

**AVALIAÇÃO DAS ORESTES *NO MATCH* GERADAS
PELO PROJETO GENOMA HUMANO DO CÂNCER
(LICR/FAPESP-HCGP)**

ROGÉRIO DA SILVA FONSECA

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

Área de concentração: Oncologia

Orientadora: Dra. Helena Paula Brentani

São Paulo

2005



FICHA CATALOGRÁFICA

**Preparada pela Biblioteca do Centro de Tratamento e Pesquisa
Hospital do Câncer A.C. Camargo**

Fonseca, Rogério da Silva

Avaliação das *ORESTES NO MATCH* geradas pelo Projeto Genoma Humano do Câncer (LICR/FAPESP-HCGP) – São Paulo, 2005.

44p.

Dissertação(mestrado)-Fundação Antônio Prudente.

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

Orientadora: Helena Paula Brentani

Descritores: 1. ESTS. 2. MRNA/análise. 3. BIOINFORMÁTICA.

Esperamos a luz, mas contemplamos a escuridão.

Isaías 59:9

É melhor acender uma vela do que praguejar contra a escuridão.

Adágio

DEDICATÓRIA

Dedico este trabalho à três Marias:

À **Maria Emília**, quem me gerou e criou, me transmitiu bons valores e junto a meu pai, mesmo diante à dificuldades, sempre me deu suporte para a continuidade dos estudos.

À **Maria Valéria**, quem me tirou da rotina de estudos e me manteve ocupado entre tapas e beijos.

E por último, mas não menos importante, à **Maria Lídia**. Quem ao longo de cinco anos, com amor e profissionalismo, tem me ajudado a lidar com as marés.

AGRADECIMENTOS

AGRADEÇO:

...ao **Dr. Saul G. Jachieri** (*in memorian*), com quem convivi por um curto período, e que, ao me aceitar como aluno, mostrou-me que devemos crer sempre em um futuro melhor, mesmo quando vemos tempestades a nossa frente.

...ao **Dr. Luiz Fernando Lima Reis**, por ser um exemplo de homem de ciência.

...à **Dra. Helena Brentani**, por me acolher quando “órfão” e por presentear, a mim e aos que vivem ao seu redor, com bom humor e belos sorrisos.

...à **CAPES**, à **FAP** e ao **Instituto Ludwig** por todo suporte humano e material.

...à todos que me ajudaram por amizade ou companheirismo

...e também, a todos que, não podendo ou não querendo me ajudar, não me atrapalharam.

RESUMO

Fonseca RS. **Avaliação das ORESTES *NO-MATCH* geradas pelo Projeto Genoma Humano do Câncer (LICR/FAPESP-HCGP)**. São Paulo; 2005. [Dissertação de Mestrado-Fundação Antonio Prudente].

O Projeto Genoma Humano do Câncer gerou aproximadamente um milhão de Etiquetas de Sequências Expressas (*ESTs* - *Expressed Sequence Tags*) pela metodologia ORESTES, almejando principalmente obter dados sobre o câncer. Deste total, 341.680 não mostraram similaridade com sequências depositadas nos bancos de dados públicos de transcritos, sendo por isto chamadas de “sem-pares” ou “*no-match*”. É plausível que uma parte deste conjunto de dados represente transcritos humanos de baixa expressão ou de difícil detecção. Neste trabalho, criamos um protocolo de bioinformática para identificação de sequências *no-match* que com maior probabilidade representem transcritos humanos. Este foi aplicado a 14.908 agrupamentos de sequências, os quais continham ao menos uma sequência *no-match* de próstata ou mama. Os critérios de seleção do protocolo foram: 1) Usar apenas sequências com alta identidade com o genoma humano; 2) *clusters* que contenham sequências provindas de diferentes bibliotecas; 3) *clusters* que apresentem sequências com *splicing* e 4) destes, apenas os que apresentam sítios canônicos de *splicing* em suas sequências. O total de 2.226 *clusters* passou pelos dois primeiros critérios, sendo que 8.4% deste montante possuía sequências com *splicing* quando alinhadas contra o genoma. Deste percentual, 94 *clusters* possuíam sequências com sítios canônicos de *splicing* e foram então submetidos à inspeção manual para confirmação de seus alinhamentos, o que indicou que sessenta destes *clusters* possuíam sequências novas, dos quais oito mostraram alinhamento sobre genes preditos *ab initio* sendo selecionados para validação experimental. RT-PCR seguido de sequenciamento foi feito para validar a maior sequência com *splicing* de cada *cluster* selecionado, resultando no total de cinco sequências confirmadas como verdadeiros fragmentos de transcritos humanos. Dois destes mostraram-se diferencialmente expressos entre tecidos normal e tumoral em análise de expressão *in silico*.

SUMMARY

Fonseca RS. [Mining ORESTES no-match database] São Paulo; 2005. [Dissertação de Mestrado-Fundação Antonio Prudente].

The Human Cancer Genome Project generated about 1 million Expressed Sequence Tags (ESTs) by the ORESTES methodology, principally aiming to obtain data from cancer. Of this total, 341,680 show no similarity with sequences in the public transcript databases, therefore referred to 'no-match'. Some of these sequences may represent low abundance or hard to detect human transcripts. Hereby, we create a bioinformatics pipeline to identify no-match sequences, which may represent human transcripts, and apply it with 14,908 sequence clusters, which contains at least one no-match sequence derived from prostate or breast tissues. The selection criteria of such pipeline were: 1) sequences with high similarity to the human genome; 2) clusters originating from different libraries; 3) spliced clusters; and 4) clusters whose spliced sequences show canonical splicing sites. A total of 2,226 clusters passed by the two first criteria and 8.4% of these are spliced upon genomic alignment. Of this remainder, 94 show canonical splice sites, and were subjected to a manual inspection regarding genomic hits. Six of these showed to be clusters with new sequences and eight which align in *ab initio* predicted genes, were selected for experimental validation. RT-PCR followed by sequencing was performed to validate the largest spliced sequences of each selected cluster, resulting in the confirmation of five sequences as true human transcript fragments. Two of them showed differentially expressed between tumor and normal tissue by *in silico* analysis.

LISTA DE FIGURAS

- Figura 1** Fluxograma dos procedimentos com dados de próstata e mama.
- Figura 2** Esquema demonstrando as três etapas da confirmação do transcrito (gel com produto de PCR / seqüenciamento / interface gráfica mostrando alinhamento BLAT sobre genoma humano).
- Figura 3** Esquema mostrando os tipos de alinhamentos encontrados na inspeção manual.
- Figura 4** Fluxograma dos procedimentos com o conjunto de dados do restante das *no-match*.
- Figura 5** Distribuição por tamanho dos *clusters no-match* sem *splicing*.
- Figura 6** Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* de tamanho três.
- Figura 7** Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* de tamanho quatro.
- Figura 8** Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* com tamanho de 5 a 10.
- Figura 9** Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* com tamanho dez ou maior.

LISTA DE TABELAS

- Tabela 1** Sequência dos *primers* desenhados para alinharem com os segmentos contíguos das ORESTES selecionadas.
- Tabela 2** Tamanhos esperados (em pares de bases) dos produtos de amplificação dos pares de *primers* desenhado para cada ORESTES selecionada.
- Tabela 3** *Clusters* selecionados do conjunto de dados de próstata e suas respectivas ORESTES levadas à validação experimental.
- Tabela 4** *Clusters* selecionados do conjunto de dados de mama e suas respectivas ORESTES levadas à validação experimental.
- Tabela 5** Perfil dos alinhamentos genômicos da inspeção manual das 94 ORESTES resultantes do *pipeline*.
- Tabela 6** Informações sobre os cinco RNAm associados às seqüências validadas.

LISTA DE ABREVIATURAS

BLAT	Basic Local Alignment Tool
BLAST	Basic Local Alignment Search Tool
CGAP	Cancer Genome Anatomy Project
ESTs	Expressed Sequence Tags
HCGP	Human Cancer Genome Project
IMAGE	Integrated Molecular Analysis of Genomes
MGI	Merck Gene Index
RefSeq	Reference Sequences
RITE	Regions of Increased Tumor Expression
SAGE	Serial Analysis of Gene Expression
TFU	Transcript Fragments Units

ÍNDICE

1	INTRODUÇÃO	1
1.1	As ORESTES do HCGP	8
2	OBJETIVOS	11
3	MATERIAIS E MÉTODOS	12
3.1	Procedimentos Computacionais	12
3.2	Validação Experimental	15
3.2.1	Extração de RNA	15
3.2.2	Avaliação da Quantidade de RNA	16
3.2.3	Produção de cDNA e RT-PCR	17
3.2	Análise de Expressão <i>in silico</i>	20
4	RESULTADOS	21
5	DISCUSSÃO	28
6	CONCLUSÕES	39
7	REFERÊNCIAS BIBLIOGRÁFICAS	40

ANEXOS

Anexo I Géis de poliacrilamida mostrando os produtos de PCR das ORESTES selecionadas.

Anexo II Lista com as informações da inspeção manual dos 94 clusters selecionados pelo pipeline.

Anexo III Interfaces gráficas do UCSC Genome Browser mostrando o alinhamento das ORESTES originais e da sequência do produto da amplificação por *primers* específicos desenhados para confirmação experimental.

Anexo IV Tabela com o resultado da busca de expressão *in silico*.

1 INTRODUÇÃO

Um dos passos para o entendimento das bases genéticas de um organismo consiste na completa determinação de todos os genes de seu genoma. Neste sentido, empreendimentos como os projetos genomas começaram a surgir na década de 1980. A concepção de tais projetos só foi possível graças ao advento da tecnologia de seqüenciamento (MAXAM e GILBERT 1977; SANGER et al. 1977) o que veio permitir o conhecimento da seqüência completa dos 16,5 Kb do genoma mitocondrial humano (ANDERSON et al. 1981) e dos 40 Kb do genoma do bacteriófago Lambda (SANGER et al. 1982). As primeiras experiências em seqüenciamento de grandes genomas começaram a ser feitas na década de 1990 a partir do desenvolvimento dos seqüenciadores automáticos de DNA. Exemplos de organismos cujo genoma está seqüenciado são os procariontes *Haemophilus influenzae* (FLEICHMANN et al. 1995), *Mycoplasma genitalium* (FRASER et al. 1995) e *Xillela fastidiosa* (SIMPSON 2000), o microorganismo eucarionte *Saccharomyces cerevisiae* (GOFFEAU et al. 1996), o nematódeo *Caenorhabditis elegans* (The *C. Elegans* Sequencing Consortium, 1998; BERKS 1995) e o vegetal *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 1996; DELSENY 1991), entre outros. Mais recentemente foram completados os genomas da mosca *Drosophila melanogaster* (ADAMS 2000) e do camundongo *Mus musculus* (Mouse Genome Sequencing Consortium; WATERSTON 2002). Hoje existem seqüenciados 18 genomas eucariontes, 217 procariontes e 1436 virais.

O Projeto Genoma Humano (*Human Genome Project – HGP*) foi uma iniciativa de pesquisa mundial oficialmente lançado em 1990 com o objetivo de mapear e seqüenciar todo o genoma humano. Hoje, 15 anos após seu lançamento, existem dois seqüenciamentos completos do genoma humano, oriundos do consórcio público (*International Human Genome Sequencing Consortium – IHGSC*; LANDER 2001) e da empresa privada Celera Genomics (VENTER et al. 2001), que usaram as estratégias de seqüenciamento “montagem de *contigs clone-by-clone*” e “*whole-genome shotgun*”, respectivamente.

Conforme os dados de seqüências moleculares, tanto de DNA como de proteínas, começaram a ser gerados desde o começo dos anos 80, estes foram sistematicamente armazenados em bancos de dados computacionais, e estão hoje disponíveis ao público na Internet em *sites* como GenBank do NCBI (*National Institute of Biotechnology Information*), EMBL (*European Molecular Biology Laboratory*), DDBJ (*DNA Data Bank of Japan*), PIR (*Protein Information Resource*), SWISS-PROT entre outros. Juntamente com a criação destes bancos de dados, métodos computacionais tiveram que ser desenvolvidos para recuperação e análise dos dados neles armazenados, incluindo algoritmos de busca de similaridades de seqüências, predição de genes, de estrutura protéica, etc. Estas ferramentas da biologia computacional ou bioinformática são hoje indispensáveis aos pesquisadores das áreas de genômica, transcriptômica ou proteômica.

Tendo hoje disponível a seqüência completa do genoma de vários organismos, os *softwares* de predição de genes tornam-se ferramentas de suma importância na procura por novos genes. Estes programas reconhecem padrões estruturais gênicos, representando predições *ab initio* ou podem ter suporte parcial

por seqüências de transcritos. Exemplos de padrões estruturais gênicos usados nos programas são as seqüências sinais de transcrição, regiões controladoras a montante (*upstream control sequences*) como os elementos promotores, os TATA-boxes e as ilhas CpG, sítios de poliadenilação (p.ex.: AAUAAA), seqüências sinais de tradução (códon de início e término, p.ex.: ATG e TAA respectivamente), os sítios de processamento (*splicing*) de RNAm nas bordas dos íntrons (GT..AG) e a tendência de uso de códon nos quadros abertos de leituras (ORFs - *Open Reading Frame*) dos genes (UBERBACHER e MURAL 1991; FICKETT 1996; CLAVERIE 1997). Exemplos destes programas são o GENEMARK (BORODOVSKY e MCININCH 1993) GRAIL (UBERBACHER et al. 1996) e o GENEPARSER (SNYDER e STORMO 1995). Estes softwares são continuamente melhorados, e os mais avançados tais como o GENESCAN (BURGE e KARLIN 1997) levam em conta a compatibilidade dos quadros de leitura de éxons adjacentes e as propriedades composicionais de íntrons e éxons. Parte da estratégia pode incluir comparação de seqüências de DNA de dois organismos na busca de seqüências conservadas evolutivamente (O'BRIEN et al. 1999). No entanto, os softwares de predição de genes, que primordialmente foram desenvolvidos para analisarem genomas menores e menos complexos como os de procariontes, podem falhar na predição de genes de eucariontes devido, em parte, ao aumento da complexidade destes, causado pela presença de grandes regiões não-codificantes, os íntrons, e em parte, pela degeneração das seqüências sinais que torna muitas vezes impossível distinguir um sinal funcional de um não-funcional. Foi constatada que a acurácia destes algoritmos na predição de éxons em eucariontes é de apenas aproximadamente 70% e eles

mostram dificuldade na identificação de pequenos éxons e, por isto, em determinar com exatidão uma sequência aberta de leitura (GUIGO et al. 2000).

Uma outra estratégia adotada com a finalidade de se identificar novos genes é a partir da procura de seu produto de transcrição, o RNAm. A sequência do transcrito é a prova final da existência de um gene expresso e um requerimento absoluto para a precisa identificação do início e fim de cada éxon, dos processamentos alternativos de RNAm e para definir padrões teciduais e temporais de expressão (CAMARGO et al. 2001b). Desta maneira, pode-se conseguir sucesso na determinação das regiões codificadoras com o uso de DNAs complementares (cDNA), produtos da transcrição reversa de RNAm. Estes cDNAs, representam uma fonte direta para se observar as regiões codificantes e as formas alternativas de processamento dos genes.

