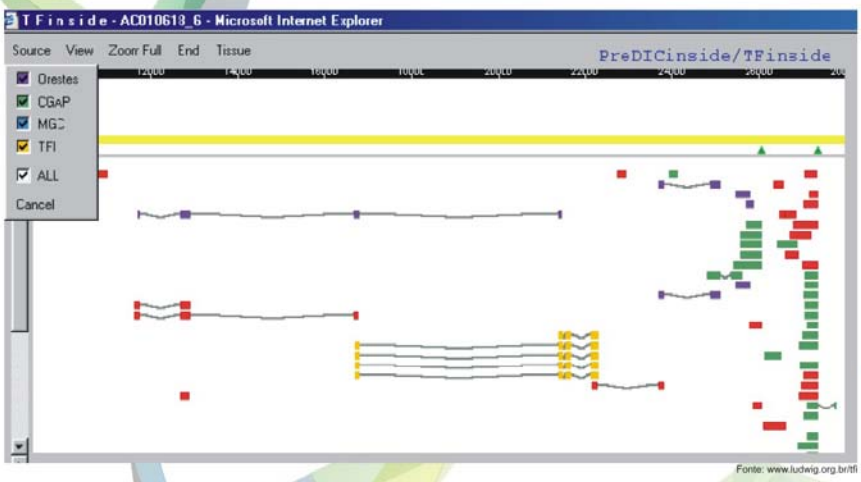
"Transcript Finishing Initiative"
Contribuição do Laboratório IL2

Adriana Camargo Ferrasi



Orientadora: Profa. Dra. Maria Inês de Moura Campos Pardini

Designer - F.A.Ferrasi

Dissertação apresentada ao Instituto de Biociências da Universidade Estadual Paulista - UNESP, Campus de Rio Claro para obtenção do título de Mestre em Ciências Biológicas - Área de Concentração em Biologia Celular e Molecular.

Rio Claro - São Paulo - Brasil
2003

“Transcript Finishing Initiative” Contribuição do Laboratório IL2

Adriana Camargo Ferrasi

Orientadora: Profa. Dra. Maria Inês de Moura Campos Pardini

Dissertação apresentada ao Instituto de Biociências da Universidade Estadual Paulista - UNESP, Campus de Rio Claro para obtenção do título de Mestre em Ciências Biológicas - Área de Concentração em Biologia Celular e Molecular.

Rio Claro - São Paulo - Brasil
2003

A parte experimental deste trabalho foi realizada no Laboratório de Biologia Molecular do Hemocentro de Botucatu - Faculdade de Medicina – UNESP, denominado IL2 pela rede virtual de laboratórios constituída para o projeto "Transcript Finishing Initiative".

Resumo

O principal objetivo na análise de um genoma é a identificação gênica. Várias ferramentas computacionais estão disponíveis para este propósito e são baseadas em similaridade (*BLAST* e *BLAT*) ou em predição de genes (*Genscan* e *Fgenes*). Entretanto, estes programas estão se mostrando ineficientes para detectar e caracterizar todos os genes presentes no genoma humano. A importância das informações de *cDNAs* tem sido reconhecida desde o início do Projeto Genoma Humano, entretanto, o seqüenciamento em larga escala de cDNAs completos ainda requer técnicas avançadas tais como a produção de bibliotecas de cDNAs enriquecidas por transcritos grandes e raros. O seqüenciamento parcial de etiquetas de seqüências expressas (*ESTs*) foi desenvolvido como uma técnica alternativa para gerar, em larga escala, vários tipos de *cDNAs*. Atualmente, a maioria das informações de *cDNAs* no *GenBank* são representadas por *ESTs* convencionais 3´ e 5´ e *ORESTES* (provenientes das porções centrais dos transcritos). Baseados nos bancos de dados gerados pelo alinhamento de todas essas seqüências com as seqüências genômicas humanas disponíveis foi proposta a estratégia "*transcript finishing*" para a caracterização e validação de novos genes humanos, como parte do consórcio entre FAPESP e Instituto Ludwig de Pesquisa sobre o Câncer. O projeto "*Transcript Finishing Initiative*" está sendo realizado por uma rede de 31 diferentes grupos de pesquisa do Estado de São Paulo. Foram selecionados pela coordenação do projeto, 602 transcritos e destes 300 (50%) foram validados. Destes transcritos, 20 foram atribuídos ao laboratório validador IL2, e destes, 11 (55%) foram validados. Utilizando ferramentas de bioinformática, o laboratório IL2 realizou uma anotação preliminar dos consensos de seus transcritos validados (disponibilizados pela coordenação do projeto) e das seqüências de *cDNA* parcial geradas durante o processo de validação, que não alinharam ao local esperado no genoma. Para os 10 transcritos com consensos montados, 6 combinaram com genes conhecidos e descritos no *RefSeq Genes* e 4 alinharam a *ESTs* ou *mRNAs*. Destes, 2 transcritos apresentaram *splicing* alternativo já representados nos bancos de dados de *ESTs* e/ou *mRNAS* humanas, vindo a confirmá-los.

Das 14 cDNAs parciais pesquisadas, 13 alinharam a genes representados no *RefSeq Genes* e 1 apresentou similaridade a seqüências de *ESTs* e *mRNAs*. Destes, observou-se três *splicing* alternativos, já representados nos bancos de dados de *ESTs* e/ou *mRNAs* humanas, confirmando-os.

.

Abstrat

A fundamental task in analyzing genome is gene identification. This is relatively straightforward for compact genome but much more challenging for complex genomes. Some computational tools are available for this purpose, but they are bases on similarity (BLAST) or prediction analysis (Genscan and Fgenes). However, these programs are inefficient to detect and characterize all genes present in the genome. The importance of cDNA information has been recognized since the beginning of the Human Genome Project, however cost-effective and hightroughput sequencing of full-length cDNA still requires technical advances such as the production of cDNA libraries enriched for large and rare transcripts. Partial sequencing of expressed sequences (EST) has been developed as an alternative approach for the generation, in large-scale, of several kinds of cDNAs. Currently, the vast majority of cDNA data in the GenBank is represented both by conventional 5´and 3´expressed sequence tags (ESTs) and by ORESTES (open reading frame ESTs), which is derived from central portions of the transcripts. Based on a database generated through alignment of all of these sequences to the available human genomic sequences, have been proposed the transcript finishing strategy for characterization and validation of new human genes, as part of the FAPESP-LICR Transcript Finishing Initiative. The strategy utilizes the ORESTES scaffold EST sequence to build primers for reverse transcription (RT) - PCR reactions in order to bridge gaps, thereby confirming the membership of ESTs to a common transcript and providing information on the intervening sequence (validation strategy). The FAPESP-LICR Transcript Finishing Initiative is being pursed by a network of 31 different research groups from the State of São Paulo (The Transcript Finishing Consortium) coordinated by 2 different laboratories, located at the São Paulo Ludwig Institute and Chemistry Institute of the University of São Paulo. To date, 210 (35%) of 597 TFI fragments have been valited. Of these sequences 20 have been attributed to IL2 validation laboratory, here was realized this present study, and of these 11 (55%) was valited. In this study, the sequences validated of IL2 group and the not specific sequences were analyzed with Bioinformatic tools in a preliminary annotation.

1. Introdução

Decifrar o genoma humano não é um desafio menor que seqüencia-lo.

A identificação de todos os genes humanos será um passo fundamental para entendermos os aspectos biológicos de nossa espécie.

Embora duas versões da seqüência preliminar do genoma humano tenham sido concluídas (LANDER *et al.*, 2001; VENTER *et al.*, 2001), ainda são incertas as estimativas sobre qual o número de genes que se expressam, ou seja, que codificam proteínas, contidos na seqüência genômica (DAS *et al.*, 2001).

Os artigos publicados pelo Consórcio de Seqüenciamento do Genoma Humano Internacional (LANDER *et al.*, 2001) e pela Celera Genomics (VENTER *et al.*, 2001) quando da divulgação da seqüência preliminar do genoma humano, estimam o número de genes entre 30.000-40.000 e 26.000-38.000, respectivamente, utilizando algoritmos computacionais para realizar a predição baseada na seqüência genômica completa.

Nosso desconhecimento é ainda maior no que diz respeito à identificação destes genes, o que eles codificam ou como funcionam (SIMPSON *et al.*, 2001).

A razão é que, embora a seqüência genômica seja essencial para uma descrição precisa e classificação dos genes humanos, ainda não é suficiente. Os genes humanos são estruturas extremamente complexas e ainda não somos capazes de predizer sua presença com alguma certeza pela inspeção da seqüência de DNA genômico. Cada gene que codifica proteínas no genoma humano é tipicamente dividido em múltiplos e relativamente curtos exons com extensas seqüências de introns (não codificantes) entre eles. Uma das maiores dificuldades é identificar o exato local de início e final de cada exon, além da ocorrência de *splicing* alternativos e diferenças nos padrões de expressão. Desta forma, torna-se cada vez mais clara a necessidade de evidências experimentais por seqüenciamento de transcritos completos

(cDNAs), ESTs (*Expressed Sequence Tags*) e até mesmo seqüenciamento de genomas de outros organismos evolutivamente próximos à espécie humana, para futuras comparações de similaridades e suporte para predições computacionais (SIMPSON *et al.*, 2001).

Várias ferramentas computacionais disponíveis são baseadas em análises de similaridades entre seqüências (por exemplo, BLAST) e alguns programas foram desenvolvidos para reconhecer padrões na estrutura gênica (p.ex. GenScan e Fgenes) tais como regiões codificadoras e seqüências sinais (elementos promotores, códons de início e parada, sítios de *splicing* e sinais de poliadenilação). Entretanto, estes programas têm se mostrado ineficientes para detectar e caracterizar todos os genes presentes no genoma humano (CAMARGO *et al.*, 2001a; SAHA *et al*., 2002).

Diversos grupos de seqüenciamento têm produzido seqüências de cDNA completas e ESTs. Hoje existem aproximadamente 14.000 seqüências completas de cDNA (não redundantes) e mais de 4,8 milhões de ESTs disponíveis nos bancos de dados públicos. A discrepância entre esses números deve-se a maior dificuldade técnica em se gerar seqüências completas de cDNA (CAMARGO, 2001b; KAN *et al*., 2001; NCBI, 2002a).

Com o aumento inigualável dos dados nestes últimos anos, faz-se necessário o desenvolvimento de ferramentas para o gerenciamento e análise desses dados. Novos programas *softwares* e *hardwares* serão necessários para o processamento eficiente, montagem e anotação, bem como predições de genes e classificações funcionais e estruturais (STERKY & LUNDEBERG, 2000).

Dentro deste contexto a FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) e o Instituto Ludwig de Pesquisa sobre o Câncer uniram-se em um consórcio e lançaram o Projeto Genoma Humano do Câncer (HCGP) com o objetivo final de gerar 1 milhão de ESTs utilizando uma nova metodologia denominada ORESTES (Open Reading Frame ESTs), capaz de gerar seqüências derivadas das porções centrais dos transcritos (DIAS-NETO *et al*., 1997; CAMARGO, 2001b; FAPESP, 2002a).

Próximo ao final do referido projeto, o grupo brasileiro depositou no GeneBank, banco de dados genômicos de acesso público, aproximadamente 700.000 ORESTES (CAMARGO, 2001c) como contribuição para o objetivo de definir os genes humanos e seus produtos, de maneira que estas seqüências, somadas às ESTs provenientes de métodos convencionais, poderiam ser utilizadas como base para gerar seqüências completas de transcritos.

Como uma extensão do HCGP, em Novembro do ano 2000, a FAPESP selecionou 29 laboratórios para compor uma nova rede virtual e desenvolver o novo projeto Genoma denominado *Transcript Finishing Initiative* (TFI), ou Projeto Transcriptoma, com início das pesquisas em Janeiro de 2001 (FAPESP, 2002b).

O projeto TFI tem como objetivo determinar a estrutura e gerar seqüências de novos genes humanos utilizando como base a seqüência do genoma humano e todas as seqüências expressas (ESTs e ORESTES) disponíveis nos bancos de dados públicos (CAMARGO *et al.*, 2001c).

Ao contrário dos projetos de produção de seqüências de cDNA completas o TFI irá gerar seqüências de fragmentos parciais de cDNA evitando assim todo o trabalho relacionado com a preparo de bibliotecas de cDNA de alta qualidade. Esses fragmentos parciais (TFs) serão gerados através de RT-PCR de forma a validar e complementar a estrutura de genes humanos parcialmente representados por ESTs (FAPESP, 2002b).

O uso da bioinfomática para identificação do gene em genomas complexos ainda não é rotineiro e essa análise tem confirmado que a identificação do gene dependeria principalmente do alinhamento de seqüências genômicas com seqüências de cDNA (FAPESP, 2002b).

O Projeto TFI encontra-se dividido em duas frentes interdependentes: o desenvolvimento de ferramentas de bioinformática (realizado por 5 grupos de bioinformática) e a validação experimental (29 laboratórios de pesquisa dentro do estado de São Paulo) e, ainda, dois laboratórios centrais de Coordenação: o Instituto Ludwig e o Instituto de

Química da Universidade de São Paulo. A lista de participantes do projeto pode ser encontrada no endereço www.compbionet.org.br/transcript/.

O produto final do projeto será um catálogo de seqüências virtuais validadas experimentalmente de novos transcritos humanos. Esse catálogo também fornecerá informações adicionais sobre a similaridade com outras seqüências disponíveis em bancos de dados públicos, formas variantes de *splicing,* presença de domínios protéicos, localização cromossômica e dados preliminares sobre padrões de expressão (CAMARGO, 2001b). O Laboratório de Biologia Molecular do Hemocentro de Botucatu (IL2) foi um dos 29 laboratórios selecionados para participar dessa rede virtual. Foi atribuído a esse laboratório até o presente momento, 20 transcritos a serem validados, bem como a padronização das técnicas a serem utilizadas para sua validação. Além dos objetivos iniciais, comuns a todos os laboratórios participantes do Projeto TFI, os membros do laboratório IL2, por iniciativa própria, aplicaram outras abordagens durante o período de desenvolvimento do projeto: a análise *in silico* dos consensos montados com as seqüências validadoras geradas pelo laboratório e as ESTs de cada transcrito e, também das seqüências obtidas durante o processo de validação, mas que não validaram o transcrito, ou seja, não alinhoram à região esperada no genoma, fornecida pela coordenação.

2. Considerações Iniciais

O DNA é formado por um longo polímero, não ramificado, composto de somente quatro subunidades: os desoxirribonucleotídeos, que contém as bases adenina (A), citosina (C), guanina (G) e timina (T), ligadas a uma pentose (desoxirribose) e um grupo fosfato; unidos por ligações fosfodiéster covalentes que ligam o carbono 5' de um grupo desoxirribose ao carbono 3' do próximo (LEHNINGER *et al.*, 1995).

Estruturalmente, o DNA é um polímero helicoidal composto por duas cadeias polinucleotídicas antiparalelas, dispostas em dupla hélice, associadas por pontes de hidrogênio, que se formam complementarmente entre A e T e entre C e G (figura 2.1) (ALBERTS *et al*., 1997a).

Em termos moleculares, o gene pode ser definido como um segmento de DNA que carrega a informação genética para produzir um produto funcional, que poderá ser tanto um RNA funcional (RNA ribossômico ou RNA transportador) quanto um polipeptídio, ou ainda tratar-se de um gene regulador da transcrição (LODISH *et al*., 2000; ALBERTS *et al*., 1997b).

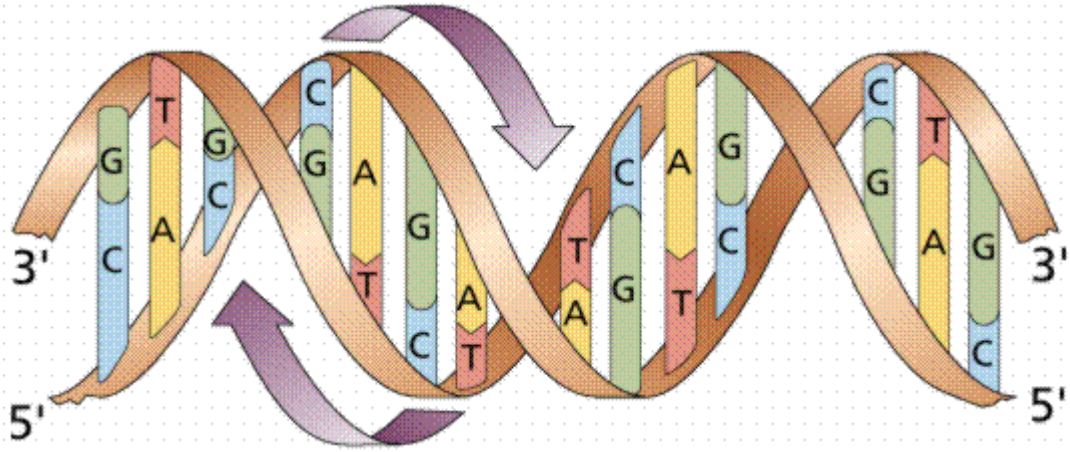


Figura 2.1- O DNA é um polímero helicoidal composto por duas cadeias polinucleotídicas antiparalelas, dispostas em dupla hélice, associadas por pontes de hidrogênio, que se formam complementarmente entre A e T e entre C e G.

Para fins didáticos, os genes são subdivididos em duas regiões básicas, a região codificante e a região não codificante. A **região codificante** contém o código genético que será lido pela maquinaria traducional no citoplasma e a **região não codificante** representa aquelas seqüências de DNA que estão embutidas entre as seqüências codificantes da grande maioria dos genes de eucariotos (mas não dos procariotos) e parecem não ter participação na função gênica ou expressão e são descartadas durante o processamento do RNA primário em RNA mensageiro. Essas seqüências são denominadas ***introns*** e as seqüências codificantes chamadas ***exons***. Também, dentre as seqüências ditas "não codificadoras", devemos incluir aquelas seqüências de DNA necessárias para a expressão da informação genética, mas que não são traduzidas em polipeptídios (KENDREW, 1999).

A região codificante de um gene é flanqueada por regiões reguladoras que controlam o início e o final da transcrição (figura 2.2), Na extremidade 5' do gene encontra-se o promotor, região onde se ligará a RNA polimerase de modo a iniciar a transcrição. O promotor consiste em dois

componentes e a combinação mais comum inclui as seqüências chamadas TATA

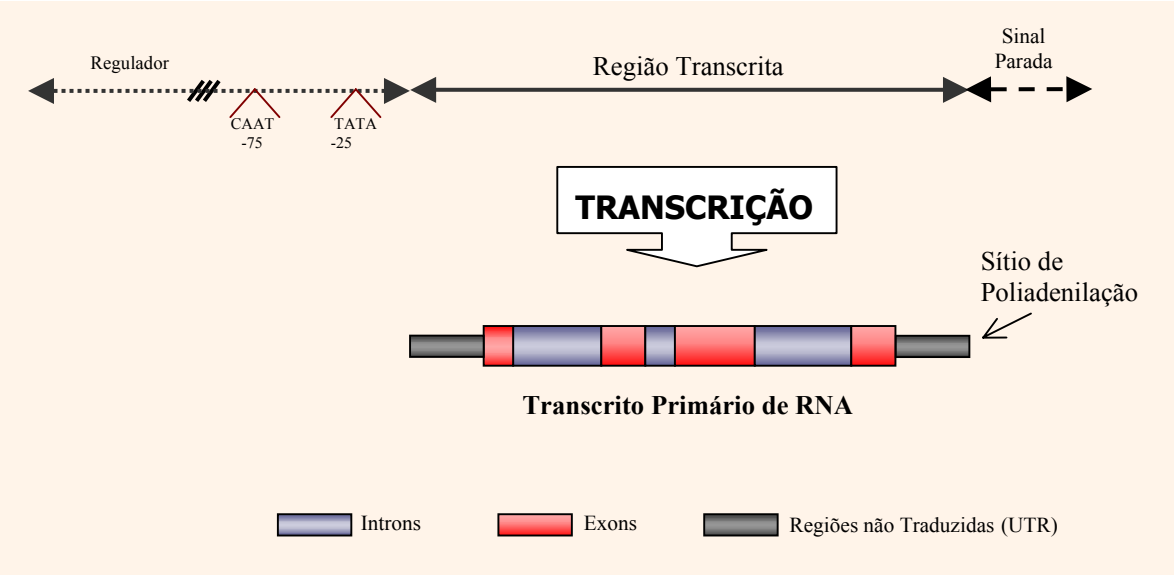


Figura 2.2 - Estrutura gênica. Diagrama esquemático da estrutura de um gene eucariótico que codifica a produção de proteínas. Modificado de Kendrew, 1999.

e CAAT. A caixa TATA localiza-se cerca de 25 nucleotídeos *upstream*, ou seja, à montante do início do codificador, no sentido da extremidade 5' e a caixa CAAT, no mesmo lado, porém um pouco mais longe, a uns 75 nucleotídeos da região codificadora. Ainda na extremidade 5', acima do promotor e muitas vezes, a milhares de nucleotídeos, estão os reguladores (amplificadores e inibidores) da expressão do gene; na extremidade 3' há uma região que sinaliza o local de término da transcrição (DE ROBERTIS, 2001; KENDREW, 1999; FARAH, 1997).

A síntese de proteínas envolve a cópia de regiões específicas de DNA (os genes) em moléculas de RNA, que então deterão toda a informação da seqüência de DNA da qual foi copiada. Esse processo é denominado de transcrição do DNA. Após a transcrição, a molécula imatura de RNA será então processada para que se torne apta a deixar o núcleo em direção ao citoplasma onde será traduzida em polipeptídios pelos ribossomos. O processamento do RNA imaturo dá-se pela remoção das regiões não codificantes (*introns*) e adição de uma guanina metilada na extremidade 5' e de uma "cauda" poliadenilada na extremidade 3' da molécula de RNA (ALBERTS *et al.*, 1997a; BROWN, 1999; STERKY & LUNDEBERG, 2000).

A seqüência de nucleotídeos na molécula de mRNA, que age como intermediário, é lida no citoplasma em série e em grupo de três. Cada trio de nucleotídeos, chamado códon, especifica um aminoácido. Em princípio, cada seqüência de RNA pode ser traduzida em qualquer uma das três formas de leitura, dependendo de onde o processo de decodificação inicia. Em quase todos os casos somente uma destas formas de leitura produzirá uma proteína funcional. Há sinais de pontuação, no início e final do mRNA, indicando onde deve começar a leitura (**códon AUG**, que também codifica para o aminoácido metionina quando no interior do mRNA) e o momento certo de parada da síntese protéica (**códons UAA, UGA E UAG**) (ALBERTS *et al.*, 1997b).

Os genes que codificam proteínas são subdivididos em duas classes, baseando-se no tipo de produto gênico:

a) genes estruturais - que codificam proteínas *housekeeping,* ou seja, aquelas utilizadas nas atividades celulares (enzimas metabólicas, proteínas de transporte, componentes do citoesqueleto, etc.);

b) genes regulatórios - os quais regulam produtos da expressão de outros genes, geralmente a níveis de transcrição.

Uma grande fração de todos os genomas de vertebrados (95 a 97%) não codifica precursores para RNA mensageiro (mRNA) ou algum outro RNA funcional. Em organismos multicelulares, este DNA não codificante contém muitas regiões que são parecidas, mas não idênticas. Nas células de eucariotos superiores, as regiões do DNA que codificam proteínas, ou seja, os genes, "misturam-se" entre essas regiões de DNA não funcional. Adicionando-se a essas regiões, encontramos outras seqüências aparentemente não funcionais no interior dos genes, os *introns* (LODISH *et al*., 2000).

Um dos maiores obstáculos, quando se pretende estudar genomas de eucariotos é o seu tamanho e complexidade, ou seja, a soma de todo esse DNA aparentemente não codificante ao DNA que codifica proteínas ou RNAs funcionais. No decorrer dos anos, pesquisadores buscaram identificar e classificar as diferentes categorias do DNA eucariótico, o qual é sumarizado no quadro 1 e, brevemente relatado abaixo:

Genes de cópia única

São genes representados somente uma vez no genoma haplóide. Acredita-se que, em organismos multicelulares, 25-50% dos genes que codificam proteínas sejam genes de cópia única (LODISH *et al*., 1999).

Genes duplicados ou divergentes

Família de genes funcionais - As famílias multigênicas são formadas por grupos de genes duplicados, que codificam proteínas com seqüências de aminoácidos similares, mas não idênticas. Originaram-se, possivelmente, pela duplicação de um gene ancestral, com os diferentes membros do grupo tendo divergido como conseqüência de mutações durante a evolução. Há dois tipos de família multigênicas: **Famílias multigênicas simples,** nas quais todos os genes presentes são aparentemente iguais (por exemplo, os genes que codificam rRNA 5S, que estão agrupados em cerca de dois mil genes no cromossomo I). **Famílias multigênicas complexas,** compostas de genes similares, porém não idênticos, mas que irão ao final produzir diferentes polipeptídios que se unirão formando uma só proteína (exemplificando, a família de genes dos polipeptídios da globina nos vertebrados) (STRACHAN & READ, 2002; COOPER, 2000).

Pseudogenes - Seqüência de DNA com alta similaridade a um gene funcional, mas que não expressa uma proteína funcional devido, provavelmente, a mutações deletérias. Esses genes degradados contêm uma ou mais mutações inativadoras, tais como uma mutação sem sentido que introduz um códon prematuro de finalização. Uma outra classe de pseudogenes (pseudogenes processados) perderam introns e também a região promotora que ficaria antecedendo a seqüência transcrita em mRNA (KENDREW, 1999; BROWN, 1999).

DNA repetitivo

DNA de seqüência simples - Consiste de seqüências curtas que são repetidas várias vezes, *in tandem,* em cópias idênticas ou relacionadas no genoma. A repetição *in tandem* dessas seqüências curtas cria uma fração com propriedades físicas distintas, que permite a sua separação da maior parte do DNA por centrifugação, por meio de um gradiente de densidade de flutuação. Uma fração desse tipo de DNA é chamada de DNA satélite. Esse tipo de DNA

repetitivo não é transcrito e contribui com a maior parte das regiões heterocromáticas do genoma, sendo encontrado nas vizinhanças dos centrômeros, telômeros e localizações específicas nos cromossomos (STRACHAN & READ, 2002; LEWIN, 1997)

DNA moderadamente repetido (elementos móveis) - Inclui os genes repetidos *in tandem* que codificam genes duplicados e os elementos móveis.

Elementos móveis - São seqüências que estão dispersas no genoma de organismos eucariotos e algumas vezes, também em procariotos. Esses elementos variam de centenas a poucos milhares de pares de base em extensão e o processo pelo qual estas seqüências são copiadas e inseridas em um novo sítio no genoma é chamada transposição. Os elementos móveis de DNA são essencialmente parasitas moleculares, que parecem não ter uma função específica na biologia do organismo hospedeiro. Os elementos móveis são classificados em duas categorias: (1) aqueles que se deslocam diretamente como DNA (transposons). (2) aqueles que se transpõem via transcrição de RNA e então se convertem em DNA dupla fita por transcrição reversa (retrotransposons) (LODISH, 2000).

(1) Transposons - Embora ocorram também em eucariotos, são mais comuns em bactérias. O deslocamento dos transposons altera a organização estrutural do genoma, afetando a expressão gênica. A importância do seu estudo está relacionada com as conseqüências que têm sobre o genoma, pois eles dão origem a mecanismos que podem ter grandes efeitos na evolução.

(2) Retrotransposons - Estão presentes nos mais diversos organismos como leveduras, moscas e mamíferos. Um excelente exemplo de retrotransposon é o elemento Ty1 de leveduras: Nestes, a primeira etapa da transposição é a transcrição completa do elemento transponível, produzindo uma cópia de RNA do elemento que possui mais de 5.000 nucleotídeos. Esse transcrito codifica uma transcriptase reversa que faz uma cópia de DNA fita dupla a partir da molécula de RNA, via um intermediário híbrido de DNA/RNA, então utilizando uma integrase, se insere em sítios aleatórios nos cromossomos. Esse processo, notavelmente similar à infecção por um retrovírus, diferencia-se pela

incapacidade da produção de uma capa protéica e, portanto, o Ty1 não pode deixar a célula hospedeira, ocorrendo o seu deslocamento somente dentro de uma única célula e sua progênie (STRACHAN & READ, LEWIN, 1997; ALBERTS *et al.*,1997a).

DNA repetido disperso - São seqüências repetidas que estão dispersas por todo o genoma. Essas seqüências são classificadas como **SINES** (elementos curtos dispersos, do inglês *short interspersed elements*) ou **LINES** (elementos longos dispersos, do inglês *long interspersed elements*). As principais SINES nos genomas de mamíferos são as seqüências *Alu*, denominadas assim porque normalmente contém um único sítio de restrição da endonuclease *Alu I.* As seqüências *Alu* tem aproximadamente 300 pares de bases e há aproximadamente um milhão destas seqüências dispersas no genoma. Embora sejam transcritas em RNA, não codificam proteínas e sua função é desconhecida. As LINES, assim como as SINES parecem ter se propagado por transposição. A LINE-1 é um tipo de retroelemento não viral, um transposon que pode se replicar e se mover no genoma por um processo que envolve transcrição reversa. Possui aproximadamente 6.000 pb e se repetem perto de 50.000 vezes no genoma (BROWN, 1999; COOPER, 2000).

CLASSIFICAÇÃO DO DNA EUCARIÓTICO

Genes que codificam proteínas
- Genes de cópia única
- Genes duplicados ou divergentes (famílias de genes codificantes e pseudogenes não funcionais)

Genes repetidos *in tandem* que codificam rRNA, 5S rRNA, tRNA, e histonas

DNA repetido
- DNA de seqüência simples
- DNA moderadamente repetido (elementos móveis)
  - Transposons
  - Retrotransposons virais
  - Elementos longos dispersos (LINES e retrotransposons não virais)
  - Elementos curtos dispersos (SINES; nonviral retrotransposons)

DNA espaçador não classificados

Tabela 2.1 - As diferentes classes de seqüências de DNA eucariótico estão resumida neste quadro. Modificado de LODISH *et al.*, 1999

Splicing Alternativo

Em organismos eucariotos os genes são normalmente monocistrônicos, ou seja, codificam um único produto gênico, enquanto que os genes bacterianos (procariotos) são, em sua maioria, policistrônicos: caso em que duas ou mais proteínas podem ser codificadas por um único tipo de mRNA, ou ainda, pela sobreposição de genes, onde uma única região de DNA pode dar origem a diferentes RNAs mensageiros (KENDREW, 1999).

Durante a transcrição, os genes eucariotos são copiados em longas moléculas precursoras de mRNA, que então, são processados por uma série de etapas para a produção de uma molécula de RNA madura. Uma dessas etapas é o *splicing* do RNA, na qual os *introns* são removidos. Apesar da natureza monocistrônica dos genes, sabe-se que em alguns casos, é comum a célula processar de formas diferentes o mesmo transcrito primário e, desta maneira, produzir diferentes cadeias polipeptídicas a partir do mesmo gene. Esse processo é denominado *splicing* alternativo (LEWIN, 1997; ROBERTS & SMITH, 2002; CARTEGNI *et al*., 2002).

O *splicing* alternativo poderá explicar a grande disparidade entre o modesto número de genes no genoma humano e a complexidade de seu proteoma. Acredita-se que pelo menos um terço dos genes humanos sofrem *splicing* alternativo. Estes devem conter múltiplos introns e em muitos casos os exons podem ser unidos em mais de uma forma para gerar múltiplos mRNAs, codificando isoformas distintas de proteínas (Figura 2.3) (ROBERTS & SMITH, 2002; CARTEGNI *et al*., 2002). Parece que o splicing alternativo de precursores de mRNAs surgiu como um novo mecanismo de modulação da função do genoma. É provável que a maioria dos genes produzam múltiplas proteínas, e tenham múltiplas atividades por intermédio deste processo. Para o estudo das funções do genoma é essencial a identificação de todos os mRNAs que ele produz e o desenvolvimento de ferramentas para descobrir sua função e monitorar sua expressão (KAN *et al*., 2001; ROBERTS & SMITH, 2002).

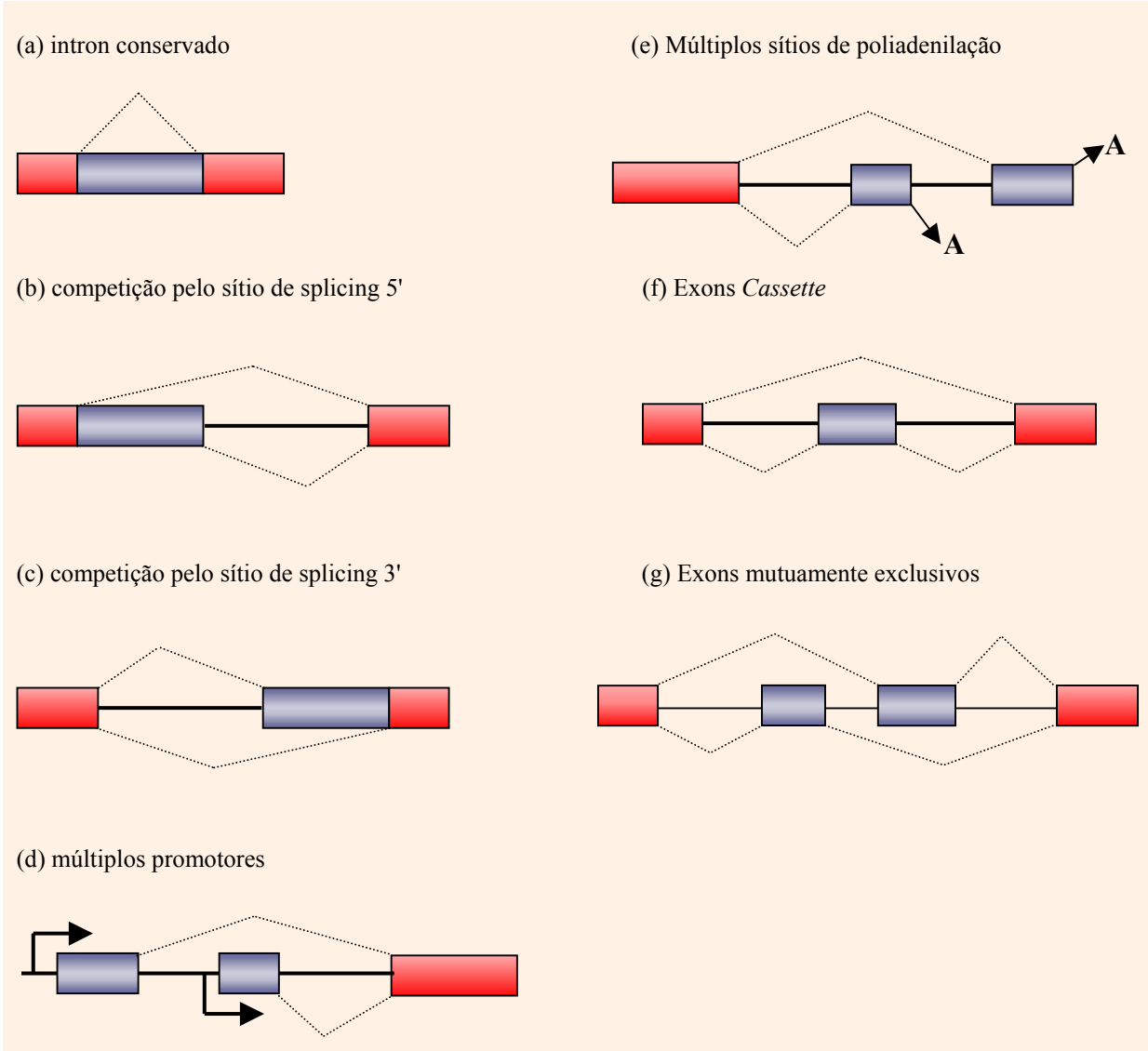


Figura 2.3 - Esta figura mostra alguns tipos de *splicing* alternativos, entre vários. Cada item mostra parte de um gene. Os introns estão representados como linhas contínuas e os exons com caixas. Os padrões de *splicing* estão indicados acima e abaixo por linhas diagonais pontilhadas. Os exons que são unidos constitutivamente após o *splicing* estão em vermelho, enquanto os segmentos que são tratados como exons ou introns, dependendo do padrão de *splicing,* estão representados por caixas cinzas. **(a)** a seqüência pode ser retirada como um intron ou permanecer no mRNA. Exons individuais podem ter sítios alternativos de *splicing* 5' **(b)** ou 3' **(c)**; **(d)** Nas extremidades 3' ou 5' dos genes, pode ocorrer *splicing* alternativo em associação com a seleção de promotores alternativos, demonstrados pelas setas ou; **(e)** diferentes sítios de poliadenilação (indicado pelo "A"); **(f)** Exons *cassette* internos podem ser incluídos ou excisados independentemente de outros exons, enquanto **(g)** *splicing* mutualmente exclusivos envolvem um arranjo de dois ou mais exons alternativos, e somente um destes poderá ser incluído no mRNA maduro. Modificado de GAVIN et al., 2002.

Expressed Sequence Tags (ESTs)

ESTs (do inglês *Expressed Sequence Tags)* ou Etiquetas de Seqüências Expressas são pequenas porções de seqüências de DNA gerados por seqüenciamento de cada extremidade (5' e 3') ou ambas, ou ainda das regiões centrais (ORESTES) de um gene expresso. Como o propósito é o sequenciamento de partes de transcritos, as ESTs são sintetizadas a partir de cDNAs complementar e aos mRNAs. Inicialmente, estas etiquetas tinham cerca de 300 nucleotídeos de comprimento, hoje, porém, seqüências com mais de 700 nucleotídeos são comuns (STRAUSBERG & RIGGINS, 2001).

Open Reading Frame Expressed Sequence Tags (ORESTES) é uma modificação da técnica que gera ESTs convencionais, desenvolvida por um grupo brasileiro (DIAS-NETO *et al*., 1997), que diferem das ESTs por fornecerem seqüências da porção central dos transcritos, onde se encontra a região codante do gene, enquanto que nas ESTs ditas convencionais há uma tendência de que sejam seqüenciadas as extremidades 3' e 5' dos genes. Outra vantagem da técnica ORESTES sobre as ESTs é a normalização do seqüenciamento tanto de genes com níveis de expressão altos a moderados, quanto para transcritos raros ou aqueles que tem um nível baixo de expressão. Acredita-se, essa normalização ocorre devido ao uso de *primers* aleatórios em condições de baixa estringência (DIAS-NETO *et al*., 2000).

CAMARGO *et al.* (2001c) geraram aproximadamente 700.000 ORESTES provenientes de 24 tecidos humanos e usaram um subgrupo de 15.095 mRNAs *full-length* disponíveis nos bancos de dados públicos genômicos para avaliar a eficiência desta estratégia, concluíram que a técnica cobriu acima de 80% dos genes humanos com níveis de expressão altos e moderados e, entre 40% e 50% de genes humanos que se expressavam raramente.

Utilizando a técnica ORESTES, DE SOUZA *et al.* (2000), produziram 250.000 seqüências de genes expressos, a partir de vários tecidos de tumores humanos, que cobriam regiões centrais destes genes (ORESTES).

Estas 250.000 ORESTES foram agrupadas em 81.429 *contigs.* Destes, 1.181 (1,45%) combinavam com seqüências no cromossomo 22, com pelo menos um *contig* de ORESTES para cada um dos 162 (65,6%) genes dos 247 genes já conhecidos, para cada um dos 67 (44,6%) dos 150 genes relacionados (similares a transcritos de outros organismos ou outros genes humanos) e, para 45 dos 148 (30,4%) genes preditos por ESTs, neste cromossomo. Com o uso de critérios estringentes para validação das seqüências, foram identificados 219 outras seqüências transcritas não anotadas anteriormente no cromossomo 22. Destes, 171 foram de fato também definidos por EST ou seqüências de cDNA *full-length* disponíveis no GenBank, mas não utilizadas na anotação inicial do primeiro cromossomo humano seqüenciado. Desta forma, seqüências ORESTES identificaram mais 48 seqüências de transcritos no cromossomo 22 ainda não definidas por outras seqüências.

Anotação Gênica

A anotação de um genoma é o processo onde se toma uma seqüência de DNA bruta produzida por seqüenciamento e empregando-se a análise e interpretação necessária para extrair sua significância biológica, e a colocação dessa seqüência no contexto de nosso entendimento dos processos biológicos. A anotação de um genoma é um processo de múltiplos passos, situando-se em três categorias básicas: anotação em nível de nucleotídeos, anotação em nível de proteínas e anotação em nível de processos (STEIN, 2001).

A anotação em nível de nucleotídeos é realizada após a obtenção de dados do seqüenciamento de um genoma completo ou partes dele. É realizada a identificação das regiões gênicas que pode ser por meio de programas de bioinformática como aqueles que predizem genes buscando sinais-padrão de identificação próprios para detectar algumas regiões conservadas (junção *exon-intron*, códons de início e parada, promotores, entre outros) e que são chamados *ab initio* ou por programas que se baseiam nas

buscas de similaridade pelo alinhamento a seqüências já descritas (STEIN, 2001; BRENT, 2002).

Em genoma procariotos pequenos onde a maior parte do genoma é composta por regiões codificadoras, sua identificação é feita, basicamente, através da identificação de longas janelas abertas de leitura (ORFs, do inglês *open reading frames).* Já em genomas eucariotos, essa tarefa é mais complicada, já que apenas uma pequena porção do genoma é composta por regiões gênicas, codificadoras de proteínas. Além disso, a existência de introns que podem ter dezenas de quilobases e formas alternativas de *splicing* de mRNAs dificulta ainda mais a correta identificação dos genes. Uma forma bastante eficiente de encontrar genes, utilizada com alguma freqüência na pesquisa atual é o alinhamento das seqüências a serem pesquisadas a bancos de dados de seqüências de cDNAs (ESTs). As ESTs representam genes transcritos e então, quando há o alinhamento da seqüência pesquisada e ESTs, mesmo que de outras espécies, é uma boa evidência do fato de que aquela região contém genes (STERKY & LUNDEBERG, 2000; STEIN, 2001).

Entretanto, o pesquisador deve estar consciente de que ESTs apresentam, algumas vezes, baixa qualidade no seqüenciamento, além de que, como é sabido, ESTs são somente seqüências parciais. Então se ganha quando há uma união entre os programas de predição gênica e aqueles que se baseiam em similaridade de alinhamento de seqüências e esta é a tendência nas anotações atuais (STERKY & LUNDEBERG, 2000; RUST *et al*., 2002).

Anotação em nível de proteínas - Esse tipo de anotação busca a montagem de um catálogo das proteínas presentes no organismo, nomeá-las e associá-as a possíveis funções. Uma forma comum de se realizar a anotação de proteínas é procurar similaridade utilizando ferramentas como BLASTp ou PSI-BLAST (http:\\www.ncbi.nlm.nih.gov/BLAST), buscando informação em diferentes bancos de dados de proteínas. As coleções mais utilizadas de seqüências de proteínas são os bancos de dados SWISS-PROT e TrEMBL (ambos no site eletrônico: http:\\www.expasy.ch/sprot). O primeiro possui uma coleção de seqüências de proteínas confirmadas e extensivamente anotadas e

contém referências de outros bancos de dados de seqüência e estrutura, referências bibliográficas, identificação de famílias protéicas e descrição sobre a provável função e papel biológico da proteína. TrEMBL, contém a tradução automática das seqüências codificadoras (cds) submetidas aos bancos de dados. Uma análise complementar seria a busca de domínios funcionais e as bases de dados mais utilizadas neste processo são: PFAM (http://www.sanger.ac.uk/Software/Pfam/), PRINTS (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/), PROSITE (http://www.expasy.org/prosite), ProDom (http://www.toulouse.inra.fr/prodom.html), SMART (http://smart.embl-heidelberg.de) e BLOCKs (http://blocks.fhcrc.org), que apesar de possuírem uma nomenclatura e métodos de procura próprios, foram reunidos em um banco desenvolvido com as melhores características de cada um, o InterPro (RUST *et al*., 2002).

Anotação em nível de processos biológicos - Para esse tipo de anotação faz se necessário mais do que trabalho computacional. São utilizadas técnicas biológicas de alta resolução, análises de expressão e microarrays, ensaio de expressão em proteínas por espectroscopia de massa, entre outras. Um consórcio chamado GeneOntology foi criado para essa finalidade e tem criado um vocabulário padrão para descrever a função de genes eucariotos (STERKY & LUNDEBERG, 2000; RUST *et al*., 2002).

Ferramentas da Bioinformática

São inúmeros os softwares destinados as mais diferentes manipulações de grandes quantidades de dados gerados pelos projetos de sequenciamento, são softwares destinados a catalogar e comparar seqüências de nucleotídeos e aminoácidos, identificar e prever localização de genes, interpretar dados de perfis de expressão, procurar polimorfismos, associar mutações a doenças, entre outros. Também, os bancos de dados são essenciais para o gerenciamento e identificação de padrões delicados

encontrados pelo uso do volume de dados biológicos que crescem exponencialmente. O NCBI (do inglês *The National Center for Biotechnology Information*) nos Estados Unidos, e o EBI (*European Bioinformatics Institute*) na Inglaterra são os dois principais servidores de informações biológicas e responsáveis pelo "tratamento" do embaralhado volume de dados. Eles mantém bancos de dados e softwares analíticos que servem como ferramentas para toda a comunidade científica, que também submete seus dados, diariamente, tornando-os públicos (RASHIDI & BUCHLER, 2000).

