, response to neurotransmitters, membrane trafficking and embryonic development [18,19] and are therefore potential tumor suppressor genes.

Recently, a new member of the family, RGS14, was characterized as a p53 target gene induced in response to genotoxic stress [20]. Overexpression of RGS14 inhibits G protein coupled growth-factor-receptor-mediated activation of the MAP kinase-signaling pathway [20]. To begin to explore the possible involvement of the two identified putative RGS proteins that are encoded within HPC1, we have performed LightCycler-based Real-Time RT-PCR to measure the levels of their respective transcripts in normal prostate tissue and two prostate cancer cell lines. The expression levels were determined as a ratio between either RGSL1 and RGSL2 and the reference housekeeping genes GAPDH and β-actin to correct for variations in the amounts of starting RNA. Although expression of both transcripts was readily detected in normal prostate, it could not be detected in the prostate cancer cell lines PC3 and DU145 (Fig. 4 and table 4) suggesting that these genes might have a role in prostate cancer. We are currently verifying the expression level of both RGS genes in paired normal and tumor tissue from prostate cancer patients.

Transcript guided gene finding

A crucial step in the analysis of a disease gene candidate region is to generate a saturated transcript map. Recently, with the availability of the human genome sequence [1, 2], gene identification has been significantly accelerated by means of computationally mediated approaches. Computer programs for gene prediction, however, remain at best a partial solution due to their high false negative and false positive rates for gene identification [4-6]. A better approach appears to be the careful computational comparison of transcript sequences to the genome followed by experimental validation [7]. Using a similar approach, we recently identified 19 novel expressed genes on human chromosome 21 that had not been predicted by computer programs and that had remained entirely unannotated [9]. In contrast, Reymond *et al.* have shown that the majority of transcripts on chromosome 21 defined solely by computer prediction do not correspond to genuine genes, confirming the necessity of experimental validation as we have undertaken here [21].

Although, it is almost certain that other genes within 1q25 (not represented in the transcript and/or genomic sequences) remain to be identified we have been able to contribute evidence for the presence of 17 additional genes within this locus, thus increasing the total number of gene count in this region by approximately 60% (Fig. 1). Eight of the additional identified transcripts constitute new candidates for Hereditary Prostate Cancer based on their expression pattern.

This approach is straightforward in comparison with more laborious wet lab techniques such as cDNA selection and exon trapping and allows the identification of transcripts expressed at very low levels and represented by only a few ESTs. As opposed to computational gene prediction, transcript-guided gene finding does not require a finished genomic sequence and is not biased against genes with small ORFs

or non-coding RNAs (ncRNAs). We conclude that transcript-guided gene identification is a powerful and alternative approach for candidate-gene identification by positional cloning and is likely to prove useful for identifying gene candidates for many genetic diseases.

Materials and Methods

Identification of genomic clones from 1q25.

The physical position of the genetic markers delimiting the HPC1 locus was determined using the NCBI MapViewer (Built #2) tool (www.ncbi.nlm.nih.gov/genome/guide/human/). The genomic clones AL139344 and AL391065 were identified as containing the sequences for D1S2818 and D1S1642 respectively. All genomic clones covering the interval between the markers were identified and a subset with minimal overlap selected manually to span the intervening region. Preference was given to finished clones over unfinished clones. All selected clones were checked for the presence of genetic markers related to the 1q25 region using the NCBI e-PCR tool. The accession number of the selected genomic clones was used as the entry into our Transcriptome Database.

Transcriptome database and selection of candidates.

The genomic sequence corresponding to the selected clones was extracted from EMBL database release 66. Megablast (from the NCBI blast package) was used to identify pairwise similarities between all transcribed sequences and the genomic data. Transcribed sequence data was extracted from several sources: (i) the human EST section of EMBL release 66; (ii) human mRNA documented in the human section of EMBL release 66; (iii) ORESTES sequences from the LICR/FAPESP

Human Cancer Genome project [22, 23]; (iv) human mRNAs documented in the NCBI curated RefSeq database (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html).

The transcript sequences were filtered for contaminants, and repetitive elements were masked out using the PFP software package (Paracel, Pasadena, CA). For each pair of matching transcribed and genomic sequences, local alignments were generated using sim4 [24], with parameters W=15 R=0 A=4 P=1. The output of sim4 was filtered to eliminate all alignments that did not contain at least one matching region with at least 95% identity over 30nt. The alignments 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 clones. Membership in a cluster was determined by the coordinates of the putative exons on the genome sequence. If coordinates of at least one exon were commom to two transcripts, then these were considered to be part of the same cluster.

By querying the Transcriptome Database we were able to distinguish clusters containing at least one full-length cDNA sequence (FLs) and clusters composed only of expressed sequence tags (ESTs). A representative sequence from each cluster was selected for manual inspection in order to exclude the possibility that the selected sequence has a better match to another genomic region and to confirm the presence of splicing sites in the case of EST sequences.