Em 1991, a primeira aplicação de seqüenciamento em larga escala de clones de cDNA foi publicada (ADAMS et al. 1991). Neste trabalho clones de cDNA oriundos de cérebros humanos foram aleatoriamente selecionados e parcialmente seqüenciados. Estas seqüências parciais de cDNAs, as quais representam a expressão de genes nos tecidos em um determinado tempo foram chamadas de Etiquetas de Seqüências Expressas ou ESTs (*Expressed Sequence Tags*). Abordagens similares a esta, para diferentes tecidos e organismos, logo se seguiram (HOULGATTE et al. 1995).

Nas metodologias de predição gênica atuais, se observa uma complementação entre as técnicas de anotação de genes a partir de ESTs e os algoritmos de predição, pois as ESTs são freqüentemente efetivas na identificação das extremidades 3' dos genes, onde os softwares geralmente falham em prever, enquanto estes mesmos programas determinam relativamente bem a porção mais a 5' do gene, onde os

clones de cDNAs geralmente não alcançam devido à perda de eficiência da enzima ao longo do processo de polimerização ou à truncagens na transcrição reversa causadas por eventuais estruturas secundárias do mRNA (BAILEY et al. 1998).

A disponibilidade da sequência do genoma humano possibilitou o mapeamento das ESTs que, junto a aplicação dos programas de predição gênica, levou ao crescimento do número de genes conhecidos. Hoje temos 22.287 locus gênicos determinados (<http://www.sanger.ac.uk/Info/Press/2004/041020.shtml>), sendo 19.599 de genes que codificam proteínas, o que parece ser um grande avanço se atentarmos para o fato de que até o final da década de 1990 tínhamos cerca de 6.000 genes identificados e mapeados (JORDE et al. 1999).

As ESTs são extremamente valiosas não somente pela identificação de novos genes (ADAMS et al. 1992; LIANG 2000), mas também porque representam um quadro do nível e da complexidade da expressão gênica da amostra tecidual. As ESTs tem sido utilizadas para a caracterização de transcriptomas tecido específico (BORTOLUZZI et al. 2000; HUMINIECKI e BICKNELL 2000; PHILLIPS et al. 2000; YU et al 2001) para estudos da diferença entre tecido normal e patológico (OKUBO et al. 1992; VASMATZIS et al. 1998), para mapeamento físico (HUDSON et al. 1995), no entendimento dos limites éxon-íntron (KAN 2001), na detecção de poliadenilação e *splicing* alternativos (WOLFBERG e LANDSMAN 1997; MODREK 2001; WANG 2003) e polimorfismos de nucleotídeos (GARG 1999).

Apesar de todas as qualidades demonstradas pelas ESTs elas apresentam limitações tais como as altas taxas de erros de seqüenciamento (~3%) (HILLIER et al. 1996) e vários tipos de contaminação. Problemas típicos encontrados incluem seqüências quiméricas, DNA genômico do próprio organismo, mRNA não

processado, DNA ou RNA microbiano (SOREK e SAFER 2003). Para sanar estes vieses, um passo importante na geração de ESTs consiste no processamento destas por programas de identificação de contaminantes (*cleaning stage*), que reconhecem e retiram algumas destas seqüências (CHOU 2001; SOREK e SAFER 2003).

As ESTs representam a maior quantidade de informação possível por base seqüenciada, e a possibilidade de que a partir delas haveria uma rápida identificação de todos os genes humanos, levou ao desenvolvimento de vários projetos para geração e armazenamento destas, financiados por órgãos públicos ou privados. Exemplos destes projetos são o consórcio IMAGE (*Integrated Molecular Analysis of Genomes and their Expression*; LENNON et al. 1996) e o MGI (*Merck Gene Index*; ECKMAN 1998). Na pesquisa contra o câncer dois grandes projetos têm gerado um extenso volume de ESTs de tecidos tumorais humanos e seus correspondentes normais: o CGAP (*Câncer Genome Anatomy Project*; O'BRIEN 1997; STRAUSBERG 2001) e o HCGP (*Human Cancer Genome Project*; BONALUME-NETO 1999; CAMARGO et al. 2001a). O CGAP, lançado em 1997 pelo Instituto Nacional do Câncer dos EUA, tem utilizado seqüências oriundas das extremidades 5' e 3' de clones de cDNAs. Já o HCGP, lançado em 1998, financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo e pelo Instituto Ludwig de Pesquisa Sobre o Câncer, fez uso da técnica ORESTES, com a qual foi produzido em 2 anos, o número de 1.190.044 seqüências. As seqüências de ESTs geradas em todos estes trabalhos podem ser recuperadas no banco dbESTs (<http://www.ncbi.nlm.nih.gov/dbEST>), uma divisão do GenBank.

A maioria dos projetos acima mencionados têm produzido bibliotecas de cDNAs baseadas no uso de iniciadores para a síntese da primeira fita de DNA que

anelam especificamente na cauda poli-A do transcrito. Deste modo, as ESTs representam comumente a extremidade 3' dos RNAm. Mesmo em projetos como o MGI, no qual também foram produzidas seqüências a partir da extremidade 5' dos clones, estas não determinam a porção 5' verdadeira do transcrito, conseguindo apenas, na maioria das vezes, obter regiões codificadoras (STRAUSBERG 2001). Esta característica das ESTs permite a compilação completa apenas de pequenas seqüências transcrituais com alto nível de expressão, através da geração de *contigs* das regiões de sobreposição das extremidades destas (CAMARGO et al. 2001a).

A técnica de geração de ORESTES (*Open Reading Frame ESTs*) difere das técnicas convencionais de geração de ESTs pois faz uso de iniciadores aleatórios de 12 a 33 nucleotídeos contendo 20 a 80% de guanina e citosina que servem para a transcrição reversa dos RNAm e para a amplificação dos cDNAs (CAMARGO et al. 2001a). Isto confere algumas características notórias às bibliotecas: 1) tendência das seqüências geradas situarem-se na porção central do transcrito e 2) alta sensibilidade, capturando transcritos de baixa expressão, com performance similar a das bibliotecas normalizadas em termos de acesso a genes com baixos níveis de expressão (efeito parcialmente normalizador) (DIAS NETO 2000). Estas características se mostram vantajosas na anotação das seqüências geradas, uma vez que estas cobrem porções pouco representadas de transcritos nos bancos de ESTs e também por possuírem uma maior representatividade de transcritos raros, quando em comparação com as metodologias tradicionais. Por outro lado, devido a técnica não fazer uso de oligo-dT para a síntese de cDNA, há a tendência de uma maior quantidade de contaminação genômica quando em comparação com as metodologias 3' e 5'. Isto é agravado pelo

fato da PCR usada na metodologia ser de baixo rigor (*low stringency*), o que faz que mesmo seqüências inespecíficas (contaminantes ou não) possam ser amplificadas.

1.1 AS ORESTES DO HCGP

As 1.190.044 ORESTES do HCGP foram produzidas a partir de 5.515 mini-bibliotecas de cDNA provindos de extratos de RNA de 24 diferentes tipos de tecidos, nos estados normal e tumoral. Após a remoção dos contaminantes (DNA mitocondrial e RNA ribossômico (9,1%), DNA genômico (<0,01%) elementos repetitivos (6,5%) e seqüências altamente similares a genomas bacterianos conhecidos (4,5%)) restaram 949.685 seqüências. Destas, 717.059 foram mapeadas no genoma humano, sendo que 598.005 mostraram alta similaridade com seqüências expressas de diversos bancos de dados públicos, tais como Unigene, dbEST, Swiss Prot, Protein Data Bank, etc. O número de 351.680 não mostrou similaridade com transcritos conhecidos, sendo então classificadas como “sem-pares” ou “*no-match*”.

As ORESTES geradas pelo HCGP têm mostrado utilidade para a definição do transcriptoma humano. Exemplo disto são as 48 novas seqüências de transcritos definidas no cromossomo 22 e dos 219 *contigs* de ORESTES que possuem similaridade com regiões não anotadas do mesmo cromossomo, dos quais 38 mostram potencial de região codificadora medidas pelo ESTSCAN (DE SOUZA 2000). Dezenove adicionais transcritos não preditos para o cromossomo 21 humano foram encontrados usando as ORESTES como uma das bases de dados (REYMOND

2002). SOGAYAR e CAMARGO (2004) fecharam os intervalos (*gaps*) apresentados por 211 pares de clusters contíguos, validando o que chamaram de Unidades Terminadoras de Transcritos ou TFU (*transcript finishing units*) usando ESTs 3', 5' e ORESTES.

Recentemente foi criado um consórcio entre os dois grandes projetos de seqüenciamento de transcritos relacionados ao câncer, o HCGP e o CGAP (*HCGP/CGAP Annotation Consortium*), o qual uniu as ORESTES do primeiro projeto com as ESTs 5' e 3' do segundo, na tentativa de se gerar uma ampla base de dados de genes expressos em câncer (BRENTANI et al. 2003). Os dados dos dois projetos constituem também a base do Índice do Câncer Humano na TIGR (<http://www.tigr.org/tdb/tgi/hcgi>) e estão disponíveis no *site* <http://cgap.nci.nih.gov/>.

Pelo exposto acima fica notória a importância do HCGP e, por conseguinte das ORESTES, na descoberta e caracterização de novos genes.

Mas um sub-conjunto de ORESTES permanecia ainda pouco explorado, as *no-match*, e a busca de novos genes humanos neste conjunto de dados nos pareceu relevante, uma vez que devido à alta sensibilidade da técnica, provavelmente entre elas exista uma parcela que represente transcritos humanos de baixa expressão, fato pelo qual não se encontraram seqüências a elas similares depositadas nos bancos de dados públicos. Para averiguarmos este conjunto de dados concebemos o desenvolvimento de um conjunto de programas computacionais (*pipeline* de bioinformática) para a seleção de agrupamentos de ESTs (*clusters*), de acordo com alinhamento genômico, que possuísem maior probabilidade de representarem transcritos humanos. As seqüências selecionadas foram inspecionadas manualmente e uma porção destas levadas à validação experimental por RT-PCR. Como uma

grande fração das ORESTES é provinda de amostras tumorais, aquelas validadas podem ajudar na identificação de novos transcritos envolvidos na tumorigênese. Transcritos deste tipo podem ser usados como marcadores moleculares em estudo, prevenção, diagnóstico, prognóstico e tratamento da doença. Portanto, procedemos com uma busca *in silico* por informação sobre expressão diferencial entre tecidos normal e tumoral dos transcritos validados.

2 OBJETIVOS

- 1 Desenvolvimento de *pipeline* de bioinformática para seleção de ORESTES *no-match* do projeto HCGP que representem reais transcritos humanos.

2.1 OBJETIVOS ESPECÍFICOS

- 2.1.1 Validação experimental por RT-PCR de uma amostragem de seqüências selecionadas pelo *pipeline*.
- 2.1.2 Análise de expressão *in silico* das seqüências validadas, comparando tecido normal e tumoral.
- 2.1.3 Criação de interface *Web* para disponibilização das *no-match* selecionadas pelo *pipeline*.

3 MATERIAIS E MÉTODOS

3.1 PROCEDIMENTOS COMPUTACIONAIS

A principal fonte de dados usada foi o banco de dados em formato relacional (MySQL) *hcgp*. Este banco contém todas as ORESTES geradas no HCGP anotadas de acordo com seus alinhamentos com seqüências de bancos de dados públicos de transcritos. As seqüências anotadas como *no-match* neste banco foram então conferidas em outro banco de dados, o *map4_bh1*, o qual contém ORESTES mais seqüências do dbEST e RefSeq RNAm (UniGene versão #153) arranjadas em clusters baseados em alinhamento genômico. Foram excluídas deste banco as seqüências derivadas de bibliotecas sugeridas como contaminadas com DNA genômico ou RNAm imaturo identificadas em trabalho anterior (SOREK e SAFER 2003). Ainda mais, todas as seqüências encontradas no *map4_bh1* satisfazem nosso critério de identidade descrito logo abaixo, sendo que apenas os melhores alinhamentos destas (*best hits*) estão contemplados. Ambos bancos de dados, *hcgp* e *map4_bh1* são mantidos pelo Laboratório de Biologia Computacional do Instituto Ludwig de São Paulo e os procedimentos computacionais requeridos para acessá-los foram por nós desenvolvidos em linguagem de programação PERL.

Delimitamos o conjunto de dados selecionando apenas aqueles agrupamentos de seqüências, determinados por alinhamento genômico, que possuíssem ao menos uma seqüência *no-match* oriunda de próstata ou mama, dois tecidos com maior

incidência de câncer no Brasil (www.inca.gov.br). Uma vez definido o conjunto de dados, os seguintes critérios de seleção foram usados:

1. Fazer uso de seqüências com similaridade exônica $\geq 93\%$ e com ao menos 45% de sua extensão com similaridade total.
2. Os *clusters* devem ser formados por ORESTES oriundas de ao menos duas amostras teciduais (bibliotecas). O racional para este critério é que quanto maior for a variedade de bibliotecas que perfaz um cluster, menor é a probabilidade deste ser fruto de contaminação. Contudo, os *singletons* que apresentam *gaps* foram mantidos.
3. Os *clusters* devem apresentar interrupções (*gaps*) no alinhamento contra o genoma humano em ao menos uma de suas seqüências, o que sugere evento de *splicing*, característico de transcritos de eucariontes (HERBERT 1999).
4. As seqüências portadoras de *splicing* devem apresentar sítio canônico 'GT..AG' nas bordas de seus *introns*. Tais sítios são encontrados em cerca de 98,7% dos transcritos de mamíferos (BURSET 2000). Isto nos confere ainda mais confiança de que o *gap* é verdadeiramente reflexo de um evento de *splicing*. A verificação destes sítios foi feita sobre todas as seqüências que apresentam *splicing* dos clusters selecionados checando-se 5 pb para o interior da EST (suposto éxon) e 5 pb para o exterior (suposto íntron) na seqüência genômica. Esta margem interna e externa é necessária uma vez que

o alinhamento da EST com a sequência genômica não é perfeito, sendo que as bordas das sequências são muitas vezes deslocadas a jusante (*downstream*) ou a montante (*upstream*) de seu local real (BURSET 2000).

Para garantirmos que estávamos lidando com ORESTES que permaneciam *no-match*, logo após a aplicação do primeiro critério de seleção removemos todos os *clusters* que continham transcritos *full-length* ou ESTs oriundas de outros projetos que não o HCGP no banco map4_bh1. Outro procedimento intermediário aos critérios do *pipeline* foi alinhar as sequências que passaram pelo segundo critério de seleção contra as versões mais recentes disponíveis em nosso laboratório dos bancos Hs.contigs e Hs.RNAm (versão UniGene #160). Aquelas sequências que obtiveram identidade igual ou superior a 96% em 100 pb (e-value de -20) foram retiradas da avaliação (WISTON 2002). O Hs.contigs é uma divisão do UniGene na qual estão disponíveis as ESTs humanas de diversos projetos, RNAm e HTC's (*Human Tentative Contigs*), clusterizadas com base na identidade entre as sequências, de forma a perfazerem *contigs* não redundantes. O Hs.mRNA contém redundantes 73.559 sequências de RNAm *full-length* e *full-inserts* humanos.