Vários programas sofisticados foram criados para manipular a predição gênica em genomas eucarióticos, dentre eles o GENSCAN, Genie, Fgenes, GeneFinder, entre outros. Esses programas consistem tipicamente de um ou mais sensores que tentam deduzir a presença de um gene caracterizando motivos ou propriedades estatísticas do DNA (por exemplo, regiões ricas em C e G, sítios de início de transcrição, TATA-boxes, etc.) (STEIN, 2001).

A maior parte destas ferramentas e bancos de dados está disponível pela Internet a pesquisadores de todas as partes do mundo, fornecendo informações úteis tanto a pesquisadores quanto a clínicos. Alguns bancos de dados requerem subscrição que pode ser gratuita ou sob pagamento. O modo mais fácil para identificar bancos de dados é por *links* a bancos de dados fornecidos pelos principais bancos de dados públicos. Por exemplo, o *National Center for Biotechnology Information* (NCBI) *(http://www.ncbi.nlm.nih.gov)* fornece *links* com o Browser Entrez (que é um banco de dados que recupera informações de seqüências de DNA e proteínas dos principais bancos de dados disponíveis). O *European Bioinformatic Institute* disponibiliza dados de seqüências de nucleotídeos e proteínas de inúmeros organismos, enquanto que o *Ensembl* produz e mantém anotação automática em genomas eucariotos (BAYAT, 2002).

Uma das mais simples e conhecidas ferramentas é a chamada BLAST (*Basic local alignment search tool*), disponível no NCBI. Este programa é capaz de pesquisar em bancos de dados por genes com seqüências de nucleotídeos ou aminoácidos similares e permite a comparação de uma

seqüência de nucleotídeos ou aminoácidos desconhecida a centenas ou milhares de seqüências humanas ou de outros organismos até que uma combinação seja encontrada. Bancos de dados de seqüências conhecidas são deste modo utilizadas para identificar seqüências similares e o resultado da pesquisa é ordenado por prioridade de máxima similaridade. A seqüência com maior *score*, ou seja, a que mais se assemelha à seqüência pesquisada é mostrada em primeiro lugar, seguida das seqüências com menor grau de similaridade. Todos os dados disponíveis para tais seqüências são acessíveis através de *links* com seus respectivos bancos de dados (NCBI, 2002b; WHEELER *et al*, 2002).

Outra ferramenta que vem tornando-se popular entre os pesquisadores do mundo inteiro é a ferramenta BLAT, disponibilizada pelo UCSC *Genome Bioinformatics* da Universidade da Califórnia, Santa Cruz (UCSC, 2002; KENT, 2002).

Este site conta com as seqüências preliminares do genoma humano e do camundongo. Nele é possível realizar buscas por similaridade de seqüências de nucleotídeos e aminoácidos de forma rápida e fácil aos principais bancos de dados disponíveis. Esta ferramenta tem se sobressaído pela sua rapidez, estabilidade e facilidade de manipulação dos dados e também por permitir o alinhamento de várias seqüências de uma única vez. Como resultado de uma busca por seqüências similares a seqüência desconhecida, obtém-se um *browser* que disponibiliza todas as informações encontradas que possam estar relacionadas à seqüência pesquisada (figura 2.4). Este browser mostra montagem de *contigs* e *gaps*, alinhamento com bancos de mRNAs e ESTs (humanos e não humanos), genes preditos, SNPs, transposons e diversas outras informações, das quais algumas são descritas a seguir:

RefSeq Genes - Mostra genes conhecidos que codificam proteínas, buscando tais informações de seqüências de mRNAs compiladas no LocusLink. Deve haver um alinhamento com pelo menos 98% de identidade para que a seqüência seja mostrada. Clicando-se sobre o nome do gene mostrado, são

fornecidos diversos links que levará a informações como a proteína correspondente ao gene, o mRNA, entre outros (MAGLOTT *et al*., 2000)

Human mRNA - Mostra o alinhamento entre mRNAs humanos do GenBank no Genoma para aquela região onde a seqüência pesquisada alinhou. As seqüências alinhadas (freqüentemente exons) são representadas com retângulos pretos conectados por linhas que representam os *gaps* (normalmente os introns retirados durante o *splicing)*. São alinhadas somente aquelas seqüências com pelo menos 95% de identidade.

Nonhuman mRNA - Alinha mRNAs de vertebrados não humanos depositados no GenBank. Utiliza a mesma simbologia do Human mRNA.

Human ESTs - Permite o alinhamento entre ESTs depositadas no GenBank à região da seqüência preliminar do Genoma Humano que apresentou similaridade à seqüência desconhecida pesquisada.

Spliced ESTs - Mostra ESTs depositadas no GenBank que mostras sinais de *splicing*. Isso tem reduzido drasticamente o nível de contaminação, já que requer somente ESTs onde foram identificados sinais de splicing.

STS Markers (human) - Essa opção mostra a localização de possíveis marcadores (STS) para aquela região pesquisada.

Repeat Masker - Esta opção poderá mostra seqüências repetidas naquela região (Estes elementos incluem SINEs, LINEs, micro-satélites, tRNAs, e outras famílias de DNA repetido)

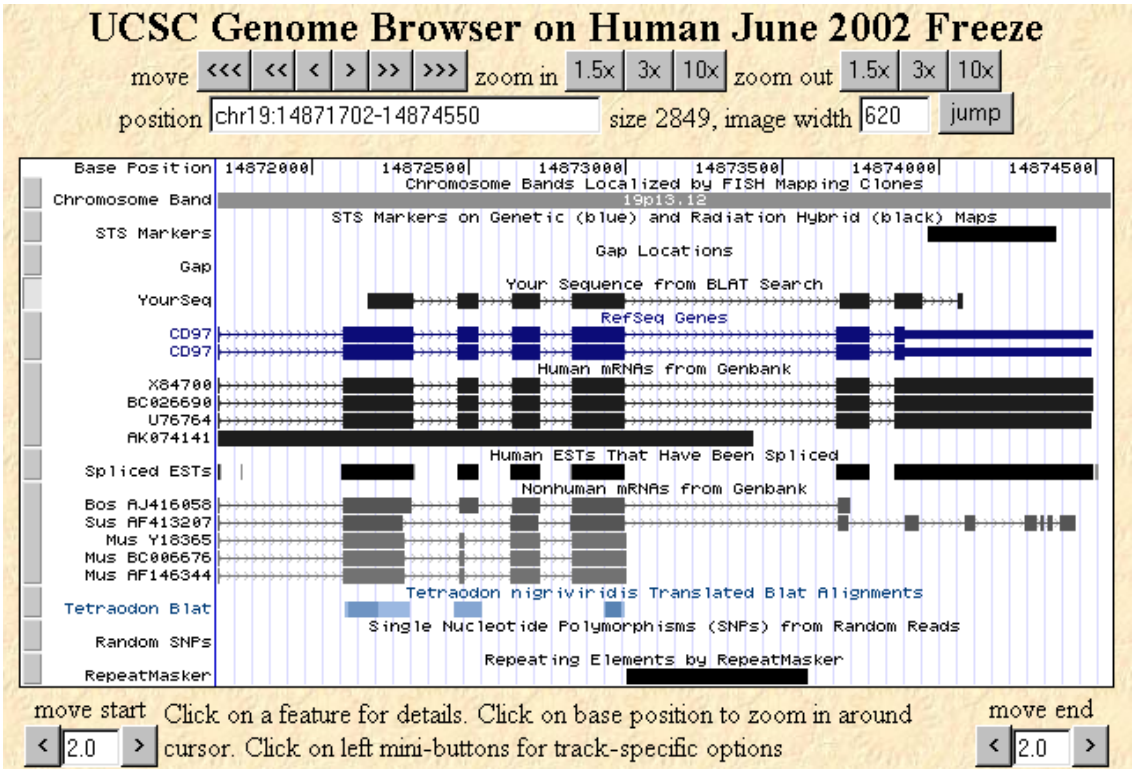


Figura 2.4 – Apresentação gráfica de um alinhamento na ferramenta BLAT. Disponível no Web Site (http://genome.ucsc.edu).

3. Objetivos

Por tratar-se de um estudo realizado no contexto de um projeto temático, os objetivos do referido projeto vão além dos objetivos deste trabalho, que por sua vez, apresenta algumas abordagens não desenvolvidas no projeto temático. Desta forma, faz-se necessária a apresentação dos objetivos separadamente, a fim de caracterizar melhor o estudo que será apresentado nos capítulos seguintes.

Objetivos Gerais do Projeto *Transcript Finishing Initiative*:

- Através de RT-PCR, clonagem e seqüenciamento, validar a estrutura e definir a seqüência de pelo menos 4.000 genes humanos *full-length* em um período de 2 anos;

- Criar um banco de dados para catalogar os genes que serão validados e também fornecer informações como localização cromossomal, padrões de expressão, formas de *splicing* alternativos, entre outros.

- Desenvolver ferramentas computacionais para a seleção dos transcritos a serem validados e avaliação da validação.

Objetivos específicos do Laboratório IL2:

- Utilizando técnicas de RT-PCR, clonagem e sequenciamento, validar os transcritos designados pela coordenação do projeto TFI ao grupo validador IL2 (*);

- Realizar, através de ferramentas de bioinformática disponíveis na *internet*, uma anotação preliminar dos transcritos validados pelo grupo IL2 (*);

- Utilizando ferramentas de bioinformática, analisar seqüências de cDNA parciais geradas por reações de PCR com o objetivo de validar transcritos (porém que não alinharam ao clone genômico informado pela coordenação) buscando possíveis seqüências inéditas nos bancos de dados genômicos, de interesse para futura validação laboratorial (**).

(*) Objetivos comuns aos 31 laboratórios de validação da Rede do Projeto *Transcript Finishing Initiative*.

(**) Objetivos específicos de iniciativa do Laboratório IL2.

4. Materiais e Métodos

4.1 Materiais

Durante o período de realização deste trabalho, foram enviados ao laboratório IL2, 20 pares de *primers* (tabela 4.1) para a validação dos TFs selecionados pela coordenação do projeto.

Para a amplificação por PCR (Polimerase Chain Reaction), foram utilizados num primeiro momento, cDNAs sintetizados a partir de RNA total extraído de tumores e de tecidos normais (Tabela 4.2) utilizando a metodologia de Colchão de Césio e mais tarde RNA poly A+ (enriquecido de mRNA) utilizando o kit comercial RNAeasy® da ProMega (Tabela 4.3). Este material foi sintetizado e distribuído aos laboratórios validadores pela coordenação do Instituto Ludwig (Dra. Anamaria Aranha Camargo) e do Instituto de Química (Dra. Mari Cleide Sogayar), São Paulo-Brasil, em gelo seco.

PRIMERS - TRANSCRIPT FINISHING INITIATIVE					
TF00035	F	gcagaaaggcaggacataaaac	TF00232	F	TCATTCAGAACAGAGTGTGGC
	R	ggcttcttcactctgttccg		R	TATTCCATACAGCCATGTCGC
TF00040	F	AGTACAAGGCCAGCCTGC	TF00041	F	GGCAGAATCGTGTCCTCG
	R	AGTGCCTCTGCTTCTCCG		R	CAGGGCATCCAGTGCAG
TF00380	F	TTCACGAGGTTCTCATATGCC	TF00404	F	TTCATGTGGTACAGGACCAGG
	R	TGTCTCCTTAGAAGGAAGGCTG		R	TTTCGTGATTTCACACTTGCC
TF01048	F	ggaaattctgcatccccc	TF01049	F	TCCCCTGAGTGAAATATGGC
	R	tctccagcagctccacg		R	TGGAAAACCACGTGACCTC
TF00072	F	ATGGTGGTCCTCCTGCTG	TF00074	F	TCCCCTGAGTGAAATATGGC
	R	CATCCAGCCAAGCCACTC		R	TGGAAAACCACGTGACCTC
TF00156	F	TGATGGCTTCACACACCTG	TF00157	F	ACGGATTCTTGCCAGTGC
	R	CAAAGCTCCAGGCCAATG		R	CAAGCGATTTCTGCAGCC
TF00193	F	TATTACCCATGAGGCCTGGAG	TF00194	F	CTCCTGGTCTGCATTCTTCAG
	R	GGTGAGGAAGCTGAGAACCAG		R	CGAATGATGCGAACACACAC
TF00308	F	TGGTGGTTGAGGTCTTACAGG	TF00309	F	AGTTTCTTGCTCCTCTCCCTG
	R	TGGAAACTCGGTTCATACCAG		R	AGCTCATGCGACAAGGATTC
TF00324	F	CTTGTTCTCTGGCTTGGAGTG	TF00325	F	TGCCTTTATCTTCCTTCTCCC
	R	GAACCTGCAAACAAATACCCC		R	CAATGTTCAAACATGAGCCTG
TF00408	F	GAGAGGGGTGTAGATTGGACAG	TF00214	F	CCCAATTGTCCTGGTCAGAG
	R	GAACACTTGATCCAGTTCTGAC		R	TGCAAGTGGAAGGAAACTCTTC

Tabela 4.1 - *Primers* enviados pela coordenação do projeto TFI ao laboratório IL2, para validação dos transcritos. A letra F representa o *primer Foward* e a letra R o *Reverse*..

BANCO DE cDNAs – RNA Total	
Linhagem Celular	**Tecidos**
H358	pulmão
MCF-7	mama
T98G	cérebro
K562	células B
SW480	cólon
HELA	útero
A172	glioblastoma, cérebro
XP (SV 40)	xeroderma pigmentoso, pele
ZR-75-1	carcinoma ductal, ascitos, glândula mamária, epitélio
Hs 578T	mama, carcinoma ductal
DU145	próstata
-	rim
-	carcinoma de tireóide, medula
-	carcinoma hepatocelular, fígado
IM-9	linfoblastos (células B)
FADu	cabeça e pescoço
Skmel-25	melanoma

Tabela 4.2 Lista de cDNAs (RNA-total) distribuídos pela Coordenação do Projeto TFI aos grupos validadores.

BANCO DE cDNAs – RNA poly A+	
Linhagem Celular	**Tecidos**
IM-9	Célula B
FADu	Cabeça/Pescoço
T98G	Cérebro, glioblastoma invasivo
A172	Cérebro, glioblastoma não invasivo
T98G	Cérebro; Glioblastoma Multiforme
SW480	Cólon
HEPG2	Fígado
U937	Linfoma histiocítico
U937	Mama
HS 795.PL	Placenta
DU145	Próstata
DU145	Próstata, carcinoma
H1155	Pulmão
-	Rim
-	Tecido Nervoso
HS1	Testículo
HELA	Útero

Tabela 4.3 Lista de cDNAs (RNA-poly A+) distribuídos pela Coordenação do Projeto TFI aos grupos validadores.

4.2 Métodos

4.2.1 Análises *in vitro*

Para todos os TFs, foram realizadas várias tentativas de amplificação como:

- variações na concentração de $MgCl_2$ e dos *primers*;
- variações nas temperaturas de anelamento e extensão final;
- substituição de enzimas de amplificação (de *Taq recombinant* para *eLongase, Platinum*).
- cDNAs provenientes de diferentes tecidos e diferentes métodos de extração (RNA total e enriquecido com mRNA (poly A+);
- variações no número de ciclos da reação de PCR;
- PCR *Nested.*

As reações padrão utilizadas são descritas a seguir:

Tubos de 200 μl foram utilizados para as reações de PCR e cada um deles foi identificado pelo número do TF (*Transcript Finishing Initiative*) seguido do número do tubo (Exemplo: TF00157/03, onde 03 significa a realização da terceira reação-teste). Cada TF a ser validado foi visualizado através da interface gráfica, de forma que se estabeleceu previamente qual o tamanho máximo da banda a ser amplificada. Essa informação foi especialmente importante, por exemplo, nos casos em que os dois *clusters* que compunham o TF estavam separados por um intervalo grande no DNA genômico. Tal constatação reforçou a importância da utilização de uma Taq polimerase de

maior fidelidade (eLONGASE Enzyme Mix, Platinum Taq Pol. High Fidelity ou AccuPrime™ Taq DNA Polumerase System – Invitrogen Life Technologies) e/ou de um maior tempo de extensão para viabilizar a amplificação (em geral, 1 minuto por kb). A escolha dos cDNAs utilizados nas reações de PCR (com relação ao tecido a partir do qual foram sintetizados) foi baseada, na maioria das amplificações, em tecidos que estavam representados por pelo menos uma EST de um dos *clusters* que compunham o respectivo TF. As reações de PCR padrão ou de PCR utilizando o protocolo da Platinum Taq, eLONGASE ou AccuPrime foram preparadas em gelo, em fluxo laminar, segundo PROTOCOLO 1:

PROTOCOLO 1 REAÇÕES DE PCR UTILIZADAS NAS AMPLIFICAÇÕES DOS TFs (*Transcript Finishing Initiative*).

1.1 REAÇÃO PCR TAQ RECOMBINATE

Enzima: Taq DNA Polymerase Recombinante (*Invitrogen-Life Technologies*)

Reagentes	**Quantidades**	Concentração Final
H_2O de ampola estéril	18,79 μL	-
10 x Taq buffer	2,5 μL	1 X
dNTP (10 mM)	0,5 μL	0,2 mM
$MgCl_2$ (50 mM)	0,75 μL	1,5 mM
Primer Forward (10 μM)	0,63 μL	0,25 μM
Primer Reverse (10 μM)	0,63 μL	0,25 μM
Taq Recombinate (5U/μL)	0,2 μL	2U
cDNA apropriado	1,0 μL	-
Total	25,0 μL/reação	-

Após a distribuição dos reagentes do PCR, os tubos foram colocados em termociclador com o seguinte programa de amplificação:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	94ºC	3 min.
Desnaturação	94ºC	45 seg.
Anelamento	55ºC	30 seg.
Extensão	72ºC	2 min.
Ext. Final	72	2 min.
4ºC	∞	

35 ciclos

1.2 REAÇÃO PCR PLATINUM Taq

Enzima: Platinum Taq DNA Polymerase High Fidelity (*Invitrogen-Life Technologies*)

Reagentes	**Quantidades**	Concentração Final
H_2O de ampola estéril	39,10 μL	-
10 x Taq buffer	5,0 μL	1 X
dNTP (10 mM)	1,0 μL	0,2 mM
$MgSO_4$ (50 mM)	1,5 μL	1,5 mM
Primer Forward (10 μM)	1,0 μL	0,25 μM
Primer Reverse (10 μM)	1,0 μL	0,25 μM
PLATINUM Taq (5U/μL)	0,4 μL	2U
cDNA apropriado	1,0 μL	-
Total	50,0 μL/reação	-

Após a distribuição dos reagentes do PCR, os tubos foram colocados em termociclador com o seguinte programa de amplificação:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	94ºC	30 seg.
Desnaturação	94ºC	30 seg.
Anelamento	55ºC	30 seg.
Extensão	68ºC	2 min.
4ºC	∞	

35 ciclos

1.3 REAÇÃO PCR eLONGASE

Enzima: eLONGase Enzyme Mix (*Invitrogen-Life Technologies*)

Reagentes	**Quantidades**	Concentração Final
H_2O de ampola estéril	34,00 μL	-
5x buffer A	5,0 μL	1,5 mM Mg^{+2}
5x buffer B	5,0 μL	
dNTP (10 mM)	1,0 μL	0,2 mM
Primer Forward (10 μM)	1,0 μL	0,2 μM
Primer Reverse (10 μM)	1,0 μL	0,2 μM
eLONGASE Mix (5U/μL)	2,0 μL	Fragm. maiores 12 Kb
cDNA apropriado	1,0 μL	-
Total	50,0 μL/reação	-

Após a distribuição dos reagentes do PCR, os tubos foram colocados em termociclador com o seguinte programa de amplificação:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	94ºC	30 seg.
Desnaturação	94ºC	30 seg.
Anelamento	55ºC	30 seg.
Extensão	68ºC	2 min.
Ext. Final	68ºC	4 min
4ºC	∞	

35 ciclos

1,4 REAÇÃO PCR AccuPrime

Enzima: AccuPrime(tm) *Taq* DNA Polymerase System (*Invitrogen-Life Technologies)*

Reagentes	**Quantidades**	Concentração Final
H_2O de ampola estéril	20,0 μL	-
10x PCR Buffer 1	2,5 μL	1x
Primer Forward (10 μM)	0,5 μL	0,2 μM
Primer Reverse (10 μM)	0,5 μL	0,2 μM
AccuPrime	0,5 μL	1U
cDNA apropriado	1,0 μL	-
Total	μL/reação	25,0 μL

Após a distribuição dos reagentes do PCR, os tubos foram colocados em termociclador com o seguinte programa de amplificação:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	94ºC	2 min.
Desnaturação	94ºC	30 seg.
Anelamento	55ºC	30 seg.
Extensão	68ºC	2 min.
Ext. Final	68ºC	4 min
4ºC	∞	

35 ciclos

Para a visualização das bandas amplificadas procedeu-se com o protocolo 2:

PROTOCOLO 2 ELETROFORESE EM GEL DE AGAROSE 0,9%

- Para um gel de 40 mL foi preparada a seguinte solução: 0,36 g de agarose mais 40 mL de TEB 0,5 X.
- No micro-ondas, a solução foi fundida por 1 minuto.
- Foram adicionados 3,0 μL de brometo de etídeo (8mg/ml) na solução.
- Após homogeneização do brometo, o gel foi colocado na cuba de eletroforese.
- A seguir, esperou-se a polimerização do gel.
- Seguiu-se com adição de 15mL de tampão TEB 0,5 X na cuba de corrida.
- Foram aplicados 4,5 μL da amostra (2,5 μL do PCR + 2,0 μL de azul de bromofenol 5 X) em cada poço.
- **Corrida** - tempo: 25 minutos; voltagem: 110 V; peso molecular: 1 kb Plus (0,7 μL – Invitrogen Life Technologies).

Bandas únicas e extremamente fortes foram purificadas de acordo com o protocolo de purificação da PCR (Concert^{TM} Rapid PCR Purification System/Gibco):

PROTOCOLO 3 PURIFICAÇÃO DO PRODUTO DE PCR

- Adicionou-se ao tampão de lavagem H2 (NaCl, EDTA e Tris-HCl) 30 mL de etanol 95 a 100%.
- Em um tubo de 2 mL foi adicionado tampão TE (10 mM Tris-HCl pH 8,0, 0,1 mM EDTA) na quantidade de 50 μL x o número de tubos para purificação. Este tubo permaneceu aquecido a 65-70ºC até o momento de uso.

- A solução H1 (400 μL: hidrocloreto de guanidina, EDTA, Tris-HCl e isopropanol) foi adicionada em tubo de 2 mL e, em seguida, 50 μL de PCR foram transferidos para o mesmo.
- A seguir, o tubo foi agitado em vortex por alguns segundos.
- O cartucho com sílica foi transferido para um tubo de 2 mL . A seguir, a solução H1 com o produto de PCR foi adicionada no centro do cartucho.
- O tubo de 2 mL mais cartucho foram centrifugados a 12.000 rpm por 1 minuto.
- O resíduo foi descartado e o cartucho foi conservado no mesmo tubo de 2 mL.
- 700 μL de H2 foram adicionados no centro do cartucho.
- Seguiu-se a centrifugação do tubo de 2 mL mais cartucho a 12.000 rpm por 1 minuto.
- O resíduo foi descartado e o cartucho foi conservado no tubo de 2 mL. Novamente, tubo mais cartucho foram centrifugados a 12.000 rpm por 1 minuto.
- A seguir, o tubo de 2 mL foi descartado e o cartucho transferido para um novo tubo de 1,5 ml.
- 50 μL de TE aquecido foram adicionados no centro do cartucho.
- O cartucho permaneceu dentro do tubo em temperatura ambiente por 1 minuto.
- A seguir, procedeu-se a centrifugação do tubo de 1,5 mL mais cartucho a 12.000 rpm por 2 minutos.
- O cartucho foi descartado. O produto purificado foi submetido à eletroforese (agarose 0,9%) para verificar a intensidade da banda. O restante do material foi conservado à -20^0 C até a etapa de clonagem.

Nos TF(s) onde houve amplificação de mais de uma banda procedeu-se eletroforese do produto de PCR em LMP para a extração das mais intensas, segundo o protocolo 4:

PROTOCOLO 4 "SIZE SELECTION" (*Low Melting Point Agarose* 1,2 %)

- Para um gel de 50 mL foi preparada a seguinte solução:
  - 0,60 g de agarose LMP
  - 1,0 mL TAE 50X
  - 49,0 mL de H_2O Milli-Q.
- A solução foi fundida em microondas (potência 50, 80 segundos - 4X de 20 seg).
- 2 μL de brometo de etídeo foram acrescentados (8mg/ml) ao gel.
- Aguardou-se a polimerização do gel.
- A seguir, o tampão TAE 1X (7 ml de TAE 50X, 343 ml de água Milli-Q) foi preparado para a corrida .
- Em cada poço do gel foram aplicados 30 μL de PCR + 6 μl de azul de bromofenol 5 X. Utilizou-se 2 μL de solução de peso molecular 1 Kb Plus (Invitrogen Life Technologies) que foi preparada (2,5 μL de marcador + 7,5 μL de água estéril) antes da aplicação.
- Corrida: tempo: 1hora e 20 minutos a 70 V.
- Os slices selecionados foram recolhidos em tubos de 2 mL.
- A seguir, este tubos foram conservados a -200 C até a purificação. Após a retirada dos slices do LMP procedeu-se com o protocolo 5 (ConcertTM Rapid Gel Extraction System/Gibco):

PROTOCOLO 5 PURIFICAÇÃO DE FRAGMENTOS DE LMP

- O tampão TE (10 mM Tris-HCl pH 8,0, 0,1 mM EDTA) foi aquecido a 65^0 a 70^0C: 50 μL X número de microtubos.
- O termobloco foi aquecido à 50^0C.
- 30 μL do tampão de solubilização L1 (perclorato de sódio, acetato de sódio, TBE) foram adicionados para cada 10 mg de gel. 3 X peso g = μl ou 1mg = 0,001g.
- Os tubos de 2 mL foram incubados a 50^0C por 15 minutos e misturados a cada 3 minutos para a dissolução do gel. Após a dissolução do gel, os tubos foram incubados por mais 5 minutos.
- Tubos de 2 mL estéreis foram identificados.
- A seguir, cartuchos com sílica foram colocados em tubos de 2 mL e, a amostra dissolvida transferida no centro destes.
- Tubos mais cartuchos foram centrifugados a 12.000 rpm por 1 minuto. Os tubos foram descartados.
- Os cartuchos foram colocados em outros tubos de 2 mL, sem tampa.
- 700 μL do tampão L2 (NaCl, EDTA e Tris-HCl) foram adicionados no centro dos cartuchos.
- Os cartuchos permaneceram dentro dos tubos por 5 minutos à temperatura ambiente.
- Seguiu-se com a centrifugação dos tubos mais cartuchos a 12.000 rpm por 1 minuto. O sobrenadante foi descartado, e os tubos mais cartuchos centrifugados novamente por 1 minuto para remover o tampão residual.
- Os cartuchos foram colocados em tubos de 1,5 mL.
- 50 μl de tampão TE aquecido foi adicionado diretamente no centro de cada um dos cartuchos. Os tubos mais cartuchos permaneceram por 1 minuto à temperatura ambiente e depois foram centrifugados a 12.000 rpm por 2 minutos.

- O produto purificado do gel foi submetido à eletroforese (agarose 0,9%) para verificar a qualidade da banda purificada. O restante do material foi conservado à -20^0 C até a etapa de clonagem (Protocolo 6).

PROTOCOLO 6 CLONAGEM DO INSERTO

Foram utilizados três tipos de produtos (Kits) para clonagem, cada um específico para o tamanho de fragmento obtido na reação de PCR e a enzima utilizada:

- TOPO TA Cloning® (Invitrogen) – Para fragmentos amplificados com enzimas Taq DNA polymerase;
- TOPO XL Cloning® (Invitrogen) – Para fragmentos entre 3 a 10Kb;
- TOPO Zero Blunt Cloning® (Invitrogen) – Para fragmentos amplificados com as enzimas de maior fidelidade, como por exemplo AccuPrime(tm) Taq DNA Polymerase System, Platinum Taq DNA Polymerase High Fidelity, etc.

As etapas de transformação e clonagem foram realizadas de acordo com o protocolo dos referidos Kits, com pequenas modificações.

Reagentes e equipamentos utilizados na transformação:

<u>Suplementos microbiológicos:</u>

- O meio sólido LB foi derretido no microondas.
- Esperou-se o meio amornar e foi adicionado: 1μL/mL de ampicilina (100μg/mL) ao meio que seria utilizado para a clonagem com os Kits TOPO TA Cloning® e TOPO Zero Blunt Cloning® ou 1μL/mL de

kanamicina (50μg/mL) ao meio que seria utilizado para a clonagem com o Kit TOPO XL Cloning®.

- 50 mL de meio com o antibiótico foram espalhados em cada placa de cultura.
- Esperou-se a solidificação do meio e as placas foram tampadas e vedadas.
- As placas foram mantidas na geladeira até o momento de uso.

Banho-maria a 42ºC.

Shanking e estufa a 37ºC.

Preparo para a transformação:

Frascos de células *E. coli* quimicamente competentes TOP 10 foram desgelados em gelo picado, e 25 μL destas células foram utilizados para cada transformação.

Execução da reação de clonagem:

TOPO TA Cloning®

Antes do início da reação TOPO, adicionou-se 3'-A nos produtos amplificados ou produtos purificados de *Low Melting* ou de PCR, como segue:

- 1U de Taq polimerase (0,2 μL) foi adicionada em cada tubo de PCR, no gelo.
- O produto de PCR foi misturado muito bem com micropipeta.
- Os tubos permaneceram em termociclador a 72^0C por 10 minutos.

A seguir, os tubos foram colocados em gelo picado e, procedeu-se imediatamente a reação TOPO:

Reagentes	Reação TOPO
H_2O estéril	3,0 μL
Solução Salina	1,0 μL
Vetor pCR®4-TOPO	1,0 μL
Produto de PCR fresco	1,0 μL
Volume Final	6,0 μL

A reação TOPO foi homogeneizada gentilmente com micropipeta e incubada por 5 minutos à temperatura ambiente (22-23^{0}C). Para produtos de PCR grandes (> 1 kb) ou clonagem de um "pool" de produtos de PCR, o tempo de incubação da reação foi de 30 minutos com o objetivo de produzir mais colônias.

A reação foi colocada em gelo picado e procedeu-se com a transformação química.

Transformação química:

- No fluxo laminar: 2 μL da reação de clonagem TOPO TA® foram adicionados dentro do frasco da *E. coli* quimicamente competente TOP10. Homogeneizou-se gentilmente sem micropipeta**.**
- Os microtubos permaneceram em gelo picado por 30 minutos.
- A seguir, um choque-térmico foi aplicado nas células por 30 segundos a 42^{0}C sem homogeneizar.
- Imediatamente os microtubos foram transferidos para o gelo picado.
- 125 μL de meio SOC a temperatura ambiente foram adicionados em cada tubo, no fluxo.

- Os tubos foram tampados firmemente e procedeu-se com homogeneização horizontal (*shaking* 200 rpm) por 1 hora a 37^0C.
- No fluxo: 150 μL de cada transformação foram espalhados em uma placa de cultura (meio LB+ampicilina).
- As placas foram incubadas a 37^0C durante uma noite.
- 24 colônias de cada TF foram escolhidas para análise.

TOPO XL PCR Cloning®

Tubos de microcentrífuga, tipo *eppendorf,* de 200μL foram devidamente identificados e colocados em gelo picado. Procedeu-se então a reação TOPO XL:

Reagentes	**Reação TOPO**
Vetor pCR®XL-TOPO	1,0 μL
Produto de PCR	1,0 - 4,0 μL
H_2O estéril	q.s.p. 5,0 μL
Volume Final	5,0 μL

A reação TOPO foi homogeneizada gentilmente com micropipeta e incubada por 5 minutos à temperatura ambiente (22-23^0C).

Passados os 5 minutos, foi adicionado 1,0 μL da solução 6X TOPO Cloning Stop e agitada por alguns segundos à temperatura ambiente.

Realizou-se a centrifugação do tubo por alguns segundos e então a reação foi colocada em gelo picado e procedeu-se com a transformação química.

Transformação química:

- No fluxo laminar: 2 μL da reação de clonagem TOPO TA® foram adicionados dentro do frasco da *E. coli* quimicamente competente TOP10. Homogeneizou-se gentilmente sem micropipeta**.**
- Os microtubos permaneceram em gelo picado por 30 minutos.
- A seguir, um choque-térmico foi aplicado nas células por 30 segundos a 42^0C sem homogeneizar.
- Imediatamente os microtubos foram transferidos para o gelo picado.
- 125 μL de meio SOC a temperatura ambiente foram adicionados em cada tubo, no fluxo.
- Os tubos foram tampados firmemente e procedeu-se com homogeneização horizontal (*shaking* 200 rpm) por 1 hora a 37^0C.
- No fluxo: 150 μL de cada transformação foram espalhados em uma placa de cultura (meio LB+kanamicina).
- As placas foram incubadas a 37^0C durante uma noite.
- 24 colônias de cada TF foram escolhidas para análise.

Zero Blunt TOPO PCR Cloning®

Tubos de microcentrífuga, tipo *eppendorf,* de 200μL foram devidamente identificados e colocados em gelo picado. Procedeu-se então a reação Zero Blunt :

Reagentes	**Reação TOPO**
Produto de PCR Fresco	1,0 μL
Solução Salina	1,0 μL
Vetor pCR®XL-TOPO	1,0 μL
H_2O estéril	3,0 μL
Volume Final	6,0 μL

A reação TOPO foi homogeneizada gentilmente com micropipeta e incubada por 5 minutos à temperatura ambiente (22-23^{0}C). Para produtos de PCR grandes (> 1 kb) ou clonagem de um "pool" de produtos de PCR, o tempo de incubação da reação foi de 30 minutos com o objetivo de produzir mais colônias.

A reação foi colocada em gelo picado e procedeu-se com a transformação química, como no item 6.2.1. TOPO TA Cloning®

PROTOCOLO 7 PCR DE COLÔNIA

Para as reações de PCR de colônia, foram utilizadas microplacas de 96 poços e cada fileira foi identificada pelo número do TF.

Previamente à reação de PCR, um esfregaço de cada colônia de bactérias (proveniente de bibliotecas de cDNA) foi transferido para cada um dos poços da microplaca de PCR (24 colônias por TF).

Simultaneamente, foi realizado um *backup* das colônias que foram palitadas para o PCR em uma microplaca de glicerol. Para isso, em cada poço da placa de glicerol foram adicionados 100 μL da seguinte solução: 250 μL ampicilina (4 mg/mL) + 10 mL de meio LB líquido. A seguir, o palito com a mesma colônia utilizada na PCR era transferido para o poço específico da microplaca de glicerol. Procedia-se com a homogeneização do palito e descarte. A microplaca com as colônias de cada TF permaneceu em estufa a 37^{0}C durante uma noite.

As colônias na microplaca de glicerol foram conservadas -80^{0}C. Para esta etapa, foi preparada uma solução contendo 8,0 mL de meio LB líquido + 125 μL de ampicilina (4 mg/mL) + 2,0 mL de glicerol 26% e, 100 μL foram adicionados em cada poço da microplaca de glicerol contendo os clones de cada TF.

Após a palitagem das colônias, um mix foi preparado em gelo, e 15 μL deste distribuído por poço na microplaca de PCR:

PCR DE COLÔNIA

Reagentes	Quantidades
H_2O estéril	3,0 μL
10x Taq Buffer	1,5 μL
dNTP (1,25 mM)	1,5 μL
$MgCl_2$ (50 Mm)	0,45 μL
Primer Reverse M13 (20pmol)	0,11 μL
Primer Forward M13 (20 pmol)	0,11μL
Taq DNA polimerase	0,11 μL
Volume Total	15,0 μL

As seqüências dos primers utilizados foram as seguintes:

- M13 Forward: 5' CGC CAG GGT TTT CCC AGT CAC GAC 3'
- M13 Reverse: 5' TTT CAC ACA GGA AAC AGC TAT GAC 3'

Após a distribuição do mix de PCR de colônia, a microplaca foi colocada em termociclador com o seguinte programa de amplificação:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	95ºC	3 min.
Desnaturação	95ºC	40 seg.
Anelamento	55ºC	40 seg.
Extensão	72ºC	55 seg.
Ext. Final	68ºC	5 min.
4ºC	∞	

35 ciclos

Para avaliação da amplificação do PCR de colônia procedeu-se com um em um gel de agarose 0,9%.

PROTOCOLO 8 PCR DE SEQUENCIAMENTO

Em uma nova microplaca de PCR foram distribuídos 8 μL em cada poço da seguinte solução:

PCR DE SEQÜENCIAMENTO

Reagentes	**Quantidades**
H_2O estéril	3,6 μL
Tampão “Save Money“	2,0 μL
ABI BigDye Terminator	2,0 μL
Primer F ou R do M13 (20 pmol)	0,4 μL
Volume total	10 μL

Para cada poço da nova microplaca, foram transferidos 2,0 μL (banda com intensidade média) ou 1,0 μL (banda com intensidade forte) do PCR de colônia, homogeneizando bem com pipeta multicanal antes e depois da transferência, e obedecendo a identificação original (números dos TFs da microplaca de PCR de colônia).

A ciclagem utilizada para a reação de PCR de seqüenciamento foi a seguinte:

CICLAGEM		
	Temperatura	Tempo
Desn. Inicial	96ºC	5 min.
Desnaturação	96ºC	30 seg.
Anelamento	55ºC	15 seg.
Extensão	68ºC	4 min.

35 ciclos

PROTOCOLO 9 PRECIPITAÇÃO

A precipitação das amostras foi realizada como segue:

- Em cada poço da microplaca de PCR de seqüenciamento foram adicionados 40 μL da solução: 3.300 μL de isopropanol + 1.100 μL de água estéril.
- A solução contida na microplaca foi homogeneizada em vortex por 1 minuto.
- A seguir, a microplaca foi incubada por 15 minutos à temperatura ambiente e protegida da luz. Seguiu-se centrifugação da microplaca a 4.000 rpm por 25 minutos a 18°C.
- A microplaca foi invertida para o descarte do sobrenadante e colocada invertida sobre papel absorvente apropriado.
- A seguir, foram adicionados 130 μL de etanol 70% (10.010 μL etanol 100% + 4.290 μL de água estéril), preparado no momento de uso, em cada poço da microplaca. Novamente a microplaca foi agitada em vortex por 1 minuto e, a seguir centrifugada a 4.000 rpm por 5 minutos.
- O sobrenadante foi descartado e o excesso de solução foi retirado por "spin down" (pulso de centrifugação) com a microplaca invertida sobre papel absorvente. A microplaca foi seca em termociclador à temperatura de 90^0 C por 2 minutos. Para conservação, a placa foi mantida embrulhada com papel alumínio a -20^0 C até o momento do uso.

PROTOCOLO 10 DESNATURAÇÃO DAS AMOSTRAS

Procedeu-se com o protocolo 10 para a desnaturação das amostras:

- 2,0 μL de "loading dye" contendo formamida foram aplicados em cada poço da microplaca.
- A microplaca foi agitada durante 1 minuto no vortex.
- Seguiu-se com um "spin down" da microplaca por 2 segundos a 1000 rpm
- As amostras foram desnaturadas em termociclador a 95^0 C por 3 minutos.
- A microplaca foi mantida em gelo picado até o momento da aplicação das amostras no gel.

O seqüenciador automático ABI 377-Applied Biosystems® foi utilizado para a leitura das seqüências produzidas.

O tampão TEB 1 X foi utilizado na corrida eletroforética e o tempo de corrida foi de 10,0 horas. As amostras foram aplicadas em gel de 96 poços (comprimento da placa: 36 cm). O seqüenciamento foi realizado de acordo com o protocolo 11:

PROTOCOLO 11 SEQÜENCIAMENTO AUTOMÁTICO

- A cuba inferior foi colocada vazia no equipamento ABI/377.
- A seguir, a placa mais gel, já montada no cassete, foi colocada no ABI/377.
- O *Software Collection* foi aberto. A seguir, "clicou-se" na janela *File* → *New* e a pasta *SeqRun* foi aberta.
- A porta do ABI/377 foi fechada e esperou-se o barulho "click" do seqüenciador. Acionou-se o módulo *Plate Check*.
- Se a placa estivesse limpa, foi dado um *Cancel* → *Terminate*.

- A porta do ABI/377 foi aberta e, a cuba superior foi colocado no equipamento, apoiando-a na placa da frente. As presilhas laterais foram fechadas.
- 660 mL de H_2O MilliQ foram colocados na cuba superior.
- Na cuca inferior, foram colocados 600 mL de TBE 1X.
- As 4 presilhas do cassete foram abertas e placa térmica foi colocada apoiando-a no protetor do "laser". As mangueiras e o fio terra foram conectados.
- A porta ABI/377 foi fechada e, a seguir o módulo *PreRun* acionado. Iniciou-se a pré-corrida e, logo em seguida, abriu-se a porta do seqüenciador para a aplicação das amostras ímpares com a multicanal, em $\pm$ 15 minutos.
- Após aplicação das amostras ímpares, a porta do ABI/377 foi fechada, e clicou-se na janela *Window* $\rightarrow$ *Status*. A migração das amostras durante a pré-corrida foi de 3 minutos.
- Abriu-se a porta do ABI/377 e, as amostras pares foram aplicadas com a multicanal.
- A porta do ABI/377 foi fechada novamente e esperou-se de 2-3 minutos.
- A seguir, a porta foi aberta e o equipamento pausado. 66 ml de H_2O da cuba superior foram removidos. Seguiu-se com a adição e homogeneização de 66 mL de TEB 10X na cuba superior. A porta do equipamento foi fechada.
- Aguardou-se a *PreRun* reiniciar (barulho c*lick*).
- A pré-corrida foi cancelada (*Cancel* $\rightarrow$ *Terminate*).
- O módulo *Run* foi acionado (*Gel File:* IL2- identificação dos TFs $\rightarrow$ *Save*) para iniciar a corrida.

PROTOCOLO 12 ANÁLISE DOS RESULTADOS

O *tracking* das seqüências foi feito automaticamente pelo *Software* de análise. Erros de *tracking* foram corrigidos manualmente, reextraíndo os *lanes* novamente.

PROTOCOLO 13 "ZIPAGEM" DE ARQUIVOS

Os arquivos de cada TF foram "zipados" de acordo com o protocolo a seguir:

No "Macintosh", as melhores seqüências (aquelas que apresentaram o menor número de Ns) foram selecionadas e colocadas em arquivos datados e separados por TF.

No PC, o programa *Winzip* foi aberto: *File* → *New*.

Deu-se um nome ao arquivo: exemplo: IL2-TF01048-A.

O item *include system and hidden files* foi desmarcado. Clicou-se Ok.

As *reads* (Exemplo: IL2-TF01048-A, IL2-TF01048-B...) foram selecionadas através das teclas *shift* + ↓. A seguir, Clicou-se *Add*.

Os arquivos zipados foram submetidos à equipe de bioinformática, através das ferramentas disponíveis na página do projeto.

PROTOCOLO 14 SUBMISSÃO DOS TFs

Os arquivos "zipados" em PC foram submetidos à rede virtual através da *Homepage* do Instituto LUDWIG (http://www.ludwig.org.br/tfi), os quais eram avaliados e disponibilizados na homepage (figura 4.1) do projeto cerca de 48 horas depois, com acesso restrito aos participantes do projeto.

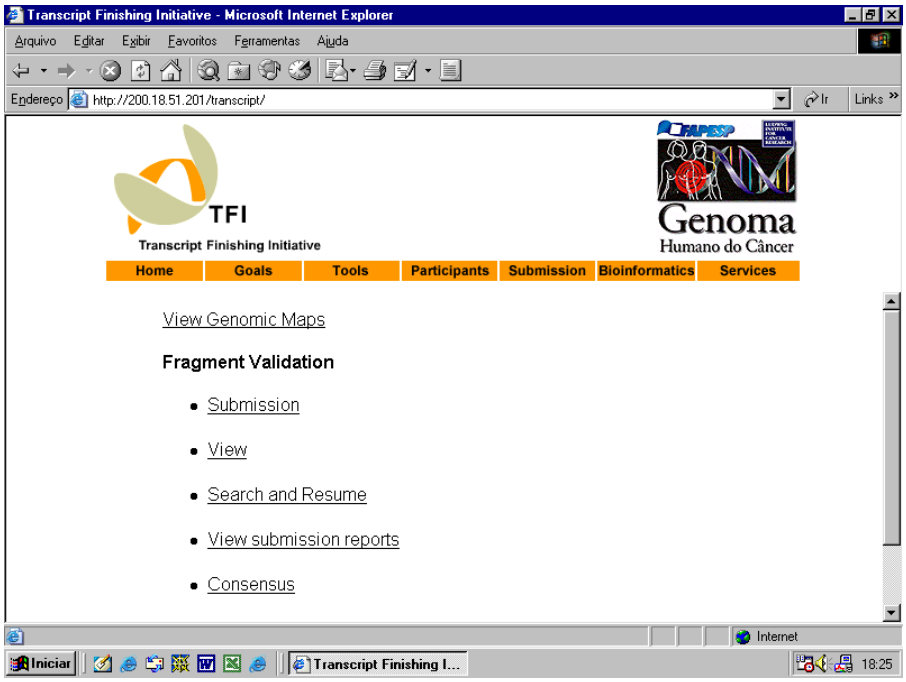


Figura 4.1 - Homepage desenvolvida para o Projeto *Transcript Finishing Initiative*.

