Prospecção Gênica: avaliação de tumores raros na descoberta de genes humanos

Ricardo Pereira de Moura

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

Área de concentração: Oncologia

Orientador: Dr. Andrew J. G. Simpson
Co-Orientador: Dr. Emmanuel Dias Neto

São Paulo

2000

Prospecção Gênica: avaliação de tumores raros na descoberta de genes humanos

Ricardo Pereira de Moura

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

Área de concentração: Oncologia

Orientador: Dr. Andrew J. G. Simpson
Co-Orientador: Dr. Emmanuel Dias Neto

São Paulo

2000

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A.C. Camargo

Moura, Ricardo Pereira

Prospecção gênica: avaliação de tumores raros na descoberta de genes humanos. / Ricardo Pereira de Moura -- São Paulo, 2000.

p. 110

Dissertação (mestrado) – Fundação Antônio Prudente.

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

Orientador: Andrew John George Simpson.

Descritores: 1.GENES. 2. EXPRESSÃO GÊNICA. 3. TUMORES/avaliação.

"Não consigo dizer se é
bom ou mau
assim como o ar me parece
vital
onde quer que eu vá
o que quer que eu faça
sem você não tem graça"

(Capital inicial)

Á minha mulher Carla.

Agradecimentos

Gostaria de expressar meu carinho e gratidão às seguintes pessoas, sem as quais o desenvolvimento deste presente trabalho não seria possível:

À toda minha família, fonte onde eu busco energia para vencer o dia a dia com mais amor, alegria e segurança.

Ao Prof. Dr. Andrew J G Simpson, por me conceder a oportunidade para o desenvolvimento deste trabalho, pela paciência e compreensão que contribuiram muito para o meu desenvolvimento pessoal e profissional.

Ao Prof. Dr. Ricardo R. Brentani, por possibilitar a criação e desenvolvimento deste trabalho em uma instituição de destaque científico.

À Catarina Simpson, pelo seu valioso apoio para a conclusão deste trabalho.

À Profª. Drª. Otávia L. Cabalero, pela amizade sincera e pelo grande apoio e estímulo.

Ao Prof. Dr. Emmanuel Dias Neto, pelas discussões durante o desenvolvimento deste trabalho, que devido à sua habilidade, foram boas oportunidades para um sólido aprendizado.

Ao Dr. Ricardo G. Corrêa, pelas boas e preciosas horas de discussão e paciência, pelo apoio e estímulo e o mais importante pela amizade.

Ao Prof. Dr Fernando Soares e ao Dr. Luiz Fernando Lopes pela ajuda na seleção de amostras para o estudo.

Ao Prof. Dr. Luis Fernando Lima Reis e todo o seu grupo, Aristóbolo, Adriana Abalen, Eduardo Abrantes, Beatriz, Ludmila, Sibele Inácio, Walesca e especialmente ao Alex pela boa vontade e responsabilidade na extração do RNA usado neste presente trabalho.

À Drª. Anamaria A. Camargo e a Drª. Eloisa Moreira pela boa vontade e preocupação em oferecer ajuda.

Ao colega Paulo Maciag, pelo auxílio nas análises estatísticas.

Ao Prof. Dr. Sandro de Souza e toda a equipe do laboratório de bioinformática, pela boa vontade, presteza e responsabilidade nas análises das seqüências de DNA geradas durante o desenvolvimento deste presente trabalho.

À Suely Francisco, Rosinéia A. Carneiro e aos demais funcionários da biblioteca da Fundação Antônio Prudente, pelo empenho na obtenção do material bibliográfico e pelo auxílio no preparo das referências citadas neste trabalho.

À Ana Maria Rodrigues A. Kuninari e a Márcia M. Hiratoni, pela amizade e pelo excelente trabalho prestado junto a Pós Graduação da Fundação Antônio Prudente.

À todos os colegas do Laboratório de Genética de Câncer, Anna Christina, Elisangela Monteiro, Ana P. Duarte, Valéria Paixão, José Cláudio, Andréia Godoy, Fabrício Falconi e Nídia A. Pinheiro por compartilhar os bons e maus momentos durante estes anos.

À todos os colegas do ILPC por este período de convivência.

À FAPESP, pelo apoio financeiro.

RESUMO

Moura RP. **Prospecção gênica: avaliação de tumores raros na descoberta de genes humanos**. São Paulo; 2000. [Dissertação de Mestrado – Fundação Antonio Prudente].