Para confirmar a originalidade de nossas sequências com dados atualizados de genoma e de transcriptoma, os clusters resultantes do *pipeline* foram manualmente avaliados, alinhando suas sequências contra a sequência do genoma humano, montagem 34 versão 2, disponível pelo NCBI e julho/2003 disponível pela UCSC, usando os programas de alinhamento BLAST e BLAT, respectivamente. Os resultados foram analisados nas interfaces gráficas NCBI Map Viewer e UCSC Genome Browser (KENT 2002). Os roteiros de anotação da UCSC (*annotation*

tracks) usados foram Genes Conhecidos Baseados em SWISS-PROT, TrEMBL, mRNA e RefSeq. As anotações de Genes Preditos por Genscan (BURSET e GUIGÓ 1996), mRNA e ESTs do GenBank foram ativadas também. Todas as seqüências alinhadas sobre genes conhecidos (*RefSeq Status Provisional*, *Reviewed* ou *Validated*) no NCBI Map Viewer, mesmo em íntrons, foram descartadas. O mesmo foi feito com aquelas alinhadas em RNAm ou ESTs outras que não ORESTES e em genes preditos com suporte em seqüências de transcritos (*RefSeq Status Predicted* ou *Model*). Apenas os clusters cujas seqüências alinhadas confirmaram *gaps* no alinhamento (o que sugere *splicing*) em regiões contendo predição *ab initio* no Map Viewer, em éxons preditos ou não, e que os alinhamentos foram confirmados no UCSC Genome Browser foram levadas à validação experimental.

3.2 VALIDAÇÃO EXPERIMENTAL

3.2.1 Extração de RNA

As amostras utilizadas neste estudo foram de tecidos tumorais e normais de próstata e tumorais de mama, provenientes de pacientes do Hospital do Câncer. As amostras removidas cirurgicamente eram imediatamente colocadas em tubos contendo RNA later (Ambion, The RNA Company), que consiste em uma solução aquosa fixadora e tamponante que estabiliza e protege o RNA celular *in situ*, sem afetar sua qualidade ou quantidade, conforme instruções do fabricante.

As extrações de RNA total foram feitas por dois métodos distintos: Trizol Reagente (Invitrogen Life Technologies) segundo especificações do fabricante e por ultracentrifugação isopícnic de cloreto de Césio (GLISIN et al. 1974).

Na técnica de ultracentrifugação isopícnic as amostras tumorais de próstata e mama ainda congeladas foram homogeneizadas em 9ml de solução de lise (4M isotiocianato de guanidina / 25mM citrato – pH 7,0 / 0,1M β -mercapto). Em seguida, os lisados foram transferidos para tubos de ultracentrífuga contendo 4ml de solução de CsCl (5,7M CsCl e 25mM AcNa) e então centrifugados a 150.000g por 17 horas a 20°C. Após a centrifugação, o sobrenadante de proteínas e DNA foi descartado por inversão e o *pellet* de RNA preservado. Finalmente, o precipitado de RNA foi solubilizado em 200 μ l de água DEPC. As concentrações foram dosadas em espectrofotômetro.

3.2.2 Avaliação da qualidade do RNA

A qualidade do RNA foi avaliada para verificar se houve degradação do material ou se o mesmo mostra-se contaminado com DNA genômico. Para isto, foi aplicado 1 μ g de RNA total em gel de agarose 1%. A revelação foi feita com brometo de etídeo. Foram considerados íntegros os RNAs que apresentaram as bandas correspondentes aos RNAs ribossômicos 28S e 18S na proporção 2:1. A contaminação com DNA foi verificada por PCR do gene hMLH1 (human mut-L homologue 1), utilizando-se 200ng de RNA total e 1mM dos iniciadores de polimerização (*primers*) (Forward – 5' TGG TGT CTC TAG TTC TGG 3' e Reverse – 5' CAT TGT TGT AGT AGC TCT GC 3') desenhados para os íntrons 12 e 13 do

gene, de maneira que, qualquer amplificação presente será devido à contaminação com DNA, sendo o tamanho esperado para o fragmento de 250 pb. A tentativa de visualização do produto foi feita em gel de poliacrilamida 8%, com coloração de prata. Para as amostras contaminadas, foi feito tratamento com DNase (Gibco) (1 U/2,5µg de RNAm total), segundo instruções do fabricante. Para confirmar se a contaminação foi eliminada, o teste de hMLH1 era repetido.

3.2.3 Produção de cDNA e RT-PCR (*Reverse Transcriptase-Polimerase Chain Reaction*)

Para produção de seqüências de cDNA foram utilizados 1µg de RNAm, 1µl de oligo-dT (0,5µg/µl) e 8 µl de água DEPC. A mistura foi deixada à 70°C por 5 minutos para desnaturação do RNA. Para a polimerização da primeira fita de cDNA juntamos à solução anterior 2µl de dNTPs (10mM), 4 µl de 5X *First strand buffer*, 2µl de DTT (0,1M), 200U de *RNase out* e 40U de transcriptase reversa *SuperScriptII* (Gibco®). Esta mistura foi incubada a 42°C durante 1 hora, acrescentando posteriormente mais 40U de *SuperScriptII* e deixando a 42°C por mais 1 hora.

Avaliamos a qualidade do cDNA por PCR utilizando *primers* específicos (*Forward*: 5' CTG CAC CAA CTG CTT A 3' e *Reverse*: 5' CAT GAC GGC AGG TCA GGT C 3') para os éxons 6 e 7 do gene GAPDH (gliceraldeído desidrogenase). Este gene apresenta expressão constitutiva em virtualmente todos os tecidos, de maneira que em uma boa síntese de cDNA observamos uma forte amplificação, com tamanho de 300 pb. Procedemos também com PCR para o gene NOTCH2 (*primers Forward* - 5' TGT GGC CAA CCA GTT CTC C 3' and *Reverse* - 5' GGC AGT

CAT CAA TAT TCC TC 3') com intuito de nos certificarmos de que os cDNAs haviam sido adequadamente estendidos, uma vez que os *primers* usados delimitam uma região de 300 pb do transcrito distante 11.433 pb de sua extremidade 3'.

Na PCR foi utilizado 1µl da solução de cDNA, 1,5µl de MgCl₂ (50 mM), 1 µl de dNTPs (10 mM), 2,0µl de cada *primer* (5 mM), 2µl de betaína (5M) (HENKE, 1997) e 1U de Taq DNA polimerase (Gibco), em tampão apropriado (10X). A reação era iniciada com desnaturação a 94°C por 5 minutos, seguida de 35 ciclos de: 45 segundos a 94°C, 45 segundos a 60°C e 1 minuto a 72°C com extensão final de 10 minutos a 72°C. O produto era visualizado em gel de poliacrilamida 8% e corado com prata 0,2%.

Os *primers* para validação foram desenhados a partir da sequência genômica para alinharem com dois segmentos contíguos das maiores seqüências de cada cluster selecionado, usando para isto o programa Primer3 (ROZEN e SKALETSKY, 2000). As seqüências dos *primers* podem ser conferidas na Tabela 1.

Tabela 1 - Iniciadores de polimerização usados para amplificação das ORESTES selecionadas.

GENE_ID	Forward Primer	Reverse Primer
BF371440	5' GACGAGCCCATTCCCAG 3'	5' ATGGATGGTTCGGATGTGTT 3
363833_OR	5' ACCACAGGGACCACAGAC 3'	5' GTTAGTTCCTGCACAAATCA 3'
221609_OR	5' GAATCTGGGAGGAGAAAAAG 3'	5' CCTTGTTGGTCACATGAGG 3'
AW996738	5' GCAAGAATTTGGGGACAGTG 3'	5' TTGATTCAACCCAGAGGCTT 3'
BE077462	5' GATGGCAGATTCTGGATG 3'	5' GTTGCTCCAGATGCTC 3'
AW999051	5'ACTTGGTGGGTGAAATAACG 3'	5' GTGAAGTAGTGCTGCTGAG 3'
BE062313	5' CAGTGACATTGCTCCAGCT 3'	5' CCACAGGATTTACGGAGTG 3'
475371_OR	5' TACTCATCCTTCCACTGGGG 3'	5' GGCTTGTCAGGTGTTTGGTG 3'

As mesmas condições de PCR foram utilizadas para amplificação das seqüências dos clusters selecionados pelo *pipeline*, diferenciando apenas a temperatura de anelamento dos *primers*. Após tentativa de amplificação das seqüências selecionadas, os produtos que apresentaram bandas de tamanho próximo ao esperado (Tabela 2 e Anexo I) foram corridos em gel de agarose/TBE 1% e purificadas com o kit Concert™ Nucleic Acid Purification System (GibcoBRL), de acordo com as especificações do fabricante, e levados a seqüenciamento. A reação de seqüenciamento foi feita usando-se 1 µl de produto do PCR purificado, 1 µl (5 mM) de *primer* (forward ou reverse) e 2 µl de Big-Dye Terminator Kit (Applied Biosystems). O produto foi então carregado em um seqüenciador ABI 3100 (Applied Biosystems). Após seqüenciamento, as seqüências em formato FASTA foram alinhadas contra o genoma humano, usando o BLAT, para checagem do alinhamento (*hit*).

Tabela 2 - Tamanhos de bandas esperados da amplificação de cada seqüência selecionada.

ORESTES	Tamanho de Banda Esperado (pb)
BF371440	164
363833_OR	248
221609_OR	218
AW996738	162
BE077462	178
AW999051	232
BE062313	261
475371_OR	188

3.3 ANÁLISE DE EXPRESSÃO *IN SILICO*

Para verificação de expressão diferencial dos fragmentos transcrituais por nós validados, assumimos que estes são provindos dos genes preditos nos quais alinharam. Assim, geramos SAGE (*Serial Analysis of Gene Expression*; VELCULESCU et al. 1995) *tags* virtuais de 10 bases a jusante (*downstream*) do sítio de corte da enzima *Nla*III (5' CATG 3') mais a 3' dos RNAm preditos correspondentes e posteriormente verificando a frequência destas *tags* em banco de dados local de SAGE.

Com o mesmo intuito usamos as coordenadas genômicas destes genes preditos para quantificar as ESTs de origem tumoral e normal alinhadas nestas regiões. As contagens foram feitas usando-se o map4_bh1.

Uma outra tentativa de se acessar uma possível expressão diferencial foi verificar se as marcas citogenéticas nas quais os transcritos se localizam estão descritas como RITE (*chromosomal Regions of Increased Tumor Expression*; ZHOU et al. 2003), ou seja, se são regiões cromossômicas que mostram aumento de expressão em tumor.

4 RESULTADOS

Em nosso trabalho avaliamos um número limitado de ORESTES *no-match*, cerca de 18% do total destas, as quais somam 62.165, provindas de bibliotecas de próstata e mama. Deste montante, 23.189 encontravam-se armazenadas no banco map4_bh1, o que significa dizer que foram mapeadas no genoma humano atendendo ao nosso critério de identidade e agrupadas juntamente com outras ORESTES e ESTs 3' e 5' de acordo com o alinhamento genômico. Estas seqüências se encontravam distribuídas em 14.908 *clusters*. Entre estes, identificamos e removemos 938 (1.676 seqüências) que apresentam transcritos *full-length* incorporados. Dos 13.970 *clusters* restantes retiramos 1.467 *clusters* mistos, os quais são caracterizados por conterem ESTs de outros projetos que não o HCGP. Ambos números sustentam a noção de que as *no-match* possuem potencial de representantes de transcritos humanos. Após, retiramos 1.953 *clusters* cujas seqüências provinham todas de apenas uma biblioteca e 8.286 *singletons* que não apresentavam *gaps* em seus alinhamentos.

Todas as seqüências restantes foram submetidas a alinhamento do tipo BLAST contra o banco Hs.contigs e de Hs.RNA_m (e-value -20). O resultado do alinhamento foi a identificação de 354 seqüências similares (96% de similaridade em 100pb) a um ou ambos bancos e a retirada dos 38 *clusters* aos quais elas pertenciam, restando 2.226 *clusters*. Destes, 131 são *singletons* com *splicing*, 2.039 são *clusters* sem *splicing* e 56 são *clusters* com *splicing*. Os *clusters* cujas seqüências processadas não apresentaram sítios canônicos e aqueles sem *splicing* não foram levados adiante

em nossa análise, embora acreditamos que parte deles também possa representar transcritos humanos, sobretudo aqueles com maior tamanho e com maior variedade de bibliotecas. Estes serão levados a validação experimental por *microarray* em um projeto subsequente.

Dos 131 *singletons* com *splicing*, 65 apresentaram sítios canônicos (GT..AG). Já dos 56 clusters com *splicing*, 29 os possuíam (Anexo 2).

Alinhamos as seqüências portadoras de *splicing* de cada *cluster* assim como os *singletons* selecionados contra o genoma humano usando o BLAST para visualizarmos o *hit* e a região genômica no Map Viewer do NCBI e, adicionalmente, usamos o BLAT no *Genome Browser* da USCS (Anexo 3). Após a inspeção manual dos 94 clusters selecionados constatamos que 60 eram clusters com novas seqüências, sendo que aqueles que alinharam dentro de genes preditos *ab initio*, total de oito, foram levados à validação experimental (Tabelas 3 e 4). O fluxo de dados do pipeline até a inspeção manual pode ser conferido na Figura 1. Para estas ORESTES selecionadas foram desenhados *primers* que anelam com dois de seus segmentos contíguos (potenciais éxons) para procedermos com as PCRs. Os produtos da amplificação foram submetidos à eletroforese em gel de poliacrilamida 8% e à revelação por nitrato de prata. Obtivemos bandas de tamanhos próximos aos esperados para cinco das oito ORESTES testadas (Tabela 2 e Anexo 1). Estes produtos foram corridos em gel de agarose 1%, as bandas purificadas e o produto seqüenciado. Os arquivos FASTA dos seqüenciamentos foram então submetidos a alinhamento do tipo BLAT contra o genoma humano. As cinco seqüências alinharam exatamente na região esperada confirmando a especificidade dos cinco fragmentos amplificados (Figura 2 e Anexo 3).

Tabela 3 – Informações gerais sobre os três clusters selecionados do conjunto de dados de próstata.

Cluster	Tamanho	Bib	Maior sequência com <i>splicing</i>	Tecido	Resultado do alinhamento BLAT no UCSC Genome Browser	Resultado da validação experimental
105202	2	2	BF371440	T/N	Dois hits em dois éxons preditos - NT_008183.629 crom.8	Confirmado
242293	1	1	363833_OR	T	Dois hits em um íntron de gene predito – NT_011387.258 crom. 20	Confirmado
9307	1	1	221609_OR	T	Três hits, dois em dois éxons contíguos preditos – NT_022184.134 crom.2	Não confirmado

Tabela 4 - Informações gerais sobre os cinco clusters selecionados do conjunto de dados de mama.

Cluster	Tamanho	Bib	Maior sequência com <i>splicing</i>	Tecido	Resultado do alinhamento BLAT no UCSC Genome Browser	Resultado da validação experimental
75485	3	2	AW996738	N	Três hits em três éxons contíguos preditos - NT_006576.797 crom.5	Confirmado
203647	1	1	BE077462	T	Dois hits em um íntron de gene predito - NT_010718.394 crom.17	Não Confirmado
42329	1	1	AW999051	N	Três hits, um em éxon predito e dois em íntron - NT_005535.5 crom.3	Confirmado
12313	1	1	BE062313	T	Dois hits, um sobrepondo parcialmente um éxon predito e outro em íntron - NT_022184.1316 crom.2	Confirmado
98150	1	1	475371_OR	T	Três hits em um íntron de gene predito - NT_007592.4 crom.6	Não confirmado

Legenda: T=tumor; N=normal. Bib= número de bibliotecas.