PROTOCOLO 15 PREPARO DAS SOLUÇÕES:

15.1 MEIO LB

32 g de LB agar.
8 g de Bacto agar.
1 L de água Milli-Q.
A solução foi autoclavada a 121^{0}C por 15 minutos.

15.2 AMPICILINA (4mg/mL)

40 mg de ampicilina + 10 mL de água de ampola
1 mL da solução foi transferido para tubos de 2 mL que foram armazenados a –20^{0}C.

15.3 GLICEROL 26%

29,9 mL de glicerol 87%.
100 mL de água estéril.

15.4 TEB 10X (1 L)

108 g de Tris HCl.
56 g de ácido bórico.
8 g Na_2EDTA

O volume de 1 L foi completado com água Milli-Q. A seguir, a solução foi filtrada utilizando membrana de 0,22 μm. Mediu-se o pH que permaneceu entre 8,2 a 8,8.

15.5 TAMPÃO "SAVE MONEY"

2 mL Tris HCl pH 9,0 (1 M).

1 mL de $MgCl_2$ da Gibco (50 mM).

10 mL de H2O estéril de ampola.

15.6 LOADING BUFFER

Solução Y: 500 μL de EDTA 0,5 M pH 8,0 + 9,5 mL de água Milli-Q autoclavada.

Solução X: 0,5 g de Blue Dextran foram colocados em um frasco de 50 mL e a solução Y foi adicionada até o volume de 10 mL.

Para 5 partes de formamida colocou-se 1 parte da solução X. Exemplo: 1500 μl de formanida para 300 μl de X.

A solução foi distribuída em tubos de 2 mL. Os tubos permaneceram na geladeira até o momento de uso.

4.2.2 ..... Análises in silico

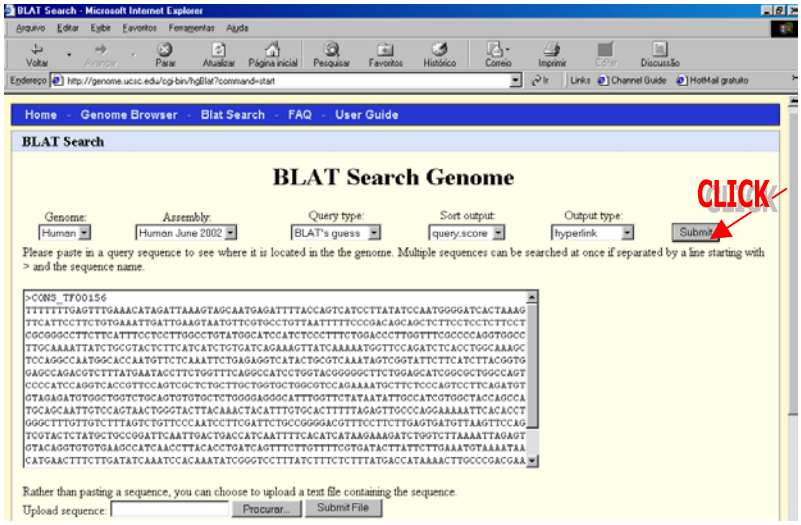
Foram realizadas análises *in silico* das seqüências dos consensos (ANEXO 1) dos TFs validados pelo laboratório IL2 disponibilizados pela coordenação do projeto TFI e das seqüências geradas durante o processo de validação mas que não alinharam ao ID fornecido pela coordenação. Essa análise tinha como objetivos, encontrar novas seqüências, ainda não representadas nos bancos de dados públicos. Optou-se por realizar uma anotação preliminar destas seqüências, utilizando as ferramentas BLAT ou BLAST, a fim de selecionar seqüências que pudessem ser interessantes para confirmação laboratorial e possíveis estudos futuros.

Foram utilizadas as seqüências de boa qualidade (ANEXO 2) avaliadas pelo grupo de bioinformática do projeto TFI.

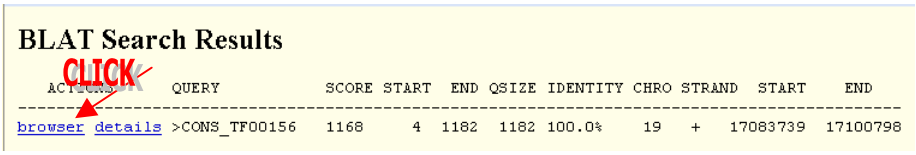
Tal análise foi realizada segundo os protocolos 16 e 17.

PROTOCOLO 16 ANÁLISE BLAT

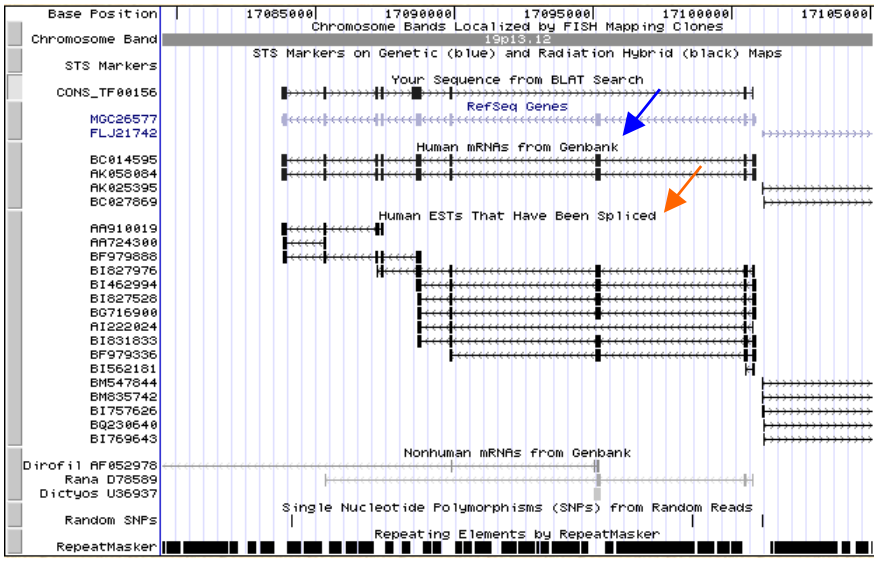
➢ A seqüência de nucleotídeos obtida nos respectivos laboratórios validadores, foi copiada da homepage do projeto TFI (www.ludwig.org.br/tfi) e submetida à ferramenta BLAT, disponível no endereço eletrônico http://genome.ucsc.edu/, como exemplo a seguir:



➢ Resultando na tela abaixo, onde se escolheu a seqüência de maior identidade, clicando sobre ela.

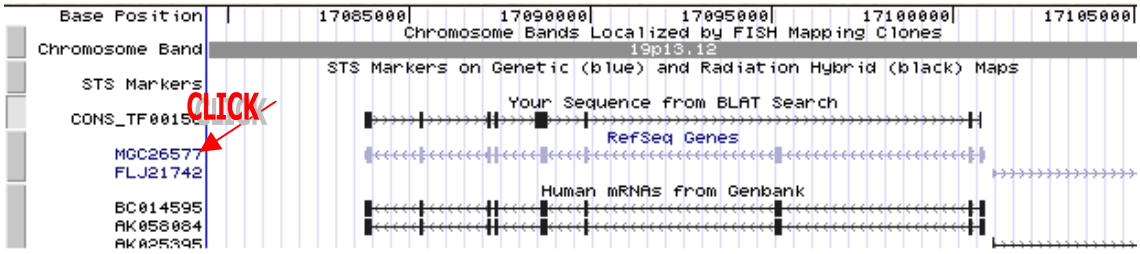


➢ Foi exibido um gráfico como o abaixo:

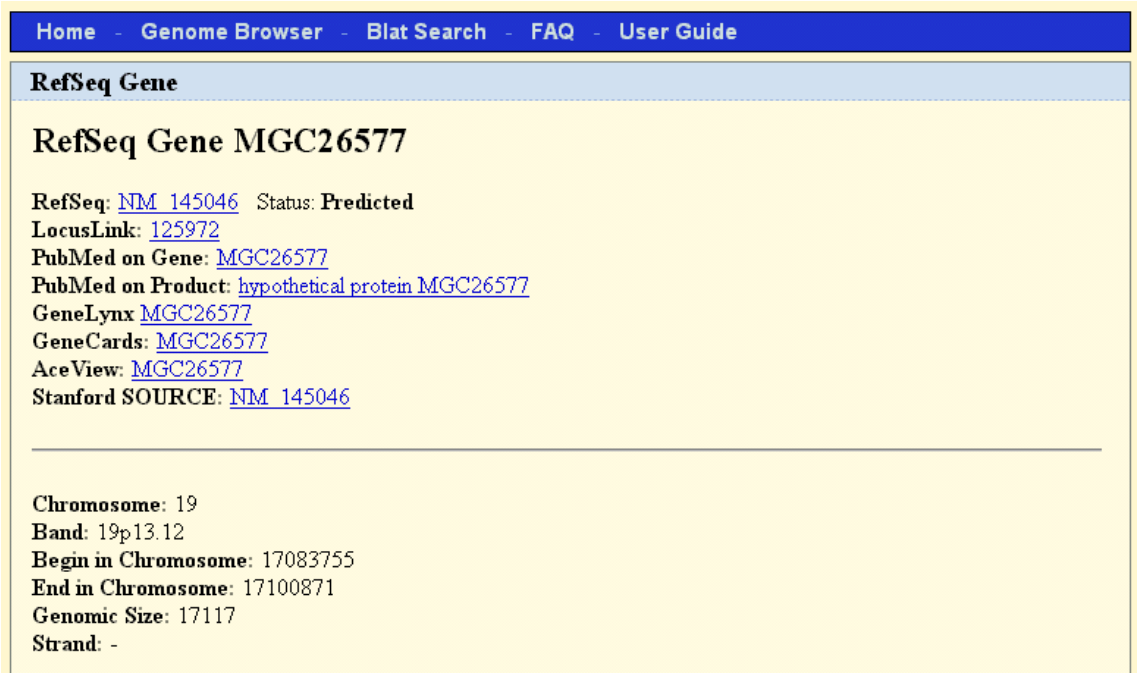


➢ Verificou-se a presença de splicing alternativo, comparando a seqüência pesquisada ao banco de dados das ESTs (seta laranja) e mRNAs (seta azul).

➢ Os dados para a anotação preliminar foram adquiridos clicando-se sobre o nome do gene, como demonstrado abaixo:



➢ Dependendo dos dados disponíveis de cada TF, esta tela poderia ser ou não visualizada.

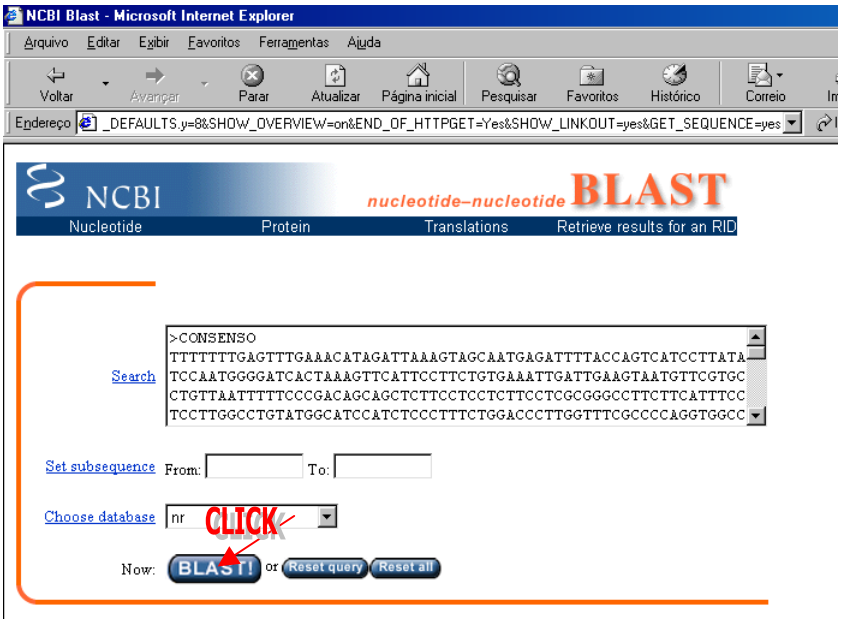


➢ A tela acima apresenta links para diferentes bancos de dados, onde foi possível aquisição das informações disponíveis, necessárias para uma anotação preliminar.

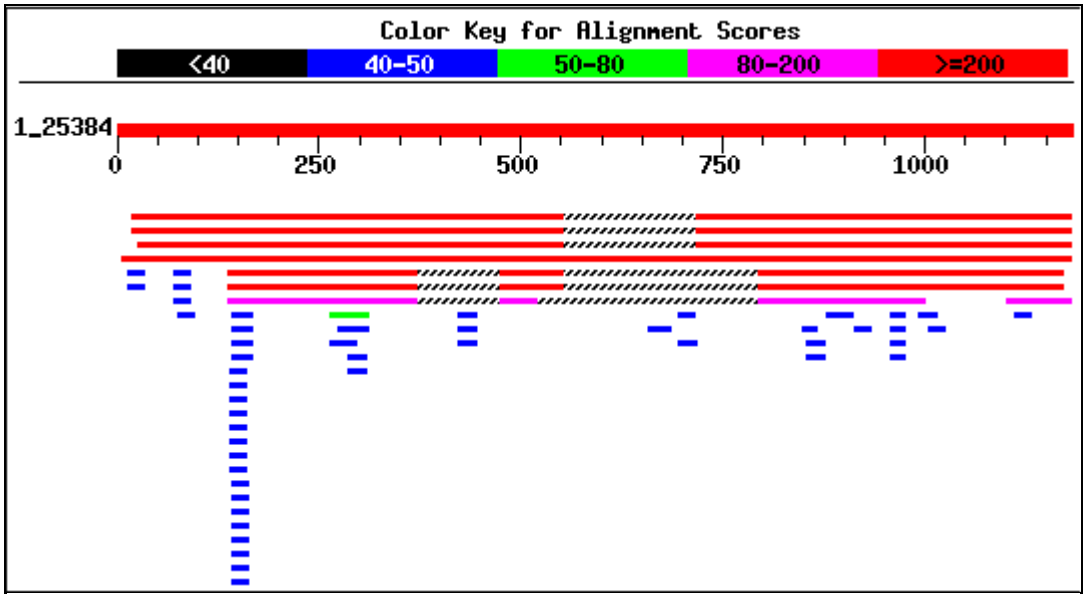
Quando nenhum alinhamento significante foi encontrado para a seqüência pesquisada pela ferramenta BLAT. Então, partiu-se para a análise utilizando outra ferramenta de bioinformática, a ferramenta BLAST, como demonstrado no protocolo 17.

PROTOCOLO 17 ANÁLISE BLAST

➢ A seqüência de nucleotídeos foi copiada da homepage do projeto TFI (www.ludwig.org.br/tfi) e submetida à ferramenta BLAST (opção BLASTn), disponível no endereço eletrônico http://www.ncbi.nlm.nih.gov, como exemplo a seguir:



➢ Após a submissão, foi exibido um gráfico como o abaixo, onde os alinhamentos por similaridade de seqüências são organizados por ordem de identidade e *score*:



Sequences producing significant alignments:		Score (bits)	E Value	
gi|21450883|ref|NM_145046.1|	Homo sapiens calreticulin 2 (C...	1027	0.0	L U
gi|15779043|gb|BC014595.1|BC014595	Homo sapiens, Similar to...	1027	0.0	L U
gi|16554106|dbj|AK058084.1|	Homo sapiens cDNA FLJ25355 fis,...	1015	0.0	L U
gi|21743748|gb|AC008764.9|	Homo sapiens chromosome 19 clone...	708	0.0	
gi|21624618|ref|NM_028500.1|	Mus musculus RIKEN cDNA 170003...	270	4e-69	L U
gi|12839755|dbj|AK006582.1|	Mus musculus adult male testis ...	270	4e-69	L U
gi|12857889|dbj|AK018263.1|	Mus musculus adult male medulla...	182	7e-43	L U

➢ Optou-se pela seqüência que apresentava alinhamento com o maior *score* e melhor E-value, ou seja mais próximo a zero. Clicando-se sobre o número do *score* (seta laranja), é visualizado o alinhamento da combinação entre a seqüência pesquisada e a seqüência depositada no GenBank. Ao clicar na caixas com as letras "L", "U" ou sobre a identificação genômica (seta vermelha), foram obtidos dados para a anotação preliminar da seqüência pesquisada.

➢ O mesmo procedimento foi adotado para a opção BLATx, que traduz a seqüência de nucleotídeos em aminoácidos e pesquisa nos bancos de dados de proteínas disponíveis por seqüências similares.

PROTOCOLO 18 CÁLCULO DE TAMANHOS MÍNIMO E MÁXIMO DO "GAP" ENTRE AS ESTs DOS TFs A SEREM VALIDADOS

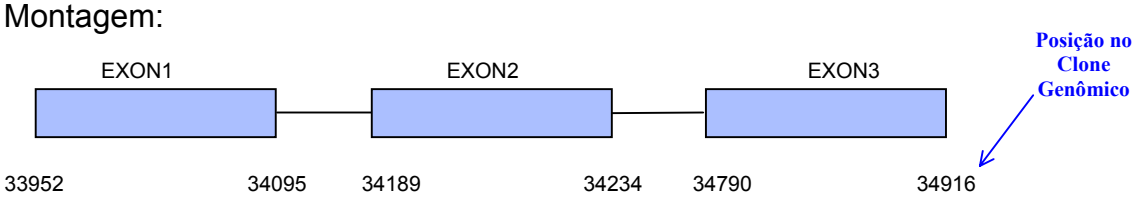
Para o cálculo dos tamanhos mínimos e máximos dos *gaps* entre as ESTs 1 e 2 dos TFs, ou seja, o tamanho provável do fragmento a ser amplificado, seguiu-se o seguinte protocolo:

- 1)Foi realizado um alinhamento no *Blast 2 sequences* (www.ncbi.nlm.nih.gov/blast/bl2seq.html/) entre a EST1 e o Clone genômico, para verificar a posição da EST em relação ao clone.
- 2) O mesmo procedimento do item 1 foi realizado para a EST2.
- 3) Exemplo com a EST1 do TF00072:

Resultado Blast 2 sequences: EST1(AW504573) x Clone genômico (AC006942)

```
Score =  277 bits (144), Expect = 7e-72
Identities = 144/144 (100%)
 Strand = Plus / Plus

Query:                  1     cacggggtatggtccccgggtccgagggcccggcccgcgccgggagcgtggtggccgacg 60
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                  33952 cacggggtatggtccccgggtccgagggcccggcccgcgccgggagcgtggtggccgacg 34011

Query:                  61    tggtgtttgtgattgagggtacggccaacctgggaccctacttcgaggggctccgcaagc 120
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                  34012 tggtgtttgtgattgagggtacggccaacctgggaccctacttcgaggggctccgcaagc 34071

Query:                  121   actacctgctcccggccatcgagt 144
                              |||||||||||||||||||||||||
Sbjct:                  34072 actacctgctcccggccatcgagt 34095

Score = 89.1 bits (46), Expect = 4e-15
Identities = 46/46 (100%)
 Strand = Plus / Plus

Query:                  143   gtattttaatggtggtcctcctgctgagacggacttcgggggagac 188
                              ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                  34189 gtattttaatggtggtcctcctgctgagacggacttcgggggagac 34234

Score =  244 bits (127), Expect = 5e-62
Identities = 127/127 (100%)
 Strand = Plus / Plus

Query:                  189   tatgggggacccagtacagcctcgtggtgttcaacacagtggactgcgctcccgagtcc 248
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                  34790 tatgggggacccagtacagcctcgtggtgttcaacacagtggactgcgctcccgagtcc 34849

Query:                  249   tacgtacaatgtcacgctcccaccagcagcgcctatgagtttgtcacctggctcgatggc 308
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                  34850 tacgtacaatgtcacgctcccaccagcagcgcctatgagtttgtcacctggctcgatggc 34909

Query:                  309   attaagt 315
                              |||||||
Sbjct:                  34910 attaagt 34916
```

Montagem:



➢ 4) Foi realizado um alinhamento no *Blast 2 sequences* (www.ncbi.nlm.nih.gov/blast/bl2seq.html/) entre o *Primer Foward* e o Clone genômico, para verificar a posição deste em relação ao clone.

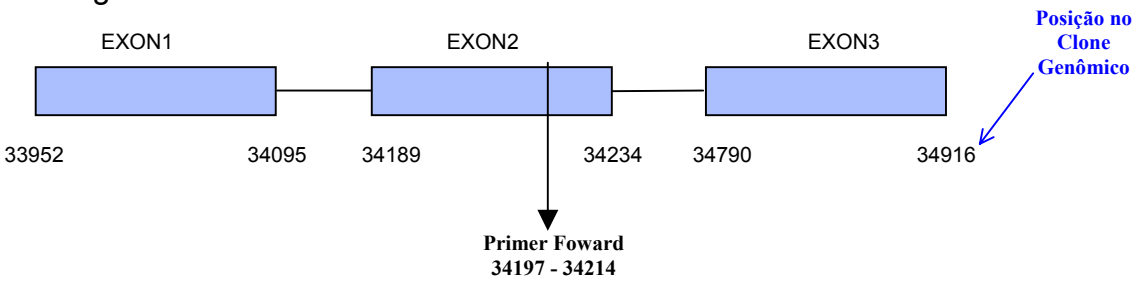
➢ 5) O mesmo procedimento do item 4 foi realizado para o *Primer* Reverse.

➢ 6) Exemplo com o *Primer Forward* do TF00072:

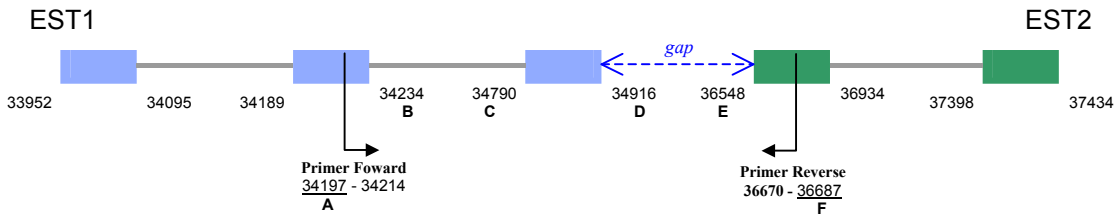
Resultado Blast 2 sequences: *Primer Foward* x Clone genômico (AC006942)

```
Score = 35.3 bits (18), Expect = 0.17
Identities = 18/18 (100%)
 Strand = Plus / Plus
                                                          |
                                                          |


Query:                    1     atggtggtcctcctgctg 18
                                |||||||||||||||||
Sbjct:                    34197 atggtggtcctcctgctg 34214
```

Montagem:



➢ 7) Ao final teremos as seguintes coordenadas para o TF00072:



- 8) Cálculo dos possíveis tamanhos do produto de PCR:

Assumindo a ausência de *introns* *(a região entre as ESTs corresponde totalmente a um exon)*

Tam. Máximo = (B-A) + (D-C) + (E-D) + (F-E)

Tam.Máximo = (34234-34197) + (34916-34790) + (36548-34916)+(36687-36548) = **1.934 pb**

Assumindo a existência de *introns* *(a região entre as ESTs corresponde totalmente a um intron)*

Tam. Mínimo = (B-A) + (D-C) + (F-E)

Tam. Mínimo = (34234-34197) + (34916-34790) + (36687-36548) = **302 pb**

5. Resultados e Discussão

Até o presente momento, foram atribuídos ao laboratório IL2, 20 TFs para validação. Destes, 11 (55%) foram validados experimentalmente (Gráfico 5.1). Os dados da validação dos TFs foram resumidos na tabela 5.2. Neste período, padronizou-se as técnicas de PCR, clonagem e seqüenciamento e também, de análises *in silico (*via microcomputadores) de seqüências para futuras anotações.

A eficiência de validação foi variável entre os 31 grupos que compuseram a rede. A porcentagem variou entre 7% a 79% (até Outubro de 2002), como demonstrado no gráfico 5.2.

Aos transcritos validados, após a disponibilização de consensos montados pela coordenação, foi realizada uma análise preliminar (*ïn silico*) com ferramentas de bioinformática, detalhados no item 5.2.1.

Dos transcritos que não foram validados, ou seja, que a seqüência não alinhou com o clone genômico fornecido pela coordenação, realizou-se uma análise a partir de ferramentas de bioinformática disponíveis, em busca de novas seqüências, detalhados no item 5.2.2.

VALIDAÇÃO DE TRANSCRIPT FINISHING (TFs)	
TFs VALIDADOS	**TFs NÃO VALIDADOS**
TF00035	TF00072
TF00040	TF00074
TF00041	TF01048
TF00156	TF01049
TF00157	TF00193
TF00194	TF00408
TF00308	TF00324
TF00309	TF00325
TF00380	TF00214
TF00404	
TF00232	

Tabela 5.2 TFs validados e não validados pelo Laboratório IL2.

Grafico 5.1 Dos 20 TFs recebidos pelo grupo IL2, 11 TFs foram validados (55%)

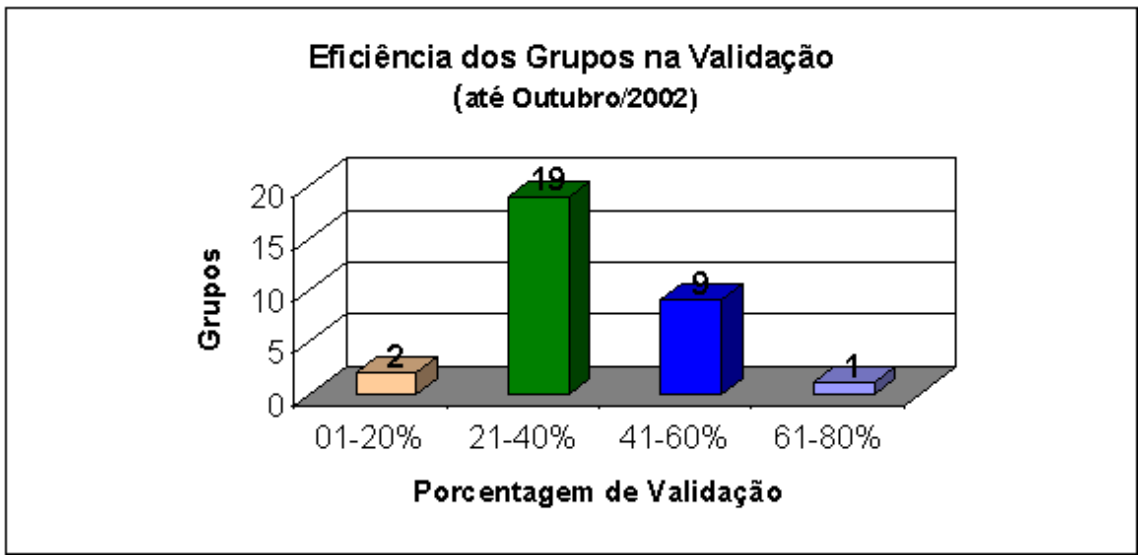


Gráfico 5.2 – Eficiência de validação dos laboratórios validadores do Projeto TFI até Outubro de 2002. O Laboratório IL2 está entre os nove laboratórios que apresentam eficiência entre 41 e 60%.

5.1 Validação

5.1.1 Transcript finishing (TFs) validados

As figuras nas páginas seguintes são os mapas virtuais na interface gráfica do Projeto Transcript Finishing Initiative, dos TFs recebidos e validados pelo grupo IL2.

A cor das ESTs representa a origem de cada uma delas: ORESTES em roxo, CGAP (*Cancer Genome Anatomy Project*) em verde, MGC (*Mammalian Gene Collection*) em azul, entre outras.

Abaixo de seu respectivo mapa virtual encontra-se a foto do gel de agarose, evidenciando a banda gerada pela reação de PCR que validou o transcrito.

Durante a execução do projeto, foram realizadas inúmeras reações (tabela 5.3) até que se chegasse ao resultado desejado, ou seja, a amplificação de uma banda provável que pudesse ser clonada e seqüenciada para a validação do transcrito.

TF00035 - ID: AC003967_1

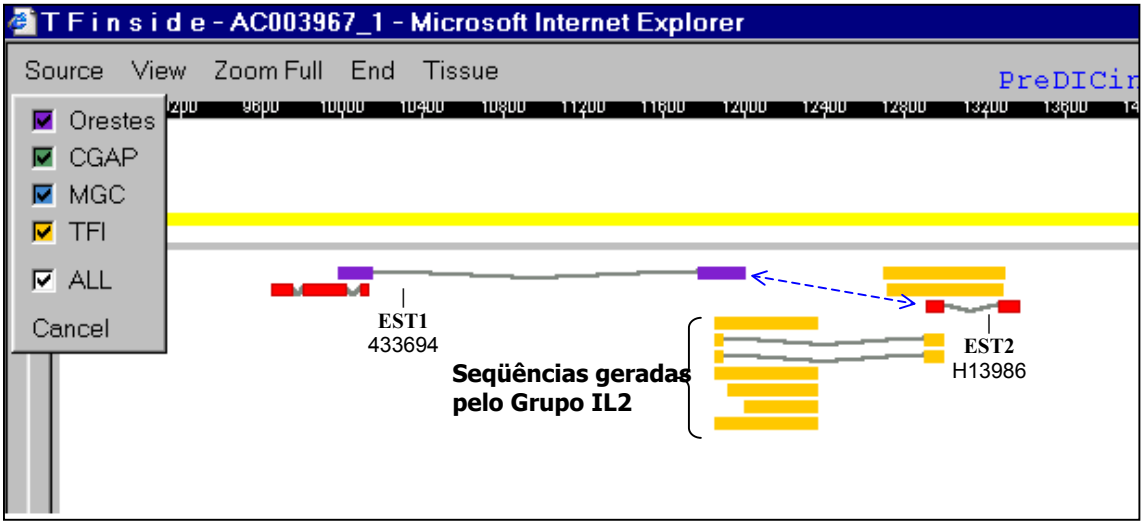


Figura 5.1 Interface Gráfica do Projeto TFI, mapa virtual do **TF00035**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado

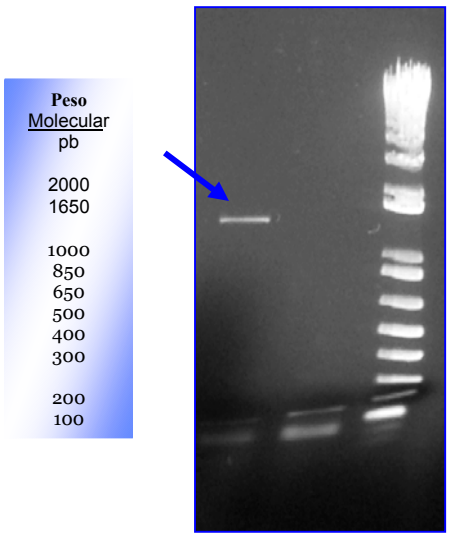


FIGURA 5.2. Foto da amplificação do TF00035/06 em gel de agarose 0,9%. Banda de Aproximadamente 1400pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00040 - ID: AC004490_1

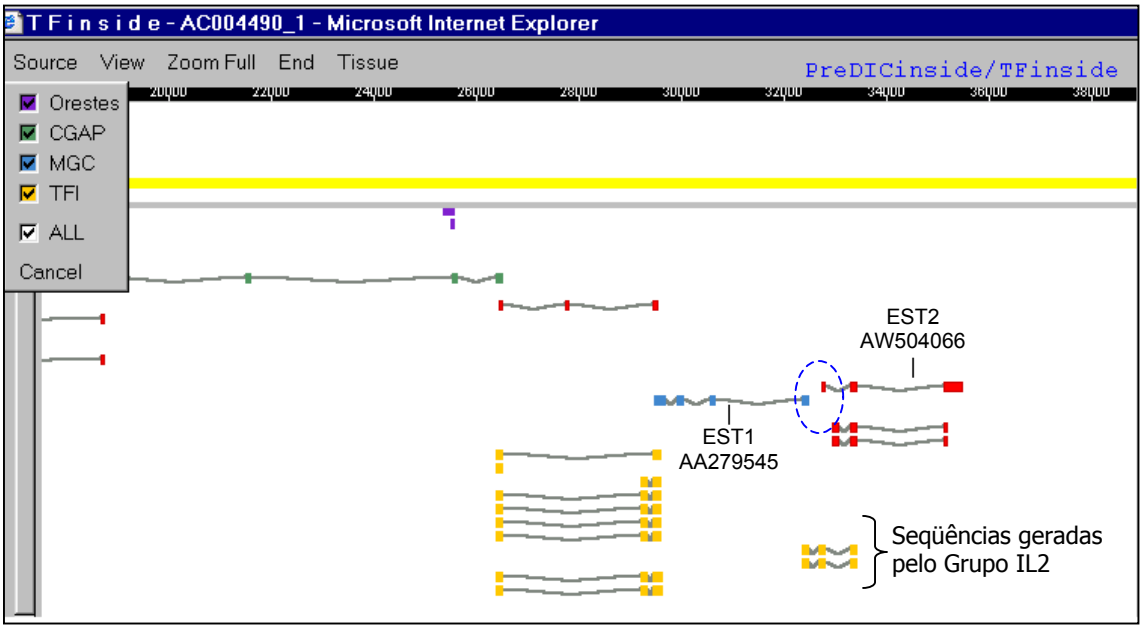


Figura 5.3 Interface Gráfica do Projeto TFI, mapa virtual do **TF00040**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras . O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado

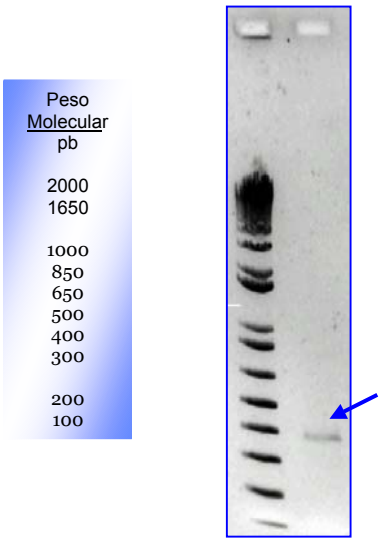


FIGURA 5.4. Foto da amplificação do TF00040/19 em gel de agarose 0,9%. Banda de aproximadamente 400pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00041 - ID: AP004490_1

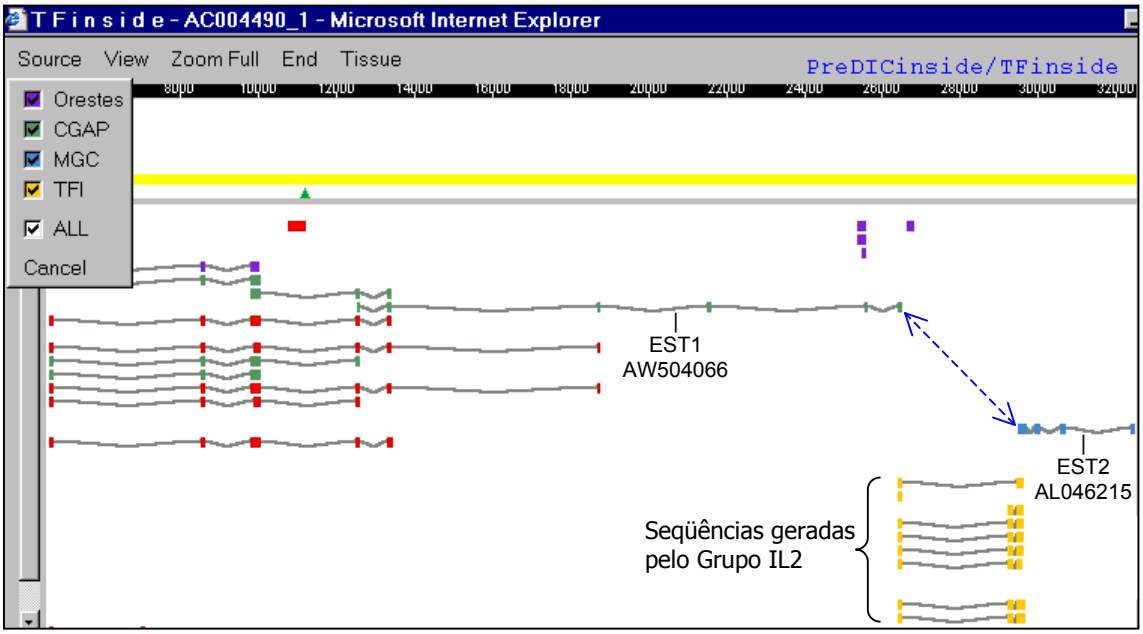


Figura 5.5 Interface Gráfica do Projeto TFI, mapa virtual do **TF00041**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

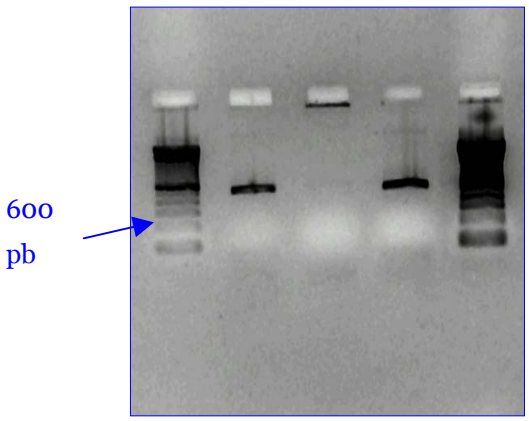


FIGURA 5.6 Foto da amplificação do TF00041 em gel de agarose 2,0%. Banda de aproximadamente 600pb. Peso molecular Ladder 100 pb Invitrogen Life *Technologies*.

.

TF00156 - ID: AC008764_1

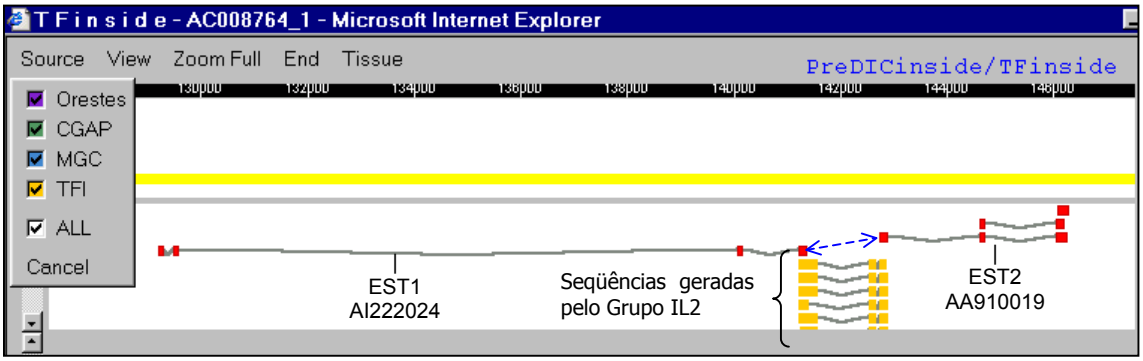


Figura 5.7 Interface Gráfica do Projeto TFI, mapa virtual do **TF00156**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

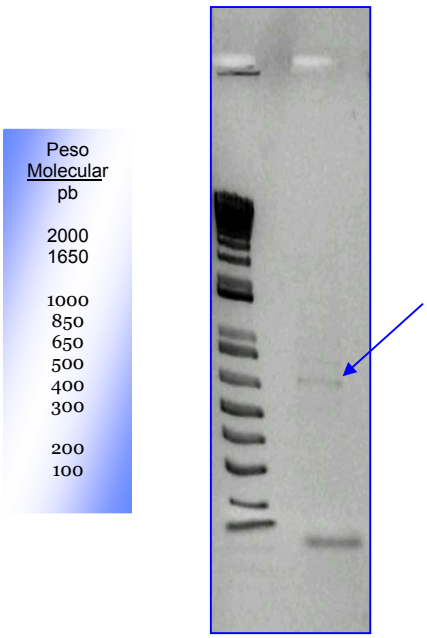


FIGURA 5.8. Foto da amplificação do TF00156_ em gel de agarose 0,9%. Banda de aproximadamente 650 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00157 - ID: AC010618_6

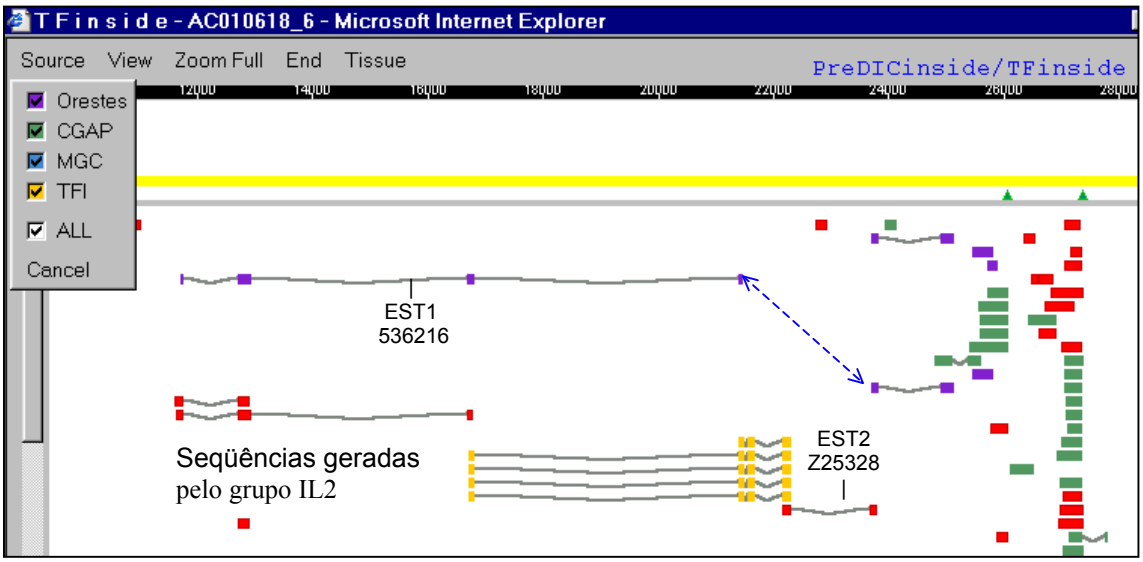


Figura 5.9 Interface Gráfica do Projeto TFI, mapa virtual do **TF00157**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

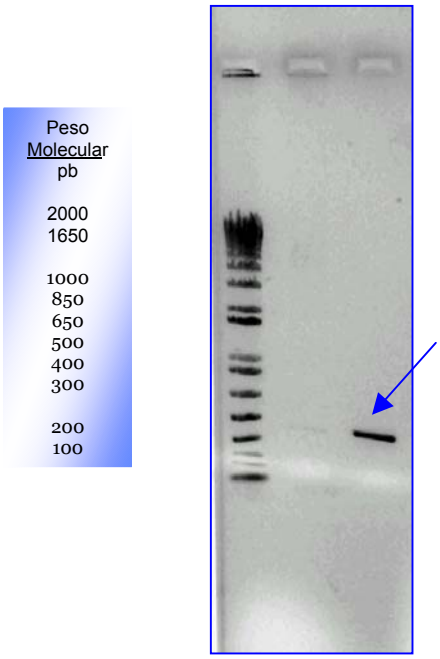


FIGURA 5.10 Foto da amplificação do TF00157/04 em gel de agarose 0,9%. Banda de aproximadamente 300 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00194 - ID:AC05382_1

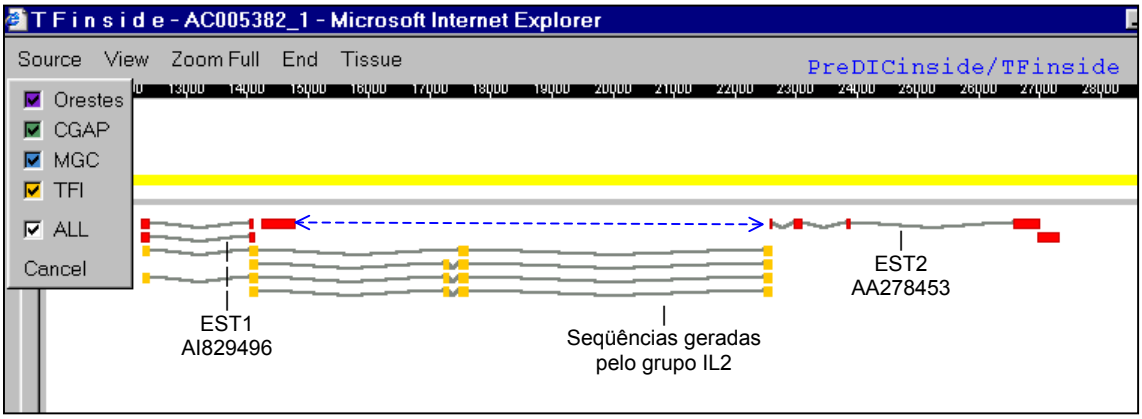


Figura 5.11 Interface Gráfica do Projeto TFI, mapa virtual do **TF00194**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

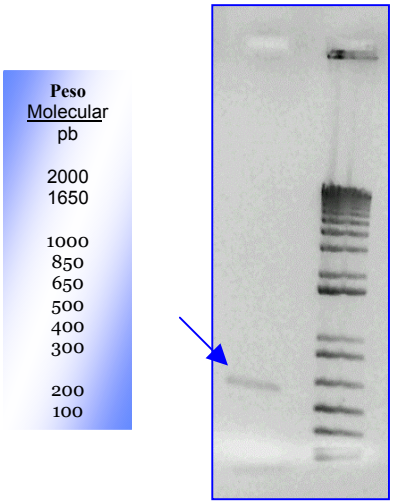


FIGURA 5.12 Foto da amplificação do TF00194/04 em gel de agarose 0,9%. Banda de aproximadamente 600 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00232 - ID:AC007773_1