Expression pattern analysis

The expression of the selected transcripts was tested by RT-PCR. Total RNA derived from normal prostate and 19 other normal human tissues (testis, lung, small

intestine, heart, uterus, bone marrow, placenta, colon, fetal brain, liver, fetal liver, thymus, salivary gland, spinal cord, kidney, spleen, skeletal muscle, trachea and adrenal gland) was purchased from Clontech laboratories (Palo Alto, CA) and used for cDNA synthesis. The quality of total RNA was tested by PCR using MLH1 primers located at intronic sequences flanking exon 12 (*Forward* – 5' TGG TGT CTC TAG TTC TGG 3' and *Reverse* – 5' CAT TGT TGT AGT AGC TCT GC 3'), as an indicator of possible genomic DNA contamination. If necessary, the total RNA was treated with DNAse.

Reverse transcription was carried out using the Superscript First Strand Synthesis Kit according to manufacture's recommendations (Life Technologies). A nested-PCR approach was adopted since the expression level of the newly identified transcripts was expected to be low. RT-PCR and nested-PCR conditions, primer sequences and the expected size of PCR fragments are available upon request from A.A.C. (anamaria @compbio.Ludwig.org.br). All PCR amplifications were carried out using 2μl of cDNA reaction as a template to the final volume of 50μl and 1X buffer, 1.5mM $MgCl_2$, 0.2mM dNTP, 0.2uM of each specific primer and 0.025 U/μl of Taq DNA polymerase (Life Technologies). The following cycling protocol was used: initial denaturation of 94°C for 4 min; 94°C for 30 s; 55°C for 45 s; 72°C for 1 min for 35 cycles; along with a final extension at 72°C for 7 min. All PCR products were cloned and sequenced to verify amplification specificity.

Sequencing of IMAGE clones.

cDNA clones corresponding to selected ESTs were obtained from the Resource Center of the German Human Genome Project (RZPD) (AF508902,

AF508903, AF508904, AF508905, AF508906, AF508907, AF508908, AF508909, AF508910, AF508911). DNA sequencing reactions were performed using the "DYEnamic™ ET terminator cycle sequencing" kit (Amersham Life Science) and an ABI377 sequencer (Applied Biosystems-Perkin Elmer).

Rapid amplification of cDNA ends (RACE).

5'-RACE was performed using the Marathon-Ready RACE system (Clontech) according to the manufacturer's instructions. A Marathon-Ready cDNA library derived from testis was used as a template for RACE reactions. Adaptor primer AP1 was used in the primary RACE reaction with each 5'gene specific primer. The primary reaction was carried out using 2 ng of template cDNA with 200nM of each external forward and reverse primer, 200nM dNTPs, 1X cDNA PCR reaction buffer and 1X Advantage cDNA Polymerase Mix (Clontech) in a final reaction volume of 50µl. All PCRs were carried out in a Thermocycler model 9700 (Perkin-Elmer) using the following cycling conditions: initial denaturation at 94°C for 2 min; five cycles of 30s at 94°C and 2 min at 68°C; 25 cycles of 30s at 94°C, 30s at 60°C and 2 min at 72°C; followed by a final extension at 72°C for 10 min. PCR. RACE reaction products were used as templates for secondary RACE reactions using the AP2 adaptor primer and nested gene specific primers. RACE products were extracted from agarose gels using QIAquick Gel Extraction kit (Qiagen) and were cloned using the TOPO TA cloning kit (Invitrogen) according to the manufacturer's recommendations. The clones were sequenced as described above.

Cell Culture and RNA Extraction

The prostate cancer cell lines PC3 and DU145 were obtained from American Type Culture Collection and were grown under specified conditions. Total RNA was extracted using the acid phenol/guanidium method and the quality of the RNA samples was determined by electrophoresis through agarose gels and staining with ethidium bromide. Relative intensities of the 18S and 28S RNA bands were evaluated under UV light.

Real Time PCR

Quantitative PCR was performed with the LightCycler™ (Roche Diagnostics, Mannheim, Germany) apparatus in a total reaction volume of 10 µl per capillary. The 10-µl reaction mixture contained 0.75 U of Platinum Taq DNA Polymerase (Gibco BRL, Life Technologies), 1µl of the supplied 10x buffer, 5 mM $MgCl_2$, 0.2 mM dNTPs (Promega), 0.4 µM each primer, 5% DMSO and 0.1 µl of SYBR® Green I (Sigma) working dilution (1:100) and 1ul of cDNA prepared as described previously. Experiments were repeated twice and performed in duplicates for each data point.

The amplification conditions for LightCycler consisted of an initial 2 min denaturation step at 95°C, followed by 45 cycles of denaturation at 94°C for 15s, annealing at 62°C for 20 s and extension at 72°C for 30 s. Detection of the fluorescent product was carried out at the last step of each cycle at a temperature 3-4°C below the product melting temperature (Tm). Primers were designed in different exons and sequences are available upon request from A.A.C. (anamaria @compbio.Ludwig.org.br). To confirm amplification specificity, PCR products were subjected to a melting curve analysis at the end of each PCR reaction by cooling the

samples from 65°C and then increasing the temperature to 95°C at 0.1 °C/s, with continuous acquisition of the fluorescence.