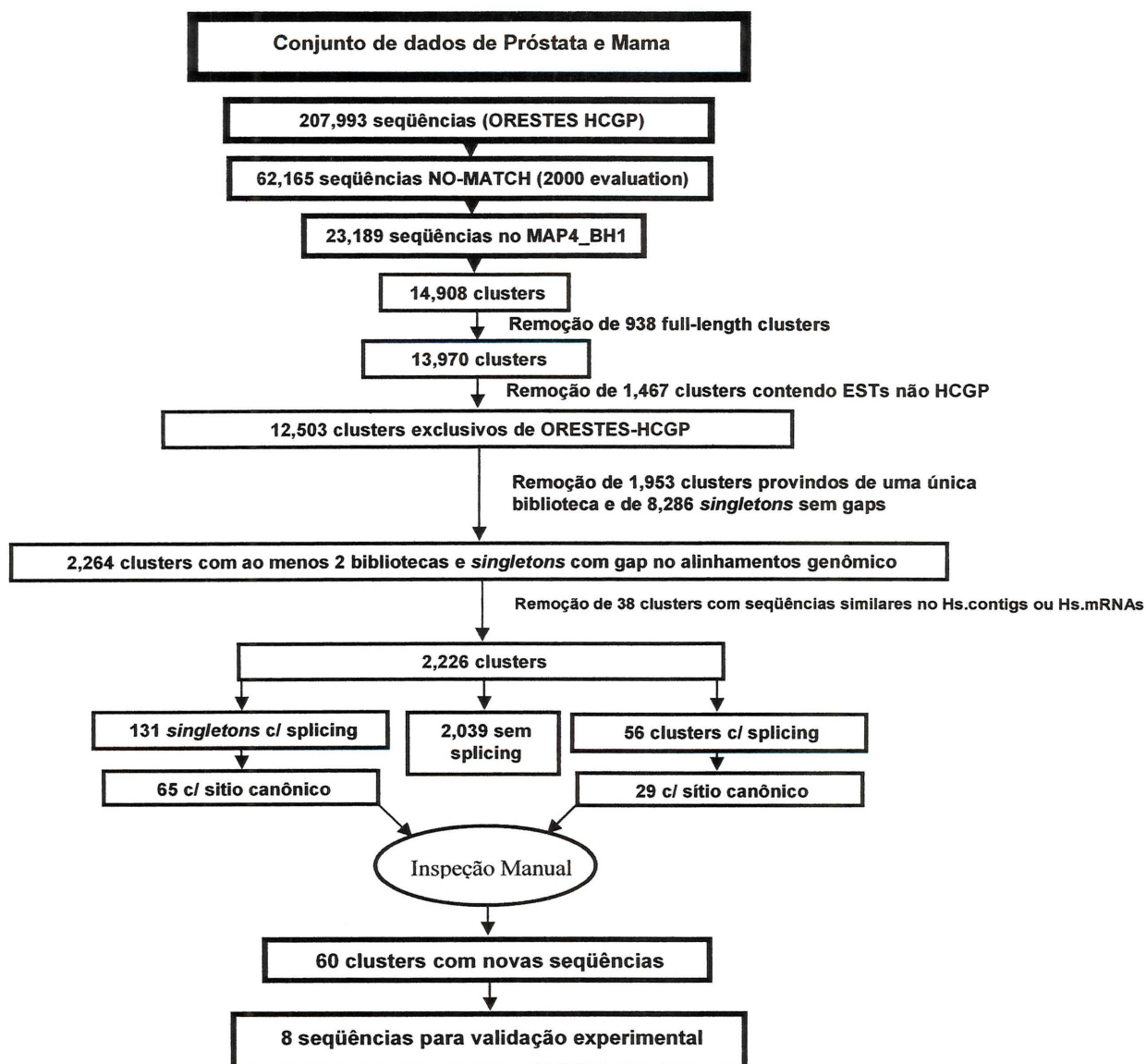
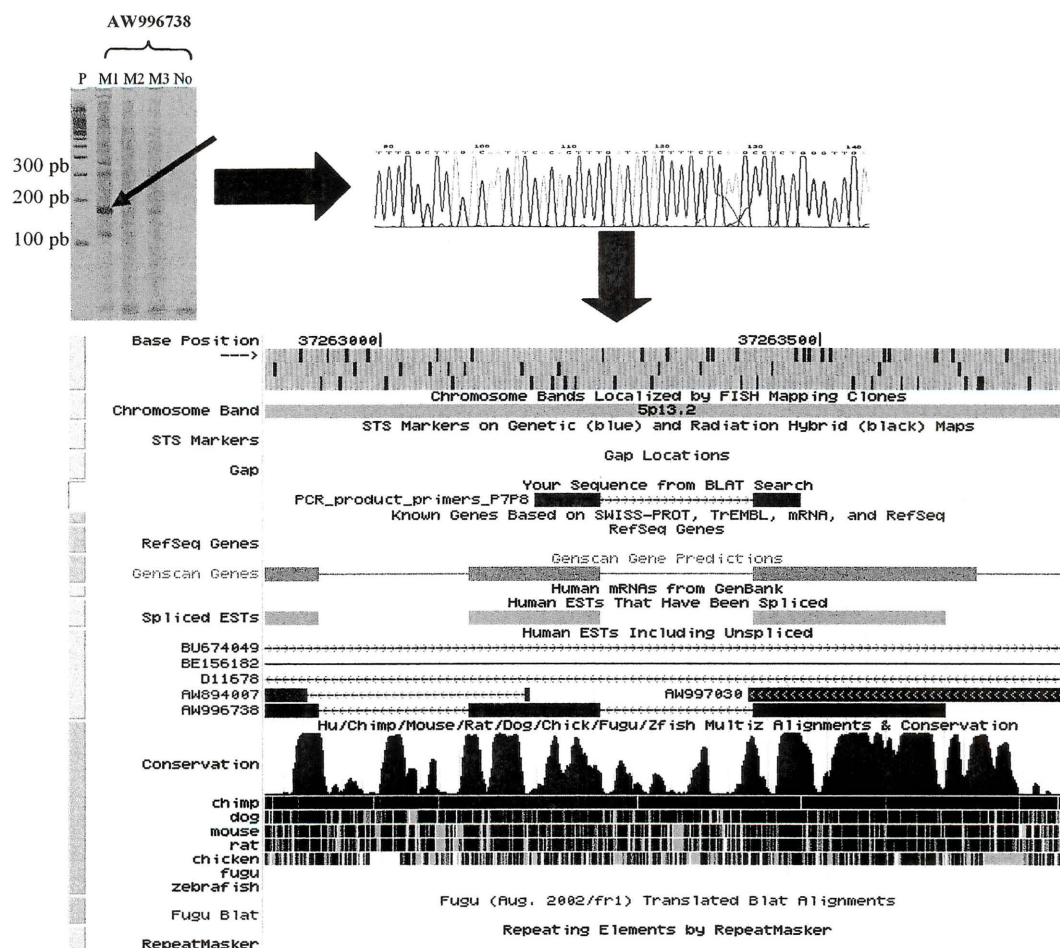


Figura 1 - Fluxograma mostrando passos do *pipeline* com números do conjunto de dados de próstata e mama.



Legenda: M1, M2, M3 = amostras de RNA de tumor de mama. P = padrão de corrida (Ladder 100 bp Invitrogen). PCR_product_primers_P7P8= sequência do produto de PCR dos primers P7P8.

Figura 2 - Esquema demonstrando as três etapas para validação das ORESTES selecionadas (gel do produto de amplificação/ sequenciamento/ interface gráfica do alinhamento BLAT da sequência amplificada sobre o genoma humano). No exemplo a ORESTES AW996738.

Foi analisada a expressão, por SAGE e por quantificação de ESTs, para os genes preditos (NT_011387.258, NT_006576.797, NT_005535.5, NT_022184.1316 e NT_008183.629), dos quais assumimos derivarem as ORESTES 363833_OR, AW996738, AW999051, BE062313 e BF371440, respectivamente. Para a

BF371440, a *tag* foi extraída do RNAm portador de cauda poli-A que cobre a extremidade 3' do seu respectivo gene predito. O número de acesso do GenBank para este RNAm é BC017966. A *tag* provinda do RNAm predito da sequência BF371440 foi encontrada 269 vezes no banco de dados de SAGE, distribuídas em 61 bibliotecas, 238 vezes (40 bibliotecas) de cérebro, dos quais 223 (34 bibliotecas) de tumor de cérebro. A *tag* referente à AW996738 foi encontrada 31 vezes no banco de SAGE, distribuídas em 17 bibliotecas, 27 vezes (13 bibliotecas) de tecido tumoral e, destes, 13 (6 bibliotecas) de tumor de mama. A *tag* do outro RNAm predito, correspondente a sequência AW999051 foi encontrada 24 vezes em 18 bibliotecas, 18 vezes (15 bibliotecas) de tecido tumoral. A análise de expressão por SAGE mostrou resultados escassos para 363833_OR e BE062313. O resultado da contagem da população de ESTs presentes na região delimitada pelos transcritos preditos podem ser conferidos na Tabela 6. As análises estatísticas realizadas em ambos os casos, SAGE *tags* e conteúdo de ESTs, foi a Bayseana (TOWNSEND 2002; VÊNCIO 2004), na qual o erro de Bayes deveria ser menor que 0,3 para considerarmos que um transcrito se mostra diferencialmente expresso. Em outras palavras, a área de sobreposição das distribuições das duas populações de frequências das tags, normal e tumoral, deveria ser menor que 30% (www.lbc.ludwig.org.br/sage/betabin). Foi encontrada diferença estatisticamente significativa entre o total de SAGE *tags* de origem tumoral e normal para duas sequências, AW996738 ($E=0,2$) e para BF371440 ($E=0,2$). Não foi encontrada diferença de expressão estatisticamente significativa no resultado de contagem de ESTs para nenhum dos genes preditos.

Quanto a localização citogenética, verificamos que as seqüências BF371440 e AW999051 estão localizadas em RITEs de tecidos do mesmo tipo dos quais elas provêm, respectivamente, próstata e mama. A BE062313 esta localizada em RITE correspondente a cérebro, fígado e pâncreas. As outras duas seqüências não estão localizadas em RITE.

As informações de todas as nossas 94 seqüências inspecionadas manualmente estão disponível no endereço www.lbc.ludwig.org.br/nomatch, pelo qual se consegue *link* direto com *browser* genômico para observação do alinhamento das mesmas.

5 DISCUSSÃO

A primeira avaliação das ORESTES foi feita há cerca de quatro anos atrás, em dezembro/2000, quando o HCGP foi terminado. Nesta avaliação as seqüências foram anotadas de acordo com o seu alinhamento contra seqüências depositadas em bancos de transcritos. Ao decorrer do tempo, novas seqüências de transcritos de vários outros projetos foram geradas e publicamente disponibilizadas, o que fez com que 12.173 *clusters* de ORESTES anteriormente definidas como *no-match* deixassem esta qualidade, incorporando *full-length* e/ou novas ESTs, o que nos dá uma taxa de 3.000 clusters/ano. Esta taxa nos fornece uma idéia de quão substancial é o conjunto de dados das ORESTES *no-match* confirmando nossa hipótese de que estes dados são em parte constituídos por transcritos de baixa abundância ou de difícil detecção e capturados por uma metodologia sensível, por isto, apenas gradualmente ao longo do tempo vão sendo obtidos por outras metodologias de geração de ESTs. Por outro lado, percebemos que uma grande fração deste conjunto de dados é possivelmente DNA/RNA contaminante de outras espécies, quimeras ou ainda seqüências com alto grau de erro, uma vez que apenas cerca de um terço das que avaliamos em nosso trabalho (23.189 / 65.165) atingiu nosso critério de identidade genômica.

O perfil das seqüências selecionadas que foram levadas à inspeção manual nos mostra que 60 dos 94 clusters possuem seqüências novas *bona fide*. 13,8% das seqüências anotadas no map4_bh1 como portadoras de *splicing* não apresentavam *gaps* no alinhamento genômico (Figura 3 e Tabela 5). Para este fato pensamos ser estas anotações errôneas devido a falsos *gaps* mostrados no primeiro alinhamento,

uma vez que nesta ocasião a qualidade de montagem de partes do genoma humano era inferior à atual. Mostra-nos também que 32,9% dos *hits* foram em região intragênica (íntrons ou éxons) de genes conhecidos (*RefSeq Status Provisional, Validated e Reviewed*). Para os intragênicos intrônicos com ou sem *splicing*, com *splicing* expandindo exôn e exônicos-intrônicos, que perfazem 24,4%, especulamos que possam ser prováveis formas alternativas de baixa expressão destes genes já caracterizados e que podem ser levadas à validação por RT-PCR. Uma parte destas pode ser também provinda de outros genes sobrepostos, ou ainda RNAs não codificantes (ncRNA) (KAPRANOV et al. 2002; REIS 2004), principalmente no caso das alinhadas apenas em íntrons (22,3%). Não obstante, entre as intragênicas intrônicas sem *gaps*, que perfazem 5,3%, uma parcela pode ser contaminação genômica ou provinda de RNAm imaturo. Todas estas especulações são igualmente pertinentes para os *hits* em genes preditos (39,3%), mantendo-se as devidas proporções (Tabela 5). Já para os *hits* na região intragênica exônica de genes caracterizados (10,5%) duas implicações podem ser consideradas: a primeira é que o número crescente de seqüências depositadas nos bancos de dados públicos faz com que haja uma continua defasagem de dados nos bancos locais “congelados”. Isto fica claro em nosso estudo, pois havíamos tentado retirar todos os clusters que possuísem transcritos já validados, os *full-lengths* e ESTs outras que não ORESTES. A outra implicação é de que nosso *pipeline* parece ser eficaz, pois obtivemos de imediato que dez por cento do que foi selecionado é real transcrito humano. Focando apenas a relação entre o número de clusters levados a validação e o número dos que foram confirmados (5/8), temos 62,5% de eficácia. As cinco seqüências por nós confirmadas, embora denotem a existência de cinco novos transcritos, são apenas

fragmentos, se fazendo necessário a extensão dos mesmos, por técnicas como RACE, por exemplo, para que sejam plenamente caracterizados. Apesar de serem fragmentos, a confirmação da existência destes transcritos, já permite uma avaliação experimental quantitativa de sua expressão nos estados normal e tumoral, seja por PCR ou *microarray*, por exemplo. Já a busca por função, por homologia, por exemplo, se torna inteiramente viável somente após a extensão destes transcritos.

Com intuito de determinarmos se foi vantajosa a aplicação do *pipeline* para explorar as *no-match*, comparamos o grupo de 94 *clusters* selecionados pelo *pipeline* com grupos de 94 *clusters* pegos aleatoriamente entre aqueles removidos ao decorrer do processo. Estes grupos de *clusters* removidos são os representantes de uma única biblioteca, os *singletons* sem *splicing* e os *clusters* sem *splicing*. Esta comparação nos mostrou que o grupo selecionado pelo *pipeline* tem maior quantidade de *hits* exônicos em genes conhecidos e preditos (40,3 contra 1,4% dos grupos aleatórios) e menos alinhamentos intrônicos e intergênicos (50,7 contra 98,6%), o que sugere maior contaminação genômica nos grupos removidos e que a aplicação do *pipeline* para explorar as *no-match* é desta forma vantajosa.