A razão principal para o sequenciamento completo do genoma humano é a identificação de todos os genes expressos. Uma poderosa ferramenta que permite identificar a porção expressa do genoma é o seqüenciamento de ESTs (*Expressed Sequence Tags*). Porém, para possibilitar a identificação de genes pouco expressos, é fundamental a normalização da população dos transcritos, diminuindo a representação dos mais abundantes e enriquecendo a presença dos mais raros. Aproximadamente 75% dos genes humanos são pouco expressos sendo desejável o uso de bibliotecas de cDNA normalizadas para permitir a sua identificação. Recentes estudos sobre expressão gênica, sugerem que os genes mais expressos possuam cerca de 12.000 cópias transcritas por célula, enquanto a grande maioria dos genes, apresenta expressão extremamente baixa, da ordem de 10 transcritos por célula. A maior parte dos genes humanos já identificados, é derivada de genes expressos em abundância em tecidos freqüentemente estudados. Sendo assim, acreditamos que um número de genes ainda não conhecidos, expressos em raras condições patológicas, ainda permanecem por serem identificados. Desta forma, padronizamos uma metodologia para geração de mini-bibliotecas de cDNA parcialmente normalizadas dirigidas para a porção 3' dos transcritos, a partir de tumores raros, tecidos patológicos nem sempre disponíveis em grandes quantidades. Os resultados gerados indicam que a estratégia é adequada em termos de normalização, uma vez que grande parte das ESTs obtidas representa genes com pequeno ou médio nível de expressão. Os tecidos PNET, leiomiossarcoma, Wilms/DenisDrash, hibernoma e leiomioma mostram um alto percentual de genes novos. Nestes tecidos 4,4% de um total de 2,114 ESTs compreendem transcritos ainda não conhecidos, contra 2,1 % de um total de 467 ESTs nos tumores usados como controle (mama e cólon).

Descritores: Genes. Expressão gênica. Tumores.

SUMMARY

Moura RP. **Prospecção genica: avaliação de tumores raros na descoberta de genes humanos**. São Paulo; 2000. [Dissertação de Mestrado – Fundação Antonio Prudente].

The primary reason for sequencing the complete human genome is the identification of all expressed genes. Current single pass cDNA sequencing (ESTs) is of considerable value for identifying the expressed portions of the human genome. However, normalization of transcript populations is fundamental for increasing the coverage of rare transcripts. Approximately 75% of human genes have low expression levels and the discovery of these porly expressed genes favored by the use of normalized libraries where, ideally, all expressed genes are equally represented. A commonly accepted mRNA species abundance classification estimates that highly expressed genes are represented by around 12,000 mRNA copies per cell while most genes, produce less than 10 transcripts per cell. Because a large fraction (10,000 full length cDNAs) of all human genes have already been identified, redundant identification of genes that are expressed in multiple tissues is difficult to avoid, even using normalized libraries. Conversely an unknown number of genes expressed at low levels in rare pathological conditions remain to be identified. In an attempt to improve the representation of mRNAs in pathological tissues available in small quantities, we developed a simple methodology for generating 3` ESTs. We apply a variant of RNA fingerprinting to construct cDNA mini-libraries from colon carcinomas and breast carcinoma as controls; as well as five rare tumors: PNET, Wilms (associated with Denis Drash syndrome), leiomyossarcoma, hibernoma and leiomyoma to evaluate the technique in terms of gene discovery and mRNA normalization. The results obtained with RNA fragments are adequate in respect to normalization, since most of the ESTs generated are from genes with medium or low levels of expression. The tissues PNET, leiomiossarcoma and Wilms/DanisDrash, showed a greater percentage of new genes, 4,14% in a total of 2,114 ESTs, against 2,1% of the 467 ESTs in control tissues (Breast and Colon carcinomas).

Descriptors: Genes. Genes Expression. Tumors.