Quantitative values were obtained at the point during cycling when amplification of the PCR product was first detected, rather than the amount of PCR product accumulated after a fixed number of cycles. The parameter CP (crossing point) is defined as the fractional cycle number at which the fluorescence generated passes a fixed threshold above baseline. The quantification data were analyzed with the LightCycler analysis software as described previously [25]. The target message in unknown samples was quantified and then divided by the endogenous reference (β-actin and GAPDH) to obtain a normalized target value, which compensates for variability in RNA amount and integrity [26]. Testis cDNA was used as positive control for RGSL1 and RGSL2 expression and a negative amplification control was also used (a capillary without any cDNA).

TABLE 1: GenBank accession number of genomic clones mapped to 1q25 region.

AC019075	AL121996	AL118512
AC015683	AC027182	AL135797
AL139344	AC027690	AL133553
AL138776	AL356273	AL162722
AL353778	AL096819	AL096803
AL158215	AL136086	AL049198
AL355999	AL109956	AL033533
AL157943	AL109865	AL022241
AL158011	AL136133	AL022147
AL354657	AL078644	AL049797
AL445228	AL117347	AL096802
AL078645	AL133383	AL138928
AL358152	AC023275	AC025702
AC074116	AL135796	AL391065
AL391824	AL133515	

TABLE 2: Sequence features and prostate expression profile of novel full-length cDNA sequences mapped to the 1q25 region.

FL	Accession no.	Unigene cluster ID	ORF	Splicing	PolyA signal	Anotation	Prostate expression
1	AL136902	Hs.307080 (Size 1)	YES	YES	YES	RGS domain	YES
2	AK023174	Hs.333467 (Size 2)	NO	NO	NO	NO	YES
3	AF075081	None	NO	NO	NO	NO	YES
4	AK001442	Hs.140402 (Size 3)	YES	YES	YES	Hypothetical Protein	NO
5	AF142063	Hs.283759 (Size 1)	YES	NO	NO	OCLM oculomedin	YES
6	AK022377	None	NO	NO	NO	NO	YES
7	AK023363	Hs.300827 (Size 4)	NO	NO	YES	NO	YES

TABLE 3: Sequence features and prostate expression profile of novel ESTs mapped to the 1q25 region

EST	Accession no.	Unigene cluster ID	PolyA signal	RZPD identifier	Prostate expression
1	AA757244	Hs.121200 (Size 1)	YES	IMAGp998A223330Q2	YES
2	BE566336	None	NO	IMAGp958G08365Q2	NO
3	AI187857	None	NO	IMAGp998D064417Q2	Pseudogene
4	AI018001	Hs.131220 (Size 5)	YES	IMAGp998N074130Q2	Pseudogene
5	AW205088	Hs.116070 (Size 19)	YES	IMAGp998C217408Q2	YES
6	AI201365	Hs.344374 (Size 2)	NO	IMAGp998O054461Q2	NO
7	BE906931	None	NO	IMAGp998O089705Q2	NO
8	AI243690	None	NO	IMAGp998F114536Q2	NO
9	AA460184	None	NO	IMAGp998F071962Q2	NO
10	AI025599	None	NO	IMAGp998G244172Q2	NO

TABLE 4: Real Time PCR data

Gene	Samples	Medium CP	Anneling/aquisition
GAPDH	Normal	19,10	62°C / 85°C
	DU-145	14,99	
	PC3	14,51	
	Testis	26,60	
	No	-	
β-actin	Normal	14,88	62°C / 82°C
	DU-145	12,10	
	PC3	11,94	
	Testis	22,94	
	No	-	
RGSL2	Normal	29,91	62°C / 82°C
	DU-145	-	
	PC3	-	
	Testis	35,41	
	No	-	
RGSL1	Normal	38.23	65°C / 80°C
	DU-145	-	
	PC3	-	
	Testis	39.84	
	No	-	