A comparação entre o número médio da cobertura de ESTs para genes conhecidos de 65 EST/cluster (SOGAYAR 2004) (~136 EST/cluster em análise feita sobre os *full_lengths* contidos no map4_bh1) e a quantidade de ESTs encontrada nas regiões delimitadas pelos genes preditos (Tabela 6) não corrobora a expectativa de que seriam pouco abundantes ou de difícil detecção os transcritos para os quais atribuímos derivarem as ORESTES validadas. Assim, é plausível que estas sequências sejam *no-match* não pelo fato da baixa abundância de ESTs que representam os transcritos dos quais são provenientes, mas pelo fato de que a maioria

das ESTs se encontra preenchendo outras regiões do gene, principalmente as extremidades, não sendo possível sobreposição com estas ORESTES. Isto sugere que uma parte do conjunto de dados *no-match* sejam *singletons* ou *clusters* pequenos (3,6 ESTs/*cluster*), quando em comparação com a média de 5,9 seqüências para *clusters* compostos por ESTs de diversos projetos (SOGAYAR 2004), por representarem regiões de difícil acesso para outras metodologias. Contudo, também é possível que os éxons representados por elas sejam providos de transcritos alternativos pouco expressos, o que os faria éxons raros (SAKABE 2003). Em ambos casos, nos parece que este conjunto de dados possa ajudar na complementação de vários genes e suas formas variantes. Quanto ao gene predito correspondente a ORESTES 363833_OR, sua região mostra-se pobremente representada, com apenas quatro ESTs, denotando uma real baixa expressão ou difícil detecção.

Embora os resultados de SAGE e da quantificação de ESTs não mostrem diferença estatisticamente significativa entre tecido normal e tumoral para as ORESTES validadas AW999051 e BE062313, as marcas citogenéticas nas quais alinham são apontadas como RITEs, o que nos indica uma possível sobre-expressão em tumor. O mesmo acontece para a ORESTES BF371440, se encontrando em RITE de próstata. Para esta seqüência a análise de SAGE também sugere uma tendência de maior expressão em tumor, destacando-a como candidata a marcador tumoral.

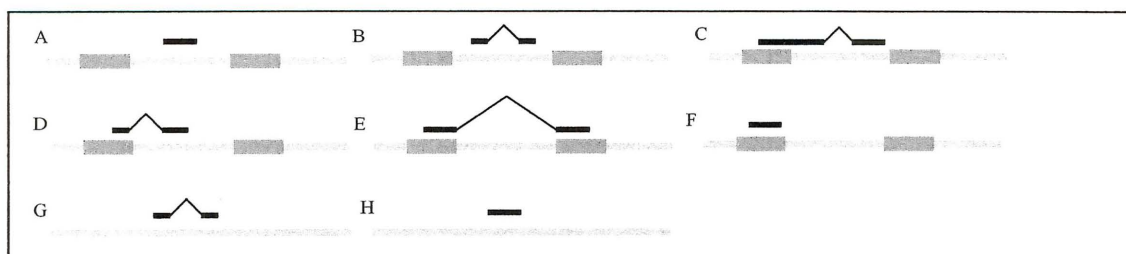
Os resultados do *pipeline* obtidos para o conjunto de dados de próstata e mama nos encorajaram a repassá-lo para o total das ORESTES *no-match*, objetivando a obtenção de um grande número de seqüências para validação (Figura 4). O número por nós obtido desta vez foi de 335 clusters, para os quais há necessidade ainda de uma inspeção manual. Considerando os resultados anteriores,

esperamos que deste montante cerca de 131 destes clusters alinhem em genes preditos ainda não validados, e que possam ser estendidos e caracterizados com o auxílio destas ORESTES.

Ao repassarmos o pipeline para o restante das ORESTES *no-match* tivemos acesso ao número total destas que não atendem ao nosso critério de identidade, 207.206 (58%). Isto representa 17,4% do total das ORESTES geradas no HCGP. Desta forma, se somarmos estas àquelas identificadas como contaminantes na primeira avaliação (20.1% do total HCGP; CAMARGO et al. 2001a) temos que 37,5% do total das ORESTES são seqüências espúrias. A título de comparação podemos citar HILLIER (1996), que relatou 28% de contaminação em bibliotecas 5' e 3'.

Com o objetivo de maximizar a probabilidade de serem as seqüências *bona fide*, um grande número de clusters sem *splicing* foi descartado, resultando em uma pequena quantidade de seqüências para validação. Por isto, não foi possível explorar todo o potencial das ORESTES *no-match*, conseqüentemente deixando de lado uma grande fração destas que representam transcritos humanos. Portanto, para aumentar o número de seqüências *no-match* levadas a validação pensamos em utilizar os *clusters* com *splicing* cujas seqüências não mostraram sítios canônicos de processamento, assim como os maiores *clusters* sem *splicing*, sobretudo aqueles com maior diversidade de bibliotecas (Figuras 5,6,7,8,9) e com ESTscan *score* positivo (~1/3 do total). A tentativa de validação para estes será feita por *microarray*. Acreditamos que esta abordagem de validação em larga escala poderá contribuir na identificação de um número significativo de novos transcritos humanos.

Esta avaliação revelou o potencial das ORESTES *no-match* para identificação de novos genes, novas isoformas e novos marcadores tumorais.



Legenda: As barras representam DNA genômico. Em cinza claro estão os íntrons (com exceção para G e H, onde representam regiões intergênicas), em cinza escuro estão os éxons e em preto as ESTs alinhadas. Traços finos representam os *gaps*. A) Intrônico sem splicing; B) Intrônico com splicing; C) Éxon expandido com splicing sobrepondo íntron; D) Splicing intrônico-Exônico; E) Exônico com splicing; F) Exônico sem splicing; G) Intergênico com splicing e H) Intergênico sem splicing.

Figura 3 - Tipos de alinhamentos encontrados na inspeção manual.

Tabela 5 - Resultados gerais dos alinhamentos, classificados de acordo com suas categorias e subcategorias.

	Tipos de hits genômicos	Números de hits	Porcentagem de seqüências no-match	Comentários
INTRAGÊNICO (<i>RefSeq Status: Reviewed, Provisional ou Validated</i>)	Intrônico sem splicing	5	5.3%	Prováveis formas alternativas de <i>splicing</i> / genes sobrepostos / ncRNA e contaminação genômica
	Intrônico com splicing	16	17%	
	Éxon expandido com splicing sobrepondo íntron	1	1%	
	Splicing exônico-intrônico	1	1%	
	Exônico c/ splicing	8	8.5%	Genes recentemente identificados
INTRAGÊNICO (genes preditos – <i>RefSeq Status: Model ou Predicted</i>)	Intrônico sem splicing	3	3,2%	Prováveis formas alternativas de <i>splicing</i> dos genes preditos / genes sobrepostos / ncRNA
	Intrônico com splicing	4	4.2%	
	Splicing exônico-intrônico	5	5.3%	
	Exônico c/ splicing	16	17%	Corroboração predição gênica
	Exônico sem splicing	1	1%	
Predição <i>ab initio</i>	Intrônico c/ splicing	3	3,2%	Corroboração predição gênica
	Éxon expandido com splicing sobrepondo íntron	1	1%	
	Splicing exônico-intrônico	1	1%	
	Exônico c/ splicing	3	3,2%	
INTERGÊNICO (região sem anotação)	Com splicing	3	3,2%	Contaminação genômica / genes não preditos / ncRNA
	Sem splicing	4	4.2%	
Início ou término alternativo	Com splicing	0	0%	Prováveis formas alternativas
	Sem splicing	0	0%	
Hits opostos	-	17	18%	Quimera
Vários hits locais e globais	-	2	2,1%	
TOTAL	-	94	100%	-

Tabela 6 - Informações sobre o cinco mRNA associados com as seqüências validadas.

ORESTES	Tamanho do mRNA predito (b)	Número de ESTs de tecido normal	Número de ESTs de tecido tumoral	Total de ESTs contidas na região dos genes preditos	E
BF371440	2,499	26	36	62	0,6
363833_OR	1,590	3	1	4	0,6
AW996738	4,461	73	59	132	0,9
AW999051	2,532	106	67	173	0,4
BE062313	6,369	195	161	356	0,9

Legenda: E = erro de Bayes.

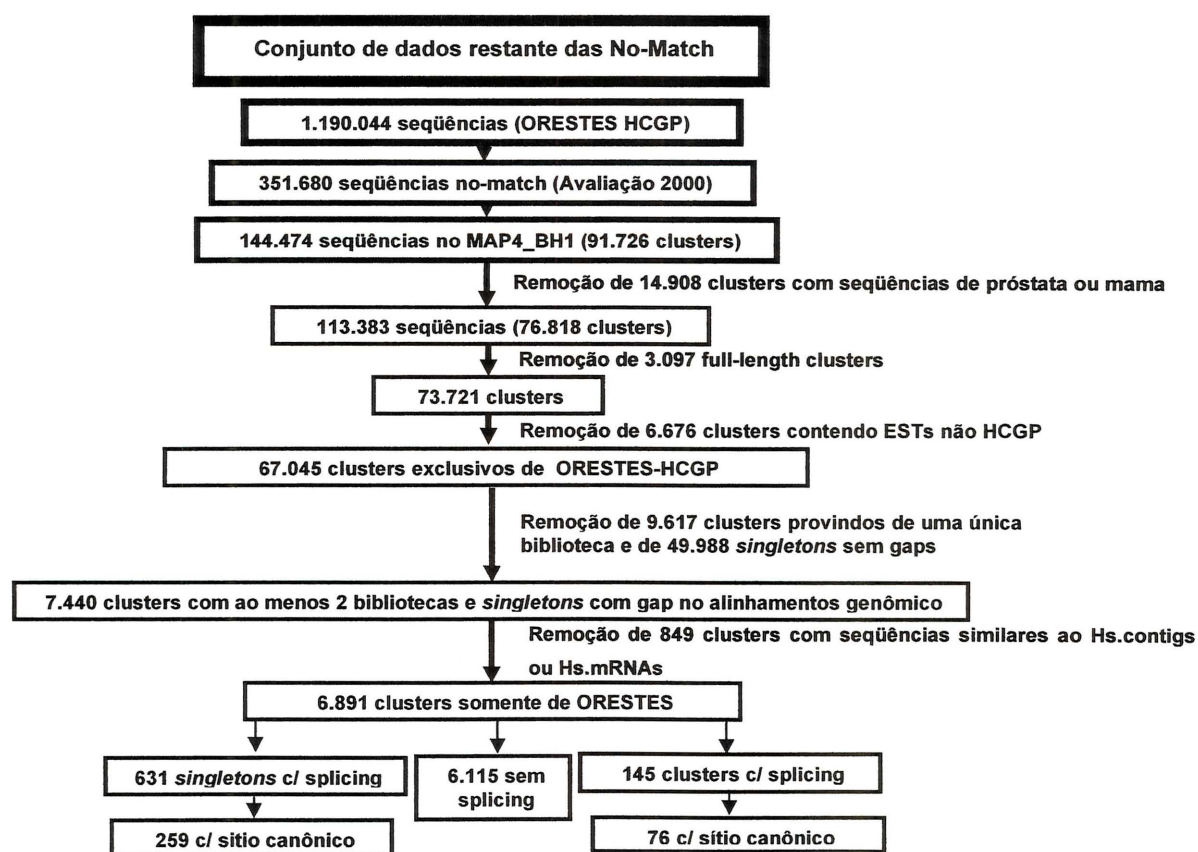


Figura 4 - Fluxograma do *pipeline* com o número restante das *no-match*.

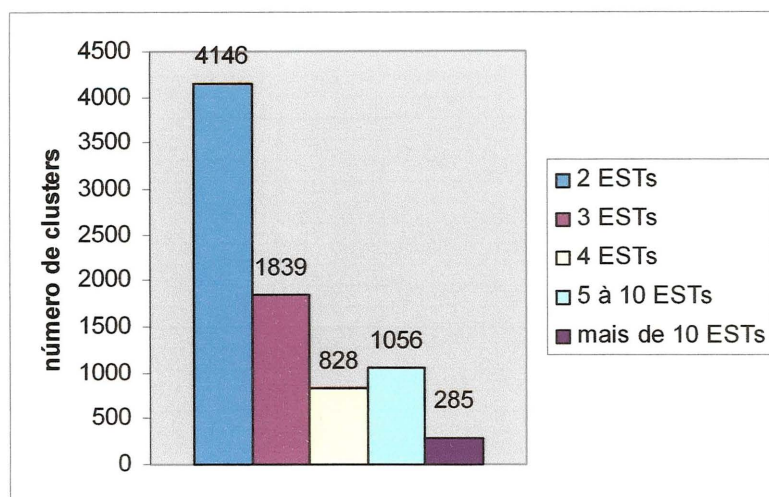


Figura 5 - Distribuição por tamanho dos *clusters* de *no-match* sem *splicing*.

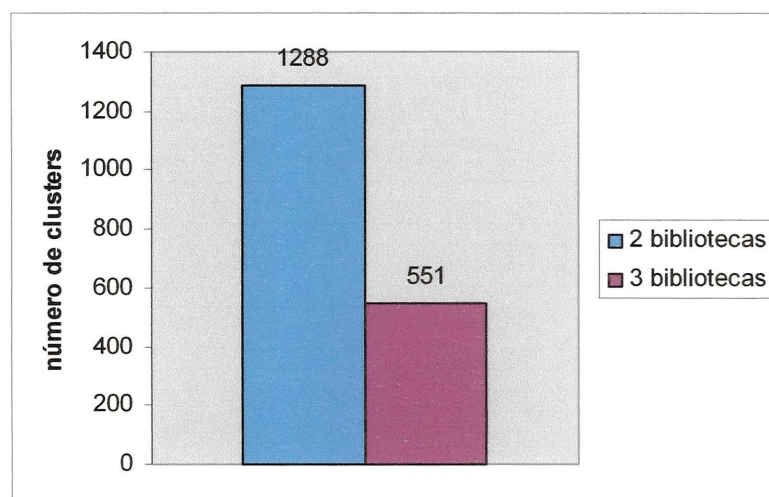


Figura 6 –Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* de tamanho três.

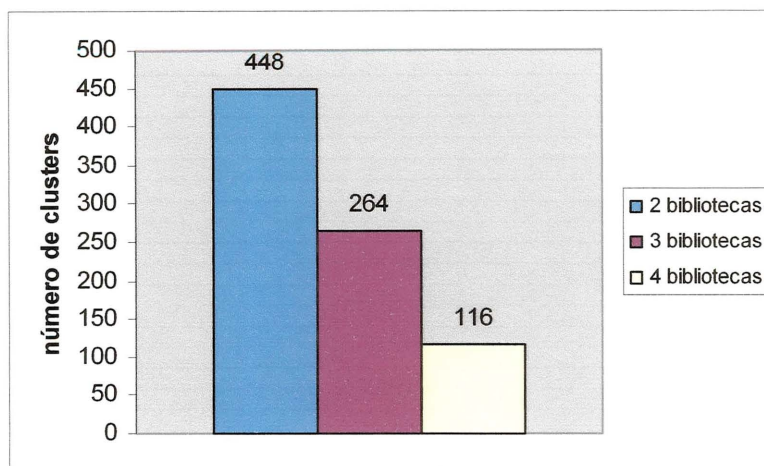


Figura 7 – Distribuição da quantidade de bibliotecas dos *clusters* sem *splicing* de tamanho quatro.