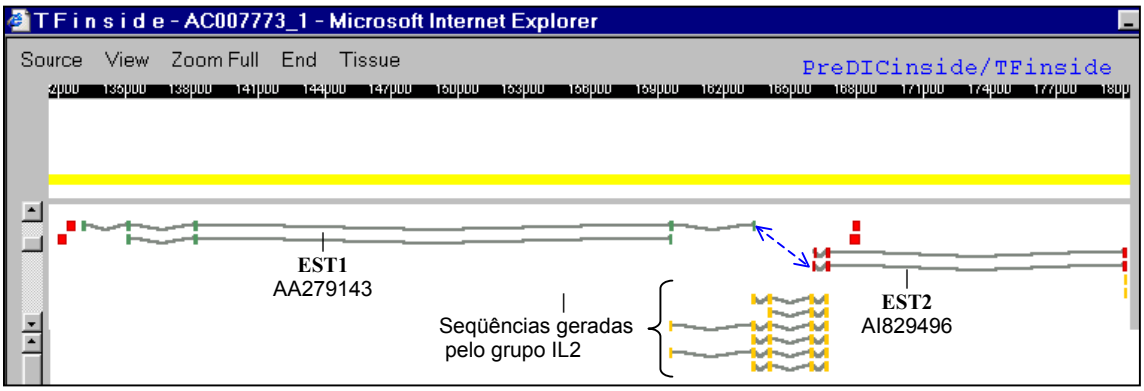


Figura 5.13 Interface Gráfica do Projeto TFI, mapa virtual do **TF00232**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

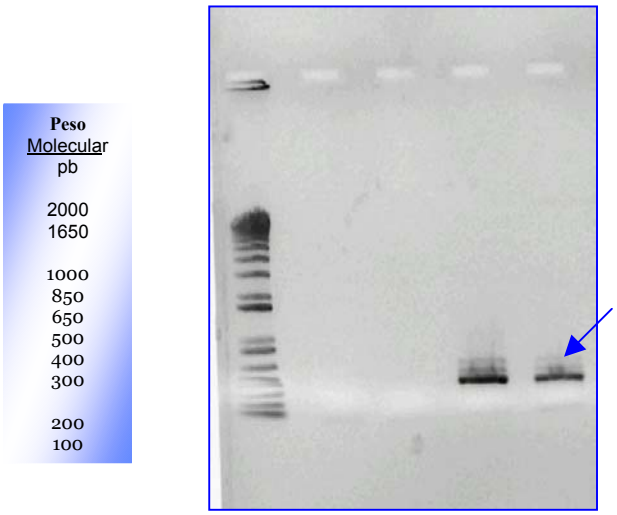


FIGURA 5.14 Foto da amplificação do TF00232/03 em gel de agarose 0,9%. Banda de aproximadamente 500 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00308 - ID:AC068139_4

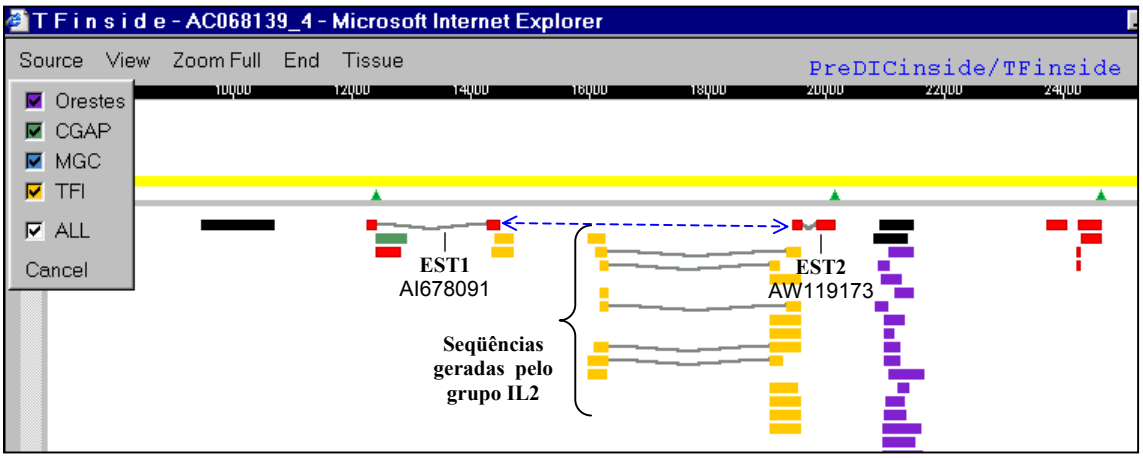


Figura 5.15 Interface Gráfica do Projeto TFI, mapa virtual do **TF00308**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

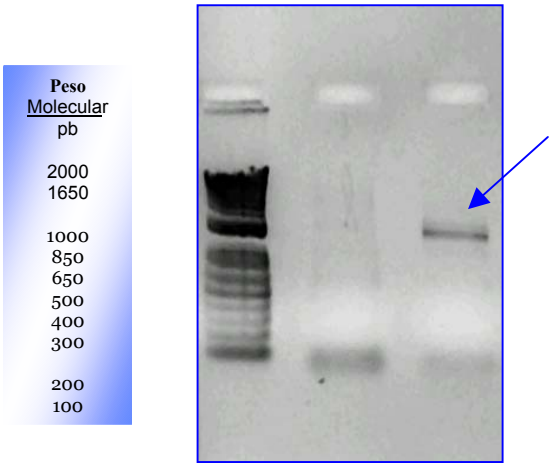


FIGURA 5.16 Foto da amplificação do TF00308/06 em gel de agarose 0,9%. Banda de com mais de 1000 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00309 - ID:AC009060_8

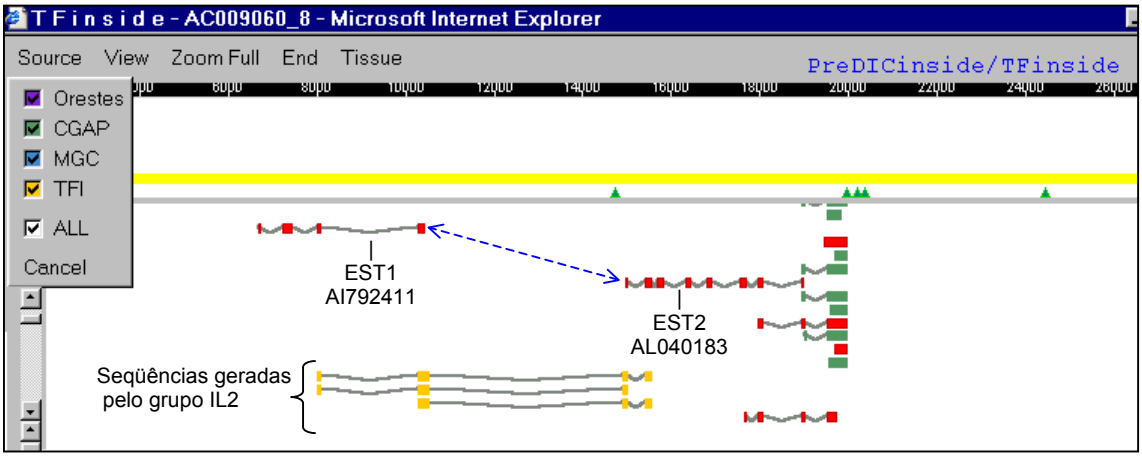


Figura 5.17 Interface Gráfica do Projeto TFI, mapa virtual do **TF00309**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

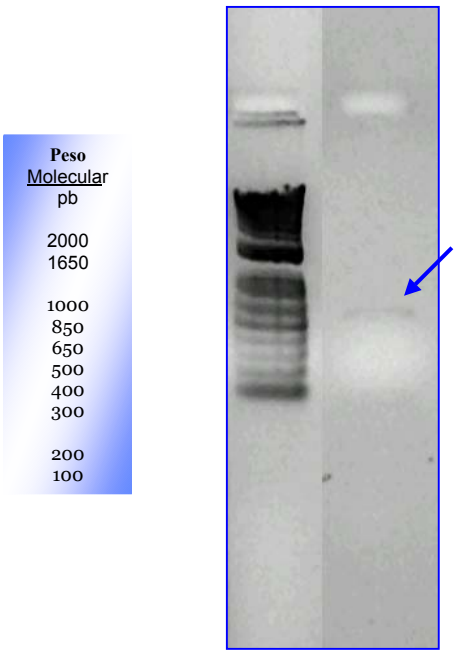


FIGURA 5.18 Foto da amplificação do TF00309/03 em gel de agarose 0,9%. Banda de aproximadamente 600 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00380 - ID:AC004827

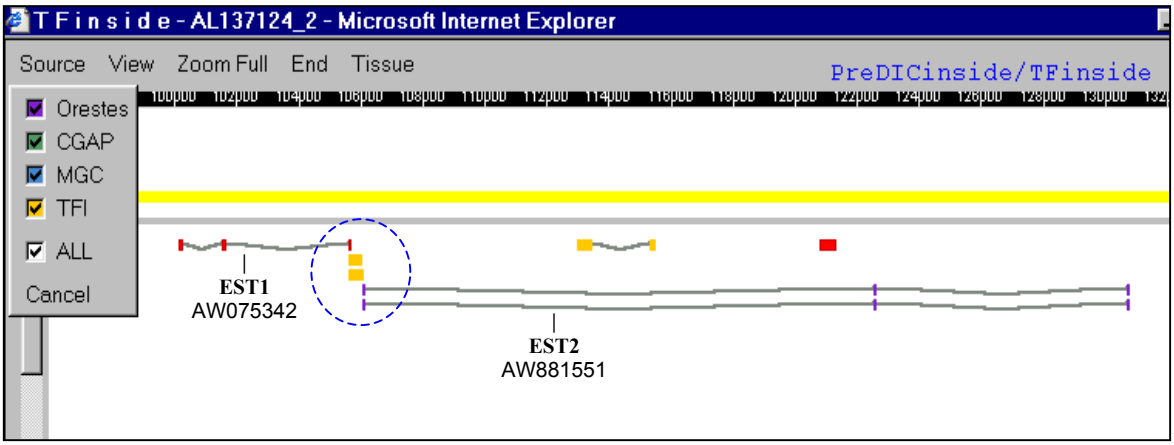


Figura 5.19 Interface Gráfica do Projeto TFI, mapa virtual do **TF00380**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

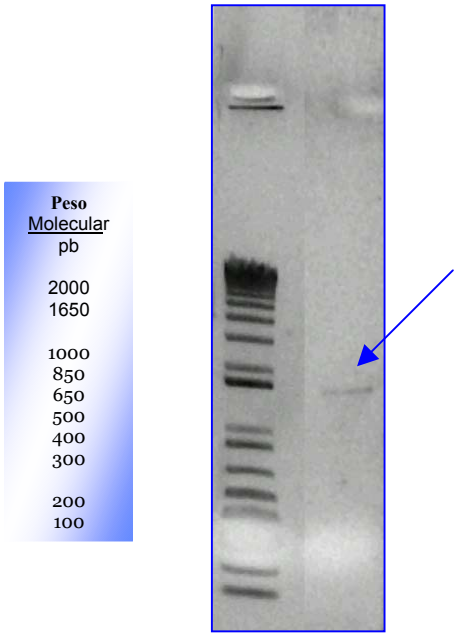


FIGURA 5.20 Foto da amplificação do TF00380/03 em gel de agarose 0,9%. Banda com mais de 1000 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

TF00404 - ID:AC004827

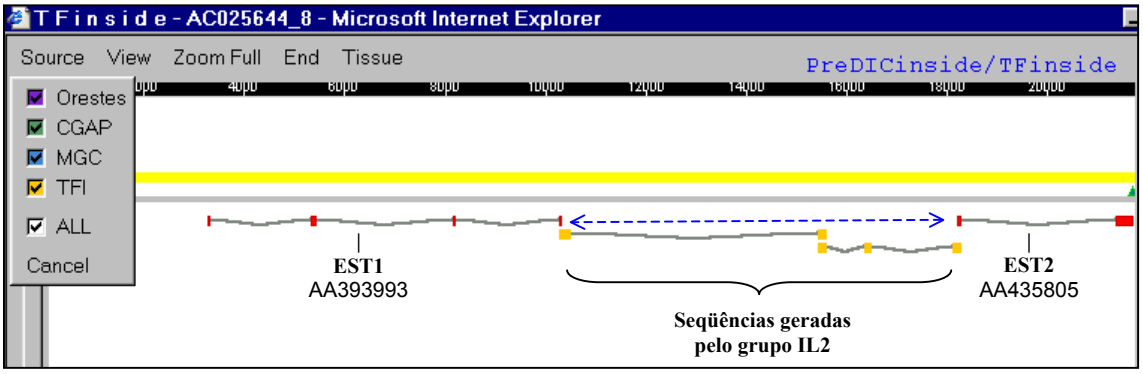


Figura 5.21 Interface Gráfica do Projeto TFI, mapa virtual do **TF00404**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. O transcrito em amarelo representa a seqüência gerada pelo grupo IL2 e indica que o gap (seta pontilhadas azul) entre dois clusters foi fechado, ou seja, o TF foi validado.

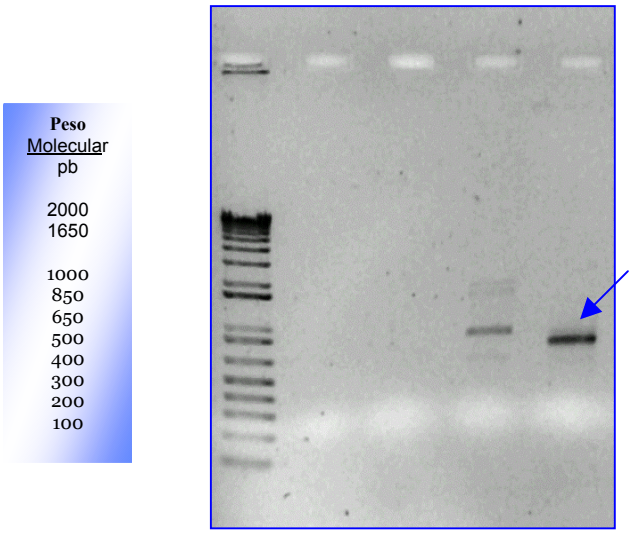


FIGURA 5.22 Foto da amplificação do TF00404/13 em gel de agarose 0,9%. Banda de aproximadamente 800 pb. Peso molecular 1 Kb *Plus Invitrogen Life Technologies*.

5.1.2 Transcript finishing (TFs) não validados

Abaixo de seu respectivo mapa virtual encontra-se a foto do gel de agarose, evidenciando a(s) banda(s) gerada(s) pela reação de PCR que não alinhou com o clone genômico fornecido pela coordenação, ou seja, não validou o transcrito.

Embora exaustivas tentativas de amplificação (tabela 5.3), clonagem e sequenciamento tenham sido realizadas para estes TFs, ou não se obteve bandas ou as bandas amplificadas não alinharam com a região esperada para o fechamento do *gap*.

TF00072 - ID:AC006942_1

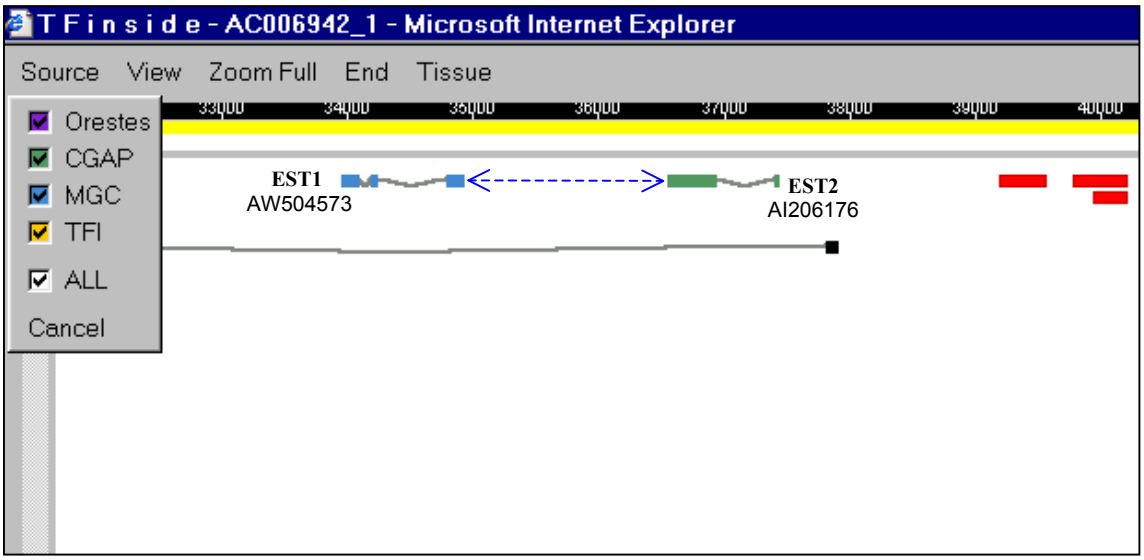


Figura 5.23 Interface Gráfica do Projeto TFI, mapa virtual do **TF00072**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os gaps que deveriam ser fechados.

Para este TF, foram realizadas 52 tentativas de amplificação e submetidas 13 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

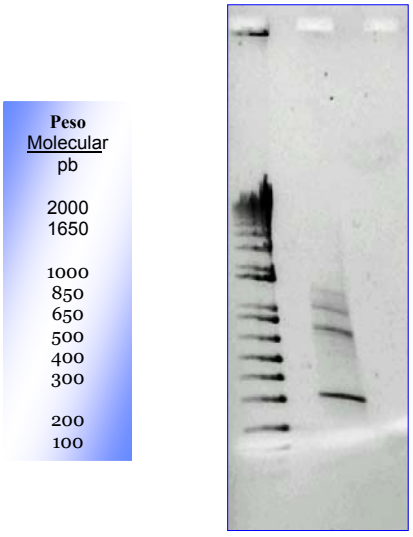


FIGURA 5.24. PCR usando cDNA de célula B para a amplificação do TF00072. Gel de agarose 0,9%. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF00074 - ID:AC005232_1

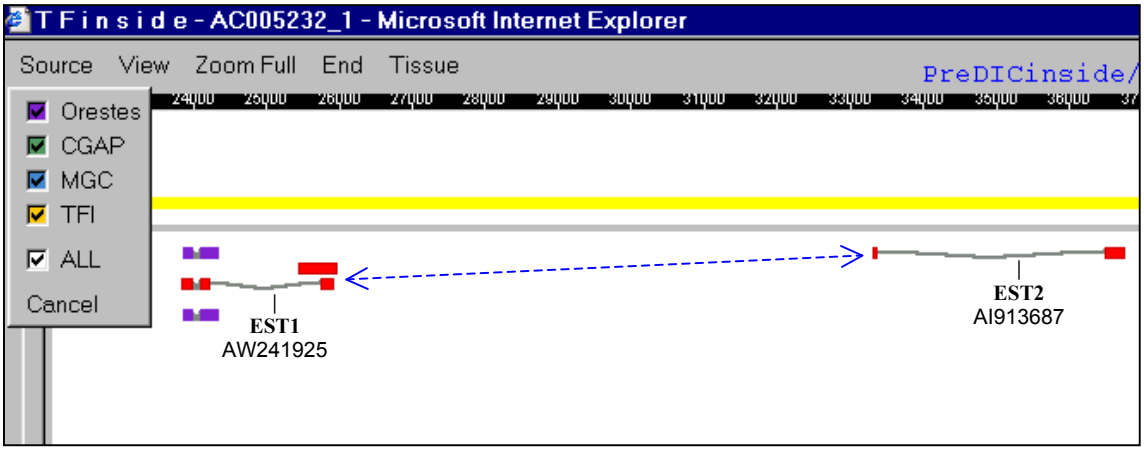


Figura 5.25 Interface Gráfica do Projeto TFI, mapa virtual do **TF00074**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os gaps que deveriam ser fechados.

Para este TF, foram realizadas 42 tentativas de amplificação e submetidas 4 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

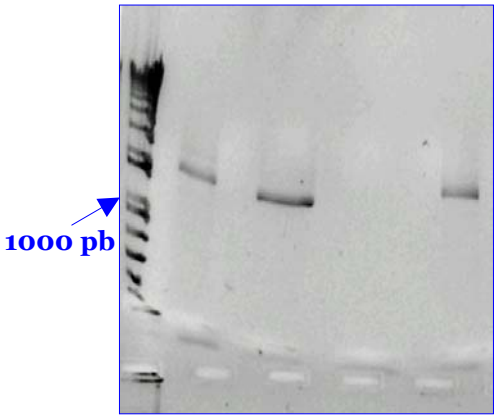


FIGURA 5.26 PCR usando cDNA de cabeça e pescoço para a amplificação do TF00074. Gel de agarose 0,9% após clonagem do inserto. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF00193 - ID:AC005023_1

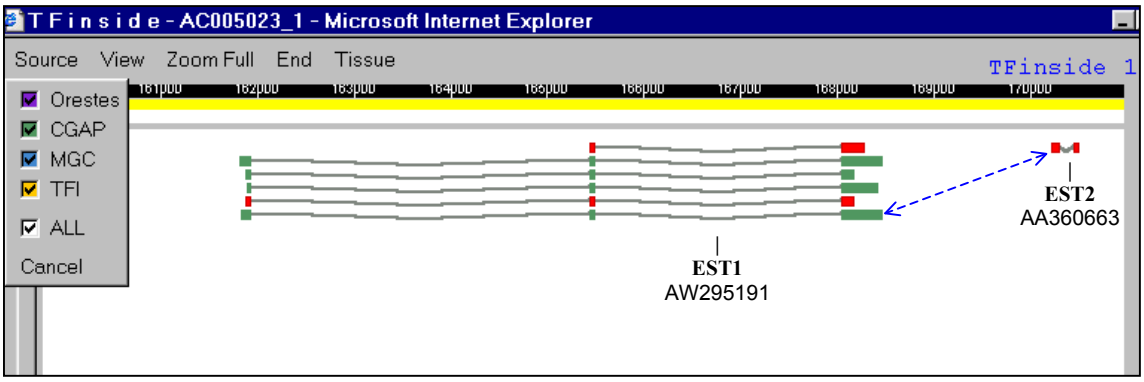


Figura 5.27 Interface Gráfica do Projeto TFI, mapa virtual do **TF00193**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 44 tentativas de amplificação e submetidas 7 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

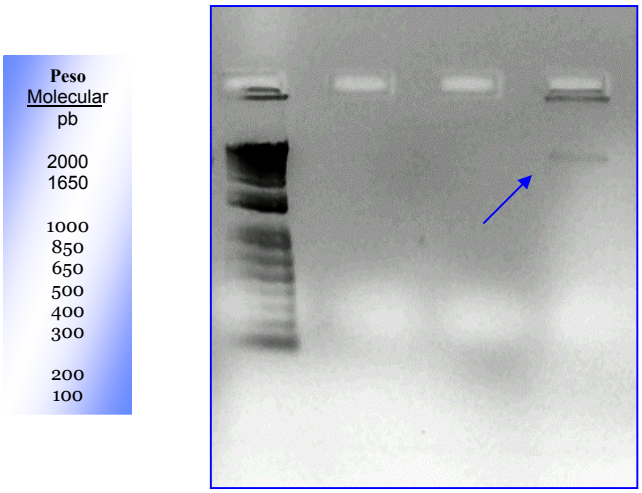


FIGURA 5.28 PCR usando cDNA de útero-poly A+ para a amplificação do TF00193. Gel de agarose 0,9%.. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF01048 - ID:AP000136_1

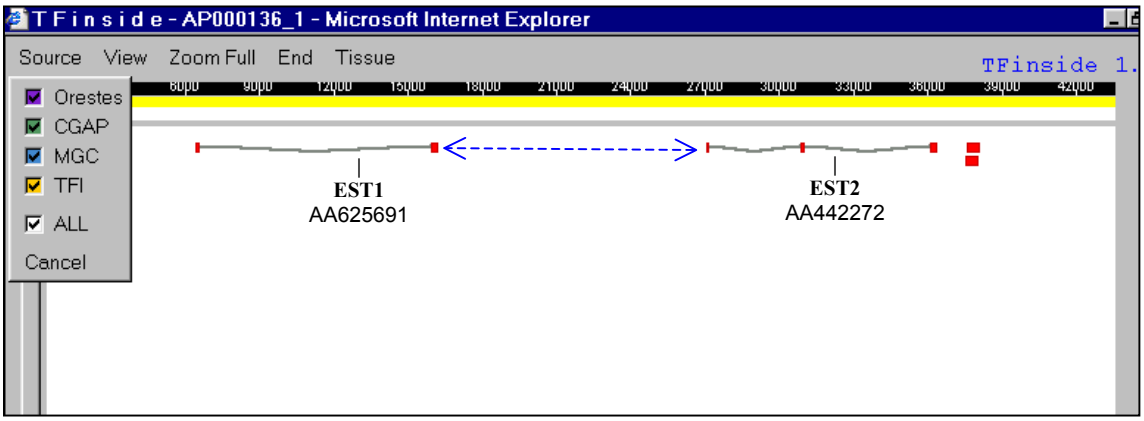


Figura 5.29 Interface Gráfica do Projeto TFI, mapa virtual do **TF01048**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 45 tentativas de amplificação e submetidas 10 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

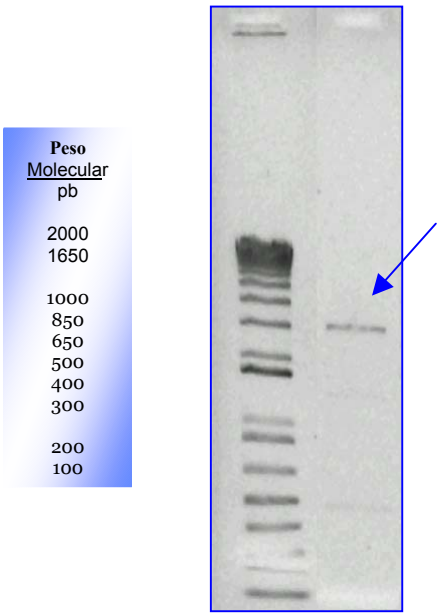


FIGURA 5.30 PCR usando cDNA de próstata para a amplificação do TF01048. Gel de agarose 0,9%.. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF01049 - ID:AI359399_1

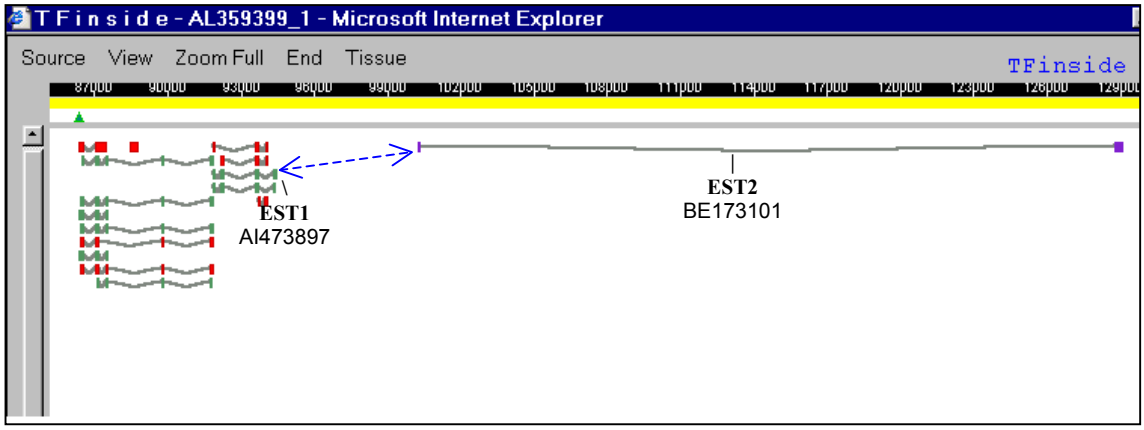


Figura 5.31 Interface Gráfica do Projeto TFI, mapa virtual do **TF01048**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 53 tentativas de amplificação e submetidas 2 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

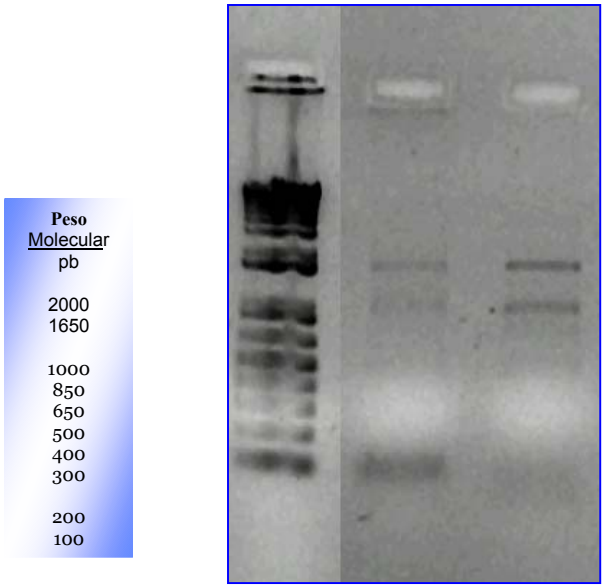


FIGURA 5.32 PCR usando cDNA de melanoma para a amplificação do TF01049. Gel de agarose 0,9%.. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF00214 - ID:AC038958_6

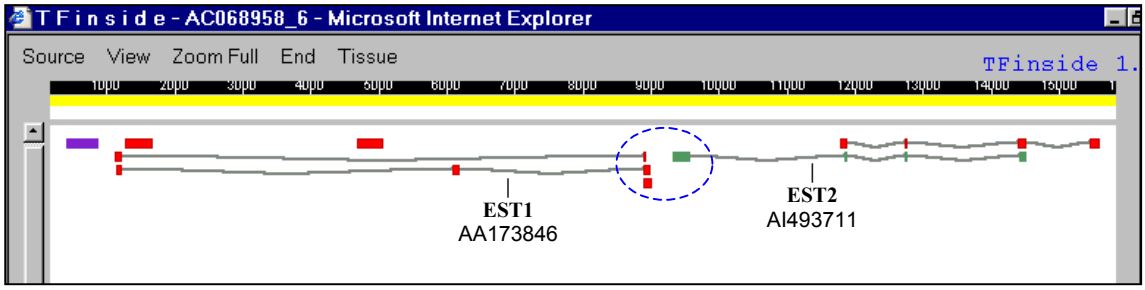


Figura 5.33 Interface Gráfica do Projeto TFI, mapa virtual do **TF00214**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 22 tentativas de amplificação e não se conseguiu amplificação de nenhuma banda.

TF00325 - ID:AI355773_1

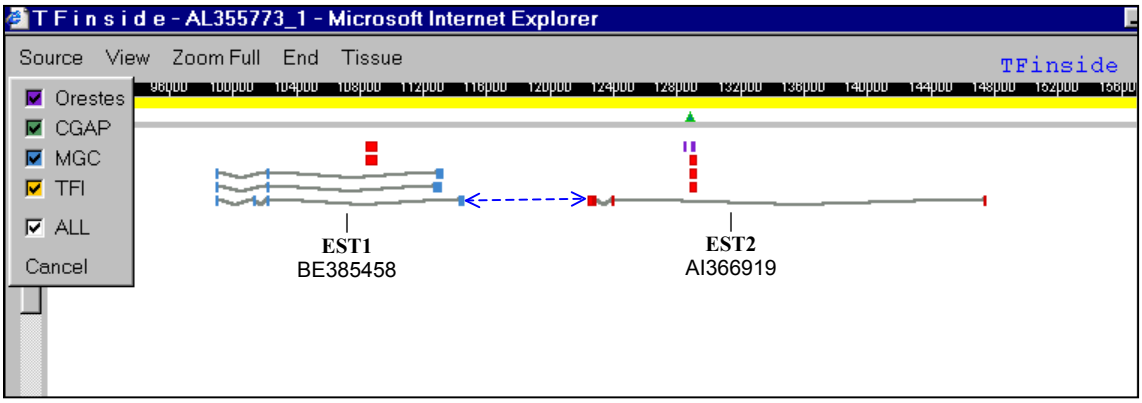


Figura 5.34 Interface Gráfica do Projeto TFI, mapa virtual do **TF00325**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 19 tentativas de amplificação e não se conseguiu amplificação de nenhuma banda.

TF00324 - ID:AC024237_14

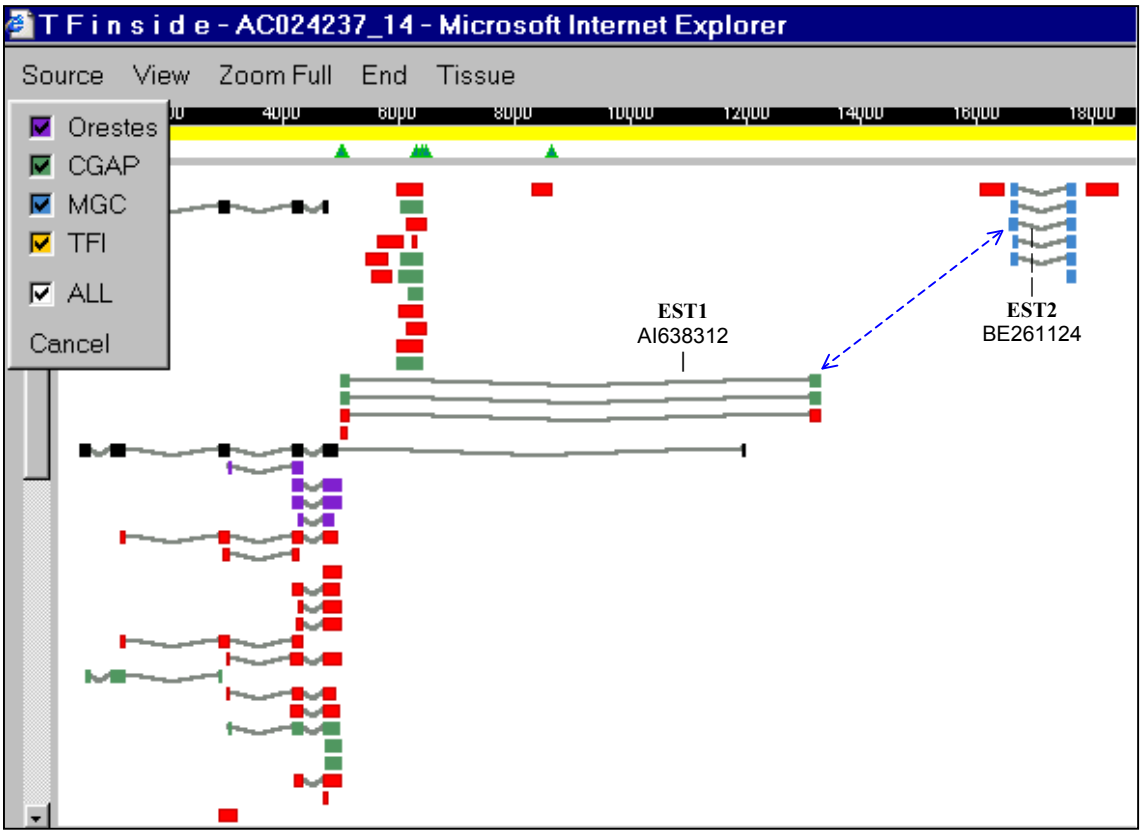


Figura 5.35 Interface Gráfica do Projeto TFI, mapa virtual do **TF00324**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 26 tentativas de amplificação e somente em uma destas reações foi positiva, uma banda maior que 12 Kb, que não foi reproduzida

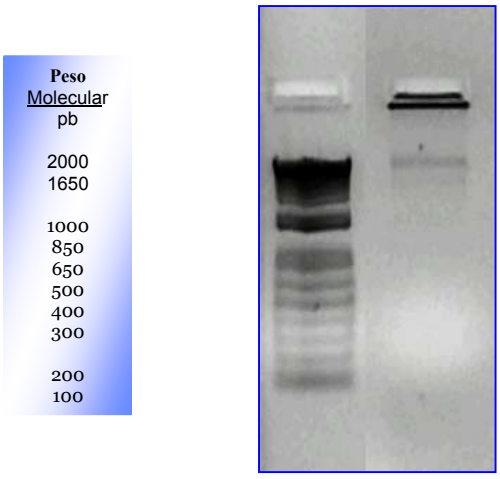


FIGURA 5.36 PCR usando cDNA de útero para a amplificação do TF00324. Gel de agarose 0,9%.. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

TF00408 - ID:Al161670_1

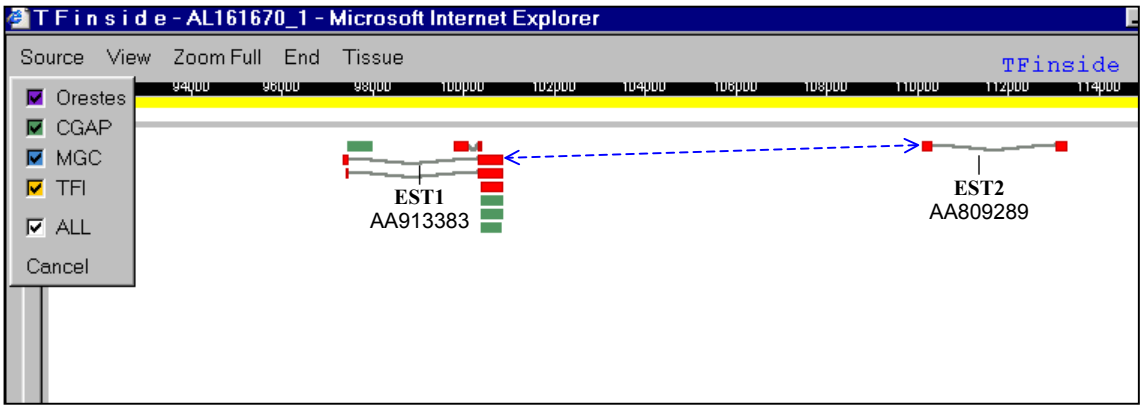


Figura 5.37 Interface Gráfica do Projeto TFI, mapa virtual do **TF00408**. A linha amarela superior representa o *draft* da seqüência do genoma humano. Os retângulos coloridos representam os *exons* e as linhas cinzas representam os *introns* nas ESTs. A cor de cada EST indica a origem de cada uma delas: ORESTES em roxo, CGAP (Cancer Genome Anatomy Project) em verde, MGC (Mammalian Gene Collection) em azul, entre outras. A seta pontilhada azul entre os dois clusters representam os *gaps* que deveriam ser fechados.

Para este TF, foram realizadas 43 tentativas de amplificação e submetidas 3 seqüências à rede virtual através da *Homepage* do Instituto LUDWIG (Projeto TFI: *Transcript Finishing Initiative*: http://www.ludwig.org.br) nenhuma delas alinharam com o *genomic_id*.

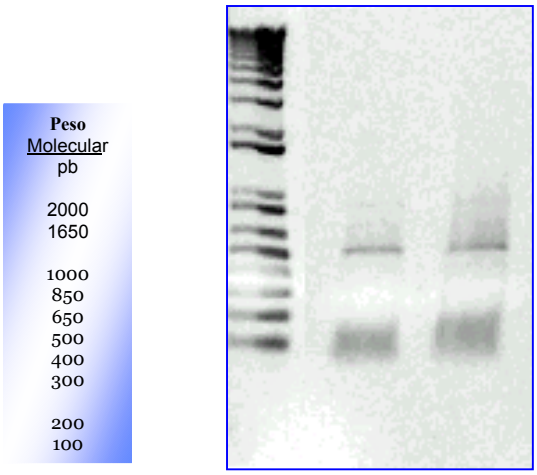


FIGURA 5.38 PCR usando cDNA de próstata para a amplificação do TF00408. Gel de agarose 0,9%.. Peso molecular 1 Kb Plus *Invitrogen Life Technologies*.

GRUPO IL2 TRANSCRIPT FINISHING	Expressão Gênica EST1/EST2	Qtde. EST/Cluster Cl 1	Cl 2	Localiz. Cromos.	Tam. Mín/Máx do *gap* (pb)	Testes realizados	Tecidos utilizados	Enzimas* utilizadas	Resultado Final	Reação Validadora Tecido	Enzima
TF00040	cólon e mama	1	3	19	315 / 600	31	cólon, mama, célula B,	Rec, eLon, Plat, Accu	validado	célula B polu A+	eLongase
TF00041	célula B, cérebro	11	4	19	1.087 / 4.081	19	cólon, mama, cérebro	Rec, eLon, Plat	validado	cérebro	eLongase
TF00035	ovário e placenta	2	1	19	-	23	útero, cérebro, rim, placenta	Rec, eLon, Plat	validado	útero	eLongase
TF00156	pool de células, pulmão	1	3	19	317 / 1.559	10	Cólon e pulmão	Rec, eLon, Plat	validado	pulmão	Platinum
TF00157	testículo e músculo da perna	5	53	19	-	8	próstata e cabeça-pescoco	Rec, Plat	validado	Próstata	Taq Recombinante
TF00380	cólon	1	3	X	-	15	útero, cólon, placenta	Rec, Plat, Accu	validado	útero polu A+	Accu Prime
TF00194	útero, célula B	3	2	19	260 / 3.125	4	célula B, útero	Rec, Plat	validado	célula B	Platinum
TF00232	amídala, célula B e útero	4	4	19	424 / 3.064	5	útero e célula B	Rec, Plat	validado	útero	Platinum
TF00308	pulmão, testículo, célula B	3	+100	10	257 / 5.217	21	pulmão, testículo, mama, linfoma	Rec, eLon, Plat, Accu	validado	testículo	Taq Recombinante
TF00309	testículo, próstata	35	26	16	348 / 5.354	3	testículo, próstata	Rec	validado	testículo	Recombinante
TF00404	testículo	1	1	-	222 / 8.117	14	testículo	Rec, Accu, Plat	validado	testículo	Accu Prime

Tabela 5.2 Dados resumidos dos resultados obtidos no processo de validação dos TFs enviados pela Coordenação do Projeto TFI ao Grupo IL2, no período de março de 2001 a Outubro de 2002.
* ENZIMAS: Rec = Taq Recombinante, eLon = Elongase, Plat = Platinum HiFi, Accu = Accu Prime (Invitrogen Technologies®)
* * Cálculo de Tamanhos Mínimo e Máximo, vide protocolo 18

GRUPO IL2 TRANSCRIPT FINISHING	Expressão Gênica EST1/EST2	Qtde. ESTs/Cluster Cluster 1	Cluster 2	Localiz. Cromos.	Tam. Mín/Máx do *gap* (pb)	Testes realizados	Tecidos utilizados	Enzimas* utilizadas	Resultado Final	Reação Validadora Tecido	Reação Validadora Enzima
TF00324	célula B, cérebro	+100	8	-	-	26	cérebro, cólon, pulmão, tec.nervoso, útero, cél.B	Rec, eLon, Plat	não validado	-	-
TF00325	útero, pele	5	6	14	248 / 8.180	19	útero, melanoma, xeroderma pig, colon, placenta, próstata	Rec, eLon, Plat, Accu	não validado	-	-
TF00214	testículo e cérebro	6	2	3	139 / 592	22	cérebro, glioblastoma, próstata, testículo	Rec, Plat, Accu	não validado	-	-
TF00408	pulmão, célula B, testículo	8	1	14	327 / 9.631	43	testículo, próstata, célula B, pulmão, linfoma	Rec, eLon, Plat, Accu	não validado	-	-
TF00072	células B, linfoma	1	1	19	302 / 1.934	52	célula B, colon, próstata, linfoma, ca	Rec, eLon, Plat, Accu	não validado	-	-
TF00074	pulmão, testículo, célula B, rim	3	1	7	234 / 7.639	42	pulmão, test., cél.B, rim, próstata, cab.e pescoço	Rec, eLon, Plat, Accu	não validado	-	-
TF00193	cólon, pulmão, célula T	6	1	X	250 / 2.019	44	cólon, pulmão, célula B, útero, cólon, linfoma	Rec, eLon, Plat	não validado	-	-
TF01048	testículo	1	1	21	309 / 11.169	45	cólon, cérebro, próstata, testículo, célula B,	Rec, eLon, Plat, Accu	não validado	-	-
TF01049	rim, cabeça e pescoço	1	1	14	334 / 34.372	53	glioblastoma, útero, cabeça e pescoço, melanoma,	Rec, eLon, Plat	não validado	-	-

Tabela 5.2 Dados resumidos dos resultados obtidos no processo de validação dos TFs enviados pela Coordenação do Projeto TFI ao Grupo IL2, no período de março de 2001 a Outubro de 2002.
* ENZIMAS: Rec = Taq Recombinante, eLon = Elongase, Plat = Platinum HiFi, Accu = Accu Prime (Invitrogen Technologies[R])
* * Cálculo de Tamanhos Mínimo e Máximo, vide protocolo 18

Durante a primeira fase do Projeto TFI, muitos protocolos experimentais e computacionais foram estabelecidos e alguns obstáculos técnicos tiveram que ser superados. Neste aspecto, a experiência prévia de vários membros do projeto TFI auxiliou na resolução precoce destes problemas. Mais uma vez, tornou-se claro a importância dos projetos realizados por consórcios de redes virtuais, onde a experiência de cada um dos membros é somada, tornando possíveis projetos como o TFI, que dificilmente seria executado por apenas um laboratório de pesquisa. É importante ressaltar que a criação de um grupo de discussão, via internet, possibilitou a troca de experiências entre os participantes do projeto, auxiliando principalmente, os laboratórios que estavam iniciando na área de estudos genômicos.