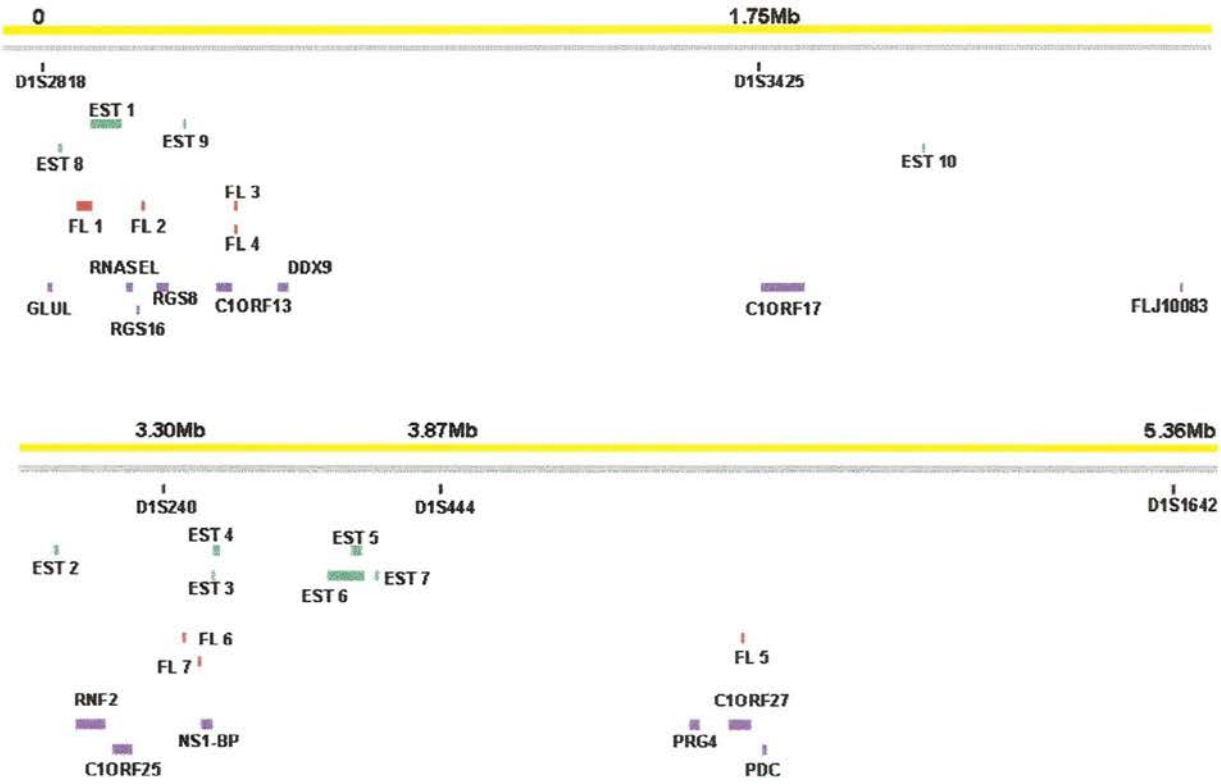


FIG. 1. Diagram of the 1q25 region. Genetic markers D1S2818, D1S3425, D1S240, D1S444, D1S1642 (black) and transcripts already mapped to 1q25 (purple) are given as reference points. ESTs and FL candidates identified in this work are represented in green and red respectively. The scale is given in Megabases (Mb).