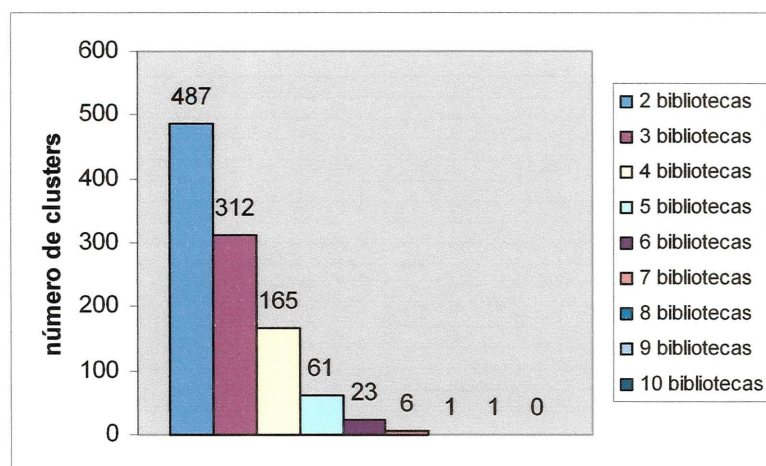


Figura 8 – Distribuição da quantidade de bibliotecas dos *clusters no-match* sem *splicing* com tamanho de 5 a 10.

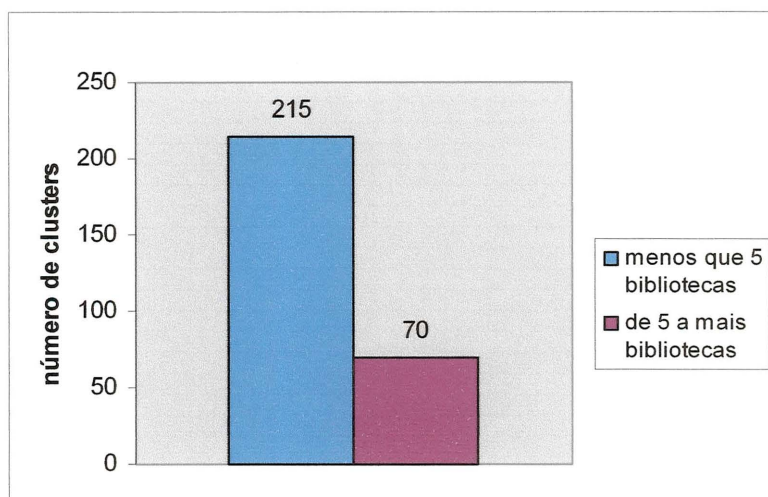


Figura 9 – Distribuição da quantidade de bibliotecas dos clusters *no-match* sem *splicing* de tamanho maior que dez.

6 CONCLUSÕES

- 1 As ORESTES representam um modo alternativo e em alguns aspectos vantajosos para geração de ESTs, no entanto, uma alta porcentagem delas (37,5%) parece ser constituída de seqüências espúrias.
- 2 É vantajosa a aplicação do *pipeline* de filtragem das ORESTES *no-match* para buscar novos genes ou novas isoformas nesta base de dados. A afirmação é baseada no fato de que o grupo de clusters selecionados apresentou menor proporção de alinhamentos intrônicos e integênicos quando comparado com os grupos removidos ao decorrer do processo de filtragem, o que sugere menor contaminação genômica no primeiro. A porcentagem de 62,5% de eficácia de validação experimental com esta aplicação também dá suporte a afirmativa.
- 3 O conjunto das ORESTES *no-match* é uma base de dados importante na identificação e caracterização de transcritos humanos. Com o resultado do *pipeline* e da validação experimental percebemos que as *no-match* podem contribuir como suporte experimental da existência de éxons preditos, na descoberta de genes não conhecidos e de formas variantes de genes já conhecidos ou preditos. Ademais, transcritos provindos desta base de dados parecem poder ajudar na identificação de novos marcadores tumorais.

7 REFERÊNCIAS BIBLIOGRÁFICAS

Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. **Science** 1991; 252:1651-6.

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

Adams MD, Celniker SE, Holt RA, et al. The genome sequence of *Drosophila melanogaster*. **Science** 2000; 287:2185-95.

Anderson S, Bankier AT, Barrell BG, et al. Sequence and organization of the human mitochondrial genome. **Nature** 1981; 290:457-65.

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

Berks M. The *C. elegans* genome sequencing project. *C. elegans* Genome Mapping and Sequencing Consortium. **Genome Res** 1995; 5:99-104.

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

Borodovsky M, McIninch J. Recognition of genes in DNA sequence with ambiguities. **Biosystems** 1993; 30:161-71.

Bortoluzzi A, d'Alessi F and Danieli A. A novel resource for the study of genes expressed in the adult human retina. **Inv Ophthalmol & Visual Science** 2000; 41:3305-9.

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

Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. **Nucleic Acids Res** 2000; 28:4364-75.

Brentani H, Caballero OL, Camargo AA, et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. **Proc Natl Acad Sci U S A** 2003; 100:13418-23.

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

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

Camargo AA, Samaia HP, Dias-Neto E, et al. The contribution of 700000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci U S A** 2001a; 98:12103-8.

Camargo AA, de Souza SJ, Brentani RR, Simpson AJG. Human gene discovery through experimental definition of transcribed regions of the human genome. **Curr Opin Chem Bio** 2001b; 6:13-16.

Chou H, Michael HH. DNA sequence quality trimming and vector removal. **Bioinformatics** 2001; 17:1093-104.

Claverie JM. Computational methods for the identification of genes in vertebrate genomic sequences. **Hum Mol Genet** 1997; 6:1735-44.

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

Delseny M, Raynal M, Laudie M. Sequencing and mapping the Arabidopsis genome: a weed for real crops. **Symp Soc Exp Bio** 1996; 50:5-9.

Eckman BA, Aeronson JS, Borkowski JA, et al. The merck gene index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. **Bioinformatics** 1998; 14:12-13.

Fickett JW. Finding genes by computer: the state of the art. **Trends Genet** 1996; 12:316-20.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD. The minimal gene complement of Mycoplasma genitalium. **Science** 1995; 270:397-403.

Fleischman RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science** 1995; 269:496-512.

Garg K, Green P and Nickerson, DA. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. **Genome Res** 1999; 9:1087-92.

Glisin V, Crkvenjakov R, Byus C. Ribonucleic acid isolated by cesium chloride centrifugation. **Biochemistry** 1974; 13:2633-7.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H. Life with 6000 genes. **Science** 1996; 274:546, 563-7.

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

Herbert A, Rich A. RNA processing and the evolution of eukaryotes. **Nat Genet** 1999; 21:265-9.

Henke W, Herdel K, Jung K, Schnorr D, Loening SA. Betaine improves the PCR amplification of GC-rich DNA sequences. **Nucleic Acids Res** 1997; 25:3957-8.

Hillier L.D., Lennon G, Becker M. Generation and analysis of 280,000 human expressed sequence tags. **Genome Res** 1996; 9:807-28.

Houlgatte R, Mariage-Samson R, Duprat S, et al. The Genexpress Index: a resource for gene discovery and the genic map of the human genome. **Genome Res** 1995; 5:272-304.

Hudson TJ, Stein LD, Gerety SS, et al. An STS-based map of the human genome. **Science** 1995; 270:1945-54.

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

Jorde LB, Carey JC, Bamshad MJ, White RL. *Genética Médica*. Granabara Koogan 2º ed. 1999; pág 3.

Kan Z, Rouchka EC, Gish WR, States DJ. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. **Genome Res** 2001; 11:889-900.

Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. **Science** 2002; 296:916-9.

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

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

Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. **Genomics** 1996; 33:151-2.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. An optimized protocol for analysis of EST sequences. **Nucleic Acids Res** 2000; 28:3657-65.

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

Maxam AM, Gilbert W. A new method for sequencing DNA. **Proc Natl Acad Sci U S A** 1997; 74:560-4.

Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. **Nucleic Acids Res** 2001; 29:2850-59.

O'Brien C. Cancer genome anatomy project launched. **Mol Med Today** 1997; 3:94.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, et al. The promise of comparative genomics in mammals. **Science** 1999; 286: 458-62, 479-81.

Okubo K, Hori N, Matoba R, et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. **Nat Genet** 1992; 2:173-9.

Phillips RL, Ernst RE, Brunk B, et al. The genetic program of hematopoietic stem cells. **Science** 2000; 288:1635-39.

Reis EM, Nakaya HI, Louro R, et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. **Oncogene** 2004; 23:6684-92.

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

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. **Methods Mol Biol** 2000; 132:365-86.

Sakabe NJ, de Souza JE, Galante PA, et al. ORESTES are enriched in rare exon usage variants affecting the encoded proteins. **C R Biol** 2003; 326:979-85.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci U S A** 1997; 74:5463-7.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. **J Mol Biol** 1982; 162:729-73.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS, Barros MH, Bonaccorsi ED, Bordin S, Bove JM, Briones MR, Bueno MR, Camargo AA, Camargo LE, Carraro DM, Carrer H, Colauto NB, Colombo C, Costa FF, Costa MC, Costa-Neto CM, Coutinho LL, Cristofani M, Dias-Neto E, Docena C, El-Dorry H, Facincani AP, Ferreira AJ, Ferreira VC, Ferro JA, Fraga JS, Franca SC, Franco MC, Frohme M, Furlan LR, Garnier M, Goldman GH, Goldman MH, Gomes SL, Gruber A, Ho PL, Hoheisel JD, Junqueira ML, Kemper EL, Kitajima JP, Krieger JE, Kuramae EE, Laigret F, Lambais MR, Leite LC, Lemos EG, Lemos MV, Lopes SA, Lopes CR, Machado JA, Machado MA, Madeira AM, Madeira HM, Marino CL, Marques MV, Martins EA, Martins EM, Matsukuma AY, Menck CF, Miracca EC, Miyaki CY, Monteriro-Vitorello CB, Moon DH, Nagai MA, Nascimento AL, Netto LE, Nhani A Jr, Nobrega FG, Nunes LR, Oliveira MA, de Oliveira MC, de Oliveira RC, Palmieri DA, Paris A, Peixoto BR, Pereira GA, Pereira HA Jr, Pesquero JB, Quaggio RB,

Roberto PG, Rodrigues V, de M Rosa AJ, de Rosa VE Jr, de Sa RG, Santelli RV, Sawasaki HE, da Silva AC, da Silva AM, da Silva FR, da Silva WA Jr, da Silveira JF, Silvestri ML, Siqueira WJ, de Souza AA, de Souza AP, Terenzi MF, Truffi D, Tsai SM, Tsuhako MH, Vallada H, Van Sluys MA, Verjovski-Almeida S, Vettore AL, Zago MA, Zatz M, Meidanis J, Setubal JC. The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. **Nature** 2000; 406:151-7.

Snyder EE, Stormo GD. Identification of protein coding regions in genomic DNA. **J Mol Biol** 1995; 248:1-18.

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

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

Strausberg RL. The cancer genome anatomy project: new resources for reading the molecular signatures of cancer. **J Pathol** 2001; 195:31-40.

Townsend JP, Hartl DL. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. **Genome Biol.** 2002; 3(12): RESEARCH0071.

Uberbacher EC, Xu Y, Mural RJ. Discovering and understanding genes in human DNA sequence using GRAIL. **Methods Enzymol** 1996; 266:259-81.

Uberbacher EC, Mural RJ. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. **Proc Natl Acad Sci U S A** 1991; 88:11261-5.

Vasmatzis G, Essand M, Brinkmann U, Lee B, Pastan I. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. **Proc Natl Acad Sci U S A** 1998; 95:300-4.

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

Vêncio RZ, Brentani H, Patrão DF, Pereira CA Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). **BMC Bioinformatics** 2004; online: 1-13.

Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. **Science** 2001; 292:1838.

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

Waterson RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. **Nature** 2002; 420:520-62.

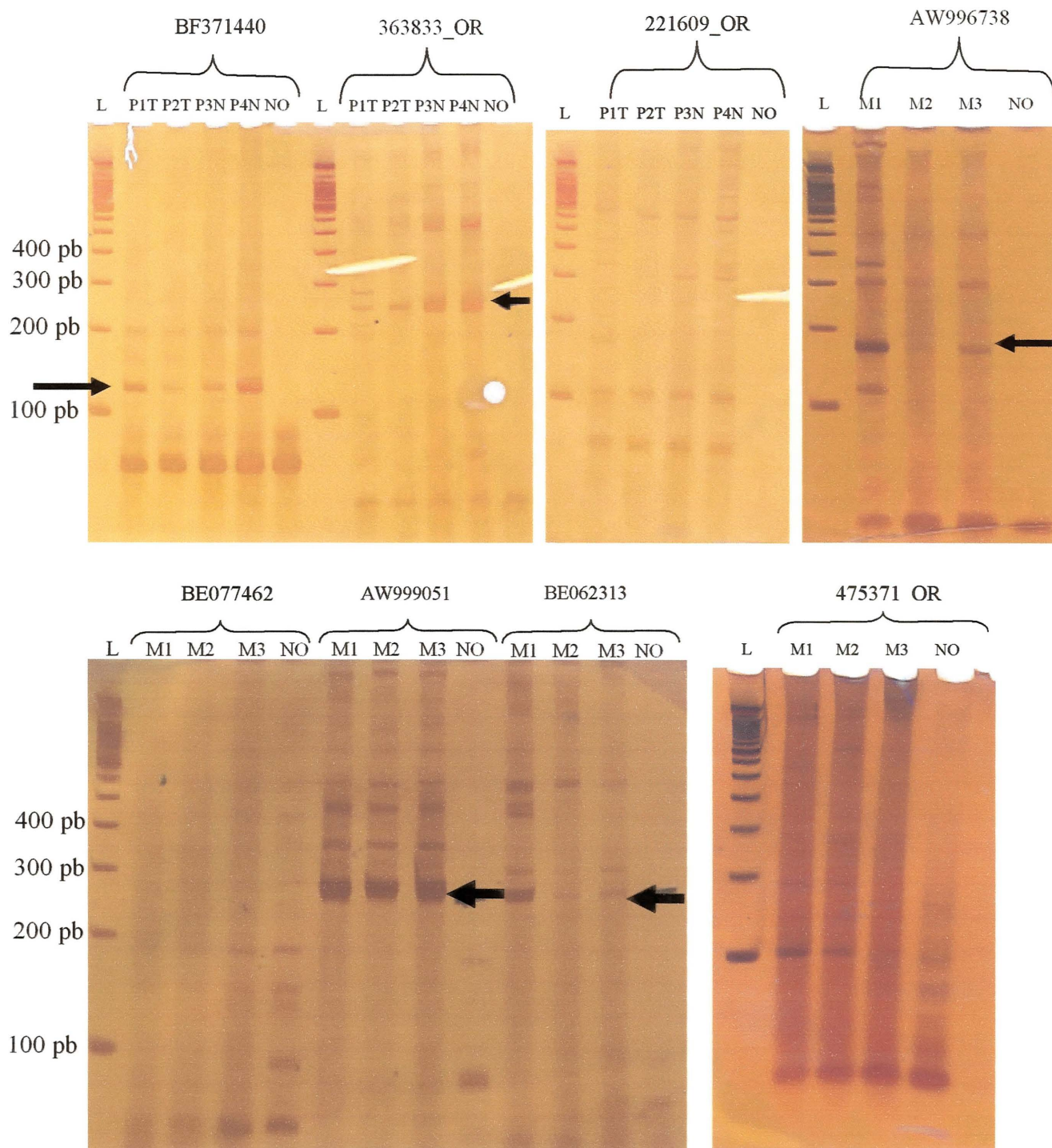
Wistow G, Bernstein SL, Touchman JW, et al. Grouping and identification of sequence tags (GRIST): bioinformatics tools for the NEIBank database. **Mol Vis** 2002; 8:164-70.