Índice:

Lista de Tabelas:

Lista de Figuras:

Lista de Abreviaturas:

μg - Micrograma

μl - Microlitro

μM – Micromolar

μj – Microjaule

3´UTR – porção não traduzida, situada na extremidade 3´do gene (do inglês *Untranslated region*)

5´UTR - porção não traduzida, situada na extremidade 5´do gene (do inglês *Untranslated region*)

A - Adenina

AP-PCR –reação em cadeia da polimerase utilizando iniciadores randomicos (do inglês *Arbitrarily Primed – PCR*)

ATP – Trifosfato de adenosina

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

Blast - *Basic Local Alignment Search Tool*

BSA – Soro albumina bovina (do inglês *Bovine serum albumine*)

C - Citosina

cDNA - DNA complementar

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

Crom- Cromossomo

dbEST - Banco de dados de ESTs

ddH_2O - água destilada e deionizada

ddNTPs - dideoxinucleotídeos

DDRT-PCR – Amostragem diferencial de mRNAs por RT-PCR (do inglês *Differential Display Reverse Transcriptase – Polymerase chain reaction*)

DEPC - Di-etil pirocarbonato

DMSO – dimetilsulfóxido

DNA - Ácido desoxirribonucléico

DNase - Desoxirribonuclease

DO - Densidade óptica

DOE – *Department of Energy Of USA*

EDTA – Ácido Etilenodiaminotetracetico disódio

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

EUA – Estados Unidos da América

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

G - Guanina

GAPDH - Gliceraldeído-3 fosfato desidrogenase

H_2O-DEPC - água tratada com 0.1% de dietilpirocarbonato

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

ILPC – Instituto Ludwig de Pesquisas sobre o Câncer

IPTG - iso – propil-β-D-tiogalactopiranosídeo

LB – Luria Bertain (meio para crescimento de bactérias)

LINEs - Elementos interespaçados de repetição longa (do inglês *Long Interspersed Nuclear Element*)

mRNA - RNA mensageiro

mtDNA - DNA mitocondrial

NCBI – *National Center for Biotechnology Information*

nm – nanômetros

Blast nr – *Basic Local Alignment Search Tool - Non redundant*

ORESTES – *Open Reading Frame ESTs*

PACs – Cromossomo artificial derivado do elemento P1 (do inglês *P1 Artificial Chromosomes*)

pb - pares de bases

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

PGH – Projeto Genoma Humano

pH – Potencial hidrogeniônico

PNET – Tumor neuroectodérmico primitivo

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

pUC – plasmídio para clonagem universal (do inglês *Plasmid for Universal Cloning*)

RAPD – amplificação randômica de polimorfismos de DNA (do inglês *Randomly Amplified polymorphic DNA*)

RAP-PCR –amplificação do RNA por PCR usando iniciadores randômicos (do inglês *RNA Arbitrarily Primed – PCR*)

RNA - Ácido ribonucléico

RNAse - Ribonuclease

rpm – rotações por minuto

rRNA - RNA ribossomal

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

SAGE – Análise seriada da expressão gênica (do inglês *Serial Analysis of Gene Expression*)

SDS – Dodecilsulfato de sódio

SINES - Elementos interespaçados de repetição curta (do inglês *Short Interspersed Element*)

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

T – Timina

TGC – Tumores de células germinativas

THCs - montagem de consensos de ESTs na tentativa de gerar seqüências completas dos transcritos humanos (do inglês *Tentative Human Consensus*)

TIGR – *The Institute for Genomic Research*

UV – Ultravioleta

WAGR – *Wilms tumor, Aniridia, Genitourinary malformation, mental Retardation*

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

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

1 INTRODUÇÃO:

1.1 Anatomia do genoma humano:

Todos os organismos possuem um elaborado manual de instruções contendo informações necessárias para o seu desenvolvimento e para a manutenção de sua vida. Este manual de instruções, que é passado através das gerações (hereditariedade), está presente na forma de uma molécula química chamada DNA (ácido desoxirribonucléico). A molécula de DNA é um polímero linear constituído por uma seqüência de quatro unidades repetitivas chamadas nucleotídeos: adenina (A), citosina (C), guanina (G) e timina (T). Os nucleotídeos são compostos por um grupo fosfato, uma molécula de açúcar (pentose) e uma base nitrogenada e são unidos entre sí através de uma ligação fosfodiéster. As moléculas de DNA encontram-se enoveladas e organizadas em estruturas denominadas cromossomos. O núcleo de cada uma das células somáticas que compõem o corpo humano contém 22 pares de cromossomos autossômicos e dois cromossomos sexuais, cuja combinação XX determina o sexo feminino e XY o sexo masculino. Os cromossomos são compostos por duas fitas complementares de DNA ligadas por pontes de hidrogênio, onde cada nucleotídeo tem o seu respectivo ligante, Adenina/Timina e Citosina/Guanina que formam o que chamamos de pares de bases (pb). Além disso, moléculas protéicas denominadas histonas têm um importante papel na compactação desse material nucléico. O DNA é composto de cerca de três bilhões de nucleotídeos na espécie humana, onde apenas 3% é considerado funcional. Todo este conjunto, considerado o patrimônio genético de um organismo, é denominado genoma (STRACHAM e READ 1996a).