Até o final do mês de Outubro de 2002, 597 TFs foram selecionados pela coordenação e entregues aos 31 grupos validadores. Destes, 210 (35%) foram validados.

Os protocolos estabelecidos para amplificação, clonagem e seqüenciamento parecem estar adequados e a eficiência de amplificação dos TFs está em uma faixa prevista, com base na eficiência de amplificação por RT-PCR de 27% obtida no trabalho de DAS *et al*. (2001), voltado para a validação de novos transcritos localizados no cromossomo 22.

O grupo IL2, validou 11 dos 20 TFs enviados pela coordenação do projeto, alcançando a eficiência de 55% de validação e, até esta fase do projeto, está entre os 9 grupos validadores que apresentam eficiência na validação entre 41 e 60% (Gráfico 5.2).

Quanto à eficiência na amplificação por PCR, observou-se que não houve nenhuma diferença entre as enzimas utilizadas para a amplificação dos fragmentos, já que os 4 tipos utilizados validaram cada uma, em média, 3 das 11 TFs. Também não foi possível correlacionar a validação destes TFs ao tecido utilizado para extração do RNA, pois aquele que validou um determinado TF em certa ocasião, não funcionou para o outro TF que também apresentava expressão para o mesmo tecido.

Embora a distância entre a EST1 e EST2 seja um fator influente na eficiência da validação, já que uma maior distância entre os “clusters” diminui a probabilidade dos mesmos pertencerem a um mesmo transcrito e também diminui a eficiência de amplificação da RT-PCR. Isto porque em um

maior intervalo genômico espera-se um maior número de exons presentes. Também, não foi possível correlacionar o tamanho do “gap” com validação, pois os cálculos de seu tamanho (protocolo 18) fornecem um intervalo extenso entre os valores máximo e mínimo prováveis.

Questiona-se a existência de uma possível correlação entre o número de ESTs que representam cada um dos dois clusters de cada TF e a eficiência de validação. Se assim fosse, clusters com mais ESTs apontariam para transcritos mais expressos. Nossos dados não apresentaram resultados consistentes sobre essa correlação.

A não validação de TFs pode ser atribuída aos limites de amplificação da técnica, descrita por DAS *et al*. (2001) pois, somente os transcritos com grandes chances de validação foram escolhidos pela coordenação do projeto.

A execução de projetos como o TFI deve ser realizada em um ritmo acelerado e acompanhada de uma constante atualização dos bancos de dados locais utilizados para a seleção dos TFs. Trata-se de uma área de interesse mundial e a cada momento, novas seqüências são geradas por outros grupos de pesquisa. Verificou-se que, entre a escolha de alguns TFs e a sua validação, o transcrito foi validado por outro projeto, perdendo seu caráter de inediticidade.

Fez-se imprescindível a integração entre as ferramentas da bioinformática já disponíveis e a validação laboratorial para a identificação de todos os genes do genoma humano, bem como o desenvolvimento de novas ferramentas computacionais e implementação das técnicas laboratoriais.

Deve-se salientar, que a amostra utilizada para análise da eficiência de validação é pequena, tratando-se apenas de uma pequena porção (20) dos 597 TFs distribuídos aos grupos validadores. As correlações com o total dos TFs distribuídos, serão realizadas após o término do projeto e certamente serão mais esclarecedoras por tratar-se de um número maior de transcritos.

5.2 Análise preliminar *in silico*

5.2.1 Transcript Finishing (TFs) validados

A análise preliminar *in silico* foi realizada de acordo com os protocolos 16 e 17, apresentados no subitem métodos de Material e Métodos.

As figuras abaixo, representam o gráfico resultante do alinhamento virtual do consenso montado pela coordenação do Projeto TFI das EST1, EST2 e seqüência validadora com o *Draft* do genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros; através da ferramenta BLAT (disponível em http://genome.ucsc.edu). As tabelas apresentam a análise preliminar "in sílico", resumida, das informações disponíveis nos bancos de dados para as seqüências. Somente foram analisados os TFs validados que tiveram o consenso disponibilizado pela coordenação. De todos os TFs validados pelo grupo IL2, somente o consenso do TF00035 ainda não havia sido disponibilizado, devido a problemas de montagem. Para dados resumidos de todos os TFs, vide tabela 5.12, ao final deste item.

TF00040 e TF00041

Genes anotados para oTF00040 e TF00041	
 *H. sapiens*	Consensos **EST1 + IL2-TF00040-G10F + EST2** (1141 bases) **EST1 + IL2-TF00041-H04F + EST2** (1146 bases)
Cromossomo/citogenética	19p13.3
Gene anotados - símbolo	DOT1L
Número de Exons	24
Produto Gênico	*histone methyltransferase DOT1L*
RefSeq.	NM_032482 (Status: Provisório)
LocusLink	84444
Homologia	*Drosophila melanogaster*
Observações	Gene conhecido, sem splicing alternativo. Consultar figura 5.39, observar que estes dois TFs fazem parte de um mesmo gene.

Tabela 5.3 - Dados resumidos da anotação preliminar dos consensos do TF00040 e TF00041. Ferramentas utilizadas: BLAT e/ou BLAST.

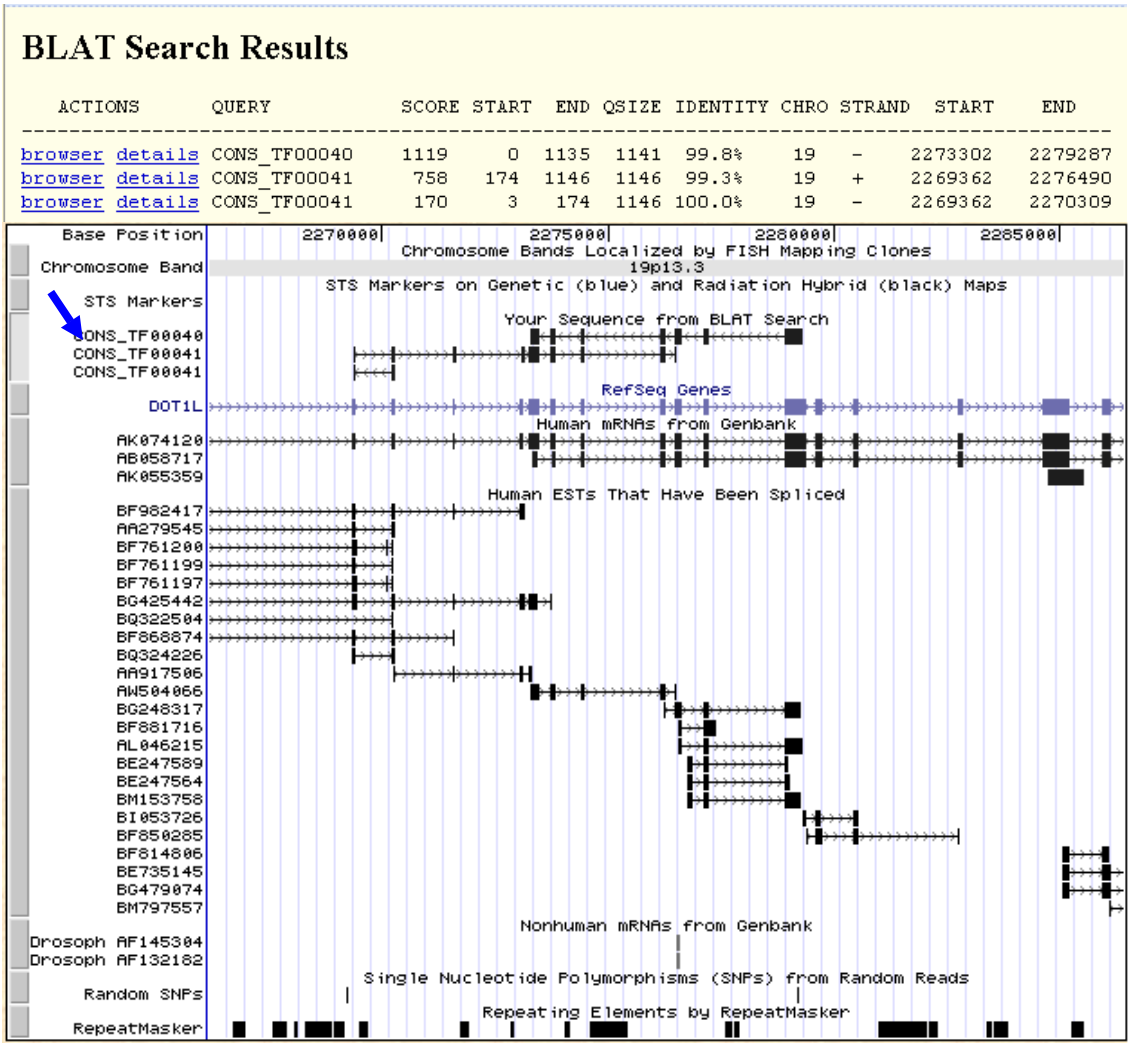


Figura 5.39 - Alinhamento dos consensos do TF00040 e TF00041 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta BLAT.

TF00156

Genes anotados para o TF00156	
H. sapiens	Consenso **EST1 + IL2-TF00156-H04R + EST2** (1181 bases)
Cromossomo/citogenética	19p13.12
Gene anotados - símbolo	CRT2
Número de Exons	9 (Exon Skipping)
Produto Gênico	*hypothetical protein MGC26577*
RefSeq.	NM_145046 (Status: Predito)
LocusLink	125972
Homologia	*Rana rugosa, Dirofilaria immitis*
Observações	Gene conhecido, possui splicing alternativo (*exon skipping*) já representado no banco de dados de ESTs (Figura 5.40).

Tabela 5.4 - Dados resumidos da anotação preliminar do consenso do TF00156. Ferramentas utilizadas: BLAT e/ou BLAST.

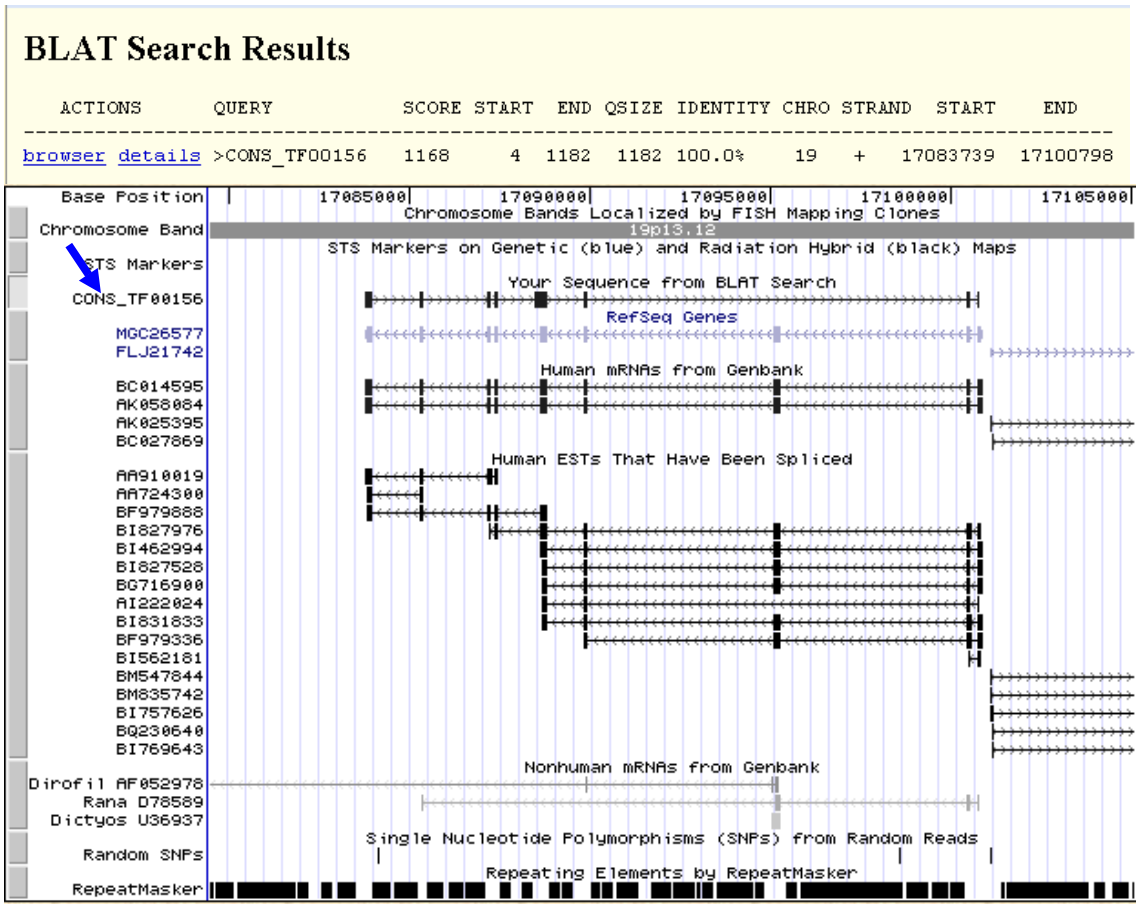


Figura 5.40 - Alinhamento do consenso do TF00156 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta BLAT

TF00157

Genes anotados para o TF00157	
 *H. sapiens*	Consensos **EST1 + IL2-TF00157-A03F + EST2** (888 bases) 
Cromossomo/citogenética	19p13.12
Gene anotados - símbolo	FLJ22329
Número de Exons	5
Produto Gênico	*hypothetical protein FLJ22329*
RefSeq.	NM_024656 (Status: Predito)
LocusLink	125972
Homologia	*não*
Observações	Transcrito conhecido no banco de ESTs, porém, alinha de forma estranha ao gene acima descrito e ao banco de dados dos mRNAs (Figura 5.41).

Tabela 5.5 - Dados resumidos da anotação preliminar do consenso do TF00157. Ferramentas utilizadas: BLAT e/ou BLAST.

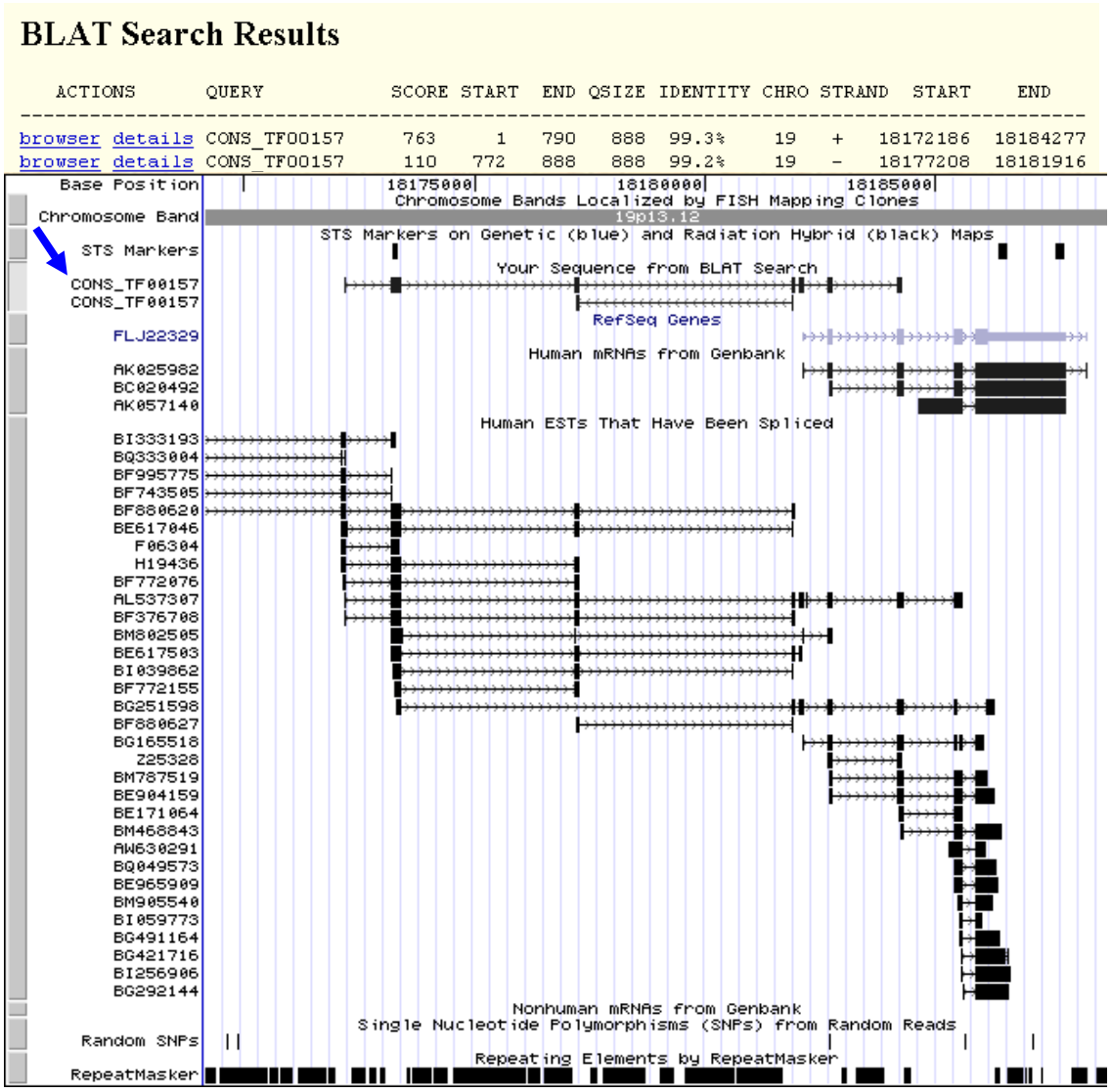


Figura 5.41 - Alinhamento do consenso do TF00157 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta BLAT

TF00194

Genes anotados para o TF00194	
H. sapiens	Consensos **EST1 + IL2-TF00194-G05R + EST2** (1410 bases)
Cromossomo/citogenética	19p13.11
Gene anotados - símbolo	_____
Número de Exons	_____
Produto Gênico	_____
RefSeq.	_____
LocusLink	_____
Homologia	*Mus musculus* (BLASTx)

OBSERVAÇÕES:

Pesquisa pelas ferramentas:

BLAT (Figura 5.42)
sem similaridade com genes conhecidos

BLASTn (Figura 5.43)
Similaridade com:

>gi|21734362|emb|AL833713.1|HSM805026 *Homo sapiens* mRNA; cDNA DKFZp667O169 (from clone DKFZp6670169); Extensão = 5994.

SCORE = 1806 bits (911); e-value = 0.0; Identidade = 925/932 (99%).

BLASTx: (Figura 5.44)
Similaridade com:

>gi|23617915|ref|XP_133389.2| Similar to *C. elegans* DPY-19 protein (corresponding sequence F22B7.10) [*Caenorhabditis elegans*] [*mus musculus*] extensão = 453.

SCORE = 427 bits (1099); e-value = e-118; Identidade = 215/276 (77%), Positivos = 234/276 (84%); Frame = +3.

Tabela 5.6 - Dados resumidos da anotação preliminar do consenso do TF00194. Ferramentas utilizadas: BLAT e/ou BLAST.

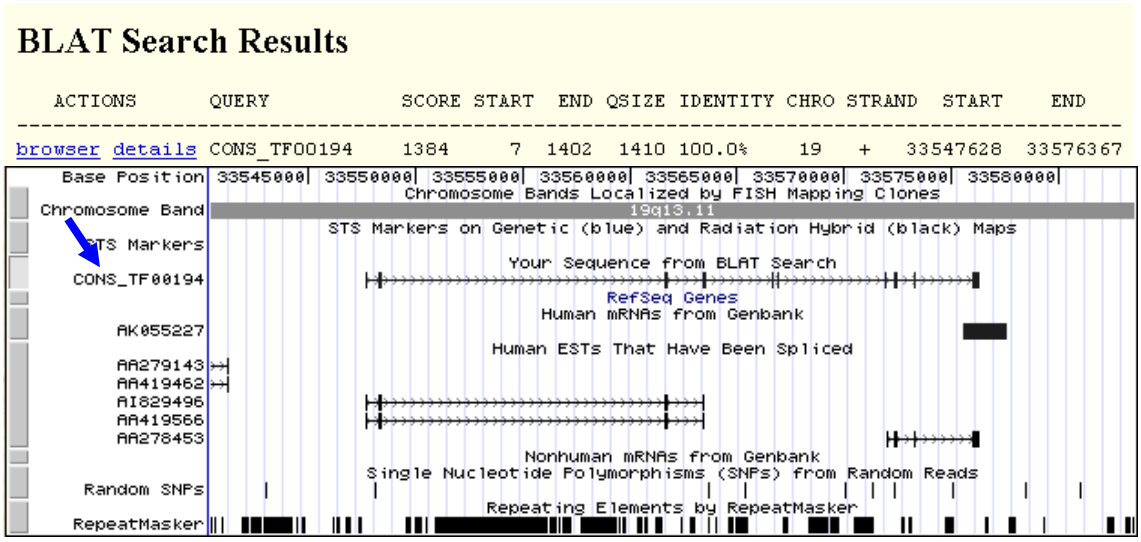
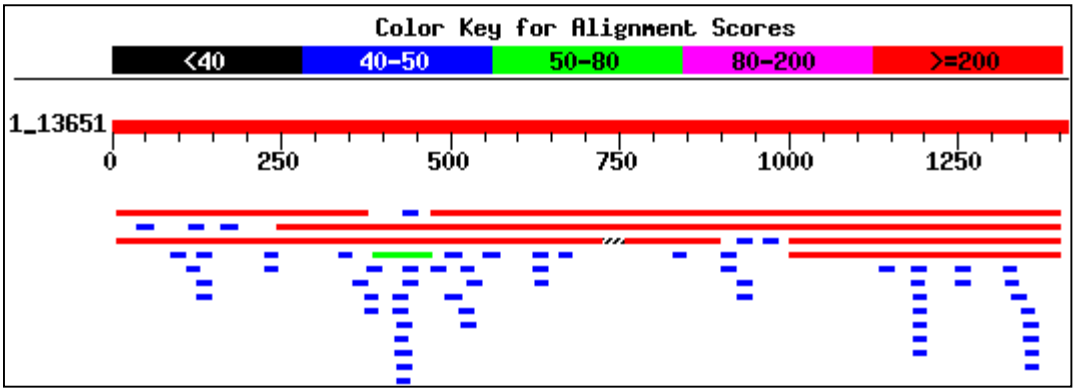


Figura 5.42 - Alinhamento do consenso do TF00194 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

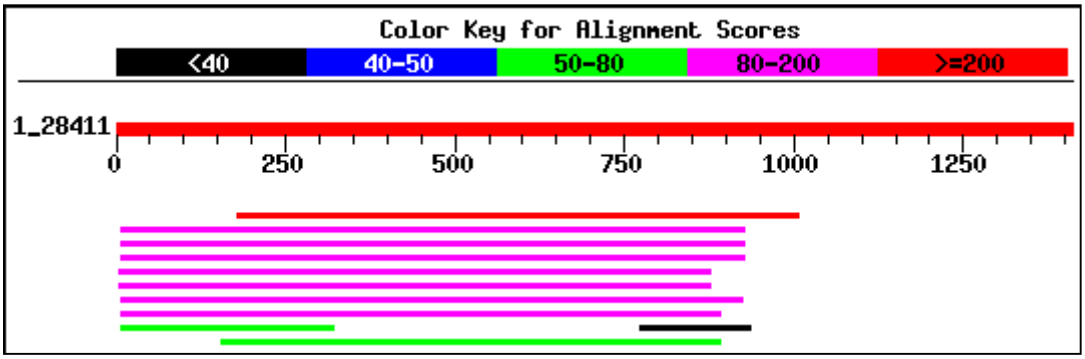
Nenhum alinhamento foi encontrado para o consenso do TF00194 pela ferramenta BLAT, com genes ou mRNAs humanos dos bancos de dados disponíveis. Apresenta alinhamentos parciais com ESTs. Então, partiu-se para a análise utilizando outra ferramenta de bioinformática, a ferramenta BLAST. As figuras abaixo mostram os alinhamentos mais importantes encontrados nos bancos de dados de nucleotídeos (BLASTn) e aminoácidos (BLASTx).



Sequences producing significant alignments: (bits) Value

gi|21734362|emb|AL833713.1|HSM805026	Homo sapiens mRNA; cDN...	1806	0.0	U
gi|18590534|ref|XM_085993.1|	Homo sapiens LOC147991 (LOC147...	751	0.0	L
gi|3386587|gb|AC005382.1|AC005382	Homo sapiens chromosome 1...	751	0.0	
gi|16549910|dbj|AK055227.1|	Homo sapiens cDNA FLJ30665 fis,...	743	0.0	U
gi|23617914|ref|XM_133389.2|	Mus musculus similar to C. ele...	634	e-179	L

Figura 5.43 - Alinhamento do consenso do TF00194 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTn**



Sequences producing significant alignments:		Score (bits)	E Value	
gi|23617915|ref|XP_133389.2|	similar to C. elegans DPY-19 p...	427	e-118	L
gi|25051287|ref|XP_146665.3|	similar to KIAA0877 protein [H...	98	2e-19	
gi|20541809|ref|XP_029942.3|	similar to F22B7.10.p [Homo sa...	96	2e-18	L
gi|4240243|dbj|BAA74900.1|	KIAA0877 protein [Homo sapiens]	96	2e-18	L
gi|630612|pir||S44629	F22B7.10 protein - Caenorhabditis ele...	94	5e-18	
gi|17552842|ref|NP_498909.1|	F22B7.10.p [Caenorhabditis ele...	94	5e-18	L

Figura 5.44 - Alinhamento do consenso do TF00194 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTx**

TF00232

Genes anotados para oTF00232	
H. sapiens	Consensos **EST1 + IL2-TF00232-G07R + EST2** (1196 bases)
Cromossomo/citogenética	19p13.11
Gene anotados - símbolo	_____
Número de Exons	_____
Produto Gênico	_____
RefSeq.	_____
LocusLink	_____
Homologia	_____

OBSERVAÇÕES:

Pesquisa pelas ferramentas:

BLAT (Figura 5.45)
sem similaridade com genes conhecidos

BLASTn (Figura 5.46)

similaridade com:

>gi|23397480|ref|NM_153220.1| *Homo sapiens hypotetical protein* MGC35440, mRNA; Extensão = 2298.
SCORE = **533** bits (269), E-value = e-148; Identidade = 276-277 (99%), Gaps = 1/277 (0%).

BLASTx (Figura 5.47)

>gi|23617915|ref|XP_133389.2| similar to *C. elegans DPY-19 protein (corresponding sequence* F22B7.10) [*Caenohabditis elegans*] [*Mus musculus*]; Extensão = 453.
SCORE = **144** bits (364), e-value = 2e-33, Identidade = 74/87 (85%), Positivos = 79/87 (90%); Frame = +2.

Tabela 5.7 - Dados resumidos da anotação preliminar do consenso do TF00232. Ferramentas utilizadas: BLAT e/ou BLAST.

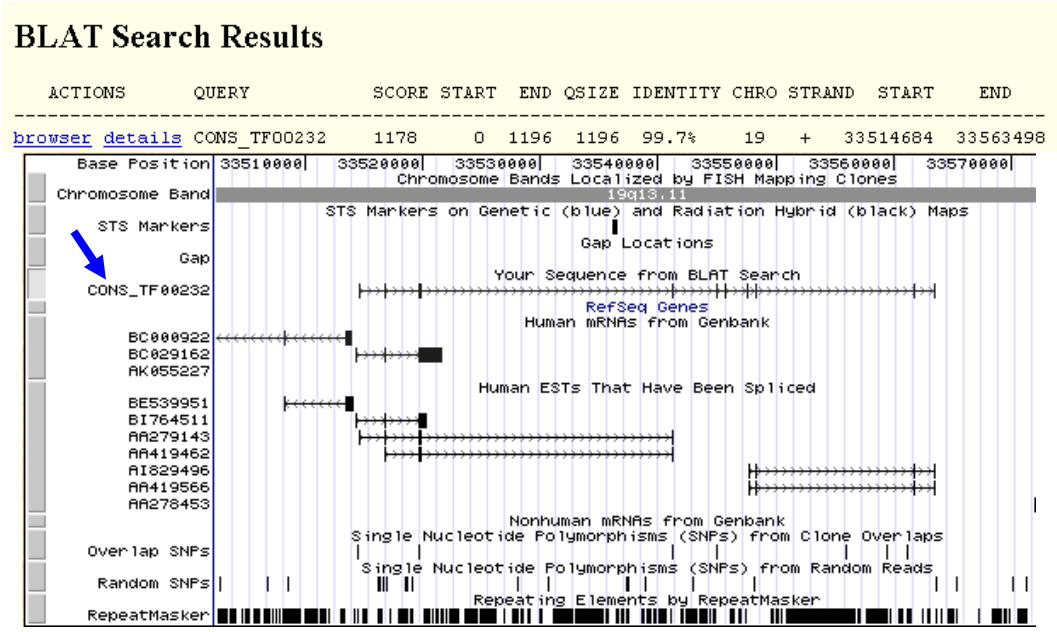
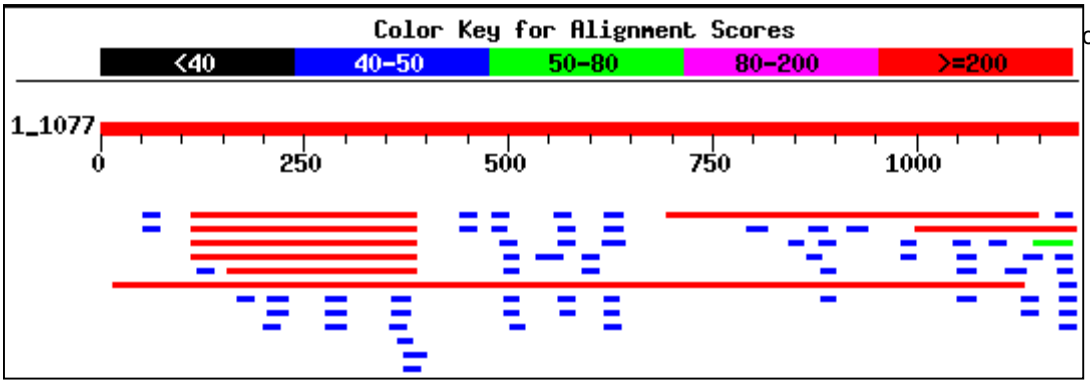


Figura 5.45 - Alinhamento do consenso do TF00232 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

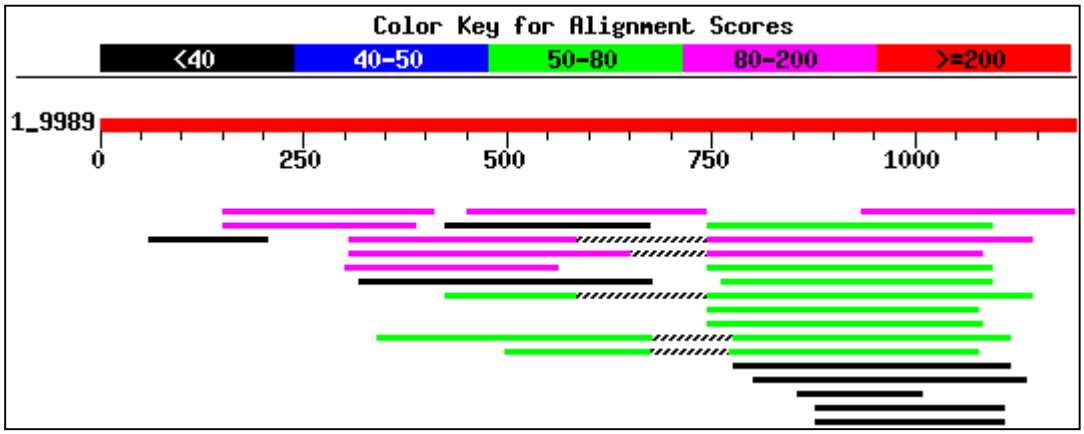
Nenhum alinhamento significante com genes ou mRNAs nos bancos de dados disponíveis foi encontrado para o consenso do TF00232 pela ferramenta BLAT, apenas alinhamentos com ESTs. Então, partiu-se para a análise utilizando outra ferramenta de bioinformática, a ferramenta BLAST. As figuras abaixo mostram os alinhamentos significantes encontrados nos bancos de dados de nucleotídeos (BLASTn) e aminoácidos (BLASTx).

os de



Sequences producing significant alignments:		Score (bits)	E Value	
gi|23397480|ref|NM_153220.1|	Homo sapiens hypothetical prot...	533	e-148	L
gi|22052063|ref|XM_097358.2|	Homo sapiens LOC147990 (LOC147...	533	e-148	L
gi|20809804|gb|BC029162.1|	Homo sapiens, LOC147990, clone M...	533	e-148	L U
gi|16304934|emb|AL390123.14|	Human DNA sequence from clone ...	470	e-129	
gi|23617914|ref|XM_133389.2|	Mus musculus similar to C. ele...	448	e-123	L
gi|22051643|ref|XM_171371.1|	Homo sapiens LOC147990 (LOC255...	406	e-110	L
gi|5032322|gb|AC007773.1|AC007773	Homo sapiens chromosome 1...	299	5e-78	

Figura 5.46 - Alinhamento do consenso do TF00232 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTn**



Sequences producing significant alignments:		Score (bits)	E Value	
gi|23617915|ref|XP_133389.2|	similar to C. elegans DPY-19 p...	144	2e-33	L
gi|18590533|ref|XP_097358.1|	hypothetical protein XP_097358...	109	7e-23	L
gi|22051644|ref|XP_171371.1|	hypothetical protein XP_171371...	103	4e-21	L
gi|17552842|ref|NP_498909.1|	DumPY : shorter than wild-type...	90	5e-17	L

Figura 5.47 - Alinhamento do consenso do TF00232 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTx**

TF00308

Genes anotados para o TF00308	
H. sapiens	Consensos **EST1 + IL2-TF00308-D10R + EST2** (1310 bases)
Cromossomo/citogenética	10q22.3
Gene anotados - símbolo	_____
Número de Exons	_____
Produto Gênico	_____
RefSeq.	_____
LocusLink	_____
Homologia	_____

OBSERVAÇÕES:

Pesquisa pelas ferramentas:

BLAT (Figura 5.48)
sem similaridade com genes conhecidos

BLASTn (Figura 5.49)

similaridade com:

>gi|16073659|emb|AL359195.24| Human DNA sequence from clone RP11-36D19 on chromosome 10, complete sequence; Extensão = 171148.

SCORE = **1168** bits (589), E-value = 0.0; Identidade = 589/589 (100%)

BLASTx
A seqüência pesquisada não apresentou similaridade significativa, apresentando *scores* muito baixos (menores que 100).

Tabela 5.8 - Dados resumidos da anotação preliminar do consenso do TF00308. Ferramentas utilizadas: BLAT e/ou BLAST.

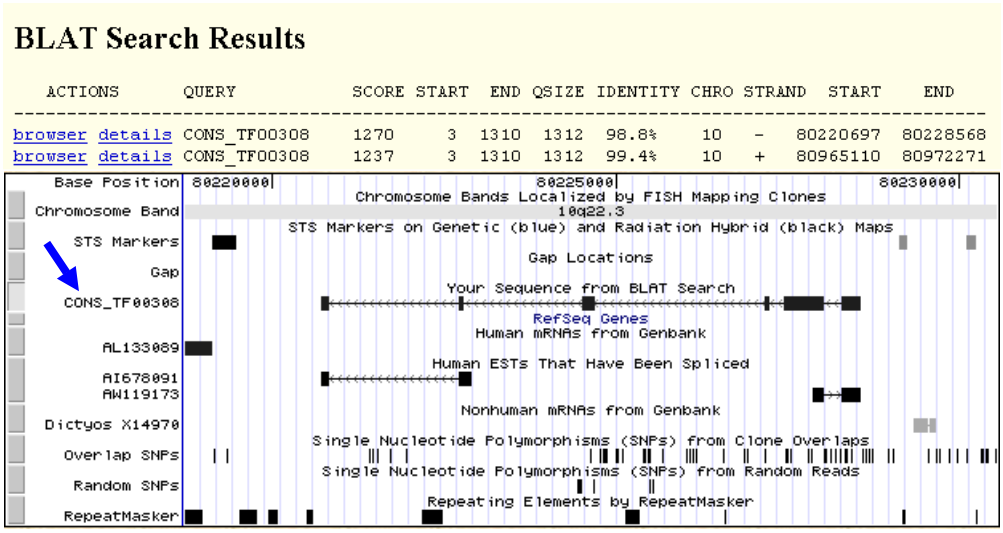
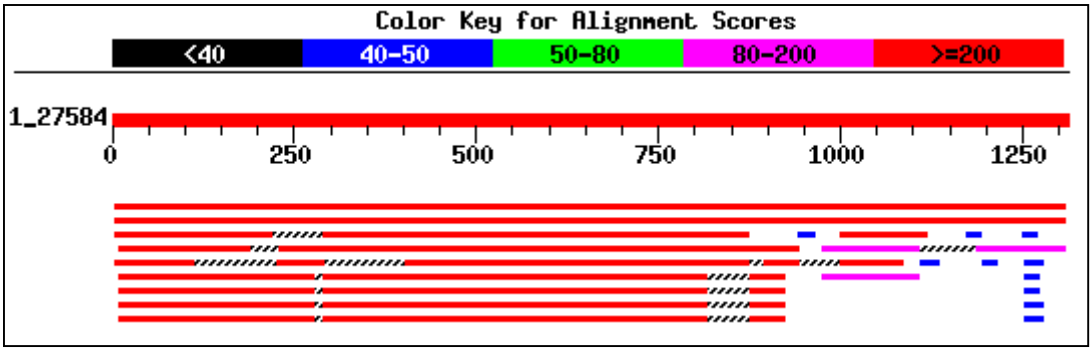


Figura 5.48 - Alinhamento do consenso do TF00308 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

Nenhum alinhamento significante com genes ou mRNAs nos bancos de dados disponíveis foi encontrado para o consenso do TF00308 pela ferramenta BLAT. Então, partiu-se para a análise utilizando outra ferramenta de bioinformática, a ferramenta BLAST, somente 2 ESTs das regiões nas extremidades. As figuras abaixo mostram os alinhamentos significantes encontrados nos bancos de dados de nucleotídeos (BLASTn) e aminoácidos (BLASTx).



Sequences producing significant alignments:		Score (bits)	E Value
gi|16073659|emb|AL359195.24|	Human DNA sequence from clone ...	1168	0.0
gi|20334534|gb|AC068139.6|	Homo sapiens chromosome 10 clone...	1128	0.0
gi|13443261|gb|AC073270.6|	Homo sapiens BAC clone RP11-468B...	428	e-117
gi|20514788|gb|AC092562.4|	Papio hamadryas BAC RP41-285I13 ...	385	e-103

Figura 5.49 - Alinhamento do consenso do TF00308 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTn**

TF00309

Genes anotados para o TF00309	
 *H. sapiens*	Consensos **EST1 + IL2-TF00309-F01R + EST2** (1366 bases) 
Cromossomo/citogenética	16q22.1
Gene anotado - símbolo	_____
Número de Exons	_____
Produto Gênico	_____
RefSeq.	_____
LocusLink	_____
Homologia	_____

OBSERVAÇÕES:

Pesquisa pelas ferramentas:

BLAT (Figura 5.50a e 5.50b)
sem similaridade com genes conhecidos

BLASTn (Figura 5.51)

similaridade com:

>gi|22070270|ref|XM_170811.1| *Homo sapiens* similar to RIKEN cDNA 1200009H11 (LOC255565mRNA); extensão: 2372.

SCORE: **2418** BITS (1220), e-value = 0.0, identidade = 1331/1355 (98%), gaps = 10/1355 (0%).

BLASTx (Figura 5.52)

>gi|22760438|dbj|BAC11199.1| unnamed protein product (Homo sapiens); extensão: 446.

SCORE: **411** bits (1056), e-value = e-172, identidade = 193/217 (88%), positives = 195/217 (89%), Frame = +1

Tabela 5.9 - Dados resumidos da anotação preliminar do consenso do TF00309. Ferramentas utilizadas: BLAT e/ou BLAST.

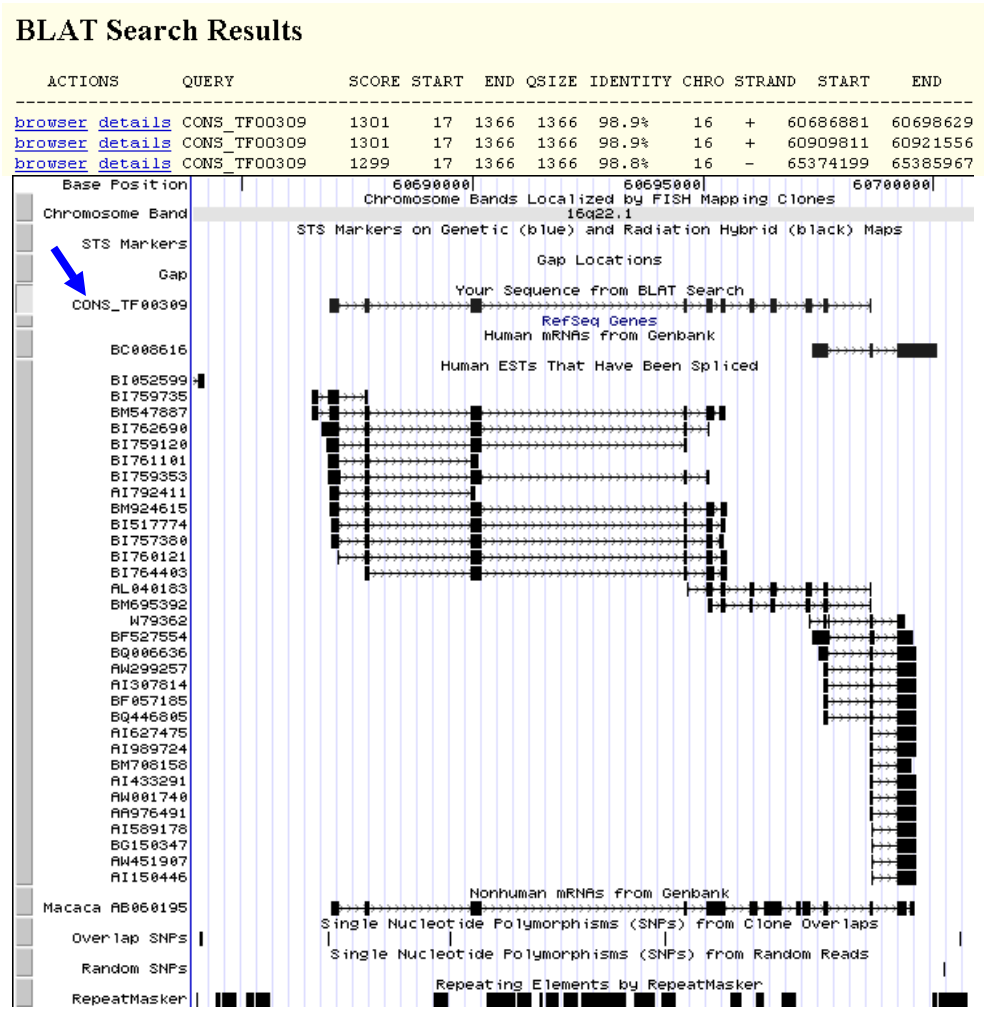


Figura 5.50a - Alinhamento do consenso do TF00309 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

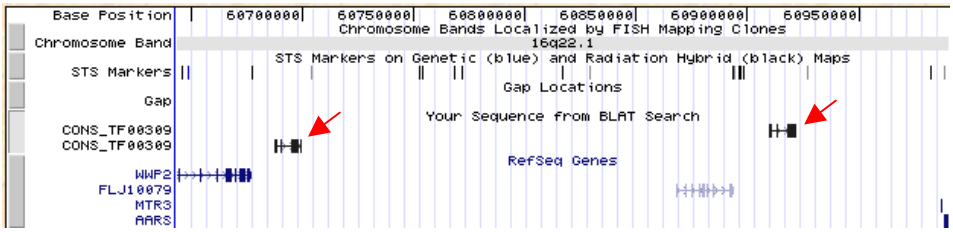
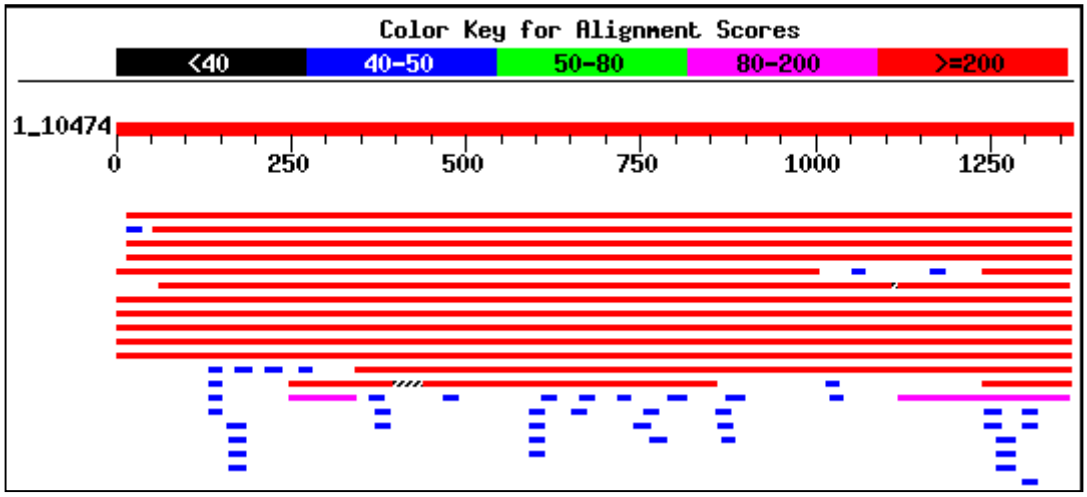


Figura 5.50b - Detalhe do alinhamento anterior mostrando que a seqüência pesquisada *match* em dois locais distintos (setas vermelhas) no *Draft* do Genoma Humano.