Wolfsberg TG, Landsman D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. **Nucleic Acids Res** 1997; 25:1626-32.

Yu Y, Zhang C, Zhou G et al. Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. **Genome Res** 2001; 11:1392-403.

Zhou Y, Luoh SM, Zhang Y, et al. Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. **Cancer Res** 2003; 63:5781-4.

Anexo I – Géis de poliacrilamida mostrando os produtos de amplificação dos *primers* desenhados para as ORESTES selecionadas.



Legenda: PT- próstata tumor; PN – próstata normal; M1,M2 e M3 – mama tumor. L – 100 bp Ladder Invitrogen. As setas indicam as bandas esperadas.

Anexo II - Lista com as informações da inspeção manual dos 94 clusters selecionados pelo pipeline.

Cluster	Largest Spliced ORESTES	Crom	Marca Citogenét	Contig Gen	Hit	RefSeq Status
>1331	375349_OR	1	1q25.1	NT_004487.16	E	predicted
>4809	AW935873	1	1p36.13	NT_004610.16	E	model
>5339	BE081366	1	1q32.3	NT_021877.16	B	reviued
>253	443249_OR	1	1p32-p3	NT_032977.6	B	reviewed
>1757	299676_OR	1	1p13.1	NT_004754.15	A	predicted
>2185	238431_OR	1	1q43	NT_004836.15	Q	-
>11119	BF331779	2	2q35	NT_005403.14	E	predicted
>9351	BF816101	2	2p24.1	NT_022184.13	E	provisional
>9307	221609_OR	2	2p23.3	NT_022184.13	G-P-D-V	ab initio
>8773	540115_OR	2	2p13	NT_022184.13	B	reviewed
>12699	BE008890	2	2q11.2	NT_022171.13	B	reviewed
>9071	AW999890	2	2p23.2	NT_022184.13	H	-
>12313	BE062313	2	2p11.2	NT_022184.13	G-P-C-V	ab initio
>138202	BE086696	2	2q21.2	NT_005058.14	A	validated
>9789	BF331301	2	2q37.1	NT_005120.14	E	model
>11199	BF748905	2	2q37.1	NT_005403.14	Q	-
>39739	BF749100	2	2q35	NT_005403.14	Q	-
>55469	BE089058	3	3q21.2	NT_005612.14	Q	-
>55743	BE066140	3	3p26.1	NT_022517.16	B	provisional
>42329	AW999051	3	3q29	NT_005535.15	G-P-E-V	ab initio
>58683	BE068995	4	4q32.2	NT_016606.16	Q	-
>58547	392448_OR	4	4q22.1	NT_016354.16	Q	-
>75485	AW996738	5	5p13.2	NT_006576.14	G-P-E-V	ab initio
>76477	BF750352	5	5q13.3-q14.1	NT_006713.13	B	provisional
>74225	BE081711	5	5q23.2	NT_034772.5	B	provisional
>97370	479506_OR	6	6p21.33	NT_007592.13	G	-
>97946	BF350531	6	6p22.3	NT_007592.13	Q	-
>97964	BE005626	6	6p22.3	NT_007592.13	D	model
>96974	BF330647	6	6p21.32	NT_007592.13	B	reviewed
>98726	BF740910	6	6q22.31	NT_025741.13	G	-
>95860	BF744369	6	6q23.3	NT_025741.13	E	validated
>95930	406073_OR	6	6q24.1	NT_025741.13	E	provisional
>98366	BE092837	6	6p25.3	NT_034880.3	B	reviewed
>95528	AW992696	6	6q27	NT_007422.12	B	provisional
>96854	BF330199	6	6p21.31	NT_007592.13	E	validated
>98150	475371_OR	6	6p24.3	NT_007592.13	G-P-B-V	ab initio
>96652	217436_OR	6	6p21	NT_007592.13	E	provisional
>99316	BF325801	7	7q36.3	NT_079597.1	A	model
>100902	BF371227	7	7q36.1	NT_079596.1	B	predicted
>100112	BE073317	7	7q34	NT_079596.1	Q	-
>101770	BE092109	7	7p21.1	NT_079592.1	D	model
>105286	218403_OR	8	8p22	NT_030737.8	E	validated
>105202	BF371440	8	8q21.13	NT_008183.17	G-P-E-V	ab initio
>134414	497147_OR	8	8q24.3	NT_008046.14	B	reviewed
>105112	BF330594	8	8p23.1	NT_077531.3	E	model
>134332	406555_OR	8	8q21.11	NT_008183.17	Q	-
>134642	BF329643	8	8p21.2	NT_023666.16	B	predicted
>136476	BF328518	9	9p24.2	NT_008413.16	C	provisional
>136226	BF751669	9	9q22.33	NT_008470.16	Q	-
>136056	385338_OR	9	9q34.11	NT_00847616	E	predicted
>138008	BF373406	9	9q21.31	NT_023935.16	F	model
>136260	BF858487	9	9q22.33	NT_008470.16	E	predicted
>134788	BI046360	9	9q34.3	NT_024000.15	D	provisional
>136394	AW996810	9	9p24.1	NT_008413.16	D	model
>41893	AW998378	9	9q31.2	NT_008470.16	E	predicted
>139226	BF371181	10	10q26.11	NT_030059.11	Q	-
>139118	BE005443	10	10q26.2	NT_035040.4	H	-
>140354	BE079594	10	10q24.32	NT_030059.11	Q	-
>166981	BE060994	11	11p13	NT_009237.16	B	reviewed

>166933	BF332512	11	11q13.1	NT_033903.6	B	predicted
>173173	BE093511	12	12q13.2	NT_029419.10	Q	-
>172775	BF332254	12	12q14.1	NT_029419.10	E	predicted
>173027	BF900722	12	12q12	NT_029419.32	E	model
>171397	BE072957	12	122q12-q13.11	NT_029419.10	E	model
>171797	249693_OR	12	12q24.31	NT_009755.16	H	-
>190921	BE068988	13	13q32.3	NT_009952.14	A	provisional
>190495	BE070021	13	13q13.3	NT_024524.13	E	model
>194005	AW993564	14	14q22.3	NT_026437.10	E	validated
>194551	BF748287	14	14q32.32	NT_026437.10	A	provisional
>194089	238644_OR	14	14q22.1	NT_026437.10	Q	-
>197187	BF823136	15,Y,Mit		NT_029419.32	Q	-
>196841	BF366684	15	15q22.32	NT_010194.16	E	model
>199457	BG010766	16	16q23.2	NT_024797.14	E	provisional
>198513	BE065591	16	16p13.3	NT_037887.3	A	reviewed
>200769	AW993120	16	16p11.2	NT_024812.10	B	provisional
>202287	BF371036	17	17q25.3	NT_010663.14	B	reviewed
>202317	BF855686	17	17q25.3	NT_0224871.11	E	model
>203937	376565_OR	17	17p13.1	NT_010718.14	Q	-
>203647	BE077462	17	17p22	NT_010718.14	G-P-B-V	ab initio
>233659	BE062197	17	17q23.3	NT_010783.14	G	-
>203449	474845_OR	17	17p13.1	NT_010718.14	B	reviewed
>170221	BE009356	17	17q24.1	NT_010783.14	E	model
>203817	BE073375	17	17p11.2	NT_010718.14	A	reviewed
>235519	176494_OR	18	18q23	NT_025004.13	H	-
>235551	BQ307142	18	18q21.1	NT_010966.13	D	model
>237085	382629_OR	19	19q12	NT_011109.15	D	model
>238377	BE090820	19	19q13.31	NT_011109.15	Q	-
>242293	363833_OR	20	20p12.2	NT_011387.8	G-P-B-V	ab initio
>243129	BF326892	20	20q11.23	NT_011362.8	E	predicted
>243281	BE080754	20	20q12-q13.1	NT_011362.8	B	reviewed
>245725	AW992824	22	22q12.2	NT_011520.9	A	predicted
>247057	BE085704	22	22q13.31	NT_011523.9	B	model
>270473	BE008446	X	Xq11.2	NT_011669.14	Q	-
>287	547196_OR	45 hits			Q	-

Legenda: A=Intrônico sem *splicing*; B=Intrônico com *splicing*; C=Éxon expandido com *splicing* sobrepondo intron; D=*Splicing* intrônico-Exônico; E=Exônico com *splicing*; F=Exônico sem *splicing*; G=Intergênico com *splicing*; H=Intergênico sem *splicing*; P=predito *ab initio*; V=selecionado para Validação; Q=quimera;

Anexo III – Sequências FASTA dos produtos de amplificação e das ORESTES originais e a interface gráfica do alinhamento de ambas contra o genoma humano.

ORESTES BF371440

Sequencia FASTA extraída do cromatograma do produto de PCR dos primers P1P2 desenhados para anelarem nos dois exons da ORESTES BF371440

↳ Rogerio_P1P2-2_C10_06 sequence exported from chromatogram file

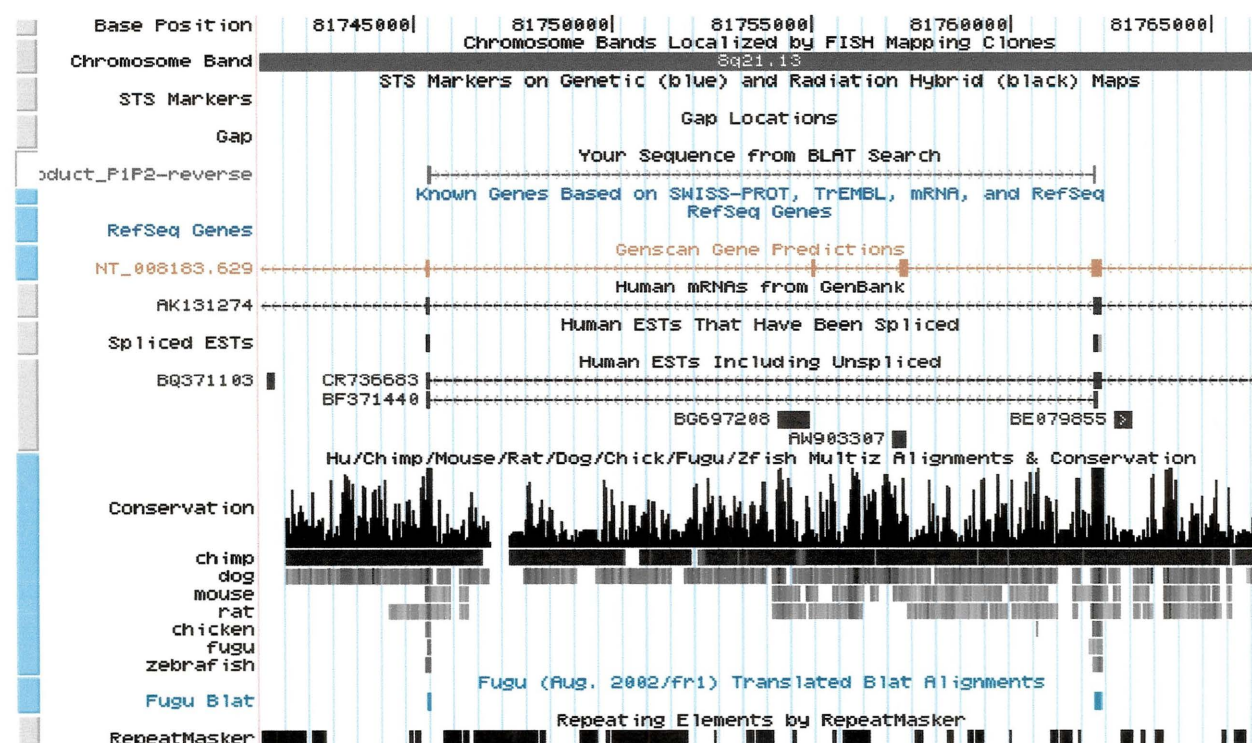
```
GNNNNNCNNGNNNNNNNNCTGGGNAATGCTGCAGTGCTCAGCAACTTCCCACANCNTNTTCCAGAGGGNNANTGAACATCACCTTNANGGA  
ATCTTTCTTTCTCTGGGAATGGGCTCGTCANNNNNNNNNNNNNNNNTTTTCCCCNNNANNNNNNAGAAAAAANGNNNNNNNNNNCNCN  
NNNNNNANNCNANAANNAGNNNNNNNNNNNNAAGNANNNNAANANNNNNNNCNAANNNNNAG
```

Sequência FASTA original

↳ BF371440

```
CTCGTACCTGCTACGCGCCCCGGCAGCCAGACGACGGTATAGACGAGGGCGGAGCCAGCAACCTGCTCTGCAGCAGAGCCATTCCCAGGA  
AAAGAAAGAATTCCATGAAGGTGATGTTCAAATGCCCTCTGGAAAACTGTGGGAAGGTGCTGAGCACTGCAGCAGGTATCCAGAAACAC  
ATCCGAACCATCCATCCGGGGCCGCTAGCA
```

Interface gráfica do alinhamento genômico de ambas sequências, a original e a provinda da amplificação pelo par de primers P1P2.