As instruções para a codificação de proteínas são transmitidas indiretamente para a maquinaría de síntese protéica através do mRNA (ácido ribonucléico mensageiro), uma molécula intermediária similar a uma fita simples de DNA, onde a base timina é substituida pela base uracila. Para que o gene seja expresso, o mRNA é produzido a partir do DNA por um

processo denominado transcrição. O mRNA é então transportado para o citoplasma da célula para servir como molde na síntese de proteínas, cujo processo é denominado tradução. Todo este processo de transferência de informações genéticas (DNA → RNA → proteína) é descrito como o dogma central da biologia molecular (VAUGHAN 1996).

As seqüências gênicas transcritas são constituidas pelos genes e representam somente 3% do genoma. A maioria dos genes é composta por exons, que são intercalados por seqüências não codificadoras, os introns. O *splicing* é um mecanismo pós-transcricional que consiste na retirada de seqüências não codificadoras (*introns*) e junção dos *exons* para a composição de um transcrito. Os pseudogenes são seqüências não transcritas originadas por retrotranscrição ou duplicação gênica (STRACHAM e READ 1996d). Os pseudogenes não produzem proteínas funcionais devido a mutações deletérias tais como falta de sinal de inicialização, ou de terminação prematura de transcrição e mutações que impedem o mecanismo de *splicing*.

Os genes humanos variam muito em tamanho, geralmente se estendendo por alguns kilobases e estão dispostos linearmente ao longo dos cromossomos. Intercalando os genes existem seqüências de DNA não codificadoras (97% do genoma), cuja função ainda não é bem conhecida. Nestas regiões, chamadas intergênicas, estão presentes seqüências responsáveis pelo controle da expressão dos genes, além de várias classes de elementos repetitivos tais como SINEs (*short interspersed repeats*) e LINEs (*long interspersed repeats*). Os maiores representantes dos SINEs em genoma de mamíferos são as seqüências Alu, que são assim denominadas por normalmente possuírem um sítio de restrição para enzima endonuclease *Alu I*. As seqüências Alu correspondem às seqüências repetitivas mais abundantes no genoma humano, com cerca de 750,000 cópias, possuindo cada uma um tamanho aproximado de 280 pb (STRACHAM e READ 1996d). Os maiores representantes dos LINEs (que pertencem à família L1) possuem cerca de 6.000 bp e são repetidos cerca de 50.000 vezes no genoma. Algumas seqüências Alu e L1 são transcritas, porém, o real papel

de Alu e L1 no funcionamento celular ainda não é bem determinado. Estes elementos também possuem a capacidade de se moverem para diferentes sítios do DNA genômico (*transposons*) e apesar de não exercerem papel conhecido no funcionamento celular, possuem importante papel evolutivo contribuindo para a diversidade genética (ALBERTS et al. 1994).

Apesar de mais de 99,9% do DNA humano ser idêntico entre diferentes indivíduos, variações sutis podem ser responsáveis tanto pela predisposição a doenças como o câncer, como pela resposta do indivíduo a toxinas, drogas e etc. As variações nucleotídicas mais comuns são as denominadas *Single Nucleotide Polymorphisms* (SNPs), que são alterações pontuais que podem ocorrer cerca de uma vez a cada 500 - 1000 pb e apresentam freqüência superior a 1% na população, fazendo com que o conjunto de SNPs de cada pessoa seja único (STRACHAM e READ 1996c). É esperado que os SNPs facilitem a associação em larga escala dos estudos genéticos, causando, desta forma, um crescente interesse em identificar e melhorar os métodos para sua detecção. O *SNP Consortium* (consórcio entre companhias farmacêutica e de bio-informática, cinco centros acadêmicos e uma entidade filantrópica) está organizando um mapa de grande densidade de SNPs no genoma humano. Estes SNPs mapeados estão sendo colocados regularmente em domínio público (http://snp.cshl.org). O objetivo inicial é produzir um mapa contendo cerca de 200,000 a 300,000 SNPs distribuídos ao longo do genoma humano. Porém, esta iniciativa provavelmente irá disponibilizar até o mês de abril de 2001 cerca de 600,000 a 800,000 SNPs (ROSES 2000).