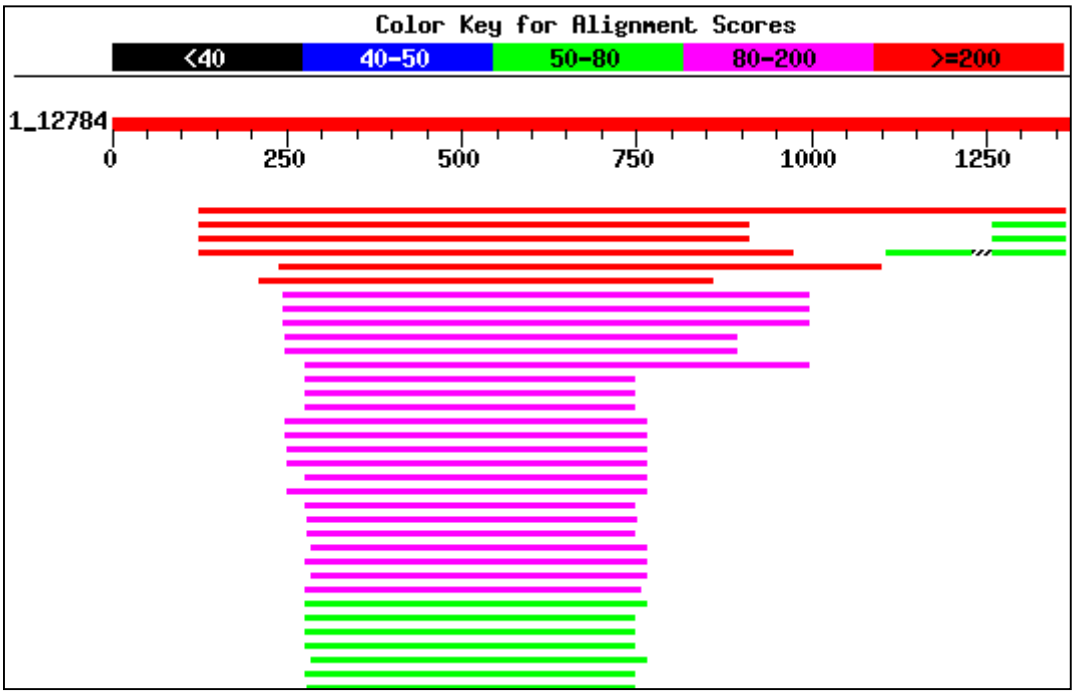
Nenhum alinhamento significante foi encontrado com genes no RefSeq para o consenso do TF00309 pela ferramenta BLAT, porém, esse transcrito é bem representado no banco de dados ESTs. Então, partiu-se para a análise utilizando outra ferramenta de bioinformática, a ferramenta BLAST. As figuras abaixo mostram os alinhamentos significantes encontrados nos bancos de dados de nucleotídeos (BLASTn) e aminoácidos (BLASTx).



Sequences producing significant alignments: (bits) Value

gi|22070270|ref|XM_170811.1| Homo sapiens similar to RIKEN ...	2418	0.0	L
gi|22760437|dbj|AK074773.1| Homo sapiens cDNA FLJ90292 fis,...	2399	0.0	
gi|24659534|gb|BC039068.1| Homo sapiens, clone MGC:34761 IM...	2014	0.0	
gi|21733974|emb|AL833339.1|HSM804652 Homo sapiens mRNA; cDN...	1871	0.0	U
gi|22067502|ref|XM_058787.4| Homo sapiens similar to RIKEN ...	1867	0.0	L
gi|13676426|dbj|AB060195.1| Macaca fascicularis brain cDNA ...	1213	0.0	
gi|23306005|gb|AC097265.4| Pan troglodytes BAC RP43-119N13...	470	e-129	
gi|18997244|gb|AC009153.10| Homo sapiens chromosome 16 clon...	462	e-127	

Figura 5.51 - Alinhamento do consenso do TF00309 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTn**



Sequences producing significant alignments:	Score (bits)	E Value
gi|22760438|dbj|BAC11199.1| unnamed protein product [Homo s...	411	e-172
gi|24659535|qb|AAH39068.1| Unknown (protein for MGC:34761) ...	342	e-133
gi|22070271|ref|XP_170811.1| similar to RIKEN cDNA 1200009H...	405	e-131
gi|22067503|ref|XP_058787.3| similar to RIKEN cDNA 1200009H...	405	e-131
gi|13676427|dbj|BAB41141.1| hypothetical protein [Macaca fa...	394	e-108
gi|20888321|ref|XP_146517.1| similar to hypothetical protei...	249	1e-73
gi|20888317|ref|XP_146516.1| similar to mannose receptor pr...	71	7e-20
gi|22761577|dbj|BAC11640.1| unnamed protein product [Homo s...	94	6e-18
gi|20555401|ref|XP_166431.1| similar to RIKEN cDNA 1200009H...	92	2e-17
gi|18490353|qb|AAH22399.1| Unknown (protein for IMAGE:46911...	92	2e-17
gi|9558454|dbj|BAB03398.1| cysteine-rich protease inhibitor...	89	1e-16

Figura 5.52 - Alinhamento do consenso do TF00309 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLASTx**

TF00380

Genes anotados para o TF00380	
 *H. sapiens*	Consensos **EST1 + IL2-TF00380-A04R + EST2** (501 bases) 
Cromossomo/citogenética	Xq23
Gene anotado - símbolo	AMOT
Número de Exons	12
Produto Gênico	*angiomotin*
RefSeq.	NM_133265 (Provisório)
LocusLink	154796
Homologia	*Mus musculus*
Observações	Gene conhecido, sem splicing alternativo. Figura 5.53.

Tabela 5.10 - Dados resumidos da anotação preliminar do consenso do TF00380. Ferramentas utizadas: BLAT e/ou BLAST

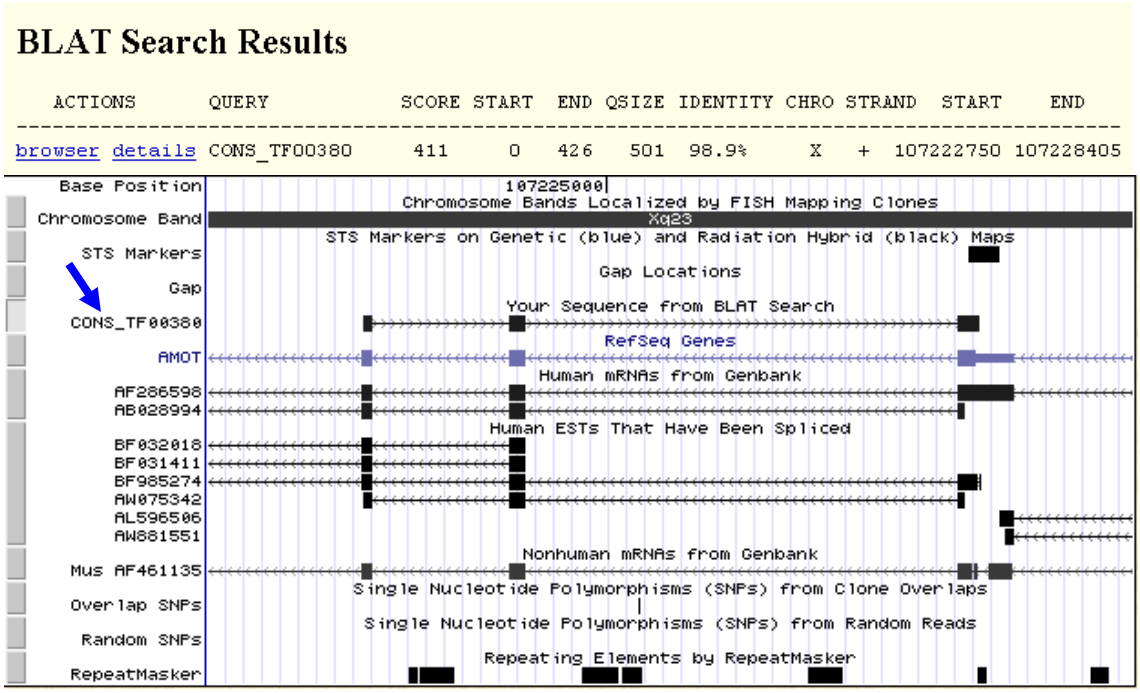


Figura 5.53 - Alinhamento do consenso do TF00380 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

TF00404

Genes anotados para o TF00404	
H. sapiens	Consensos **EST1 + IL2-TF00404-A05R + EST2** (1128 bases)
Cromossomo/citogenética	4q34.2
Gene anotado - símbolo	GPM6A
Número de Exons	7
Produto Gênico	*glycoprotein M6A*
RefSeq.	NM_005277 (Provisório)
LocusLink	2823
Homologia	*Squalus acanthias, Xenopus laevis, Mus musculus*
Observações	Gene conhecido, sem *splicing* alternativo. Figura 5.54

Tabela 5.11 - Dados resumidos da anotação preliminar do consenso do TF00404. Ferramentas utilizadas: BLAT e/ou BLAST

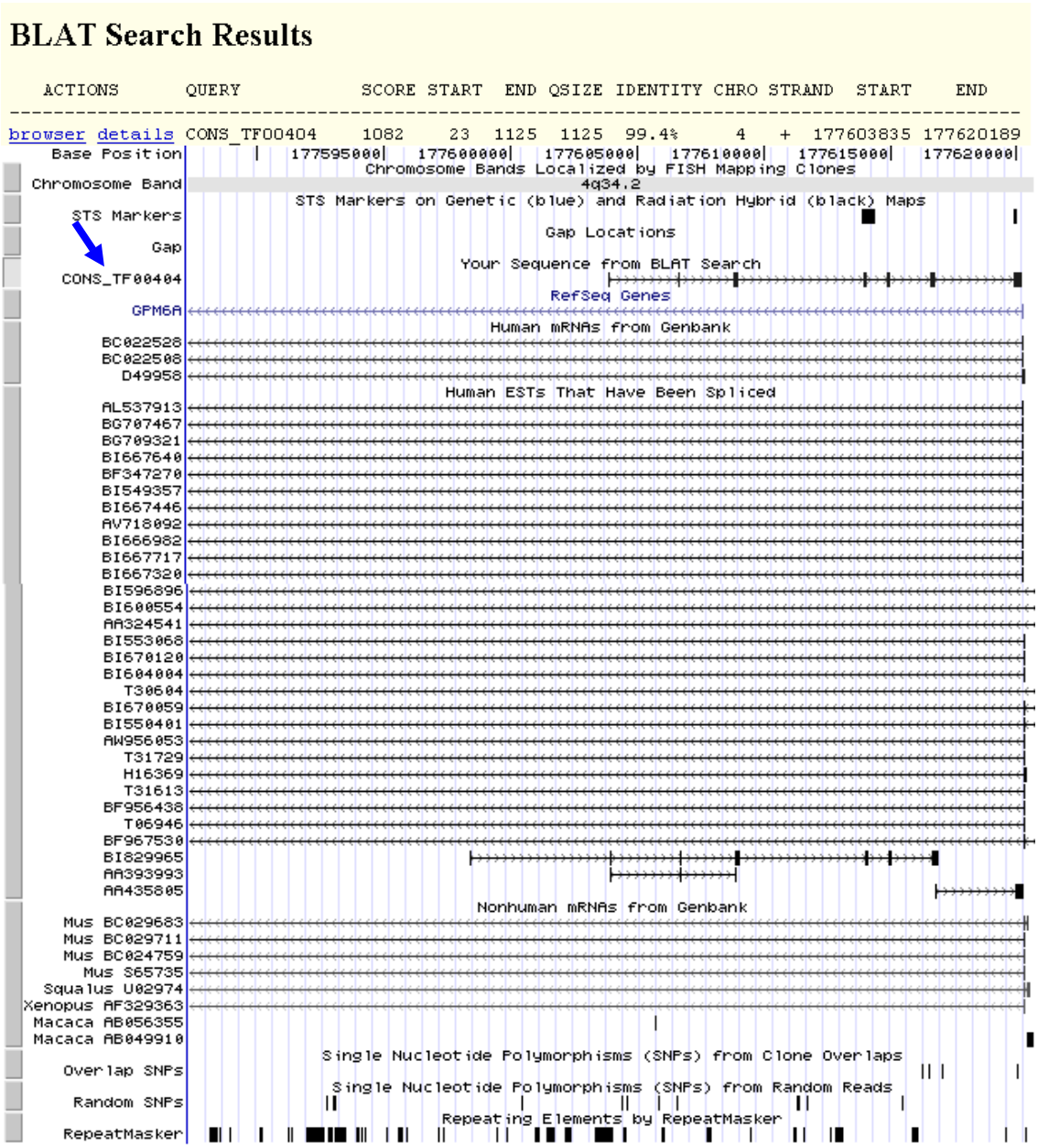


Figura 5.54 - Alinhamento do consenso do TF00404 com o *Draft* do Genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros, através da ferramenta **BLAT**

CONSENSOS TFs	Tam. Seq (bases)	BLAST	Genes ou Seqüências anotadas	Localiz. Crom.	SCORE bits	e-value	Identidade	alinh/query	Locus Link	RefSeq	Splicing Alternativo
IL2-TF00040	1141	BLAT	DOT1L - histone methyltransferase DOT1	19p13.3	1119	-	99%	1135/1141	84444	NM_032482	Não
IL2-TF00041	1146	BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF00156	1181	BLAT	CRT2 - hypotetical protein MGC26577	19p13.12	1168	-	100,0%	1178/1181	125972	NM_145046	Conhecido
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF00157	888	BLAT	FLJ22329 - hypotetical protein FLJ22329	19p13.12	763	-	99,3%	790/888	125972	NM_024656	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF00194	1410	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	Homo sapiens mRNA; cDNA DKFZp667O163;	-	1806	0.0	99,0%	925/1410	-	-	-
		BLASTx	Similar to C. elegans DPY-1-19 protein	-	427	e-118	77,0%	215/276	LOC233115	-	-
IL2-TF00232	1196	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	MGC35440: hypothetical protein MGC35440	-	533	e-148	99,0%	276/277	147990	-	-
		BLASTx	Similar to C. elegans DPY-19 protein [C. elegans][M. musculus]	-	144	2e-33	85,0%	74/87	LOC233115	-	-
IL2-TF00308	1310	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	Human DNA sequence from clone RP11-36D19 on chr. 10, compl.se	-	1168	0.0	100,0%	589/589	-	-	-
		BLASTx	Nenhum alinhamento significante (SCORE <100)	-	-	-	-	-	-	-	-
IL2-TF00309	1366	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	Homo sapiens similar to RIKEN cDNA 1200009H11	-	2418	0.0	98,0%	1331/1355	LOC255565	-	-
		BLASTx	Unnamed protein product [Homo sapiens]	-	411	e-172	88,0%	195/217	-	-	-
IL2-TF00380	501	BLAT	AMOT - *angiomotin*	Xq23	411	-	98.3%	426/501	154796	NM_133265	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF00404	1128	BLAT	G9M6A - *sglycoprotein M6A*	4q34.2	1082	-	99,40%	1102/1125	2823	NM_005277	Conhecido
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-

Tabela 5.12 Tabela Resumida da anotação preliminar dos consensos dos TFs validados pelo Grupo IL2.

5.2.2 *Transcript Finishing (TFs) não validados*

No processo de validação dos transcritos, foram geradas seqüências com os *primers* enviados pela coordenação do TFI que, após amplificação pela reação de PCR, clonagem e seqüenciamento não alinharam ao clone genômico informado, ou seja, não validaram o TF. Então, de posse destas seqüências, o grupo IL2 realizou uma análise a partir de ferramentas de bioinfomática disponíveis destas seqüências produzidas inespecificamente, avaliando se estas eram seqüências inéditas ou já conhecidas nos bancos de dados.

Essa análise preliminar *in silico* foi realizada de acordo com os protocolos 16 e 17, apresentados no item métodos em Material e Métodos.

AS figuras representam os gráficos do alinhamento das seqüências de cDNAs parciais não validadas com o *Draft* do genoma Humano, bancos de dados de ESTs, mRNAs humanos e não humanos, RefSeq, entre outros; através da ferramenta BLAT (disponível em http://genome.ucsc.edu). As tabelas apresentam a análise preliminar *in silico*, resumida, das informações disponíveis nos bancos de dados as seqüências.

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AC004490 (TF00040)

 *H. sapiens*	Genes anotados para cDNAs parcias do TF00040		
	***Read* TF40-A05R** (346 bases)	***Read* TF40-F11R** (431 bases)	***Read* TF40-G02R** (376 bases)
cDNA utilizado	Célula B	Célula B	Célula B
Cromossomo/citogenética	**17q21.2**	**1q21.2**	**1p35.3**
Gene anotado - símbolo	MRPL10	PP591	SES2
Número de Exons	05	03 *Exon Usage*	10
Produto Gênico	mitochondrial ribosomal protein L10	hypothetical protein PP591	sestrin 2
RefSeq.	NM_145255 (Status: Predicted)	NM_025207 (Status: Provisional)	NM_031459 (Status: Provisional)
LocusLink	124995	80308	83667
Homologia	*Mus Musculos*	-	-
Observações	Seqüência parcial conhecida, como exon *usage* na Read TF40 - FIIR. Observar figuras 5.55, 5.56a e b e 5.57.		

Tabela 5.13 - Genes Anotados para os cDNAs parciais não validados do TF00040

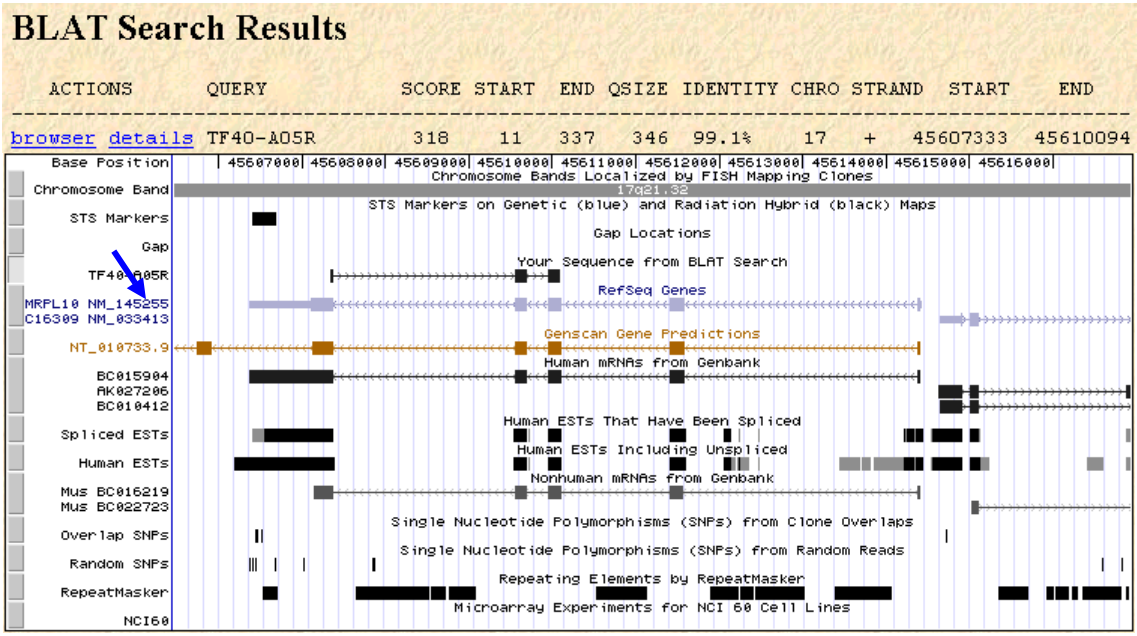


Figura 5.55 ***Read* IL2-TF00040-A05R**, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

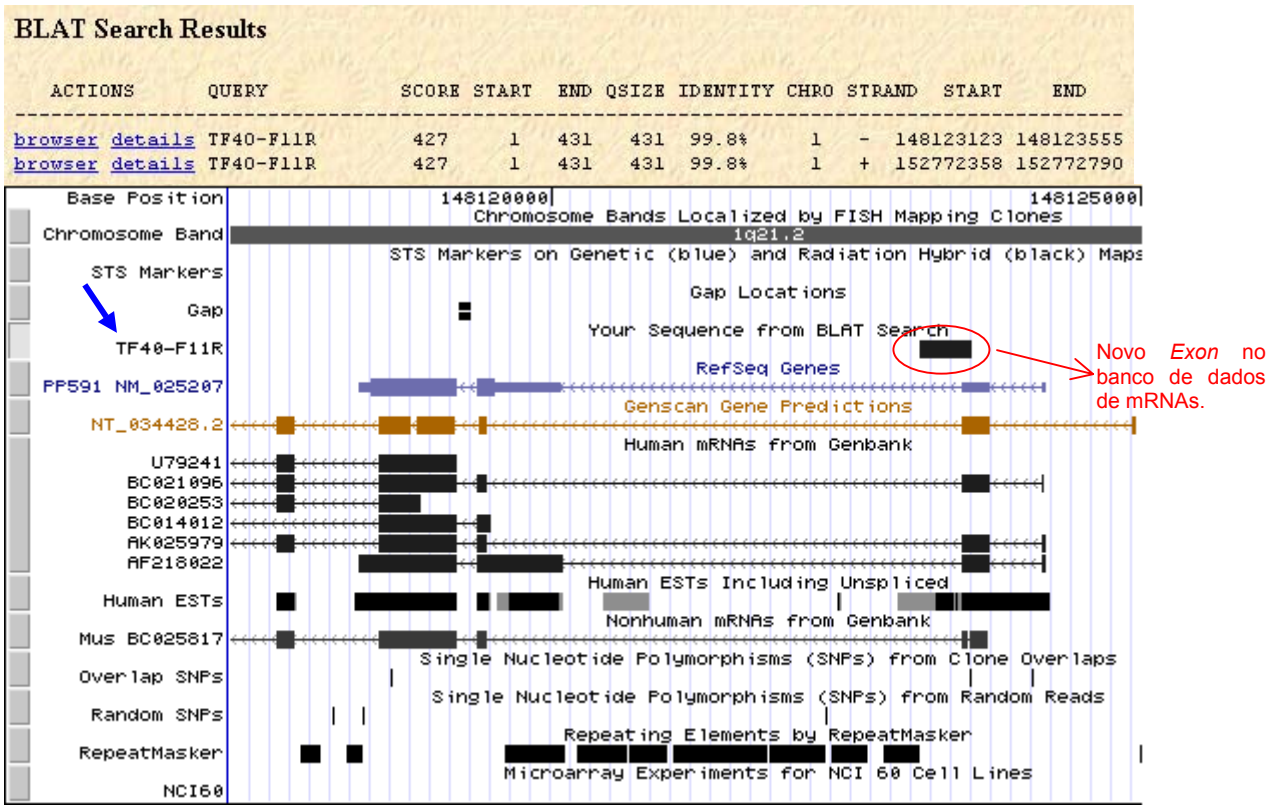


Figura 5.56a. ***Read* IL2-TF00040-F11R, *exons*** já representado no banco de *ESTs,* porém um *exon usage* no banco de dados de mRNAs.

Figura 5.56b. ZOOM da figura 5.55a., o cDNA parcial avaliado (F11R) em detalhe e o banco de dados de *ESTs. Exon* já representado no banco de *ESTs,* porém um *exon usage* no banco de dados de mRNAs (seta vermelha).

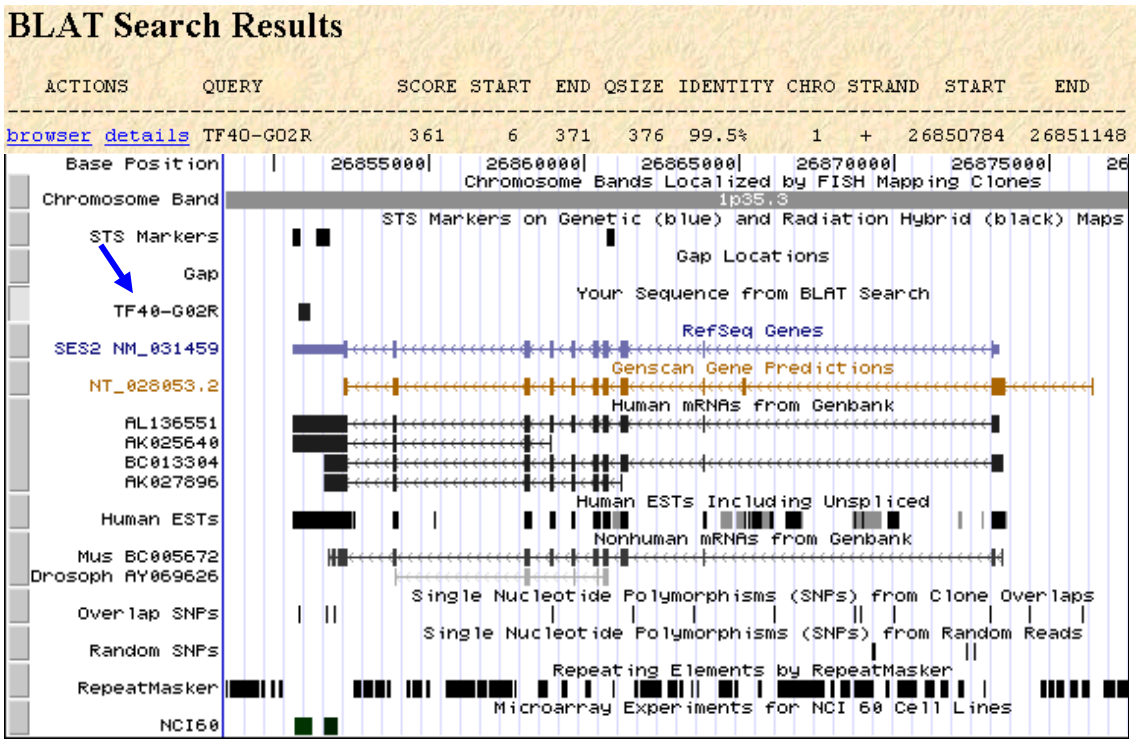


Figura 5.57 - ***Read* IL2-TF00040-G02R**, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AC006942 (TF00072)

 *H. sapiens*	Genes anotados para cDNAs parcias do TF00072	
	***Read* TF72-A01F** (281 bases)	***Read* TF72-D01F** (381 baes)
cDNA utilizado	Célula B	Célula B
Cromossomo/citogenética	22q12.2	6p22.1
Gene anotado - símbolo	**NIPSNAP1**	TRIM26
Número de Exons	10	09
Produto Gênico	nipsnap homolog 1 (*C. elegans*)	tripartite motif-containing 26
RefSeq.	NM_003634 (Status: Reviewed)	NM_003449 (Status: Reviewed)
LocusLink	8508	7726
Homologia	-	*Mus musculus*
Observações	Seqüência parcial conhecida, sem *splicing* alternativo. (Figura 5.58 e 5.59)	

Tabela 5.14 - Dados resumidos da anotação preliminar cDNAs parciais não validados do TF00072.

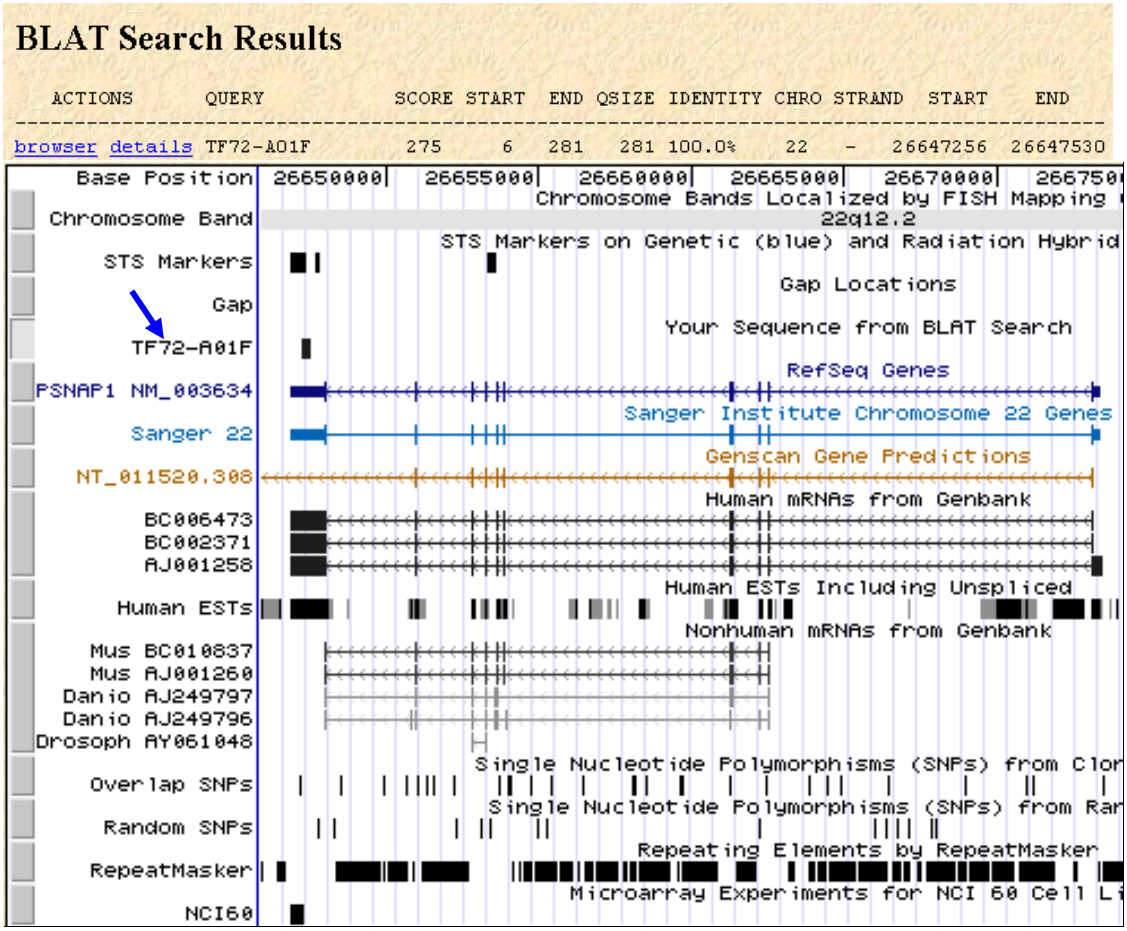


Figura 5.58 - *Read* IL2-TF00072-A01F, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

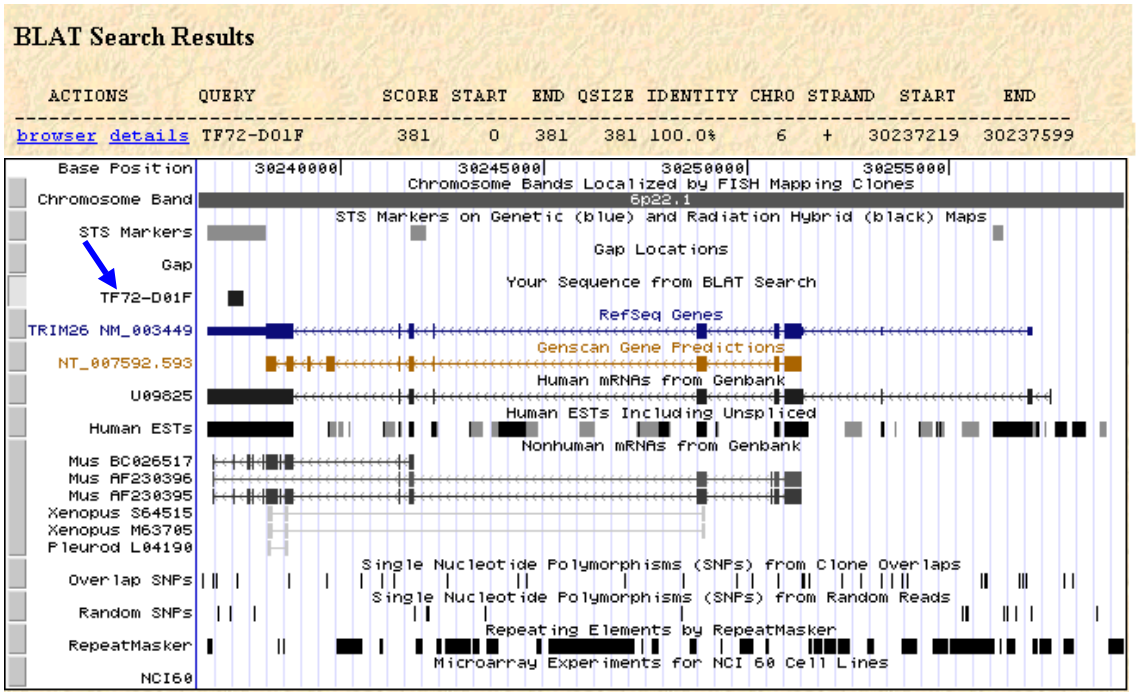


Figura 5.59 *Read* IL2-TF00072-D01F, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

"Mineração" e visualização das seqüências de cDNAs parciais que não alinharam em ID: AC005232 (TF00074)

Genes anotados para cDNAs parcias do TF00074	
H. sapiens	***Read* TF74-H02R** (308 bases)
cDNA utilizado	Cabeça e Pescoço
Cromossomo/citogenética	17_random
Gene anotado - símbolo	FLJ22865
Número de Exons	22
Produto Gênico	hypothetical protein FLJ22865
RefSeq.	NM_025109 (Status: Predicted)
LocusLink	80179
Homologia	-
Observações	Seqüência parcial conhecida, sem *splicing* alternativo(Figura 5.60)

Tabela 5.15 - Dados resumidos da anotação preliminar cDNAs parciais não validados do TF00074

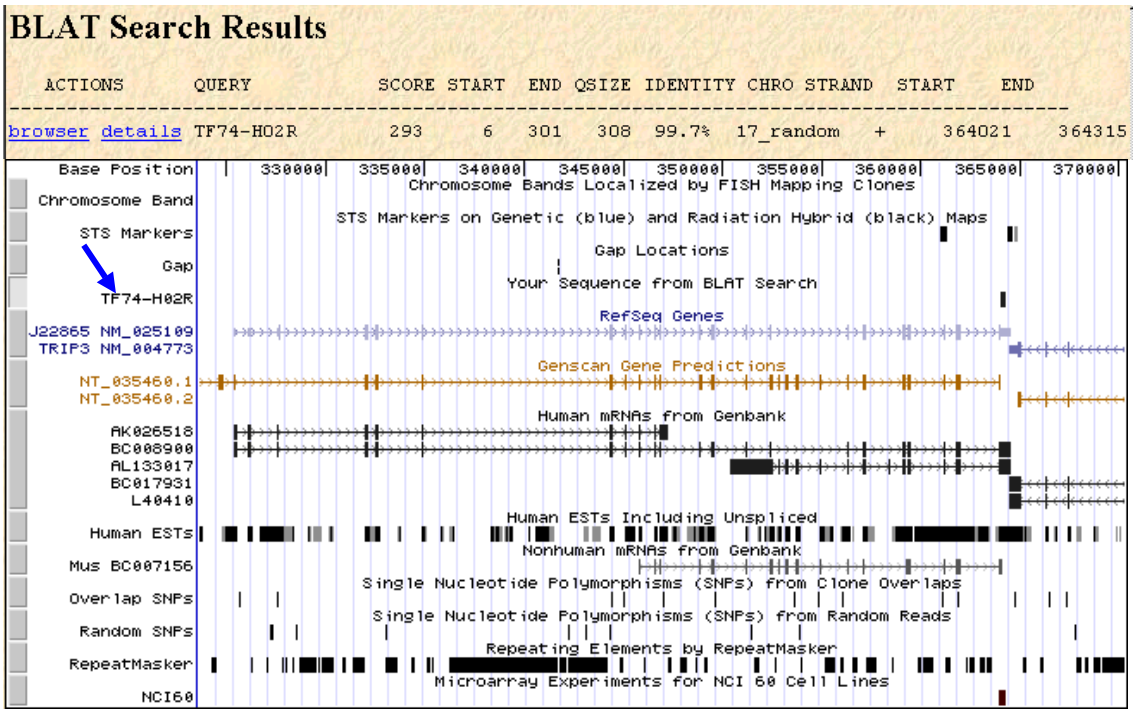


Figura 5.60 *Read* IL2-TF00074-H02R, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AC005023 (TF00193)

Genes anotados para cDNAs parcias do TF00193		
H. sapiens	***Read* TF193-A06R** (404 bases)	***Read* TF193-D07R** (468 bases)
cDNA utilizado	linfoma	linfoma
Cromossomo/citogenética	19p13.12	-
Gene anotado - símbolo	CD97	-
Número de Exons	20	-
Produto Gênico	CD97 antigen	-
RefSeq.	NM_001784 (Status: Reviwed)	-
LocusLink	976	-
Homologia	*Mus musculus* *Bos taurus* *Sus scrofa*	*Escherichia. coli*

Observações

A *read* TF193-D07R não apresentou alinhamento significante na ferramenta BLAT. Seguiu-se análise na ferramenta BLAST (BLASTn e BLASTx):

BLASTn (Figura 5.62)

>gi|1787509|gb|AE000224.1|AE000224 Escherichia coli K12 MG1655 section 114 of 400 of the complete genome, extensão = 12963.

SCORE: **914** bits (461); e-value = 0.0; identidade = 468/469 (99%), gaps = 1/469 (0%)

BLASTx (Figura 5.63)

>gi|16129217|ref|NP_415772.1 putative outer membrane protein [E. Coli K12]

SCORE: **201** bits (510); e-value = 1e-54; identidade = 100/123 (81%), Positivos = 100/123 (81%), Frame = +1.

Tabela 5.15 - Dados resumidos da anotação preliminar cDNAs parciais não validados do TF00193

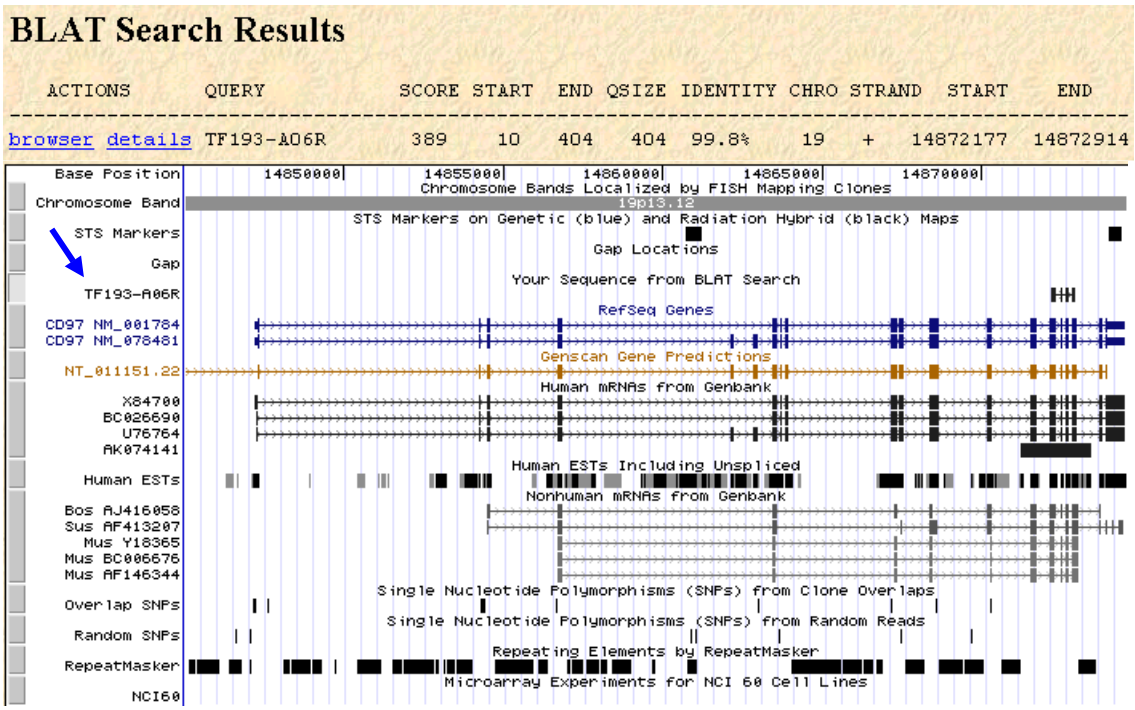
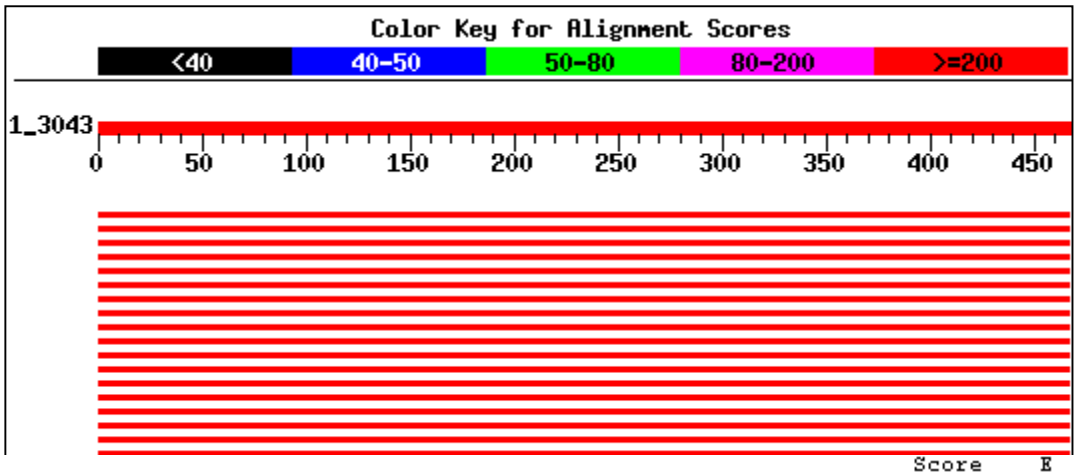
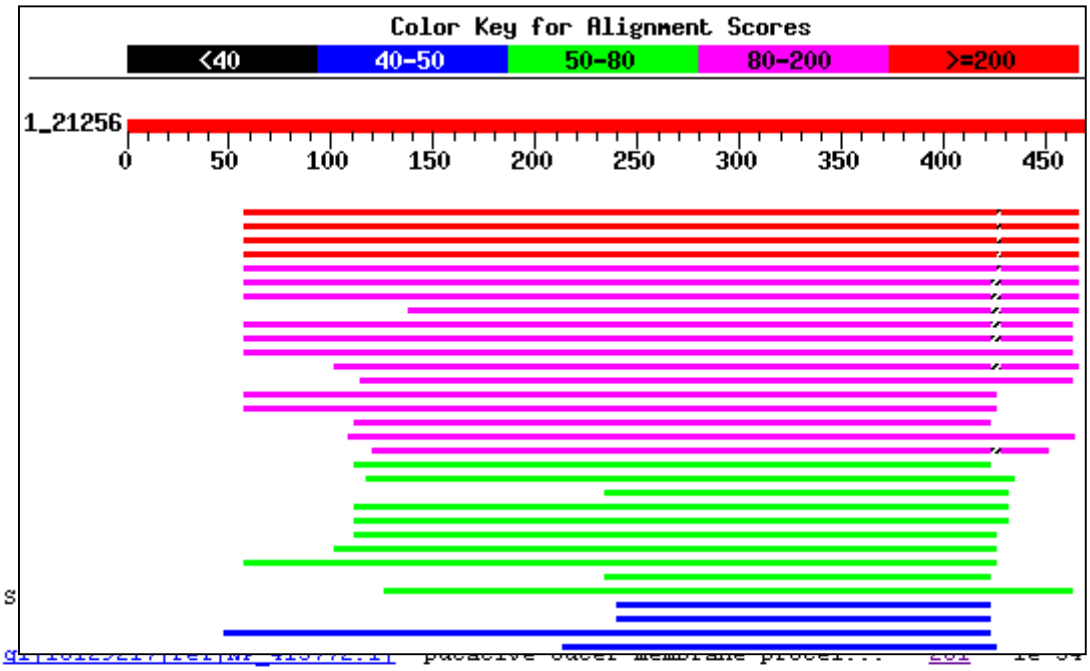


Figura 5.61 - *Read* IL2-TF00193-A06F, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.