agatgctcagtccctggtatgatgagtttctagatgaagaggactactggtttctcctttt
tacgtaactaaaatgtcctttgaggagctttgctacaagaacccaaagatggccatacag
M S F E E L C Y K N P K M A I Q
aagatcagtgatgactacaaaatatactgtgagaaagcacctaaaatagattttaaaatg
K I S D D Y K I Y C E K A P K I D F K M
gaaattatcaaggaaacaaagacagtttcacgctcaaataggaaaatgtccttgctcaaa
E I I K E T K T V S R S N R K M S L L K
agaactcttgtaaggaagccatcaatgagacccagaaacttgacagaagtcctcttaaat
R T L V R K P S M R P R **N L T E V L L N**
acacagcacttggagttcttcagggagttcctcaaggaacggaaggctaaaatcccattg
T Q H L E F F R E F L K E R K A K I P L
caatttctcacagctgtacagaagatcagtatagagaccaatgaaaagatttgcaagtct
Q F L T A V Q K I S I E T N E K I C K S
ctcatagaaaatgtaatcaagacttttcttccaaggccaactctctcctgaagaaatgctg
L I E N V I K T F F Q G Q L S P E E M L
cagtgtgatgcccctattatcaaagaaatcgcttccatgcgtcatgtcaccacaagcaca
Q C D A P I I K E I A S M R H V T T S T
ctgttaacgctccagggacatgttatgaaatctatagaagaaaagtggtttaaagattat
L L T L Q G H V M K S I E E K W F K D Y
caagacctgttcccacctcaccatcaggaggtggaagtgcaaagtgaagtacaaatttcg
Q D L F P P H H Q E V E V Q S E V Q I S
tctaggaagccctcaaagatagtgtcaacttacctacaggaatcccagaagaaaggctgg
S R K P S K I V S T Y L Q E S Q K K G W
atgagaatgatcagctttatcaggagtttttgcaagtaccgcagatttatgttgaatcct
M R M I S F I R S F C K Y R R F M L N P
agtaagcgccaggaatttgaagactatcttcaccaggaaatgcaaaatagcaaggaaaat
S K R Q E F E D Y L H Q E M Q N S K E N
ttcaccacagcacacaacacatcgggacgttcagctccaccctccacaaatgtccggagt
F T T A H N T S G R S A P P S T N V R S
gcagaccaagagaatggagaaataacccttgtaaagcgtcgtatatttggccacaggatt
A D Q E N G E I T L V K R R I F G H R I
atcactgtcaactttgcgatcaatgatctatatttctttttctgaaatggagaaatttaat
I T V N F A I N D L Y F F S E M E K F N
gatctggtcagttcagcccacatgctgcaggtcaaccgggcatataatgagaatgatgtg
D L V S S A H M L Q V N R A Y N E N D V
atcctaatgcggtccaaaatgaacattatccaaaaactcttcctgaattctgacatccct
I L M R S K M N I I Q K L F L N S D I P
ccaaagctgagggtgaatgtccctgagttccagaaggatgccatccttgctgccatcaca
P K L R V N V P E F Q K D A I L A A I T
gagggctacctagatcggagcgtcttccatggggctatcatgtctgtcttccccgttgtt
E G Y L D R S V F H G A I M S V F P V V
atgtacttctggaaaaggttttgtttctggaaggcaacccgctcttacttacagtatagg
M Y F W K R F C F W K A T R S Y L Q Y R
gggaagaagttcaaggacagaaaaagccctcctaaatctacggacaagtatcctttctcg
G K K F K D R K S P P K S T D K Y P F S
agtggaggagacaatgccatcttaaggttcaccttgctcagaggtattgagtggttgcag
S G G D N A I L R F T L L R G I E W L Q
cctcaacgggaagcaataagttcagttcaaaattcttcatcaagcaaacttactcagcca
P Q R E A I S S V Q N S S S S K L T Q P
agactcgtggtatctgccatgcagctgcatcccgtccagggacaaaaattatcctacatc
R L V V S A M Q L H P V Q G Q K L S Y I
aaaaaagagaagtaatcaagcgagacccccagcagagataaatcatctcttagaggcctc
K K E K -
CtaacactgacagaaacttttctggtcaaaaatgaaaggtcctgGtaccatcttccccag
AaacgatcaagaagggcaatgggttgacagcccagatatggcagGaccagactgcaatgg
GaatctcatacagcccattcagactaaagggtgatggaaagaagCagaggaaagcagaaa
CagcaagagtgactgagacctgctgtcatctatatagaagccttTctgatacatgtcaga
GagtgacatgaagtctcccctctttatacacacgtatgaaaataCatctatccctcattt
TgagagcctccaactgatccaagtgtcatccggaggaggaggtgGcccgtatgtgcagga
Cctcagggtgccagccctcacccatacacctccagggctgcaggtcattagccagaataa
aggatgtccacattcacaa

FIG.2. Full-length sequence of RGSL1.

Full-Length sequence of the transcript corresponding to EST1 and the predicted ORF coding for a putative protein of 500 amino acids. The RGS (Regulator of G Protein Signaling) domain is highlighted in bold.