Já as mutações patogênicas são raras e tendem a ocorrer em uma freqüência inferior a 1% na população (normalmente muito menos que 1%). Tipicamente, as mutações patogênicas são encontradas em regiões do DNA que estão direta ou indiretamente ligadas à transcrição de proteínas, provocando variações estruturais ou de níveis de expressão protéica, eventualmente causando doenças como o câncer (ROSES 2000).

1.2 Seqüenciamento do genoma humano:

Compreender a estrutura genômica, identificar todos os genes e mapear todas as suas variações são objetivos que serão atingidos com a finalização dos projetos genoma do homem. Atualmente, dentre os que mais se destacam são: *Human Genome Project* (Projeto Genoma Humano - PGH) desenvolvido pelo NIH (*National Institutes of Helth*) *Department of Energy* e *Welcome Trust*, de iniciativa pública, e o seqüenciamento do genoma humano desenvolvido pela empresa *Celera Genomics,* de iniciativa privada. O desenvolvimento do PGH foi oficialmente iniciado em 1990, com os seguintes objetivos: (i) seqüenciar o DNA humano em sua totalidade (ii) identificar todos os genes contidos no DNA humano (iii) gerar seqüências de outros organismos para o entendimento, por comparação, do funcionamento do genoma humano, (iv) aperfeiçoar as técnicas já conhecidas e desenvolver novas técnicas para a obtenção, organização e disponibilização dos dados gerados em bancos de dados públicos de acesso gratuito, (v) desenvolver tecnologias para uma rápida identificação de variações nucleotídicas no DNA (vi) explorar fatores éticos, legais e sociais que influenciam no uso, entendimento e interpretação das informações genéticas geradas, (vii) treinar e capacitar profissionais para o estudo de genomas; dentre outros (COLLINS e GALAS 1993) (http://www.ornl.gov/Tech Resources / Human_ Genome/hg5p/goal.html).

A estratégia utilizada pelo consórcio público para a decodificação do DNA humano foi baseada no seqüenciamento de clones mapeados. O mapeamento envolve a "quebra" dos cromossomos em fragmentos menores que possam ser clonados e multiplicados, para posteriormente serem ordenados de acordo com sua localização original. Na clonagem destes fragmentos são usados vetores especiais capazes de transportar fragmentos de DNA da ordem de Kilobases (Kb) até Megabases (Mb), tais como: PACs (*P1 Artificial Chromosomes*), BACs (*Bacterial Artificial Chromosomes*) e YACs (*yeast artificial chromosomes*); que transportam respectivamente 130-150 Kb, mais de 300 Kb e 0.2 a 2.0 Mb. Devido a limitações técnicas, que não permitem o seqüenciamento contínuo de mais de 500/1000

nucleotídeos, após o mapeamento físico estes fragmentos são quebrados em fragmentos ainda menores, de cerca de 500 a 1000 pb, para que o seqüenciamento do DNA possa ser realizado. Após o seqüenciamento é feita a montagem dos fragmentos menores para a recomposição dos fragmentos originais, os maiores, que devido ao mapeamento físico prévio, a localização no genoma humano, é conhecida (STRACHAM e READ 1996c). Desta forma foi realizado o seqüenciamento do cromossomo 22 (DUNHAM et al. 1999). Dentre os cromossomos autossômicos este é o segundo menor compreendendo cerca de 1,6 a 1,8% do DNA genômico. Mapeamentos diretos e indiretos sugerem que o braço longo é rico em genes quando comparado com outros cromossomos; em contrapartida não existe nenhuma evidência que indique a existência de genes no braço curto (22p). Em maio deste ano, foi também completado pelo PGH, o seqüenciamento do menor cromossomo humano, o cromossomo 21 (HATTORI et al. 2000). Uma cópia extra deste cromossomo causa a síndrome de *Down*, cuja incidência é superior a 1/700 recém-nascidos. Uma síntese da análise dos dados destes dois cromossomos seqüenciados é apresentada na tabela 01.