Sequences producing significant alignments:	Score (bits)	E Value
gi|1787509|qb|AE000224.1|AE000224 Escherichia coli K12 MG16...	914	0.0
gi|43207|emb|X13583.1|ECTRTOI E. coli DNA for intervening r...	914	0.0
gi|902477|qb|U24206.1|ECU24206 Escherichia coli K12 (yciD) ...	914	0.0
gi|902396|qb|U24197.1|ECU24197 Escherichia coli ECOR 16 (yc...	914	0.0
gi|902387|qb|U24196.1|ECU24196 Escherichia coli ECOR 4 (yci...	914	0.0
gi|902378|qb|U24195.1|ECU24195 Escherichia coli ECOR 1 (yci...	914	0.0
gi|1742031|dbj|D90763.1| E.coli genomic DNA, Kohara clone #...	914	0.0
gi|12514980|qb|AE005343.1|AE005343 Escherichia coli 0157:H7...	882	0.0
gi|13361156|dbj|AP002556.1| Escherichia coli 0157:H7 DNA, c...	882	0.0
gi|902432|qb|U24201.1|ECU24201 Escherichia coli ECOR 46 (yc...	866	0.0
gi|902414|qb|U24199.1|ECU24199 Escherichia coli ECOR 31 (yc...	866	0.0
gi|24051558|qb|AE015152.1| Shigella flexneri 2a str. 301 se...	858	0.0
gi|902468|qb|U24205.1|ECU24205 Escherichia coli ECOR 71 (yc...	858	0.0
gi|902441|qb|U24202.1|ECU24202 Escherichia coli ECOR 50 (yc...	858	0.0
gi|902423|qb|U24200.1|ECU24200 Escherichia coli ECOR 37 (yc...	858	0.0
gi|902405|qb|U24198.1|ECU24198 Escherichia coli ECOR 28 (yc...	858	0.0
gi|902450|qb|U24203.1|ECU24203 Escherichia coli ECOR 52 (yc...	842	0.0
gi|902459|qb|U24204.1|ECU24204 Escherichia coli ECOR 60 (yc...	835	0.0



gi|15801482|ref|NP_287499.1|	putative outer membrane protei...	201	1e-54
gi|902379|gb|AAB60065.1|	yciD gene product >gi|902388|gb|AA...	201	1e-54
gi|11354244|pir||T45501	hypothetical protein yciD [imported...	201	1e-54
gi|902433|gb|AAB60113.1|	yciD gene product >gi|902442|gb|AA...	199	2e-54
gi|16760149|ref|NP_455766.1|	putative outer membrane protei...	182	2e-48
gi|16765076|ref|NP_460691.1|	outer membrane protein W; coli...	182	2e-48
gi|7162107|emb|CAB76675.1|	outer membrane protein [Salmonel...	155	3e-40
gi|22125935|ref|NP_669358.1|	putative outer membrane protei...	128	1e-29
gi|16122429|ref|NP_405742.1|	putative exported protein [Yer...	128	1e-29
gi|15601622|ref|NP_233253.1|	outer membrane protein OmpW [V...	118	9e-28

Figura 5.63 - BLASTx da *Read* IL2-TF00193-D07R.

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AL161670 (TF00408)

 *H. sapiens*	Genes anotados para cDNAs parcias do TF00408 ***Read* TF408-A02F** (520 bases)
cDNA utilizado	Próstata
Cromossomo/citogenética	21q22.3
Gene anotado - símbolo	PCNT2
Número de Exons	47
Produto Gênico	pericentrin 2 (kendrin)
RefSeq.	NM_006031 (Status: Reviewed)
LocusLink	5116
Homologia	*Mus musculus*
Observações	Seqüência parcial conhecida, sem *splicing* alternativo (Figura 5.64)

Tabela 5.16 Genes Anotados para os cDNAs parciais não validados do TF00408.

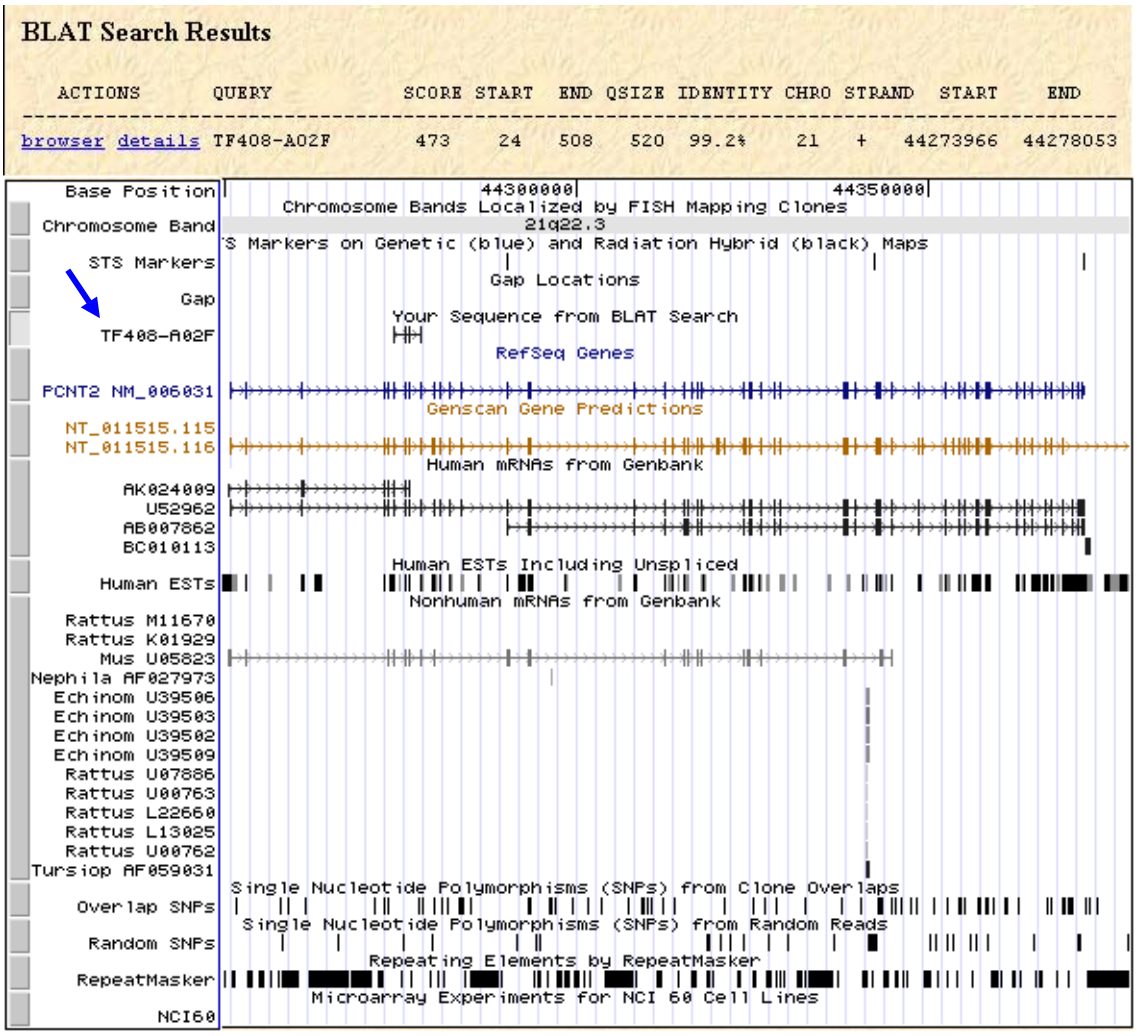


Figura 5.64 - *Read* IL2-TF00408-A02F, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs humanas.

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AP000136 (TF01048)

 *H. sapiens* — Genes anotados para cDNAs parcias do TF01048	***Read* TF1048-F03R** (209 pb)	***Read* TF1048-B02R** (422 pb)	***Read* TF1048-G07R** (383 pb)
cDNA utilizado	Cérebro	Cérebro	Próstata
Cromossomo/citogenética	5q33.1	11q23.3	17p13
Gene anotado - símbolo	**SPARC**	**RNF26**	**PFAS** (BLASTn)
Número de Exons	10	01	-
Produto Gênico	secreted protein, acidic, cysteine-rich (osteonectin)	ring finger - protein 26	phosphoribosylformylglycinamidine synthase (FGAR amidotransferase)
RefSeq.	NM_003118 (Status: Provisional)	NM_032015 (Status: Provisional)	-
LocusLink	6678	79102	5198 (BLASTn)
Homologia	*Rattus norvegicus Bovine osteonectin, Crocodylus niloticus, Mus musculus.*	*Mus musculus*	-

Observações

A *read* TF1048-G07R não apresentou alinhamento significante na ferramenta BLAT. Seguiu-se análise na ferramenta BLAST (BLASTn e BLASTx).

BLASTn (Figura 5.67)

>gi|2224662|dbj|AB002359.1| Human mRNA for KIAA0361 gene, KIAA0361 protein. Extensão = 5338.

SCORE = **731** bits (369), e-value = 0.0, identidade = 375/377 (99%)

BLASTx (Figura 5.68)

>gi|12230514|sp|O15067|PUR4_HUMAN| Phosphoribosylformylglycinamidine synthase (FGAM synthase) (FGAMS) (Formylglycinamide ribotide amidotransferase) (FGARAT); extensão = 1338.

SCORE = **263** bits (672), e-value = 4e-70, identidade = 124/125 (99%), positivos = 125/125 (100%), Frame = +1.

Para os alinhamentos dos cDNAs parciais no BLAT, figuras 5.65 e 5.66.

Tabela 5.17 - Genes Anotados para os cDNAs parciais não validados do TF01048.

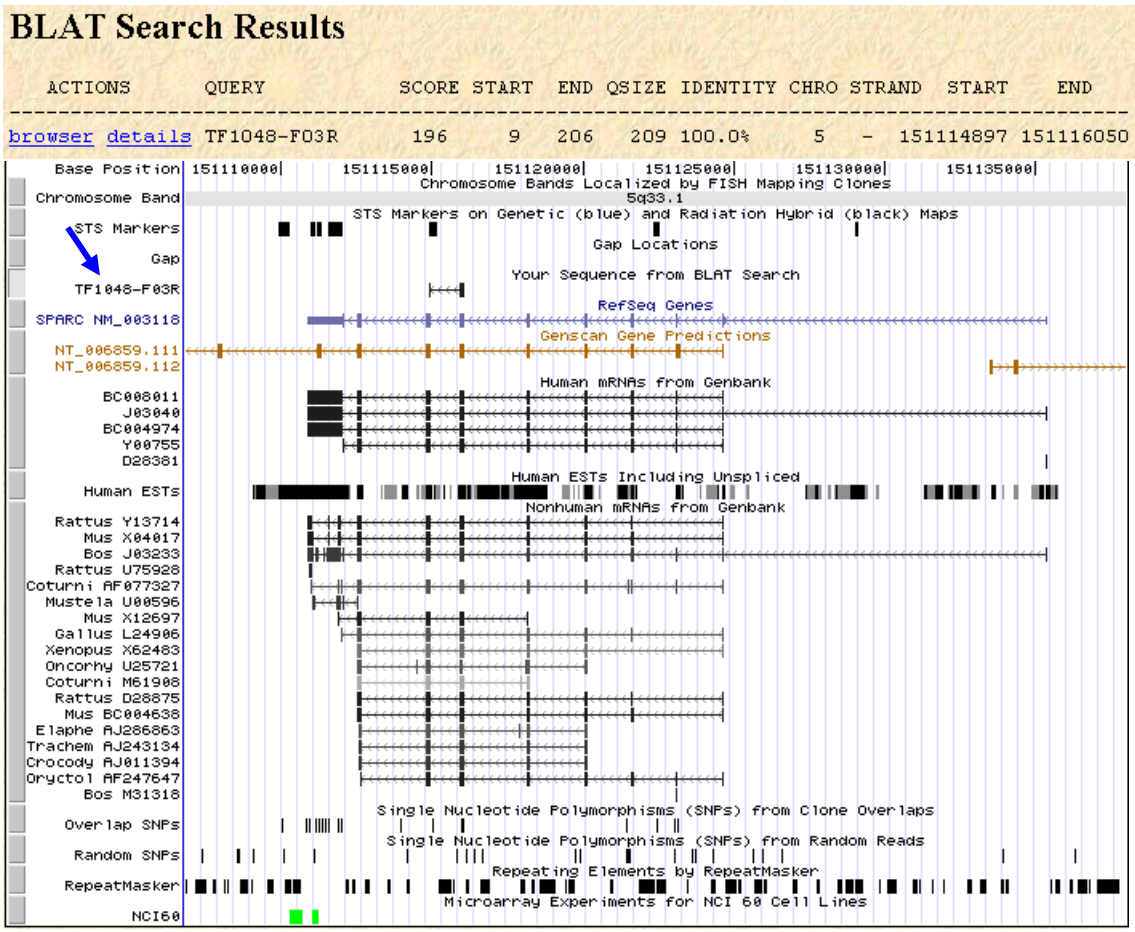


Figura 5.65 - *Read* IL2-TF01048-F03R, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.

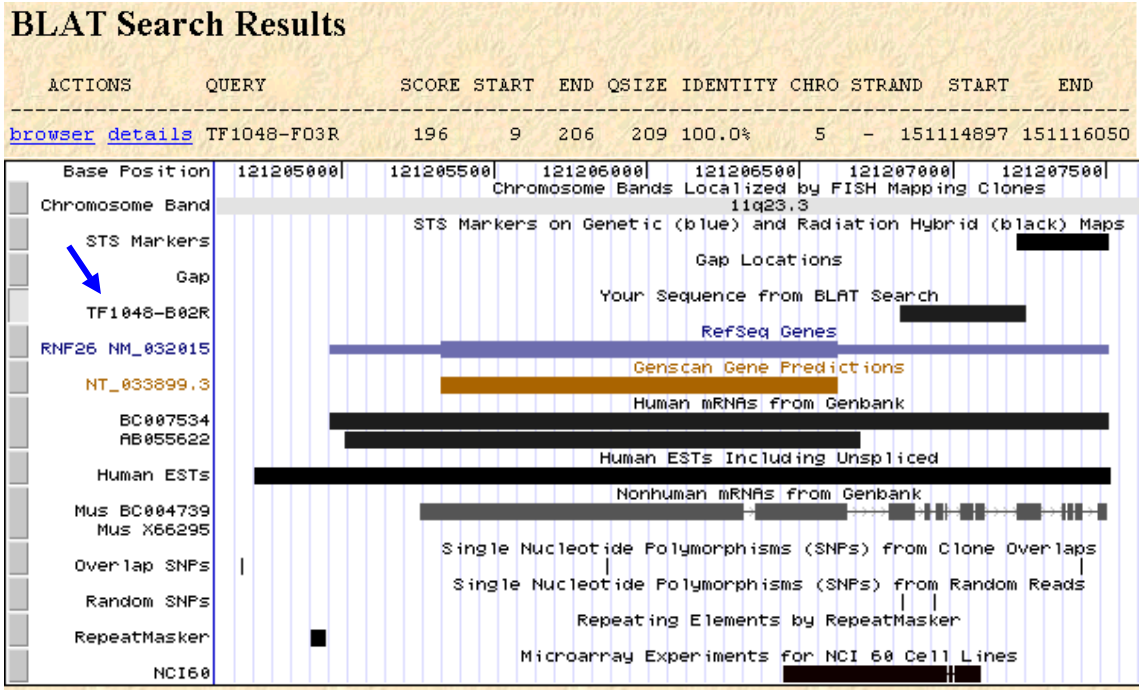
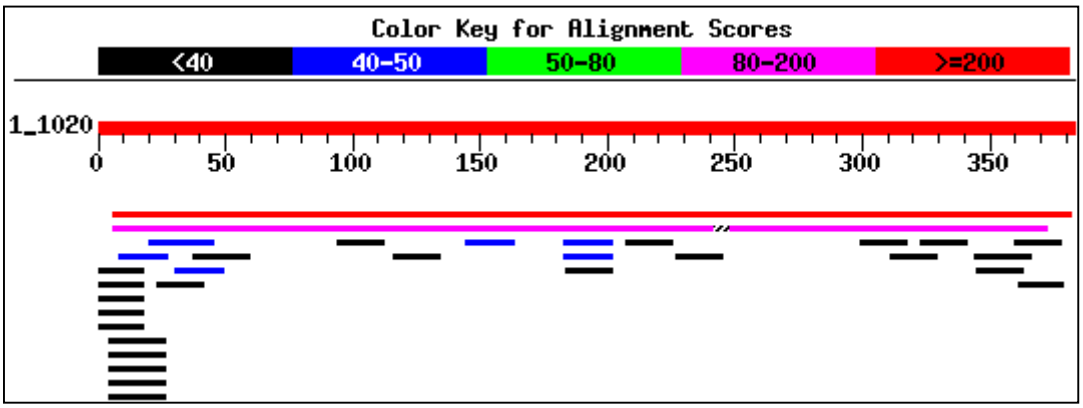
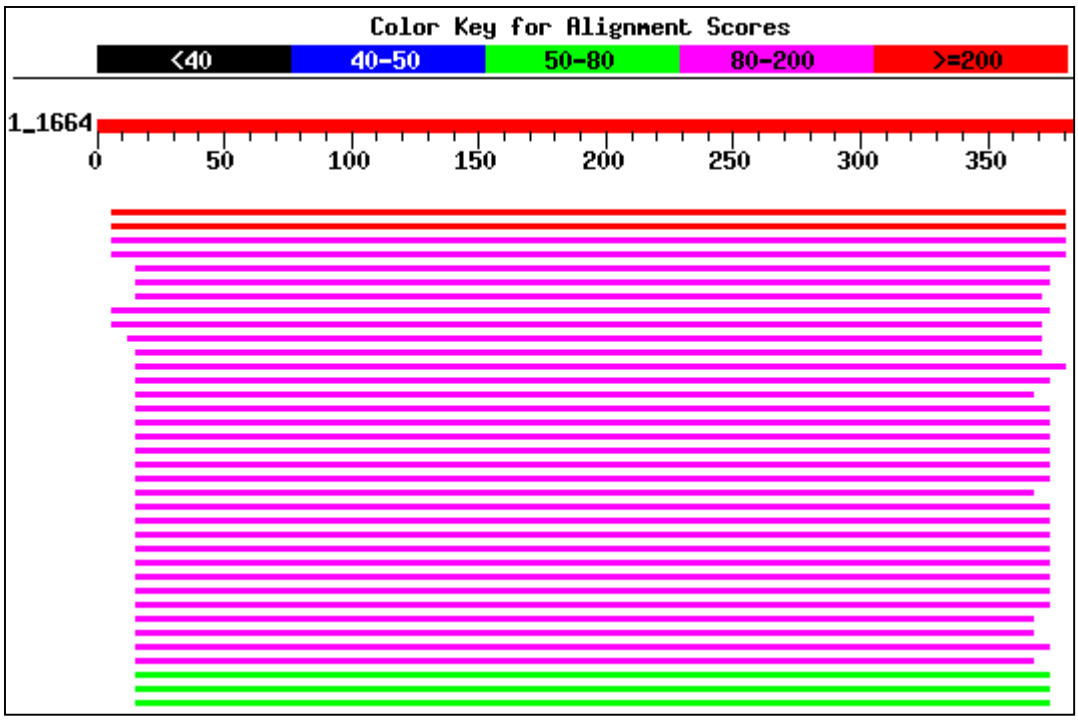


Figura 5.66 - *Read* IL2-TF01048-B02R, *exons* já conhecidos e representados no banco de dados de mRNAs e ESTs Humanas.



Sequences producing significant alignments:		Score (bits)	E Value
gi|2224662|dbj|AB002359.1|	Human mRNA for KIAA0361 gene, KI...	731	0.0
gi|20068638|emb|AL645902.6|	Mouse DNA sequence from clone R...	129	3e-27
gi|21738443|emb|AL662855.19|	Mouse DNA sequence from clone ...	44	0.14
gi|15982520|gb|AC021580.9|	Homo sapiens chromosome 9, clone...	40	2.2
gi|18693518|gb|AC015911.8|	Homo sapiens chromosome 17, clon...	40	2.2
gi|2181445|emb|Z96370.1|HS17QT037	H.sapiens telomeric DNA s...	40	2.2



Sequences producing significant alignments:		Score (bits)	E Value
gi|12230514|sp|O15067|PUR4_HUMAN	Phosphoribosylformylglycin...	263	3e-70
gi|2224663|dbj|BAA20816.1|	KIAA0361 [Homo sapiens]	263	3e-70
gi|17553022|ref|NP_497942.1|	AIR synthase related protein, ...	140	4e-33
gi|21288731|gb|EAA01024.1|	agCP12750 [Anopheles gambiae str...	137	4e-32
gi|17137292|ref|NP_477212.1|	CG9127-PA [Drosophila melanoga...	134	2e-31
gi|7438074|pir||T13363	phosphoribosylformylglycinamidine sy...	134	2e-31

Figura 5.68 - BLASTx da *Read* IL2-TF01048-G07R

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AL359399 (TF01049)

 *H. sapiens*	Genes anotados para cDNAs parcias do TF01049
	***Read* TF01049-D04R** (455 bases)
cDNA utilizado	melanoma
Cromossomo/citogenética	3p25.3
Gene anotado - símbolo	**FLJ22405**
Número de Exons	18
Produto Gênico	hypothetical protein FLJ22405
RefSeq.	NM_022485 (Status: *Predicted*)
LocusLink	64419
Homologia	*Mus musculus*
Observações	1.) 01 *exon* já representado no banco de *ESTs,* porém um *exon usage* no banco de dados de mRNAs (seta vermelha). 2.) 01*exon* já representado no banco de mRNAs, porém um *exon usage* no banco de dados de ESTs (seta azul). Figura 5.69a e b.

Tabela 5.18 Genes Anotados para os cDNAs parciais não validados do TF01049.

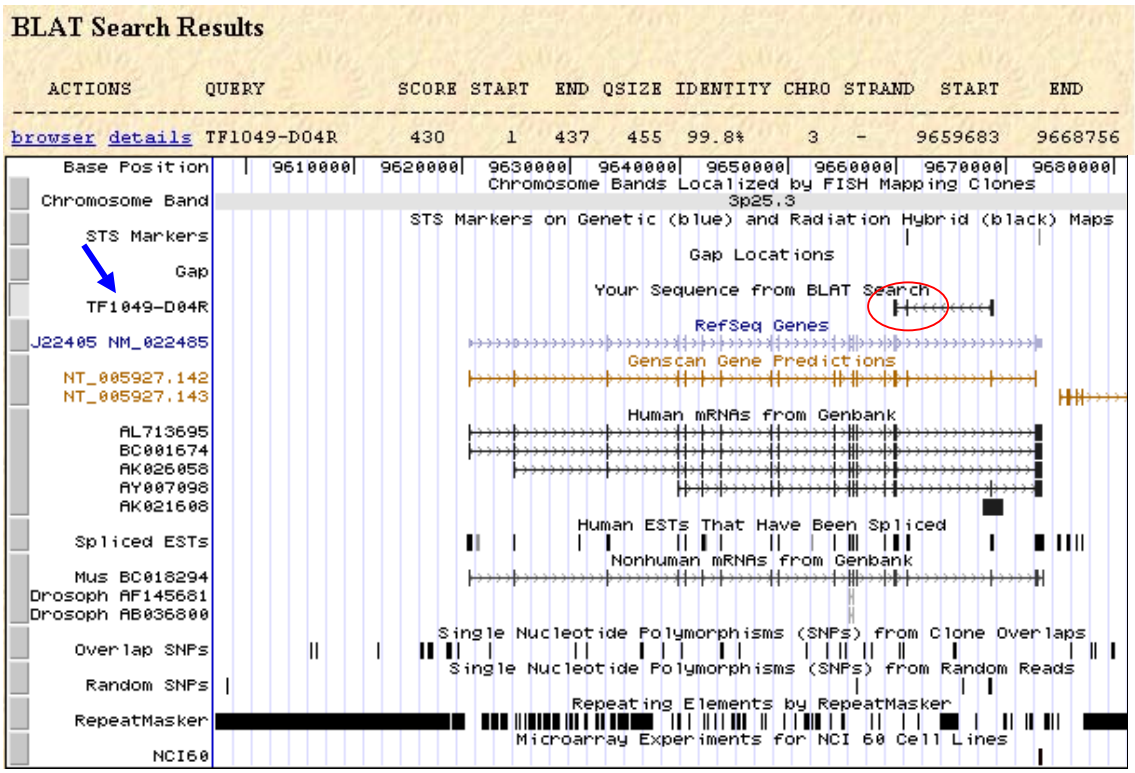


Figura 5.69a - *Read* IL2-TF01049-D04R, 1*exon* já representado no banco de *ESTs,* porém um *exon usage* no banco de dados de mRNAs (seta vermelha). 1*exon* já representado no banco de mRNAs, porém um *exon usage* no banco de dados de ESTs (seta azul).

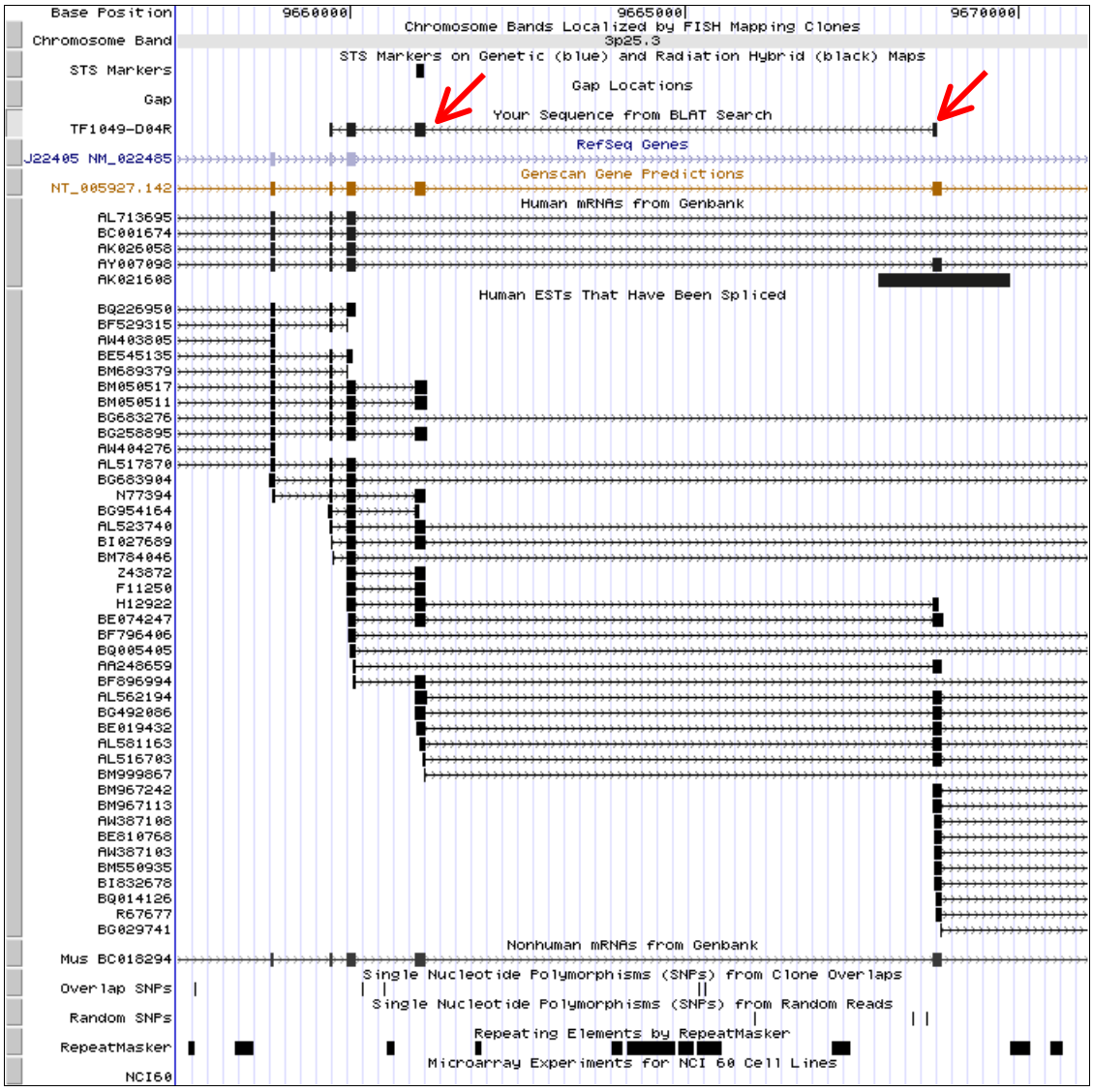


Figura 5.69b - ZOMM da figura 5.66a, 1*exon* já representado no banco de *ESTs,* porém um *exon usage* no banco de dados de mRNAs (seta vermelha).

“Mineração” e visualização das seqüências de cDNAs parciais que não alinharam em ID: AC008764 (TF00156)

Genes anotados para cDNAs parciais do TF00156	
H. sapiens	***Read* TF156-A05F** (444 bases)
cDNA utilizado	pulmão
Cromossomo/citogenética	3p21.2
Gene anotado - símbolo	**IMPDH2**
Número de Exons	-
Produto Gênico	*IMP (inosine monophosphate) dehydrogenase 2*
RefSeq.	NM_000884 (status: Provisional)
LocusLink	3615
Homologia	-

Observações

A *read* TF00156-A05F não apresentou alinhamento significante na ferramenta BLAT. Seguiu-se análise na ferramenta BLAST (BLASTn e BLASTx).

BLASTn (Figura 5.70)

>gi|15990411|gb|BCO15567.1|BCO15567| *Homo sapiens, IMP (inosine monophosphate) dehydrogenase 2*, clone MGC:20947 IMAGE:4576285, mRNA, complete cds; extensão = 1655.

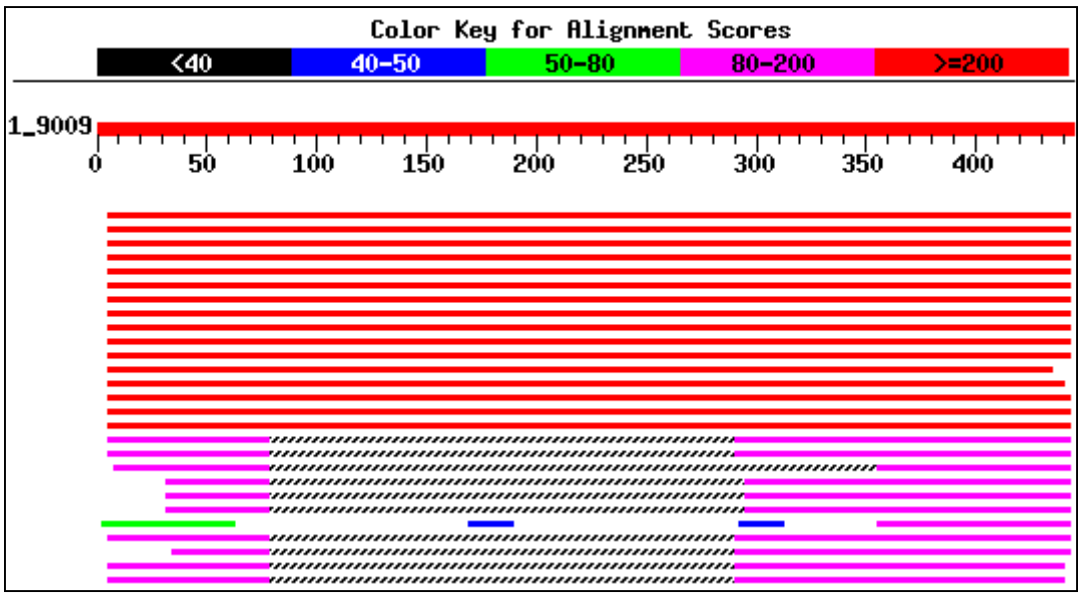
SCORE = **846** bits (427); e-value = 0.0; identidade = 436/439 (99%)

BLASTx (Figura 5.71)

>gi|124419|so|P12268|IMD2_HUMAN| *Inosine-5'-monophosphate dehydrogenase 2 (IMP dehydrogenase 2)* (IMPDH-II) (IMPD 2)

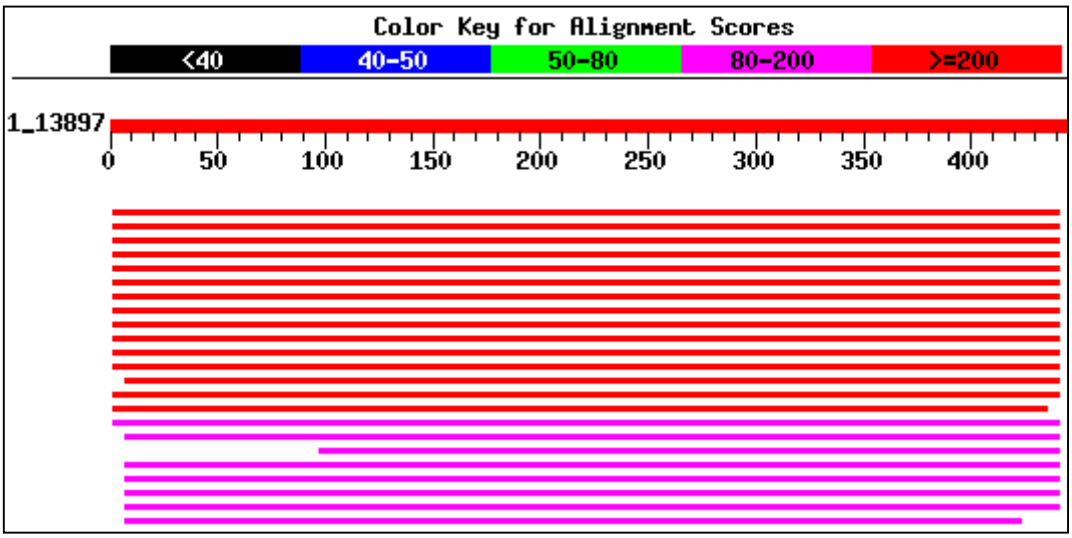
SCORE = **282** bits (722); e-value = 5e-76; identidade = 143/147 (97%), positivos = 145/147 (98%), Frame +2

Tabela 5.19 Gene Anotado para o cDNA parcial não validado do TF00156.



Sequences producing significant alignments:		Score (bits)	E Value	
gi|15990411|gb|BC015567.1|BC015567	Homo sapiens, IMP (inosi...	846	0.0	L U
gi|15277479|gb|BC012840.1|BC012840	Homo sapiens, IMP (inosi...	846	0.0	L U
gi|14714378|gb|BC009236.1|BC009236	Homo sapiens, IMP (inosi...	846	0.0	L
gi|13543972|gb|BC006124.1|BC006124	Homo sapiens, IMP (inosi...	846	0.0	L U
gi|4504688|ref|NM_000884.1|	Homo sapiens IMP (inosine monop...	809	0.0	L U
gi|186391|gb|J04208.1|HUMIMP	Human inosine-5'-monophosphate...	809	0.0	L U
gi|16307530|gb|BC010314.1|	Mus musculus, inosine 5'-phospha...	490	e-136	L U

Figura 5.70 - BLASTn da *Read* TF156-A05F



Sequences producing significant alignments:	Score (bits)	E Value	
gi|124419|sp|P12268|IMD2_HUMAN Inosine-5'-monophosphate deh...	282	5e-76	
gi|4504689|ref|NP_000875.1| IMP (inosine monophosphate) deh...	278	1e-74	L
gi|124426|sp|P12269|IMD2_MESAU Inosine-5'-monophosphate deh...	276	5e-74	
gi|6754346|ref|NP_035960.1| inosine 5'-phosphate dehydrogen...	275	6e-74	L
gi|16307531|qb|AAH10314.1| inosine 5'-phosphate dehydrogena...	275	6e-74	L
gi|124427|sp|P24547|IMD2_MOUSE Inosine-5'-monophosphate deh...	275	6e-74	
gi|20886083|ref|XP_134392.1| similar to inosine 5-phosphate...	275	6e-74	
gi|12832836|dbj|BAB22278.1| data source:MGD, source key:MGI...	273	4e-73	L
gi|25014074|sp|P20839|IMD1_HUMAN Inosine-5'-monophosphate d...	226	6e-59	

Figura 5.71 - BLASTx da *Read* TF156-A05F

cDNAs PARCIAIS	Tam. Seq (bases)	BLAST	Genes ou Seqüências anotadas	Localiz. Crom.	SCORE bits	e-value	Identidade	alinh/query	Locus Link	RefSeq	Splicing Alternativo
IL2-TF40-A05R	345	BLAT	MRPL10 - mitochondrial ribosomal protein L10	17q21.2	318	-	99,1%	326/346	124995	NM_145255	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF40-F11R	431	BLAT	PP591 - hypothetical protein PP591	1q21.2	427	-	99,8%	431/431	80308	NM_025207	exon usage
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF40-G02R	376	BLAT	SES2 - sestrin 2	1p35.3	361	-	99,5%	371/376	83667	NM_031459	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF72-A01F	281	BLAT	NIPSNAP1 - nipsnap homolog 1 (C. elegans)	22q12.2	275	-	100,0%	275/281	8508	NM_003634	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF72-D01F	381	BLAT	TRIM - tripartite motif - containing 26	6p22.1	381	-	100,0%	381/381	7726	NM_003449	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF74-H02R	308	BLAT	FLJ22865 - hypotetical protein FLJ22865	17_rand	293	-	99,7%	295/308	80179	NM_025109	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF193-A06F	404	BLAT	CD97 - CD97 antigen	19p13.12	389	-	99,8%	384/404	976	NM_001784	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF193-D07F	501	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	E. coli K12 MG1655 section 114 of 400 of the compl genor	-	914	0.0	99,0%	468/469	-	-	-
		BLASTx	putative outer membrane protein [E. Coli K12]	-	201	1e-54	81,0%	100/123	-	-	-
L2-TF408-A02F	520	BLAT	PCNT2 - pericentrin (kendrin)	21q22.3	473	-	99,2%	484/520	5116	NM_006031	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-

Tabela 5.21 – Reaulmo da anotação preliminar dos cDNAs parciais gerados pelo Grupo IL2, que não validaram TFs.

cDNAs PARCIAIS	Tam. Seq (bases)	BLAST	Genes ou Seqüências anotadas	Localiz. Crom.	SCORE bits	e-value	Identidade	alinh/query	Locus Link	RefSeq	Splicing Alternativo
IL2-TF1048-F03R	209	BLAT	SPARC - secreted protein, acidic, cysteine-rich (ostec	5q33.1	196	-	100,0%	197/209	6678	NM_003118	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF1048-B02R	422	BLAT	RNF26 - ring finger - protein 26	11q23.3	196	-	100,0%	196/196	79102	NM_032015	Não
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF1048-G07R	383	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	Human mRNA for KIAA0361 gene, KIAA0361 protein	-	731	0.0	99,0%	375/377	-	-	-
		BLASTx	PFAS - Phosphoribosylformylglycinamidine synthase	17p13	263	4e-70	99,0%	124/125	5198	-	-
IL2-TF1049-D04R	455	BLAT	FLJ22405 - hypotetical protein FLJ22405	3p25.3	430	-	99.8%	437/455	64419	NM_022485	(2)conhecidos
		BLASTn	-	-	-	-	-	-	-	-	-
		BLASTx	-	-	-	-	-	-	-	-	-
IL2-TF156-A05F	444	BLAT	Nenhum alinhamento significante no RefSeq	-	-	-	-	-	-	-	-
		BLASTn	IMPDH2 - inosine monophosphate dehydrogenase	3p21.2	846	0.0	99,0%	436/439	3615	NM_000884	-
		BLASTx	IMPDH2 - inosine monophosphate dehydrogenase	3p21.2	282	0	97,0%	143/147	3615	NM_000884	-

Tabela 5.21 – Reaulmo da anotação preliminar dos cDNAs parciais gerados pelo Grupo IL2, que não validaram TFs.

O programa BLAT foi escolhido como o primeiro a ser pesquisado devido a sua rapidez no processamento e facilidade de identificação dos *exons* e *splicing* alternativos. Esse programa realiza o alinhamento da seqüência pesquisada contra o genoma humano e de camundongo. Inclui também outras ferramentas úteis como programas de predição gênica e visualização de ESTs, fornece identificação e localização gênica com *links* para várias outras ferramentas de caracterização (KENT, *et al*; 2002).

Na análise dos consensos dos TFs validados, utilizando a ferramenta BLAT, dos 10 TFs validados que tinham consenso montado, 6 combinaram com genes conhecidos e descritos no RefSeq. Nestes, o alinhamento entre a seqüência pesquisada e a seqüência no banco de dados ocorreu, praticamente em toda a sua extensão, com *scores* altos e identidades muito próximas a 100%. No BLAT, ainda, puderam ser identificados 2 "splicing" alternativos (TF00156 e TF00404) já representados nos bancos de dados de ESTs e/ou mRNAs. Os 4 consensos restantes não alinharam em genes representados no RefSeq, porém, há ESTs e mRNAs que representam sua extensão, parcial ou totalmente, chamando a atenção para a necessidade em se manter atualizados os bancos de dados locais, pois, entre a escolha dos TFs e sua validação, estes TFs (6) foram seqüenciados por outros projetos de validação de transcritos. A estes 4 consensos, realizou-se uma busca por similaridade de seqüências na ferramenta BLAST, utilizando-se as opções BLASTn e BLASTx. Como o resultado desta busca retorna uma lista de seqüências que alinharam em maior ou menor grau de similaridade, foi escolhida somente a seqüência de "score" mais alto, desde que este fosse maior que 200 para BLASTn (nucleotídeos) e maior que 100 para BLASTx (proteínas). O e-value também foi considerado, optando-se por seqüências cujo e-value fosse igual ou muito próximo a zero.

O consenso do TF00194 apresenta 1410 bases e na opção BLASTn apresentou identidade significativa (99%), alinhando 925 das 1410 letras, com score = 1806, e-value = 0,0 à seqüência Homo sapiens mRNA, cDNA DKFZp6670169 e no BLASTx apresentou identidade de 77%, alinhando a seqüência de aminoácidos traduzida pela ferramenta (Frame +3), com *score* = 427 e e-value = e-118 à seqüência Similar to C. elegans DPY-19 protein.

O consenso do TF00232 apresenta 1196 e, no BLASTn, alinhou com maior *score* (533), e-value = e-148 e identidade de 99% a Homo sapiens hypotetical protein MGC35440 e na opção BLASTx, após tradução, alinhou a Frame +2, com 85% de identidade, *score* = 144 e e-value = 2e-33 à seqüência *Similar to C. elegans DPY-19 protein* (Resultado também obtido no TF00194.
No TF00308 (1310 bases) observou-se similaridade, no BLASTn, com a seqüência *Human DNA sequence from clone RP11-36D19 on chromossome 10, compl. Sequence*, com *score* = 1168, e-value = 0.0 e identidade de 100%. A opção BLASTx não retornou nenhuma seqüência com identidade significativa, sendo o maior *scor*e menor que 100.

Das 1366 bases do TF00309, 1331 (98%) alinharam com *score* = 2418 e e-value = 0.0 com a seqüência *Homo sapiens silimar to RIKEN cDNA 1200009H11,* utilizando a opção BLASTn. O BLASTx, alinhou a Frame+1, com *score* = 411, e-value = e-172 e identidade de 88% a *Unnamed protein product (Homo sapiens).*

Às cDNAs parciais que não validaram TFs, procedeu-se análise idêntica a realizada para os consensos dos TFs validados, levando-se em consideração que não se tratavam de seqüências completas acarretando limitações na análise.

Na ferramenta BLAT, das 14 cDNAs pesquisadas, 11 alinharam com identidade próxima a 100% a genes representados no RefSeq. Destes, observou-se um novo splicing (*exon usage*) no cDNA parcial IL2-TF40-F11R, não representado nos bancos de dados de mRNAs e ESTs. No cDNA parcial IL2-TF1049-D04R, identificou-se 01 exon alternativo já representado no banco de ESTs, porém um *exon usage* no banco de dados de mRNAs e outro exon já representado no banco de dados de mRNAs, porém um exon usage no banco de dados de ESTs.

Aos 3 cDNAs que não apresentaram alinhamento algum (no math) na ferramenta BLAT, realizou-se uma busca por similaridade de seqüências na ferramenta BLAST, utilizando-se as opções BLASTn e BLASTx. Os mesmos parâmetros para *score* e e-value para os consensos foram utilizados aqui. Destes, o cDNA IL2-TF193-D07R, alinhou 468 de suas 501 bases, com identidade de 99%, *score* = 914 e e-value = 0.0 com a seqüência *E. coli K12 MG1655 section 114 of 400 of the compl. Genome*, no BLASTn e a

tradução e busca no BLASTx retornou também uma seqüência de *Echerichia coli (Putative outer membrane protein [E. coli])*, com *score* = 201, e-value = 1e-54 e 81% de identidade.

O IL2-TF1048-G07R apresentou identidade de 99% com Human mRNA for KIAA 361 gene (no BLASTn com *score* = 731 e e-value = 0.0) alinhando 375 de suas 383 bases. E, no BLASTx, com 99% de identidade, alinhou-se a seqüência PFAS - *Phosphoribosylformyglycinamidine syntase*, com *score* = 263 e e-value = 4e-70.

De todos os alinhamentos realizados, o cDNA IL2-TF156-A05F foi o único que apresentou o mesmo resultados tanto no BLASTn quanto no BLASTx. Ambas as opções retornaram como resultado de maior *score* (BLASTn (846) e BLASTx (282)) a seqüência *IMPDH2 - inosine monophosphate dehydrogenase*, ambas com e-value = 0.0 e 99% E 97% de identidade, respectivamente.

Faz-se necessário salientar que as análises realizadas fazem parte apenas de um exercício para um maior contato com a bioinformática e algumas de suas inúmeras ferramentas para análise genômica, disponíveis gratuitamente na Internet para fins acadêmicos.