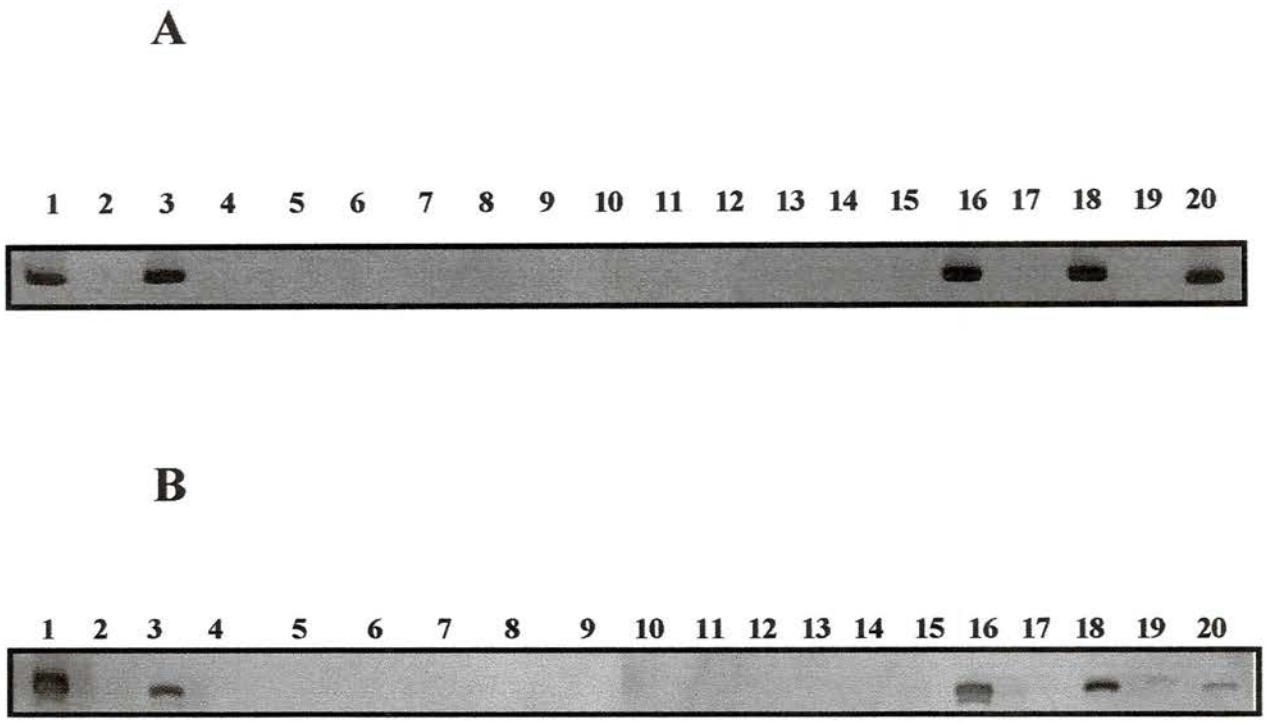
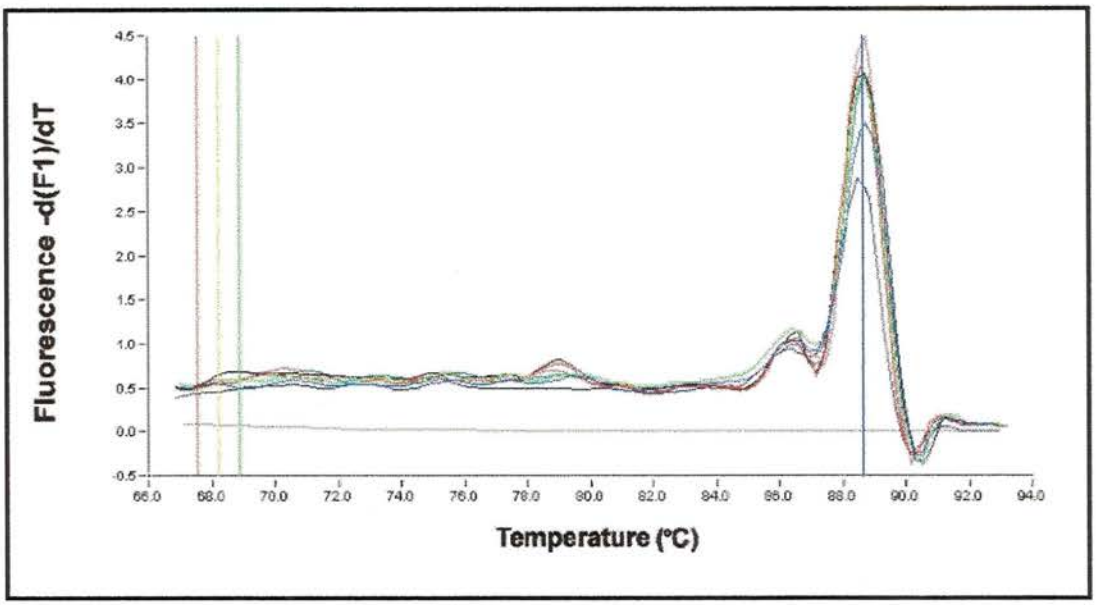
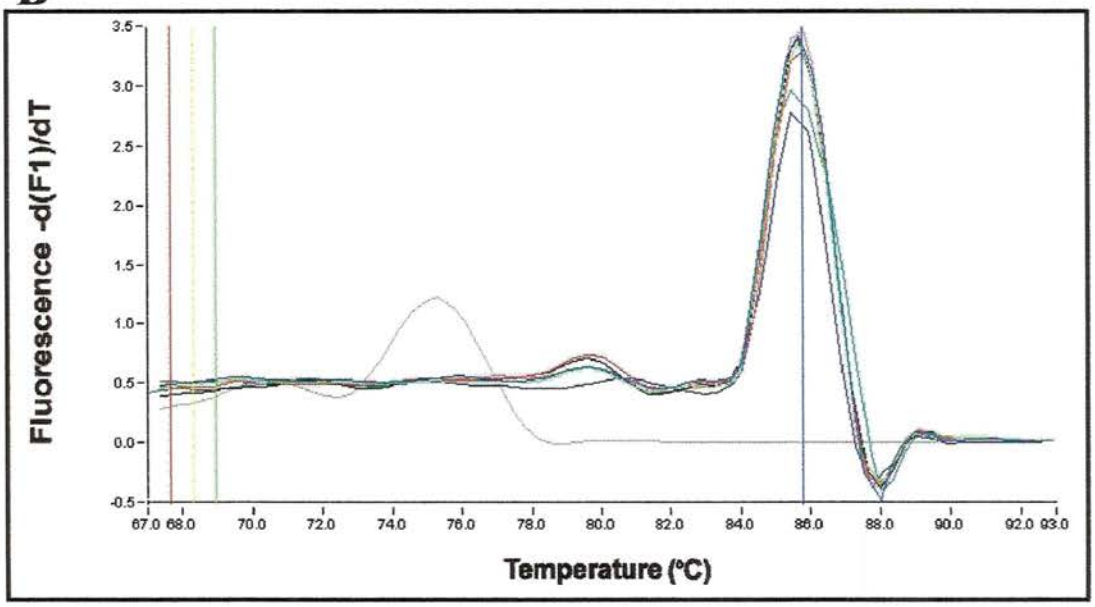