Tabela 01: Análise dos dados gerados pelo seqüenciamento dos cromossomos 21 e 22:

Class 01	Class 02	Class 03	Class 04	Class 05	Class 06	Class 07	Class 08	Class 09	Class 10	Class 11
21q	5/2000	33,65	40.06%	225	127	68	30	59	03 (100Kb)	33,546 99,7%
22q	12/1999	34,49	41.90%	545	247	148	150	134	10 (1Mb)	33,454 97%

Fonte: Dunham et al. 1999. Hattori et al. 2000.

Class 01: Cromossomo sequenciado.

Class 02: Data do término do sequenciamento.

Class 03: Tamanho do cromossomo em Mb.

Class 04: Percentual de elementos repetitivos.

Class 05: Total de genes encontrados.

Class 06: Total de genes conhecidos.

Class 07: genes identificados por predição, com base em seqüências parciais de cDNA (ESTs) em associação com análises computacionais (*in sílico*).

Class 08: genes identificados por predição, apresentando identidade parcial contra genes ou seqüências de aminoácidos identificados em humanos ou em outros organismos.

Class 09: Pseudogenes.

Class 10: Regiões não decodificadas (*Gaps*).

Class 11: Total decodificado em Kb.

Além do Consórcio Público, o genoma humano também vem sendo seqüenciado pelo setor privado, mais especificamente pela empresa *Celera Genomics* fundada em maio de 1998 (MARSHALL 1999). Esta empresa foi fundada e é presidida pelo Dr. Craig Venter, chefe do grupo que lançou o conceito de ESTs (ADAMS et al. 1991), e que em 1995 junto com o Dr. Hamilton O. Smith, completou o seqüenciamento do primeiro genoma bacteriano, o *Haemophilus influenzae* (FLEISCHMANN et al. 1995). A Celera possui centenas de seqüenciadores capilares de DNA e hoje é capaz de decodificar 140 milhões de nucleotídeos a cada 24 horas, fazendo com que este seja o centro de seqüenciamento mais poderoso do mundo (MARSHALL 1999). O grupo Celera está seqüenciando o genoma humano partindo de 5 indivíduos selecionados dos maiores grupos étnicos incluindo

3 homens e 2 mulheres. O consenso gerado pelo seqüenciamento destes 5 indivíduos poderá elucidar cerca de 20 milhões de variações nucleotídicas. Em oposição à estratégia usada no Consórcio Público, a Celera usa uma estratégia composta de duas etapas. A primeira etapa, denominada *Whole Shotgun sequencing*, semelhante à usada mais recentemente para o seqüenciamento do genoma da *Drosophila melanogaster* que contém cerca de 160 milhões de pb (ADAMS et al. 2000). Esta etapa consiste em quebrar o DNA em pequenos fragmentos, cerca de 550 pb, e após o seqüenciamento das duas fitas de DNA, fazer a remontagem considerando as extremidades de cada fragmento que se sobrepõe. Desta forma cada base é seqüênciada inúmeras vezes. A segunda etapa envolve o sequenciamento das extremidades de grandes fragmentos clonados em BACs. Estas extremidades são usadas para gerar um *Scaffold* que facilita a montagem do genoma, delimitando *Clusters* para agrupar os fragmentos menores (MYERS et al. 2000). Para reconstituir o genoma, a Celera possui um sistema de processamento de dados cuja capacidade é a segunda mais poderosa do mundo depois do Departamento de Energia do governo dos E.U.A (D.O.E), cujo equipamento é usado para simular explosões nucleares.

Após 47 anos da elucidação da estrutura fundamental do DNA pelos pesquisadores WATSON e CRICK (1955), foi anunciado no dia 26 de junho de 2000 pelo presidente dos EUA, Bill Clinton e pelo 1° Ministro Britânico, Tony Blair, o término do seqüenciamento do genoma humano. Segundo os dados divulgados pela mídia, o Consórcio Público decodificou 97% do genoma humano enquanto a iniciativa privada representada pela empresa Celera seqüenciou 99%. A entrada do setor privado no seqüenciamento do genoma humano foi muito importante, no sentido de instigar a competição. Esta competição resultou na antecipação do término do deste trabalho em pelo menos 5 anos.