A “mineração” nos bancos de dados, dos cDNAs parciais gerados inespecificamente durante o processo de validação dos TFs pode ser uma maneira de se encontrar novos *exons*. Todavia, é necessária a confirmação destas seqüências pelo re-seqüenciamento dos clones e até mesmo, reunir a estas seqüências, todas as seqüências geradas nestas condições pelos outros grupos validadores de transcritos do projeto TFI, dispostos a disponibilizá-los ao grupo IL2.

6. Conclusões

❖ Os protocolos estabelecidos para a amplificação, clonagem e seqüenciamento parecem estar adequados e a eficiência de amplificação dos TFs foi em uma faixa abaixo do previsto, mas compatível com outros trabalhos.

❖ A eficiência na validação parece estar mais relacionada à própria eficiência e limites da técnica de RT-PCR utilizada, do que outras variáveis discutidas no presente trabalho.

❖ O desenvolvimento deste projeto nos moldes de uma rede virtual, tem proporcionado a toda a equipe, um grande aprendizado, discussões e acesso à informações em um vasto campo de conhecimentos, desde técnicas laboratoriais até treinamento em bioinformática.

❖ Trata-se de uma amostragem pequena de TFs em relação ao total de todos os TFs distribuídos e certamente, esta análise aplicada aos 597 TFs será mais informativa.

❖ Os transcritos não validados constituíram um bom modelo para realização de um exercício de anotação.

7. Literatura Citada

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K., WATSON, Macromoléculas: estrutura, forma e função. In:___. **Biologia molecular da célula.** 3.ed. Porto Alegre: Artes Médicas, 1997a. p.291-334.

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K., WATSON, J.D.Tecnologia do DNA recombinante. In:___. **Biologia molecular da célula.** 3.ed. Porto Alegre: Artes Médicas, 1997b. p.291-334.

BAYAT, A. Bioinformatics. **BMJ,** v.321, p.1018-1022, 2002.

BRENT, M.R. Predicting full-length transcripts. **Trends in Biotechnology,** v.20, p.273-274, 2002.

BROWN, T. A. Genes e expressão gênica. In___: **Genética - um enfoque molecular.** Rio de Janeiro: Guanabara Koogan, 1999, p.10-55.

CAMARGO, A.A., DE SOUZA, S.J., BRENTANI, R.R., SIMPSON, A.J.G. Human gene discovery through experimental definition of transcribed regions of the human genome. **Current Opinion in Chemical Biology**, v.6, p.13-16, 2001a.

CAMARGO, A.A. Projeto: Transcript Finishing Initiative: relatório parcial período fevereiro-setembro de 2001. São Paulo: Instituto Ludwig, 2001b.

CAMARGO, A.A., SAMAIA, H.P.B., DIAS-NETO, E., SIMÃO, D.F., MIGOTTO, I.A, BRIONES, M. R. S., COSTA, F.F., NAGAI, M.A, VERJOVSKI-ALMEIDA, S., ZAGO, M.A., ANDRADE, L.E.C., CARRER, H., EL-DORRY, H.F.A., ESPREAFICO, E.M., HABR-GAMA, V., GIANNELLA-NETO, D., GOLDMAN, G.H., GRUBER, A.., HACKEL, C., KIMURA, V., MACIEL, R.M.B., MARIE, S.K.N., MARTINS, E.A.L., NÓBREGA, M.P., PAÇÓ-

LARSON, M.L., PARDINI, M.I.M.C., PEREIRA, G.G., *et al*. The contribution of 700.000 ORF sequence tags to the definition of the human transcriptome. **PNAS**, v.98(21), p.12103-8, 2001c.

CARTEGNI, L. CHEW, S.L., KRAINER, A.R., Listening to silence and understanding nonsense: exonic mutations that affect splicing. **Nature Reviews**, v.3, p.285-298, 2002

COOPER G. M. The organization of cellular genomes. In___: **The cell - a molecular approach.** Washington: Sinauer, 2000, p.137-170.

DAS, M., BURGE, C.B., PARK, E., COLINAS, J., PELLETIER, J. Assessment of the total number of human transcription units. **Genomics**, v.77, p.71-8, 2001.

DE SOUZA, S.J., CAMARGO, A.A., BRIONES, M.R.S., COSTA, F.F., NAGAI, M.A., VERJOVSKI-ALMEIDA, S., ZAGO, M.A., ANDRADE, L.E.C., CARRER, H., EL-DORRY, H.F.A., ESPREAFICO, E.M., HABR-GAMA, A., GIANNELLA-NETO, D., GOLDMAN, G.H., GRUBER, A., HACKEL,C., KIMURA, E.T., MACIEL, R.M.B., MARIE, S.K.N., MARTINS, E.A.L., NÓBREGA, M.P., PAÇÓ-LARSON, M.L., PARDINI, M.I.M.C., PEREIRA, G.G., *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed tags. **PNAS**, v.97, p.12690-3, 2000.

DE ROBERTIS, E.M.F., HIB, J. Genes e Genomas. In___: **Bases da Biologia Molecular.** Rio de Janeiro: Guanabara Koogan, 2001, p.115-123.

DIAS-NETO, E., CORREA, R.G., VERJOVSKI-ALMEIDA, S., BRIONES, M.R.S., NAGAI, M.A., SILVA, W., ZAGO, M.A., BORDIN, S., COSTA, F.F., GOLDMAN, G.H., CARVALHO, A.F., MATSUKUMA, A., BAIA, G.S., SIMPSON, D.H., BRUNSTEIN, A., OLIVEIRA, P.S.L., BUCHER, P., JONGENEEL, C.V., O'HARE, M.J., SOARES, F., BRENTANI, R.R., REIS, L.F.L, SOUZA, S.J., SIMPSON, A.J.G. Shotgun sequencing of the human

transcriptome with ORF expressed sequence tags. **PNAS**, v.97, p.3491-6, 2000.

DIAS-NETO, E., HARROP, R., CORREA-OLIVEIRA, R., WILSON, R.A., PENA, S.D.J., SIMPSON, A.J.G. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normlized libraries for the generation of ESTs from nanogram quantities of mRNA. **Gene**, v.186, p.135-142, 1997.

FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo. **Projeto Genoma Humano do Câncer**. São Paulo, 2002a. Disponível em:<http://www.fapesp.br>. Acesso em: 04 jan. 2002.

FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo. **Transcript Finishing Initiative**. São Paulo, 2002b. Disponível em:<http://www.fapesp.br>. Acesso em: 08 maio 2002.

FARAH, S. B. Da célula ao DNA. In___: **DNA - segredos e mistérios.** São Paulo: Servier, 1997, p.7-36.

KAN, Z., ROUCHKA, E.C., GISH, W.R., STATES, D.J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. **Genome Research,** v.11, p.889-900, 2001.

KENDREW J. Gene. In:___. **The Encyclopedia of Molecular Biology.** 3.ed. Porto Alegre: Blackwell Science, 1999. p.343-401.

KENT, W.J., SUGNET, C.W., FUREY, T.S., ROSKIN, K.M., PRINGLE, T.H., ZAHLER, A.M., HAUSSLER, D. The human genome browser at UCSC, **Genome Research,** v.12, p.996-1006, 2002.

LANDER, E.S., LINTON, L.L., BIRREN, B., NUSBAUM, C., ZODY, M.C., BALDWIN, J., DEVON, K., DOYLE, M., FITZhUGH, W. *et al.* Human

Genome Consortium: initial sequencing and analisys of the human genome. **Nature**, v.409, n.6822, p.860-921, 2001.

LEHNINGER, A. L., NELSON, D.L., COX, M.M. Nucleotídeos e ácidos nucléicos. In___: **Princípios de bioquímica.** São Paulo: Sarvier, 1995, p.242-68

LEWIN, B. Genes are DNA. In:___. **Genes VII**. New York: Oxford University Press, 2000. p.3-35

LODISH, H., BERK, A., ZIPURSKY, S.L.; MATSUDAIRA, P., BALTIMORE, D., DARNELL, J. E. Molecular structure of genes and chromosomes. In___: **Molecular Cell Biology.** U.S.A.: Freeman, 2000, p.294-336.

MAGLOTT, D., KATZ, K.S., SICOTTE, H., PRUITT, K.D. NCBI´s LocusLink and RefSeq, **Nucleic Acids Research**., v.28, p.126-128, 2000.

NCBI - NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. ESTs: gene discovery made easier. In___: **dbEST.** EUA, 2002a. Disponível em: <http://www.ncbi.nlm.nih.gov/dbEST/dbEST/summary>. Acesso em: 25 out. 2002.

NCBI - NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **BLAST Tutorial.** EUA, 2001b. Disponível em: < http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html>. Acesso em: 17 out. 2002.

RASHIDI, H.H., BUEHLER, L.K. **Bioinformatics basis:** aplications in biological science and medicine. Flórida: CRC Press, 2000. 185p.

ROBERTS, G.C., SMITH, C.W.J. Alternative splicing: combinatorial output from the genome. **Curr. Opinion in Chemical Biology,** v. 6, p.375-83, 2002.

RUST, A.G., MONGIN, E., BIRNEY, E. Genome annotation techniques: new approaches and challenges. **Drug discovery,** v. 7, p.70-76, 2002.

SAHA, S., SPARKS, A.B., RAGO, C., AKMAEV, V., WANG. C.J., VOGELSTEIN, B. KINZLER, K.W., VELCULESCRU, V.E. Using the transcriptome to annotate the genome. **Nature Biotechnology,** v.20, p.508-12, 2002.

SIMPSON, A.J.G., DE SOUZA, S., CAMARGO, A.A., BRENTANI, R.R. Definition od the gene content of the human genome: the need for deep experimental verification. **Comp. Funct. Genom.**, v.2, p.169-75, 2001.

STRACHAN, T., READ, A P. Organização do genoma humano. In___: **Genética molecular humana**. Porto Alegre: Artmed, 2002, p.139-168.

STRAUSBERG, R.L., RIGGINS, G.J. Navigating the human transcriptome. **PNAS,** v.98, p.11837-11838, 2001.

STEIN, L. Genome annotation: from sequence to biology. **Nature**, v.2, p.493-505, 2001.

STERKY, F., LUNDEBERG, J. Sequence analysis of genes and genomes. **Journal of Biotechnology**, v. 76, p. 1-31, 2000.

UCSC GENOME BROWSER. **User Guide**, 2002. Disponível em: <http://genome.cse.ucsc.edu/>. Acesso em: 05 out. 2002.

VENTER, J.C., ADAMS, M.D., MYERS, E.W., LI, P.W., MURAL, R.J., SUTTON, G.G., SMITH, H.O., YANDELL, M., EVANS, C.A., HOLT, R.A., GOCAYNE, J.D., AMANATIDES, P., BALLEW, R.M., HUSON, D.H., WORTMAN, J.R., ZHANG, Q., KODIRA, C.D., ZHENG, X.H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P.D., ZHANG, J., *et al*. The sequence of the human genome. **Science**, v. 291, n.5507, p.1304-51, 2001.

WHEELER D.L., CHURCH D.M., LASH A.E, LEPE D.D., MEDDEN T.L., PONTIUS J.U., SCHULER G.D., SCHRIML, L.M., TATUSOVA, T.A.,

WAGNER, L., RAPP, B.A., Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v.1, p.13-16, 2002.

ANEXO I

Seqüências dos consensos dos TFs validados utilizadas para a análise preliminar *in silico* :

>CONS_TF00040
TAGGCCAGACAGCTTCTCCAGGCGCGGCCGGCCNACGTGGCTAGGGGTACAGGGCACCNCCTCCT
GGTCCAGGGGTGAGGCCAGGTACTGGGGCGTGTTCNTGNTTCGACGAAGGCCGGCTCAGCACACC
GCAGAGCTCATAGCCAGCAGCCTGGCCGTTCATGGAGAGCTCCGGGCTCATGCTTGCTCAAGTGAG
GCAGCGAGANCTTGGTGCAGTCCAGCTCCAGGTGCAGCCGGCTGGCGTCAGGCTCCAGCTCGCGG
CCCAGGGCGGCCCTTCCCACGCAGGTGCAGGGACAGGGCGTCGTCAGGCGGCACACAGGACTTGA
GCTGCAGGAGCTCCTGCTGCCGCTGGCTCTTCTCTAGCTCCACAATGCTGATCTGCAGCTCCAGGCA
GTGCCTCTGCTTCTCCGAGATCTGGCTCTTCAGGGCCTGCTTCTCCTTCAACAGCTTCTCCAGCGACA
GCGTGGCcCAGTCCAGCTGCAGCTCCTCGCAGCGAGCCTTGAGCAGCTGCAAGCTCTGGCCGCGGA
GCGCGCGGTTGTCCTGCTCCAGCTGCTCCGACTGCTCCCGCAGCTGCTGGTTATGGGCGGAGATCT
CCTTCTGCGCTTGAATCAGGTCGTTGTAGGTCAGCGCCTTCACACCCAGCTCATCCAATTTTTGCTGA
AACAGCCTCCTGATCTCCTCCTTCTGGGCCTGGCAGTGGCTGAGGAGCTGCTGAGCCGCACCCAGG
AGCTGGGCGTTCTTCTCCTTCTCCTGGCCCAGCAGCTCCTGCAGGCTGGCCTTGTACTGGGGGGTCT
TTGTGTATGCCAGGAACTGCAGGTACTGGATCTTGAAGGACTCTAGAAGCTTCTGCAGCGCGGGCGG
GGTGGGCGCCACCAGCAGCGGGTTGGGGGAGTGCCGCTGCACGCTCGGAGGTAGCTGGTAGAACG
GGCTGTGAGGGGATCTGTAGGCATCCTGGGGTGAGGAGGCCGCCGTCTGAGACACGGTCTGAGCG
TGCAGGGCATCCAGTGCAGTTTGGTTCTTCTTGGGCTTCCGCTCGGGGTTCGCAGTGTTCATCTTCTT
GGGGCGCCCGCGCTTGCGGCCAGCCATCTTCCTCCCCTTCTTGTTTAGCTTCTTCTTGCGGGCTTTG
GAGGGAGACCTCGTG

>CONS_TF00041
GTGGGTGCGGTCGATAGTGTGCAGGTAGTAGGAGACTGGCTTCCCCGTCCACGACACCGAGCCCTT
CAGGGGCGAGAGCTCCACCACGCGCATGATGGTGCCGATGTCACTCAAGTTTCTACTGTTTATTCTG
AAGTTCAGAGGTGCAAAGGGTTTCGAGGACACGATTCTGCCGGCAGAATCGTGTCCTCGAAACCCTT
TGCACCTCTGAACTTCAGAATAAACAGTAGAAACTTGAGTGACATCGGCACCATCATGCGCGTGGTG
GAGCTCTCGCCCCTGAAGGCTCGGGTCGTGGACGGGGAAGCCAGTCTCCTACTACCTGCACACTAT
CGACCGCACCCATTAAACCCCCTGGGGGAAAAAAAAAAAAAACCCCCCAAAAATTTTTTTGTTGGTTTT
TTTACCCCCCAAAAAGGGGTTTTTCCCCCTGGGGAAAAAAAAAAAAAAAAAACCCTCCAAAACTCAA
GGGATGGACAAGGATGCAATCCCTTGGTCCTGCCATCATGTCTGATGTCAATGATCAACTGCTGCCA
CGCCCACNNTAAGGGCCCAGANGGGGCAAAGGGTGGCCGGCCCCCGGAGCCCCATGGACTCTGGT
GCTGAGGAAGAGAAGGCGGGAGCAGCCACCGTGAAGAAGCCGTCTCCCTCCAAAGCCCGCAAGAAG
AAGCTAAACAAGAAGGGGAGGAAGATGGCTGGCCGCAAGCGCGGGCGCCCCAAGAAGATGAACACT
GCGAACCCCGAGCGGAAGCCCAAGAAGAACCAAACTGCACTGGATGCCCTGCACGCTCAGACCGTG
TCTCAGACGGCGGCCTCCTCACCCCAGGATGCCTACAGATCCCCTCACAGCCCGTTCTACCAGCTAC
CTCCGAGCGTGCAGCGGCACTCCCCCAACCCGCTGCTGGTGGCGCCCACCCCGCCCGCGCTGCAG
AAGCTTCTAGAGTCCTTCAAGATCCAGTACCTGCAGTTCCTGGCATACACAAAGACCCCCCAGTACAA
GGCCAGCCTGCAGGAGCTGCTGGGCCAGGAGAAGGAGAAGAACGCCCAGCTCCTGGGTGCGGCTC
AGCAGCTCCTCAGCCACTGCCAGGCCCAGAAGGAGGAGATCAGGAGGCTGTTTCAGCAAAAATTGG
ATGAGCTGGGTGTGA

>CONS_TF00156
TTTTTTTGAGTTTGAAACATAGATTAAAGTAGCAATGAGATTTTACCAGTCATCCTTATATCCAATGGGG
ATCACTAAAGTTCATTCCTTCTGTGAAATTGATTGAAGTAATGTTCGTGCCTGTTAATTTTTCCCGACAG
CAGCTCTTCCTCCTCTTCCTCGCGGGCCTTCTTCATTTCCTCCTTGGCCTGTATGGCATCCATCTCCC
TTTCTGGACCCTTGGTTTCGCCCCAGGTGGCCTTGCAAAATTATCTGCGTACTCTTCATCATCTGTGAT
CAGAAAGTTATCAAAAATGGTTCCAGATCTCACCTGGCAAAGCTCCAGGCCAATGGCACCAATGTTCT
CAAATTCTGAGAGGTCATACTGCGTCAAATAGTCGGTATTCTTCATCTTACGGTGGAGCCAGACGTCT
TTATGAATACCTTCTGGTTTCAGGCCATCCTGGTACGGGGGCTTCTGGAGCATCGGCGCTGGCCAGT
CCCCATCCAGGTCACCGTTCCAGTCGCTCTGCTTGCTGGTGCTGGCGTCCAGAAAATGCTTCTCCCA
GTCCTTCAGATGTGTAGAGATGTGGCTGGTCTGCAGTGTGTGCTCTGGGGAGGGCATTTGGTTCTAT
AATATTGCCATCGTGGCTACCAGCCATGCAGCAATTGTCCAGTAACTGGGTACTTACAAACTACATTT

GTGCACTTTTTAGAGTTGCCCAGGAAAAATTCACACCTGGGCTTTGTTGTCTTTAGTCTGTTCCCAATC
CTTCGATTCTGCCGGGGACGTTTCCTTCTTGAGTGATGTTAAGTTCCAGTCGTACTCTATGCTGCCGG
ATTCAATTGACTGACCATCAATTTTCACATCATAAGAAAGATCTGGTCTTAAAATTAGAGTGTACAGGT
GTGTGAAGCCATCAACCTTACACCTGATCAGTTTCTTGTTTTCGTGATACTTATTCTTGAAATGTAAAAT
AACATGAACTTTCTTGATATCAAATCCACAAATATCGGGTCCTTTATCTTTCTCTTTATGACCATAAAAC
TTGCCCGACGAAAGTCTAAAATGCCCAAATCGGGAGTCATTGGTGGACTGCAACCATCGGTTTCTCCA
ATGCTCTCCGTCTAGAAATTCCTCTTGGAAATAGACGGTAGCCAGCGCCACTCGCAGCATGCATATG
GCCCAGAGCTGGACCAAAG

>CONS_TF00157
AGCTGCGTACTCCAACTTCTGGTGTGGAATGACTTCCCAGGGCTACTACAAGCGCACACCTGCCTAC
ATCCCTATTCGCTGCGAGACCGAGGGCTGCTTTCAGTTCCCATGGTGCACTCGACCTTCCTGATCGA
CCTGCGGAAGGCGGCGTCCAGGAACCTGGCCTTCTACCCACCTCACCCTGACTACACCTGGTCCTTT
GACGACATCATCGTCTTTGCCTTCTCCTGCAAGCAGGCAGAGGTTCAGATGTATGTGTGCAACAAGG
AGGAGTACGGATTCTTTGCCAGTGCCATTGCGCGCCCACAGCACCCTCCAGGATGAGGCCGAGAGC
TTCATGCATGTGCAGCTGGAGGTCATGGTGAAGCACCCGCCCGCAGAGCCCTCCCGCTTCATCTCG
GCTCCCACCAAGACACCGGACAAGATGGGCTTCGACGAGGTCTTCATGATCAACCTGAGGCGGCGG
CAGGACCGGCGGGAGCGCATGCTGCGGGCGCTGCAGGCACAGGAGATCGAGTGCCGGCTGGTGGA
GGCCGTGGACGGCAAAGCCATGAACACCAGCCAGGTGGAGGCGCTGGGGATCCAGATGCTGCCTG
GCTACCGGGACCCCTACCACGGCCGGCCCCTCACCAAGGGTGAGCTGGGCTGCTTCCTGAGCCACT
ACAACATCTGGAAGGAGGTGGTGGACCGGGGGCTGCAGAAATCGCTTGTNTTTGAGGATGACCTGC
GTTTTGAGATCTTCTTCAAGAGACGTCTGATGAACCTCATGCGGGAGATGAAGCGGGAGGGCTCTGC
GGGCGGGTGCTTCANCATGACCTNCAGCTGCACATGCATGAAGCTCTCGGCCTCATCCTGGAGGGT
GCTGTNGGCGCGCAATGGCACTNGCAA

>CONS_TF00194
GAACTAGGTTTACCATCCCACTGAGGGAGAACTGGGCGCTGCCATTCTTTGCAATTCAGATAGCAGC
AATTACATATTTCCTGAGACCAAACTTACAGCCTCTTTCTGAAAGGCTGACACTTCTTGCCATTTTCATA
TCAACTTTTCTCTTTAGTCTGACATGGCAATTTAATCAATTTATGATGCTGATGCAAGCATTAGTGCTGT
TCACACTGGACTCCCTGGACATGCTGCCAGCAGTGAAGGCGACATGGCTGTATGGAATACAGATAAC
AAGTTTACTCCTGGTCTGCATTCTTCAGGTTTTTAATTCCATGATTCTTGGATCACTGCTTATCAGTTTT
AACCTTTCAGTATTCATTGCAAGAAAACTTCAGAAAAATCTGAAAACTGGAAGCTTCCTTAATAGGCTT
GGGAAACTTTTGTTACATTTATTTATGGTTTTATGTTTGACACTTTTTCTCAACAACATAATTAAGAAAAT
TCTTAACCTGAAGTCAGATGAACACATATTTAAATTTCTGAAGGCAAAATTTGGGCTTGGAGCAACAAG
GGATTTTGATGCAAATCTCTATCTGTGTGAAGAAGCTTTTGGCCTCCTGCCTTTTAATACATTTGGAAG
GCTTTCAGATACTCTGCTTTTTTTATGCTTACATATTCGTTCTGTCCATCACAGTGATTGTAGCATTCGTT
GTTGCCTTTCATAATCTCAGTGATTCTACAAATCAACAATCCGTGGGTAAAATGGAAAAAGGCACAGTT
GACCTGAAACCAGAAACTGCCTACAACTTAATACATACCATTCTGTTTGGATTCTTGGCATTGAGTACA
ATGAGAATGAAGTACCTCTGGACGTCACACATGTGTGTGTTCGCATCATTCGGCCTATGTAGCCCTGA
AATATGGGAGTTACTTCTGAAGTCAGTCCATCTTTATAACCCAAAGAGGATATGTATAATGCGATATTC
AGTACCGATATTAATACTGCTGTATCTATGCTATAAGAACCAGAAGTCCTGACACCTGATTTCCCATCA
CTAGCAATTTTCCTGATTCACCCACCCAGGAGACAAGATTTGAATGAGCAGTAAAAATGGCCAAAGAT
GAGATGACCAAAAAAACAGTGATAGGTCTCAAACACAGCCAGAGATCAATCAGGTGCTGCTTTGATTC
TACTAGTGGTTCTTAAATAAAAGTATTATATTTTCTACGTCAGTGGAGCATACATACATTGTCATTGGTC
TTCTATGCTAATATGTGAAGTGAATTCTACCTTTGACCTTAGAATGTATATAGATATGATCAAGTCTTTT
TAGTCAACTGTCATTTGATAAAAACAATTAAGATTTAGTTAATTGTTGAATTAAATGGACTTAAGATATTA
GATAAGTGGGTAATTCAGATTGTAACA

>CONS_TF00232
GCGACCGCTCGCGTCATTGGGAGAGCTGGGGCGCGTGCCTGAACTTCCCGGCTGCCCCTGTCCTTG
GAGACCTACCTGATGGGGACGCCAGGTGTGCAGGGGCGTGGCGCGTAGGAGTGATTTGGAGAACAA
TGCATGTAAGTCTGACATCATGATGTCCATCCGGCAAAGAAGAGAAATAAGAGCCACAGAAGTTTCTG
AAGACTTTCCAGCCCAAGAAGAAAATGTGAAGTTGGAAAATAAATTGCCATCTGGTTGTACCAGTAGA
AGATTATGGAAGATTTTGTCATTGACAATTGGTGGAACATTGCCCTTTGCATTGGACTTCTTACATCTG
TCTACCTTGCCACGTTACATGAAAATGATTTATGGTTTTCTAATATTAAGGAAGTGGAGCGAGAAATCT
CATTCAGAACAGAGTGTGGCCTGTATTACTCCTACTACAAGCAGATGCTGCAGGCTCCAACCCTCGTG
CAAGGTTTTCATGGCCTAATATATGATAATAAAACTGAATCTATGAAGACAATTAACCTCCTTCAGCGA
ATGAATATTTACCAAGAGGTTTTTCTCAGTATTTTATATAGAGTTCTACCCATACAGAAATATTTAGAGC
CAGTTTATTTTTATATTTACACCTTATTTGGGCTCCAGGCGATCTATGTCACAGCTCTCTACATAACCA
GCTGGCTACTCAGTGGTACATGGCTGTCAGGACTGTTGGCAGCTTTCTGGTATGTCACAAATAGATAC
CACAAGAGTTGAGTTTACCATCCCACTGAGGGAGAACTGGGCGCTGCCATTCTTTGCAATTCAGATAG
CAGCAATTACATATTTCCTGAGACCAAACTTACAGCCTCTTTCTGAAAGGCTGACACTTCTTGCCATTT
TCATATCAACTTTTCTCTTTAGTCTGACATGGCAATTTAATCAATTTATGATGCTGATGCAAGCATTAGT
GCTGTTCACACTGGACTCCCTGGACATGCTGCCAGCAGTGAAGGCGACATGGCTGTATGGAATACAG

ATAACAAGTTTACTCCTGGTCTGCATTCTTCAGTTTTTTAATTCCATGATTCTTGGATCACTGCTTATCA
GTTTTAACCTTTCAGTATTCATTGCAAGAAAACTTCAGAAAAATCTGAAAACTGGAAGCTTCCTTAATAG
GCTTGGGAAACTTTTGTTACATTTATTTATGG

>CONS_TF00308
TTTGCGCATCTCATAGAAGCTTTTAATAGTTCATATTTACTAAAGAGTAGGAATACAGAGCGATGAAGA
TGAGCTGGAAACGACAGGTGACTTGCCAGCAGGCCAGAATGTGCTTTTTCTTTGTCCCATGGAAGGT
GTTAATTCTCTCTCCAGTTGTGAGGATCAGTTGGTTCATTTATGGGAAGGTTGTCAGGGGACCTTTGA
ATCACGGCCTTCAGATGCCACAAGGAATCCCACACAGGCCAGTGGATCACGTGCATGCATTTCTCTC
CCTTCTCGACTCAGGAAGCTTAAGATTTACTAGTGTTCAAAACATGTGAAGTAGCCAAACATCTCCTGA
CTGCAATGCCAGCCAGACTGTGTGGAAACTCGGTTCATACCAGCCGTTCTAGGGGTGATGCGAGTTG
TCATCATCCTTAGGAAAGTGTGTTGTTGTAGGATCAACCCATCCTTCAAAAGGACTGTGCCTGTTTATA
AGCTCAGCTGTTTCTGCCCTGTGAAATATGGCAAGGATATTAATTCCAGGAGAACAGAGCTTTATGAT
AAAAGATGCCCAATGAAGCATGAATTAGGGACATACTGAAAATGGGTAAGGAAATTGTCAACTCAGAA
CCCAGCAGGCATTAAGTAAAAGAGGAGGAAGCATTACAGCAACAGTTTTGATCATACTGTACTTTTATA
GCCATGTGAAATACATTTTCTATGTATAGATAGATTGTGTAAGGGTACAATTGTGAGGACAACAGGAAC
ATGGCAGATTATTTAAAATCATACTAAAGATGATGCTTTGTCTGATGAAAGTGATTCTAAACCATAGATA
AACGATTCAAGACAGACAAGAGCTgcagcagttggtagaggACTccTaCagAtGaGAGCTCTGTGCTCaagcccttg
cagagggAGAtggcagagaggaaggcTggccggcaagCAGCACAGGTCAATGTGGCTACAGGgaagccTCATCCTTT
CTCAGAATGGCCCTACTTGCCCGATGTCATGGCTGGCCCTTCAGGACCATTGATGGGCTGccAGCCG
CCTCCTCTACCTGGGTGTTGTCTGGGAACTCAAACACTCCCTCCATCTGAAGGTTTTCTGGGACCTCA
ACAAACTCCTCTACCCATTGTGGCCTGTAAGACCTCAACCACCAACTCCACATGTACCACTTGGATAA
TTTTGAAATTTGAAGAGTCTCCTGGGCCATGATATCTTGGAAAAGTCTCCACCATATTTGATGCTCCAT
GCCCCGTGCTCCCTCCTCTCTCTCTCATTTCCAGTGTTTCCCTTTAGAAACATTAGCAGTCCTAAA

>CONS_TF00309
GCACCCTTTTCCTCGTGCAGCCCAGCCTGACTCCTGGAGATTGTGAATAGCTCCATCCAGCCTGAGA
AACAAGCCGGGTGGCTGAGCCAGGCTGTGCACGGAGCGCCTGACGGGCCCAACAGACCCATGCTG
CATCCAGAGACCTCCCCTGGCCGGGGGCATCTCCTGGCTGTGCTCCTGGCCCTCCTTGGCACCACC
TGGGCAGAGGTGTGGCCACCCCAGCTGCAGGAGCAGGCTCCGATGGCCGGAGCCCTGAACAGGAA
GGAGAGTTTCTTGCTCCTCTCCCTGCACAACCGCCTGCGCAGCTGGGTCCAGCCCCCTGCGGCTGA
CATGCGGAGGCTGGACTGGAGTGACAGCCTGGCCCAACTGGCTCAAGCCAGGGCAGCCCTCTGTGG
AACCCCAACCCCGAGCCTGGCGTCCGGCCCGTGGCGCACCCTGCAAGTGGGCTGGAACATGCAGCT
GCTGCCCGCGGGCTTGGCGTCCTTTGTCGAAGTGGTCAGCCTATGGTTTGCAGAGGGGCAGCGGTA
CAGCCACGCGGCAGGAGAGTGTGCTCGCAACGCCACCTGCACCCACTACACGCAGCTCGTGTGGGC
CACCTCAAGCCAGCTGGGCTGTGGGCGGCACCTGTGCTCTGCAGGCCAGGCAGCGATAGAAGCCTT
TGTCTGTGCCTACTCCCCCAGAGGCAACTGGGAGGTCAACGGGAAGACAATCGTCCCCTATAAGAAG
GGTGCCTGGTGTTCGCTCTGCACAGCCAGTGTCTCAGGCTTGTTTCAAAGCCTGGGACCATGCAGGG
GGGCTCTGTGAGGTCCCCAGGGAATCCTTGTCGCATGAGCTGCCAGAACCACGGACGTCTCAACATC
AGCACCTGCCACTGCCACTGTCCCCCTGGCTACACGGGCAGATACTGCCAAGTGAGGTGCAGCCTG
CAGTGTGTGCACGGCCGGTTCCGGGAGGAGGAGTGCTCGTGCGTCTGTGACATCGGCTACGGGGA
GCCCAGTGCGCCACCAAGGTGCATTTTCCCTTCCACACCTGTGACCTGAGGATCGACGGATACTGCT
TCATGGTGTCTTCAAGAGGCAGACACCTATTACAGAGCCAGATGAAATGTCAGAGGAAAGCGGGGTG
CTGGCCCAGATCAAGAGCCACAAAGTGCAANGACATCCTCGCCTTCTATCTGGGCCGCCTGGAGACC
ACCAACGAGGTGATTGACAGTGACTTCTAGACCAGAACTTCTGGATCGGGCTCACCTACAAGACCGC
CAGGACTCCTTCCGCTGGGCCACAGGGGAGCACCAGGCCTTCACCAGTTTTGGCTTTGGGCAGCCT
GACAACCACGGGTTTGGCAACTGCGTGGAGCTGCAGGCTTC

>CONS_TF00380
GCAAAGAGCTGCGAGATGGTTTTTCTGGTGTCCTCTGACCCCTCATATTCCTTCTCTGCAAGCTGCTT
GTTGGCAGTCTCTAGACGCTCTCTCAGATCCCTGTTGAAATCATGCATCCTCCGAATCTCGCCCTCTA
GCTTGTTTCTCATGGCTTTCTCTAGGGCCTCTCTTTTGGAGGATGACTTCCNGAGGATCTCATATGCC
TCCGAGACGCGCTGGATTTCTGTGTCCACCTTCTGCAGTCTTGCCACCTTCTCATAGCATCCTTCCAA
CTCTTGCCTCAAGTTCCGGTTCTCGTCTGAGAGGATCTCAACCATCTGCTGGGCTCTGGAAACAATGG
CAAAAGGGTCTGCTGGCACTGGCTGATAAGAAGCAGAGGATGGCTGAGCCCGAGGCATAGCTGAAT
AGGCTCCTCCTGTGTGCTGTCCAAGCTAGCCCACATCCTCTGTGAGCCGATGTACGCACAGCGTTGT
GACTCCTCCTCTACGCATGTCCCACTCC

>CONS_TF00404
GGTTATTTTCCTCTTTAAACTCTGATCCTTTGCTCTGGAGAAGCCAGGCACCATGGCATGAGGCAGCC
CTGTGGAGAGGTTCATGtGgtacAGGACCAggcctgcagcatcacTTGagagagGCAACATAACTCTTTTACTGA
CTATCACAGATAGGACCACAAAACAGCCTTGTAAACTCATCCTAGAGGATCTAAAAGAGAACCCCCTC

ATAACAGAATGGCATTCCCAGCTTGCAGGCTCTGAGACTCCATTGAGTCATGCCACATGTATGGAAAA
TGATTTTTggAGtgttATTCTATCTAAGGAAAAACAAAGCAAATTTTCCACTTTCAAGATTAAAGGACATTC
TGTCTACATTAAGACAaGCaggtACAAAGAAGTAAGTTACTGTTGGATTAAGAATATCCAAGACAATGGT
CACAACTTGAAAGCACATGGAAAAACATGAATTATTTTTATAATGTCAGCTAATCCACAGCTGTGCCAC
CTGTTAGTATGGATGCTATCCAAAAGGATTTTTgaaacCTTGTTATCCAAATGCAGAAATGGCAGCTGAA
GAGATGATTTGTGACAGCATTTGCTGAGACgaaGTCTGTcCACTAGATGGATAAAATCAGTTCCCTTTT
CATAGCTTGATCATCTACATCTGCTAGTGACCCAGTCCCTGAGAGtcaCCAAGAATTGATTCCACTGAG
GCGATGGCAGGCGGGGCTTTGGTGTTGGAAAAGTCAGTGGTTGATATGGACAGCACAGACTCACATA
AGCTGctGGTTTATTACACTAGAAAAATAAGggctcCCTTGACCCAGCTCGGCAAGTGTgaaataacgaaATA
AATGAAGGTTAACAAAGACTGGTCTAAGGAATAGTTCACTACAATTACGCAATGCTCTTCTCACTAATA
CTCGACTCTGCATTTCGGTACTTATATCTGAAAGATTTGCAAATAATGACAACATTCGTGCCAGACATT
ACTGCAAGTGAGGAACAGACTAAAGGAGAGGAAGGAACAAAATGCCTTCCAAGAGAGAAACATTCAT
CTACCTTACTGGCAAGTGTCTAAAGCAAACAAGGAAACTGCAAAATAATTACTTAGTTACAAATAAACG
CTTTTAGCACAGT

ANEXO 2

Seqüências de cDNAs parciais (regiões de alta qualidade) que não alinharam no genômico dos TFs, informado pela coordenação do projeto TFI, utilizadas para a análise preliminar *in silico* :

TF00040

Foram selecionadas 3 seqüências que não apresentaram redundância entre as 8 seqüências submetidas ao TFI. (Analisadas: 08 - Selecionadas: 03)

>TF40-A05R

Seqüências: A04F = A05R

GTACAAGGCCAGCCTGCTGAGGATGGTGTCATCAATGCAGCCACCTAGCAGCGGCAGGAATGGCACAG
TCCTTAAGATCCGTACCATCTCCTTGACCTTGGGCTCTTCACTGACCAGCAGCATGTTGTGCCCCACAAA
AAGGGGCAGCAGATTTTGGTACTTGGAATCCTCCAGGAAGGGCTTCAGGACCCGGTTGGGGAAGACCT
TCATCAGGATCTTGTGTCTCCGCAGCTGGTGTCGCATAAGAAGCTTGTCCTCTGCACTCAGAGCCACATT
CTGGCAGACGGCTATCATTCGGTTGTCCTGGAAGACTGCTGCTATCTCCCGGCGGAGAAGCAGAGGCAC
T

>TF40-F11R

Seqüências: F11F = F11R

TGCCTCTGCTTCTCCGGGGCGCAGCGTGACGGCTGGCATCATCATTGTTGGAGATGAGATCCTTAAGGT
GTGTCTGGGACAGAAAAGGGGGGAGGGCGCTGCGTTCTCCTGTCCTTAAGGGCCTGCTGCACATCCCTC
CATGGAAGGAGGTGAAACAGGGTGGGAGCCTCTTTGCTGCAGTAGGATCCTGGAGTGAATCCAGCATA
TTGGAGGAAAATTAACCCTCATTCTTTCAATTACTCTACTGAAAAACTTGTAGCAAAGTCCCTACAATTT
TTCAACCTGAGAGAGCTGATTGAAAAAAAGAAAAATAAAAAGGTCATTGCAATTTACCAAGCAGTATG
GTAGAGGGCAGAGGAGCTACACAACAATGACAGGAAGCCCCTCGAGTAGAGGAGGTACAGGTGTGTG
TATTGTATATTTAAGAGTTT

>TF40-G02R

Seqüências: A02F = D02F = H02R = G02R
TGCCTCTGCTTCTCCGCGCCTGCACGGGTATTTAATGTGGCGCACTGGCCAGGCTCCTCTCCTGCCCCGC
CCAGCAGCAGCACGGCCCTTGCCTTTCTTCAGTGCCCTCCAAACTGGTCAGAGAGGAGGCACAGACTGC
TAGAAGAAGGGCAAGAGAAGTCGTAATTCACAGTCATTTGCTACAGAAAGTGCCTGGAATCGGTCCTT
GGCTGCTTCTGCCTGGAAGCAACCCACTTAAGCCCCAGGAGAGGGGACAGATGAGATACTGGGATGGG
GGTGGGAGTCCTCAGAAGGCACCCAGGAGACACTGGCTGCAGAAATGGGGTCCCTGATCTTGCAAGAA
AATCCGCTTGCTGTTGGCAGGCTGGCCTTGTAC

TF00072

Foram selecionadas 2 seqüências que não apresentaram redundância entre as 7 seqüências submetidas ao TFI. (Analisadas: 07 - Selecionadas: 02)

>TF72-A01F

Seqüência: A01F

GGTGGTCCTCCTGCTGGAGAAAAGCCTGGAACTGCCAAAACAAAAAAACAAATACAAACAAACCTCGT
CACCACGGTCAGGAGGAGGTTCAGACATTCTGCCCAGAGGGGCTGGGGAAAGGGATGGGAGCCAGGA
AAGATTTAGGGGCAGGAGACAGACACGTAACAGGAACTCGGAAGGTAGAACTTAGAATTCAGAACTTG
GGTTTCAATTGTAGAATTTGGAATCTGGAATACAGAATTGGGAAGAAGAACAGAGATGACAAAAGACC
ACTTGAAGAG

>TF72-D01F

Seqüências: D01F = F01R = H01R = B01F = C01F = G01R

AGTGGAACCTGGATGCTGGTAGGGCCAGAGACAGAGGCTTAACACCCTGCTGGGGAACCCGGTCAG
AACTCCCGAGGCAGGAGAGGTTCTGCTCCACTGGATGTTTGTCTTGGTGTTTTTGGATGTGCTGATCA
AGAGCAAGATGTTCTGGATTCTTAAAACTCCCCTCACAAGGACCAATCTAGAGATAATTTATTGATCAG
TGATCACAGCTTGTACCCCAAAGCCGTGTATGTCTGGATCCTTCCCTAAGACCACAGATAGCTCCAGG
GAGTCCCACCTCCTTGGCTATGGAAATATGCTCAGCCCTGGTTTCAGAGAAGCCTGGACTCCACTCT
GGACCCCATGAGATGATATGCGCTGGTACTCCAGGCTTTAAAT

TF00074

Foi selecionada 1 seqüência das 4 seqüências redundantes submetidas ao TFI. (Analisadas: 04 - Selecionada: 01)

>TF74-H02R

Seqüências: G02R = B02F = A02F = H02R

GGAAAACCACGTGACCTCTTCTGCCTTCACTGGGCTGGGGTGATCCTTGGTGCCTTTGTTTCCACAAGGC
CTTTTCCTGCCCCCTGCCTTGCCAAAGACATTTAATCAGCACACAGCTGCCAGACTATTCCCACAGTGCT
CCAAATGCACATGAACAACAGTGACGGCTCCAGCCTTCGACCCAGAGCCCCGTGCCCAGTGCGTCAGT
GGGCCTGGGGTTCCAGGCTACATCAAGCACTGATGGTGTCAGGGCTGGTAGTTACCAAATCAGGGTTA
AGAAACATCAGGGCCATATTTCACTCAGGGGA

TF00193

Foram selecionada 2 seqüências não redundantes das 4 seqüências submetidas ao TFI. (Analisadas: 04 - Selecionada: 02)

>TF193-A06R

Seqüências: A06R = A05R

ATTACCCATGAGGCCTGGAGCTCTACTTTCTTGTGGTGCGCGTGTTCCAAGGCCAGGGCCTGAGTACGC
GCTGGCTCTGCCTGATCGGCTATGGCGTGCCCCTGCTCATCGTGGGCGTCTCGGCTGCCATCTACAGCA
AGGGCTACGGCCGCCCCAGATACTGCTGGTTGGACTTTGAGCAGGGCTTCCTCTGGAGCTTCTTGGGAC
CTGTGACCTTCATCATTTTGTGCAATGCTGTCATTTTCGTGACTACCGTCTGGAAGCTCACTCAGAAGTT

TTCTGAAATCAATCCAGACATGAAGAAATTAAAGAAGGCGAGGGCGCTGACCATCACGGCCATCGCGC
AGCTCTTCCTGTTGAGCTGCACCTGGGTCTTTGGCCTGTTCATCTTCGACGATCGGAGC

>TF193-D07R

Seqüências: D07R

**GTTACTCCAATGTAGGTATATTCGTCACGTTTTTATAACCATAACGACGGAGCGGATATGAAAAAGT
TAACAGTGGCGGCTTTGGCAGTAACAACTCTTCTCTCTGGCAGTGCCTTTGCGCATGAAGCAGGCG
AATTTTTTATGCGTGCAGGTTCTGCAACCGTACGTCCAACAGAAGGTGCTGGTGGTACGTTAGGAA
GTCTGGGTGGATTCAGCGTGACCAATAACACGCAACTGGGCCTTACGTTTACTTATATGGCGACCG
ACAACATTGGTGTGGAATTACTGGCAGCGACGCCGTTCCGCCATAAAATCGGCACCCGGGCGACC
GGCGATATTGCAACCGTTCATCATCTGCCACCAACACTGATGGCGCAGTGGTATTTTGGTGATGCC
AGCAGCAAATTCCGTCCTTACGTTGGGGCAGTATTAACTACACCACCTTCTTTGATAATGGATTTAA
CGATC**

TF00408

Foi selecionada 1 seqüência não redundante das 3 seqüências submetidas ao TFI. (Analisadas: 03 - Selecionada: 01)

>TF408-A02F

Seqüências: A02F = D07F = A02R

AGAGGGGTGTAGATTGGACAGGAAGCAGCTGAGCTGAGGGAGAAGTTACAATCAGAAATGGAGAAAA
ACGCCCAGATAGTAAAGACCCTGAAGGAAGATTGTGAATCTGAAAAAGATTTATGTTTAGAAAATCTA
CGCAAAGAACTGTCTGCAAAGCATCAATCAGAAATGGAGGATTTACAAAACCAGTTTCAGAAAGAATT
GGCAGAACAGGGAGCTGAGTTGGAGAAGATTTTTCAAGACAAAAACCAGGCTGAACGGGCCCTTAGGA
ACCTGGAGAGTCATCATCAAGCAGCCATTGAGAAGTTACGTGAAGACCTGCAGTCCGAGCACGGCCGG
TGTTTAGAAGACTTGGAGTTCAAGTTCAAAGAGAGCGAGAAAGAAAAACAGCTGGAGTTAGAGAATCT
TCAAGCATCATATGAAGACCTGAAGGCACAATCACAAGAAGAGATCAGGCGCTTGCGGTCCCAGCTTG
ATTCTGCCAGGACCAGTAGACAGGAATTGAGTGAGCTACATGAGC