FIG. 3. Tissue expression pattern of the two putative RGS family members mapped to 1q25. cDNA samples from 20 different human tissues were used in nested RT-PCR reaction with primers specific for EST1 **(A)** and EST5 **(B)** candidates mapped to the 1q25 region. Lanes correspond to 1. Trachea ; 2. Fetal brain ; 3. Bone marrow ; 4. Adrenal gland. ; 5. Skeletal Muscle ; 6. Spleen ; 7. Fetal liver ; 8. Liver ; 9. Salivary gland. ; 10. Thymus ; 11. Kidney ; 12. Spinal cord ; 13. Uterus ; 14. Heart ; 15. Colon ; 16. Placenta ;17. Small intestine; 18. Testis ; 19. Lung ; 20. Prostate

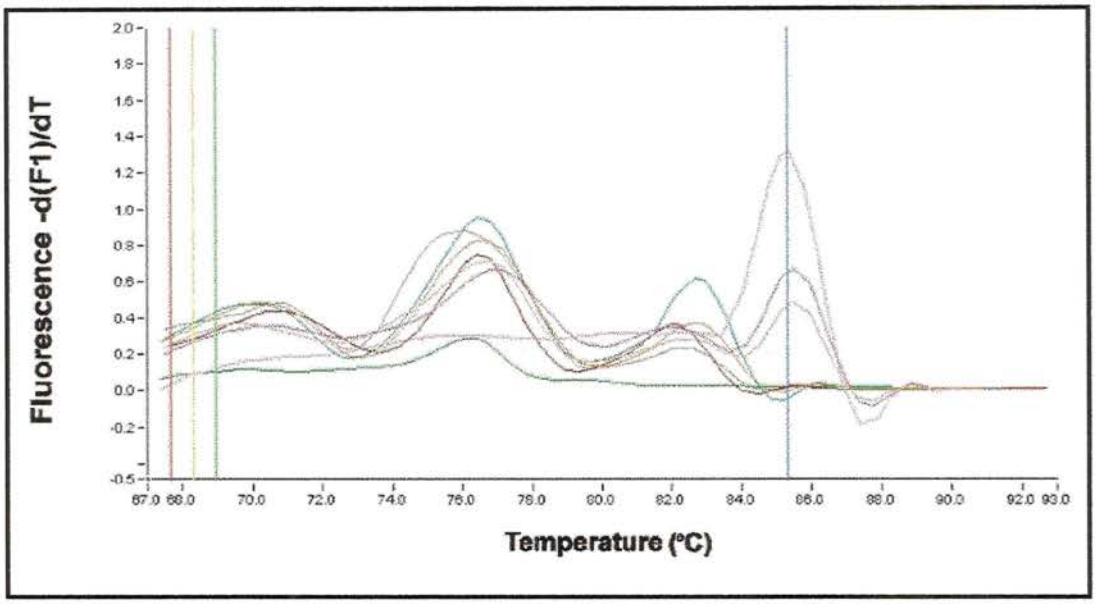
A


Fluorescence -d(F1)/dT
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
66.0
68.0
70.0
72.0
74.0
76.0
78.0
80.0
82.0
84.0
86.0
88.0
90.0
92.0
94.0
Temperature (°C)

B


Fluorescence -d(F1)/dT
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
67.0
68.0
70.0
72.0
74.0
76.0
78.0
80.0
82.0
84.0
86.0
88.0
90.0
92.0
93.0
Temperature (°C)