As informações geradas pelo seqüenciamento do DNA humano serão de importância vital para as áreas da pesquisa básica e aplicada. É esperado um grande salto na pesquisa médica do século XXI, uma vez que estes dados ajudarão não só no entendimento como também no eventual

tratamento de várias doenças (ROSES 2000). Mesmo antes do genoma humano ser totalmente seqüenciado, a indústria farmacêutica já está testando as primeiras drogas desenvolvidas com base nas informações geradas pelo seqüenciamento dos genes que estão diretamente relacionados com diversas doenças (EVANS e RELLING 1999).

1.3 Quantos são os genes humanos?

Várias são as estimativas atuais do total de genes contidos no genoma humano. Grosseiramente esta quantidade pode ser estimada pelo tamanho total do genoma. Desta forma, sendo o genoma haplóide humano algo próximo a 3 bilhões de nucleotídeos, teríamos 300,000 genes, estimando uma densidade gênica de 1/10 Kb. GILBERT (1992) usa esta estimativa como um limite superior, mas segundo este mesmo autor, uma densidade gênica mais próxima do real seria em torno de 1 gene/30 Kb, o que levaria a um conteúdo de cerca de 100 mil genes. Outro método para estimar a quantidade de genes, seria extrapolar os resultados obtidos com o seqüenciamento dos cromossomos 21 e 22 para obter uma visão da organização geral do genoma humano. A diferença da quantidade de genes encontrados nestes dois cromossomos suporta a conclusão de que o cromossomo 22 é rico em genes (545 genes) ao contrário do que ocorre no 21 (225 genes). Estes dois cromossomos juntos representam cerca de 2% do genoma humano e contêm cerca de 770 genes, o que, a grosso modo, resulta em uma densidade gênica em torno de 1 gene/87Kb. Assumindo que estes dados refletem o conteúdo gênico no genoma humano, a estimativa do total de genes humanos poderia estar perto de 40.000 (REEVES 2000).

Uma abordagem alternativa que permite a identificação e a eventual predição da quantidade de genes no genoma humano é o seqüenciamento parcial de transcritos (ESTs). Esta abordagem se baseia na retro-transcrição do mRNA em cDNA para posterior construção de bibliotecas, onde a seqüência de cada clone além de representar o fragmento de um gene, poderá ser usada para localizar este gene em um mapa cromossômico (mapa físico). Desta forma, um dos primeiros trabalhos

para a predição da quantidade de genes foi desenvolvido por FIELDS et al. (1994), que estimaram cerca de 60 a 70 mil genes para o genoma humano.

Hoje, o contexto de estimativa gênica passou a ser um assunto polêmico. No mês de junho foram publicados na revista Nature Genetics volume 25 (junho de 2000) 03 artigos usando diferentes estratégias para a predição do número de genes no genoma humano (APARICIO 2000). LIANG et al. (2000) estimaram a existência de cerca de 120 mil genes baseado em uma coleção de seqüências parciais e completas de transcritos humanos. Segundo o trabalho desenvolvido por EWING e GREEN (2000) a estimativa foi de 34,700 genes baseado na seqüência do cromossomo 22 e 33,600 genes baseado em seqüências de mRNA, resultando em uma média de 35 mil genes no genoma humano. Como os números são relativamente inferiores às estimativas anteriores, os autores propoem que a evolução nos vertebrados com relação à complexidade fisiológica deva estar diretamente relacionada à diversidade de fatores pós transcricionais tais como *Splicing* alternativo. No terceiro trabalho, CROLLIUS et al (2000) usam uma abordagem diferente baseado na seqüência genômica do *pufferfish* (*Tetraodon nigrovirids*), que é um vertebrado que possui o genoma 8 vezes mais compacto que o genoma humano, devido ao reduzido tamanho de seqüências intergênicas e regiões intrônicas. Esta abordagem, denominada *exonfish,* permite a organização dos genes humanos em grupos, chamados *ecores,* devido à identidade apresentada contra uma única seqüência codificadora do *pufferfish*. Segundo cálculos feitos no referido trabalho, cada *ecore* comporta cerca de 2,58 / 3,18 genes. Usando esta abordagem, que permite a detecção de seqüências codificadoras conservadas, eles calcularam cerca de 28,000 a 34,000 genes contidos em cerca de 1,272,3 Mb do genoma humano, que estavam disponíveis nos bancos de dados públicos em dezembro de 1999 (CROLLIUS et al. 2000).