ORESTES 363833_OR

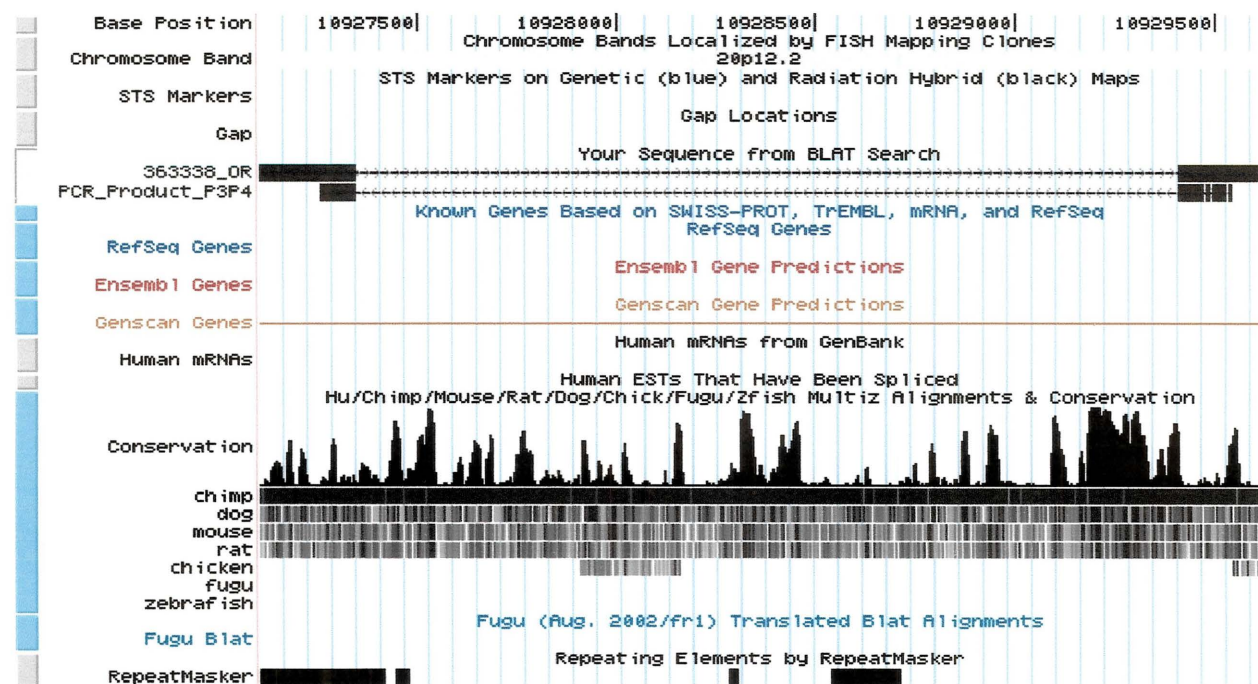
Sequencia FASTA extraída do cromatograma do produto de PCR dos primers P3P4 desenhados para anelarem nos dois exons da ORESTES 363833_OR

```
>Rogerio_P3P4-R_E10_10 sequence exported from chromatogram file - 363833_OR (seqüência errada- trocar)
TTGTCTTTTATGTCTCACTGGCGAGGAGCATTCTTATCATAAGAAAGGGAAGTGATGCTCCTTTTGCCATCCACAGAAAAAGGAAAGGAC
CATGTGGGTATAGTATGTCTTCTCCTCCCAAGTTTCAATTCTGCAGATCAGCAATTTAGGCTGGACTCAGCTGGATGTTTCTTCTTGTCTTGGC
TGGGCACCGTTATGTGTCTGTGGTCCCTGTGGTANANNNNNNGNNGNANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NCCNNNNNNNNNNNNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Sequencia FASTA

```
>363833_OR
CGGCAAAGTGTAATGTGCCTATTCTAAGTCTAGGCCTCAAGAAGTCTTGAAAGATTCTGTTCCCTCTTGGTTAGTCCACTGGAGGATGAG
AGACCCACGGCCAATAACCTCCATTGCCTCGGTCAACAGACAGCCAACCTCCAGTCTCACCATTGCCTAGACTACCACAGGGACCACAGA
CACATAACGGTGCCCAAGACAAGAAGAAACATCCAGCTGAGTCCAGCCATAAATTGCTGATCTGCAGAAATTGAAAACCTGGGAGGA
AGACATACTATACCACATGGTCCCTTTCTTTTCTGTGGATGGCAAAAAGATGCATCATTCCCGGAGCTTATGATAGAAGATGCTCCTTC
GCCAGTGTGAGACACTTAAAGGACAATGATTTGTGCAGGGAACCTAACCTGTGTGCTACGTGTCTTTTGTCTCCATTTTACTGATGTGAC
AATTGGAG
```

Interface gráfica do alinhamento genômico de ambas seqüências, a original e a provida da amplificação dos primers P3P4.



ORESTES AW996738

Sequencia FASTA extraída do cromatograma do produto de PCR dos primers P7P8 desenhados para anelarem nos dois exons da ORESTES AW996738

>Rogério_P7P8_C06_06 sequence exported from chromatogram file

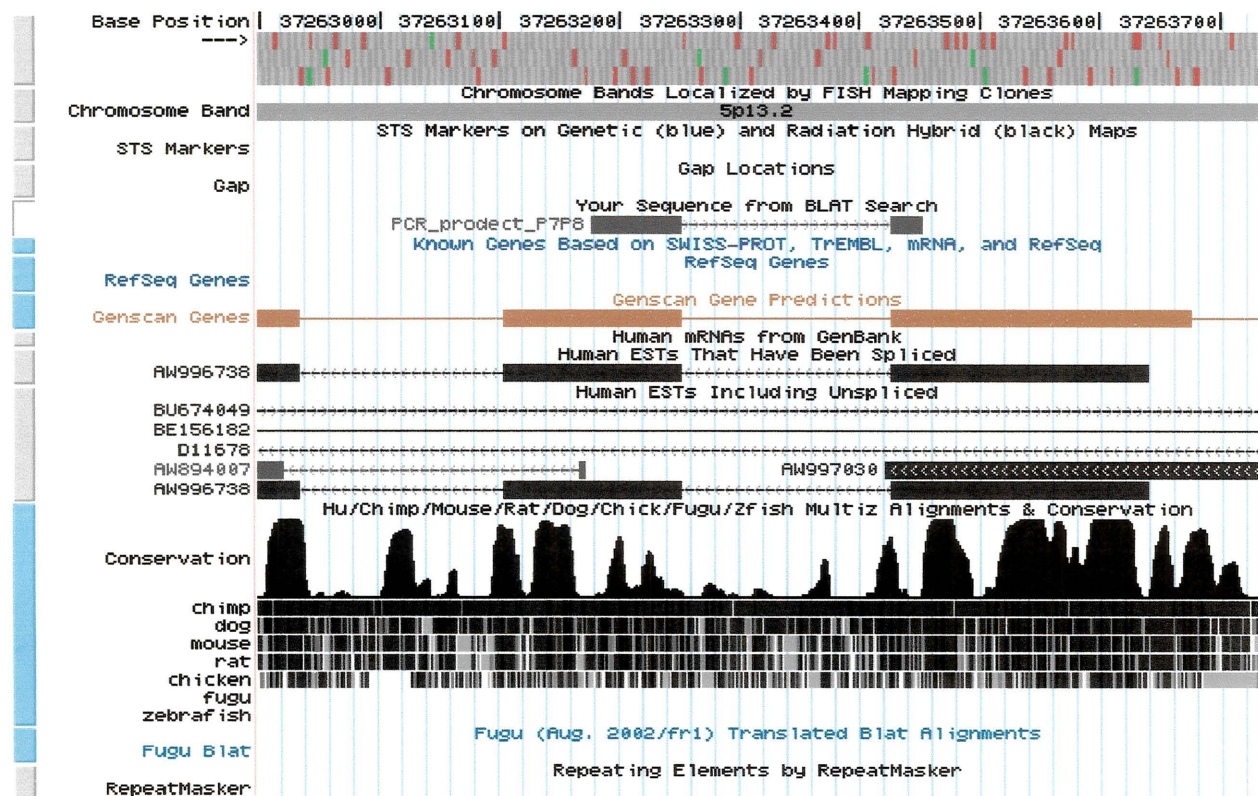
```
NNNGNNNGNAGGCGGAATTTTCATTTCCTTGGTGTTCCTTNCNNNCTAG-ANGGGAGGGAATTCAGTGATCNCNGTTCAGTCCTTTGCCTTTT
GGCTTAGACAATATCACACTTTGATATATTTTCTCAAGCCTCTGGGTTGAATCAAANA
```

Sequência FASTA original

>AW996738

```
CTACGAGATGCAATTCCAATCTTCCTTCCTTCATACTACTACACAGCACATCATCTCTTTTGTCCAAAAAGGACATAAGTTGTCTGGAAA
GTGTCTGCCTTCGTTAGTTTCTTCTGCTTCAAATCCTGAAAATCTGCTGCATTTTCATTATTGTTTCTTCTGCCTGCAAGAATTTGGGGA
CAGTGAAATCGGCTGAACTTTCATTTCCTTGGTGTTCCTAACAGGCTAGACCTTAGGGAATTCAGTGATCGCAAGTTCAGTCCTTTGCCTTT
TGGCTTAGACAATATCACACTTTGATATATTTTCTCAAGCCTCTGGGTTGAATCAAAGTAGAAGTGATCTCATGTGTACTGATGGAGATAGG
CTATCAAGAAATCGAAGGGTTGTGACCATATATCCATCAGATATAACGAGGTAGGGTAACCGTGAGTGTGCTTTTATAGAAAATCTCTGT
CTCATAGGGTCACTATCAGAAGCTGATGAATCAACAG
```

Interface gráfica do alinhamento genômico de ambas sequências, original e a provinda da amplificação pelos primers P7P8.



ORESTES AW999051

Sequência FASTA extraída do cromatograma do produto de PCR dos primers P11P12 desenhados para anelarem nos dois exons da ORESTES AW999051

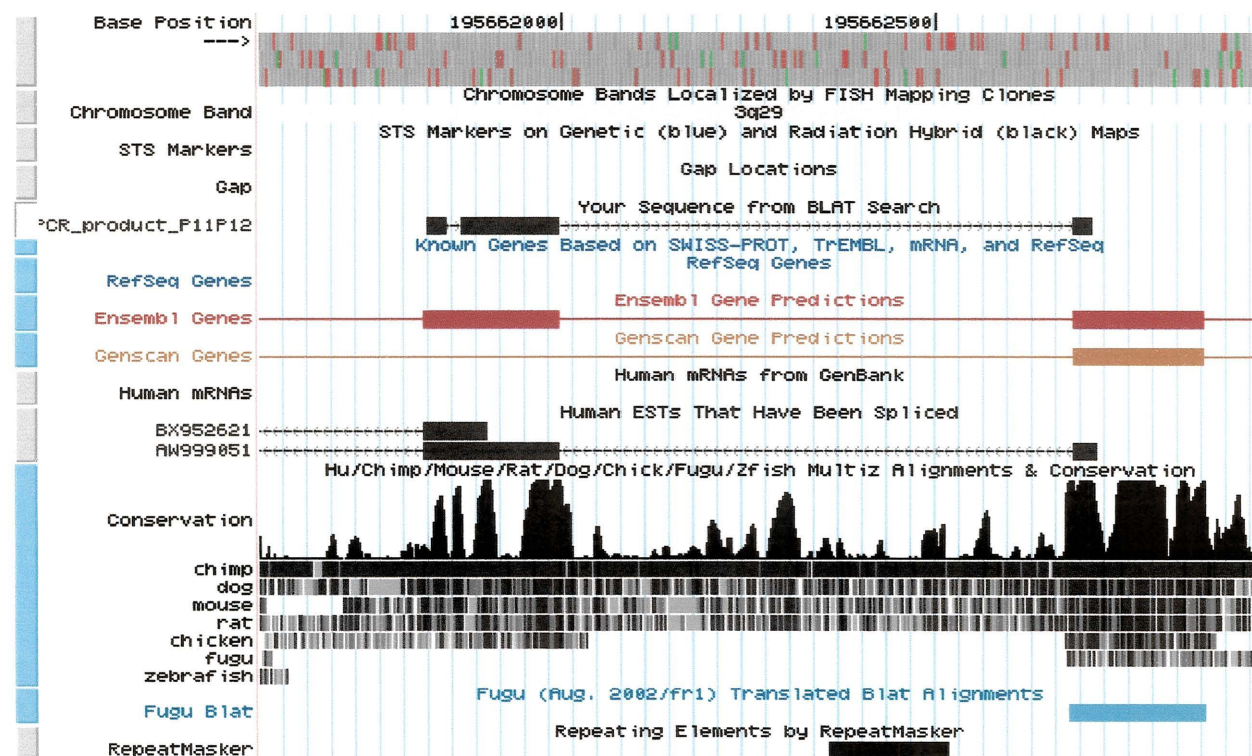
↳Rogerio_P11P12_D06_08 sequence exported from chromatogram file

```
GNNNNNNNAGNNANAGTGTGACAGTCTGTGAATATTTACTGTCNNCGTGCCGNGNNNNNTCAGTGGGATTNTCCNTTAAACAAAACGGNAT
GGCCATTGAAAAGCTTATTAGACATAGATTTTGGACTTGANACTGGGTAAAGTTTCCAAAGAAAAGAACGCGAATTTTGCACAAAACACACA
TTTGAATTCATCAGTAGTCCTCAGCAGCACTACTTCACANNNNNNNNNNNNNNNNNNNN
```

Sequência FASTA original

>AW999051

```
TTCACTTTCTACACTTGGTGGGTGAAATAACGAATCTGTTGTGATTGAGTCTGTGAATATTTACTGATCCTGTGCCTATTTTCTTCAGTGGG
ATACTCAATTAACAACTGTCATGGCCATTTGAAAAGCTTATTAGACATAGATTTTGGACTTGAAACTGGGTAAAGTTTCCAAAGAAAAGAAC
GCGAATTTTGCACAAAACACATTTTGAATTCATCAGTAGTCCTCAGCAGCACTACTTCACAGTCTTTGTGTAGAAATTGAA
```



ORESTES BE062313

Sequencia FASTA extraída do cromatograma do produto de PCR dos primers P13P14 desenhados para anelarem nos dois exons da ORESTES BE062313

>Rogério_P13P14_B06_04 sequence exported from chromatogram file

```
GGGNNACAAAGNGGGGAAAAANCANCNGNNGAATTCCNGGNAGGAAATTCNNCCTCCGNGGGGANGAAAAGNNAAGNAATTTCCCATGN
CAAGNNGAAGAAAGTAAAANTTTGGANNCCNGATGGAAAACTCCAAGACGNGTCCCCGAGAAATAANNGAAGACATNNNGGCTTCATT
ATCATGCCGCATTATCTGACNCACTNTACNTCNCACGGGAACAGACCACGGAGGGGNAANNCCGGAAAAATCNGGGGAAACANNN
AANGANAAANTNAANTNGCAAANNCGAAANGAGAGANNANACNCNGGNAANAGNNAANGCNGNNGCANGANGCGAAAAAANNNN
GANGANGGGGTGTGTTGNNNANAGNAANNNGGNAANACACGANGNGAGAGGNNAAAGNAGNCAGNNNAAGAGNNNAGAGNNNGANGGN
GGAGGGGGACGGANNNGNNAAGCNANGGNAGACGACCNNNNNGCGGANAGGAGNNGNAGAGGGNGNGNGNAGGGCGNGGGGGGGG
AGAGNANGGAGAAAGNNNGANNANTNGANNACNGGNGNGNNNANAANNAACNNAAAAGNAGGNGAAGNGNGAAGAAGNNNGAGG
AGGANNNGNCGNNGANAAACNANAGNCGNNNGCAAAGGAAGAGACNNNAANNAGGGGNGAGAAGAGGNNCGGCCGNGGGNGNAGC
GGTAGCGCNGNNGNAAAGNAGGGNGCGGGGGGNGANGNGAGAAAGGGANCCNNAGAAACGGACGGAAAAAANACAGAGNGAGNAGG
AANGNGGNGANGGGGAANGGAGAANCNAAGNNANGGNAAAAAANAAGANNAGGGAGGGGGAAGNGGNNNGAANGGAAGGNGGAN
NGGGAANAGAGACAAGNGGGGGNGCGGGGAGANAGANAANAAGANNNGGNANGAGGNCGAAGANCAACNAGGAGGAGGGGAGGA
NNGGNGAGNGNNGG
```

Sequência FASTA original

>BE062313

```
AGGAGAAGGTACACAGAGCAGTGACATTGCTCCAGCTCTAGAAGAGCACCTCTGAGATTTCCTGGCTGACACATCGCATCTCCCTCTGG
CCACAGCCAAAGCAACACCAGATGTCTGCAGTAGCAAGTAAACATTTGGATTCTCTGATGGAAAACTCCAAGACGAGCCTCTGTGAAATACCT
GCTTCATTGTGGCTTCATTTCATGCCTCATTATCTGACCCATTACTTCTTACCTCAGGGCACAGACCACGGTTGTGACACTCCGTAAAT
CCTGTGGGTATGGGGAAGAGTGAATATGACACACCGGTGGGGCAAAGGTGGCTCATGCCTGTGATCCTAGCACTTTGGGAGGCCGAGGCG
AGTGGATTACCTGAGATCAGGAGTTCGAGACCAGCTGGCCAACATGGTGAAC
```

Interface gráfica do alinhamento genômico de ambas sequências, original e a provida da amplificação pelos primers P13P14