C



D

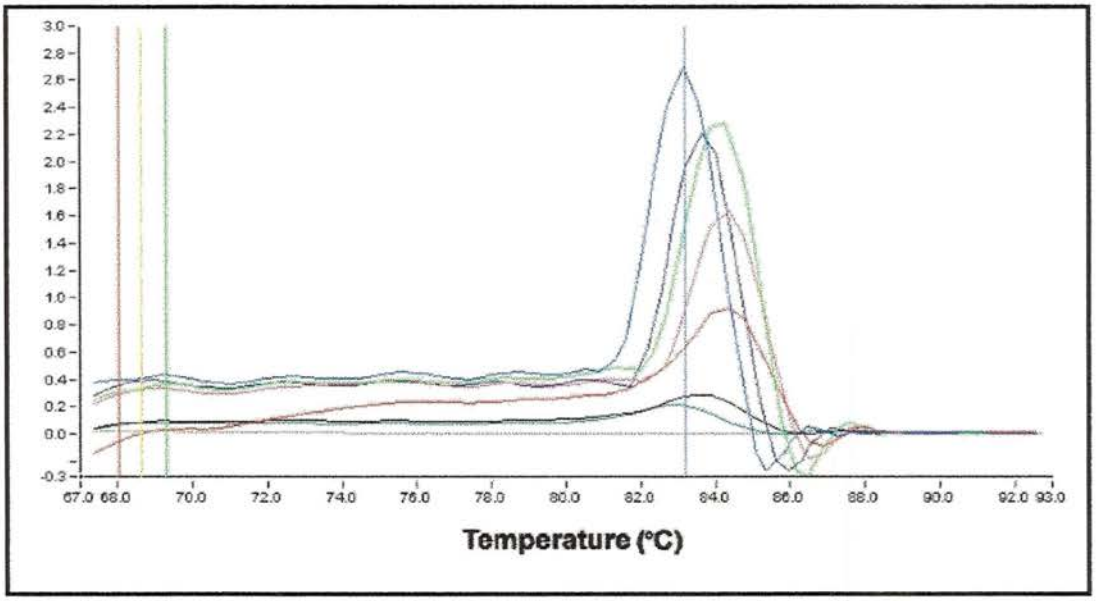


FIG. 4. LightCycler-based Real-Time RT-PCR. **(A)** A typical derivative melting curve profile for the GAPDH product, nonspecific products melted between 68°C and 85°C. **(B)** A typical derivative melting curve profile the β-actin product, nonspecific products melted between 68°C and 82°C. **(C)** A typical derivative

melting curve profile for the RGSL2 product, nonspecific products melted between 68°C and 82°C. **(D)** A typical derivative melting curve profile for the RGSL 1 product, nonspecific products melted between 68°C and 80°C.

Acknowledgments

The authors thank Ricardo Pereira de Moura, Valéria Paixão, Elisângela Monteiro, Jane Kaiano, Fabiana Bettoni and Rafael Bessa Parmigiani for outstanding technical support and Fabricio Falconi for critically reading this manuscript. This work was supported by Ludwig Institute for Cancer Research and Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) through the CEPID program. APMS is supported by a fellowship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

1. Lander, E. S., *et al.* J. Initial sequencing and analysis of the human genome. *Nature*, *409*: 860-921, 2001.

2. Venter, J. C., *et al.* The sequence of the human genome. *Science*, 291: 1304-1351, 2001.

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

4. Guigo, R., Agarwal, P., Abril, J. F., Burset, M., and Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res*, 10: 1631-1642, 2000.

5. Murakami, K. and Takagi, T. Gene recognition by combination of several gene-finding programs. *Bioinformatics.*, 14: 665-675, 1998.

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

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

8. de Souza, S. J., *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc.Natl.Acad.Sci.U.S.A*, 97: 12690-12693, 2000.

9. Reymond, A. *et al* Nineteen additional unpredicted transcripts from the human chromosome 21. *Genomics, in press*.

10. Smith, J. R., *et al.* Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science*, 274: 1371-1374, 1996.

11. Sood, R., *et al.* Cloning and characterization of 13 novel transcripts and the human RGS8 gene from the 1q25 region encompassing the hereditary prostate cancer (HPC1) locus. *Genomics*, 73: 211-222, 2001.

12. Carpten, J. D., *et al.* A 6-Mb high-resolution physical and transcription map encompassing the hereditary prostate cancer 1 (HPC1) region. *Genomics*, 64: 1-14, 2000.

13. Carpten, J., *et al.* Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. *Nat Genet*, 30: 181-184, 2002.

14. Rokman, A., *et al.* Germline Alterations of the RNASEL Gene, a Candidate HPC1 Gene at 1q25, in Patients and Families with Prostate Cancer. *Am.J.Hum Genet*, 70: 1299-1304, 2002.

15. Sato, Y., Matsuo, T., and Ohtsuki, H. A novel gene (oculomedin) induced by mechanical stretching in human trabecular cells of the eye. *Biochem.Biophys.Res Commun.*, 259: 349-351, 1999.

16. Wolfsberg, T. G. and Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res*, 25: 1626-1632, 1997.

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

18. Koelle, M. R. A new family of G-protein regulators - the RGS proteins. *Curr.Opin.Cell Biol.*, 9: 143-147, 1997.

19. De Vries, L. and Gist, F. M. RGS proteins: more than just GAPs for heterotrimeric G proteins. *Trends Cell Biol.*, 9: 138-144, 1999.

20. Buckbinder, L., *et al.* The p53 tumor suppressor targets a novel regulator of G protein signaling. *Proc.Natl.Acad.Sci.U.S.A*, 94: 7868-7872, 1997.

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

22. Camargo, A. A., *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: 12103-12108, 2001.

23. Dias, N. E., *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc.Natl.Acad.Sci.U.S.A*, 97: 3491-3496, 2000.

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

25. Morrison, T. B., Weis, J. J., and Wittwer, C. T. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques*, 24: 954-8, 960, 962, 1998.

26. Xie, D., Nakachi, K., Wang, H., Elashoff, R., and Koeffler, H. P. Elevated levels of connective tissue growth factor, WISP-1, and CYR61 in primary breast cancers associated with more advanced features. *Cancer Res*, 61: 8917-8923, 2001.