Atualmente, o termo gene é definido como uma subunidade funcional do DNA. Usualmente cada gene está relacionado a um produto gênico, porém em vários casos uma unidade de transcrição pode resultar em vários produtos gênicos (polipeptídeos ou RNAs funcionais). Em humanos e

outros eucariotos, existe uma variedade de mecanismos pós-transcricionais que podem gerar distintos produtos a partir de um único gene. Com o término do seqüenciamento do genoma humano o próximo desafio será a identificação e anotação de todos os genes. No momento as estimativas estão gerando números muito diferentes, o que não nos permite ter uma noção mais sólida da quantidade de genes que mantêm o metabolismo de nossas células. A geração de seqüências, provenientes de fragmentos de genes expressos, tanto de tecidos humanos, como de outros organismos tem demonstrado ser o caminho mais eficiente para catalogar, identificar e caracterizar todos os genes humanos.

1.4 Identificação gênica pelo seqüenciamento dos genes expressos:

O seqüenciamento de moléculas de cDNA, derivadas diretamente dos genes expressos em humanos é crucial para a correta identificação dos genes e determinação de formas alternativas de processamento (*splicing*). Embora a maior parte do genoma humano tenha sido seqüenciada, há ainda uma grande dificuldade na identificação dos genes. Em contraste com o seqüenciamento de genomas mais compactos, os programas para análises computacionais para a identificação de genes (*in silico*) no genoma humano não têm se mostrado muito eficientes. Esta ineficiência é devida à descontinuidade das seqüências codificadoras (exons), que são intercaladas pelos introns, além da existência dos pseudogenes e elementos repetitivos. Um exemplo desta ineficiência ocorreu durante o trabalho de identificação de genes no cromossomo 22 humano. O programa mais utilizado para a predição de genes, o Genscan, apresentou algumas falhas na predição de genes previamente conhecidos. Desta forma este programa foi capaz de predizer com exatidão apenas 20% do total de exons conhecidos no cromossomo 22, além disso, 16% dos exons já conhecidos não foram identificados. Outro agravante é que cerca de 40% dos genes preditos pelo Genscan não foram confirmados por outros meios, podendo incluir uma desconhecida porcentagem de falsos positivos. Desta forma foi concluído que a predição de genes *in silico* ainda não permite gerar resultados

confiáveis, porém, podem ser úteis para definir pontos iniciais para estudos experimentais (DUNHAM et al. 1999).

1.5 Seqüenciamento de genes expressos:

1.5.1 "Expressed sequence tags" (ESTs):

Em 1991, foi descrita uma abordagem que permitiu acelerar o descobrimento de seqüências de genes expressos, a geração de "ESTs" (*Expressed Sequence Tags*) (ADAMS et al. 1991). Segundo SCHULER et al. (1996) a identificação da maioria dos genes até o momento foi feita através desta técnica. As ESTs são geradas a partir do seqüenciamento de clones de bibliotecas de cDNA, produzindo fragmentos gene-específicos com algumas centenas de pares de bases (pb), que permitem identificar e "etiquetar" genes. Esta abordagem oferece a maneira mais rápida de obtenção de seqüências gênicas visto que são geradas a partir de mRNA convertidos em cDNA, ignorando desta forma as regiões não transcritas do genoma. Em 1993 foi criado dentro do GenBank um banco de dados denominado dbEST (BOGUSKI e SCHULER 1993), para a deposição das ESTs derivadas de diversos organismos. A evolução deste banco com relação à quantidade de seqüências geradas é espantosa. Em 1994, 50.000 ESTs de 23 diferentes organismos estavam disponíveis e atualmente este valor subiu para cerca de 4 milhões de seqüências disponíveis em um total de mais de 200 espécies diferentes. No dia, 28/11/2000 o número de ESTs geradas a partir de tecidos humanos era 2.757,350 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

Atualmente uma das maiores utilidades das ESTs está na montagem de consensos de alta qualidade, representando transcritos únicos tais como T.H.Cs (*Tentative Human Consensus*) ou T.Cs para outros organismos (QUACKENBUSH et al. 2000). Estas montagens podem ser usadas para anotação funcional, mapeamento gênico, identificação de *Splicing* alternativo e identificação de genes homólogos. Os genes homólogos podem ser identificados através do cruzamento dos bancos de

dados das diferentes espécies e são divididos em duas classes, as quais são: (i) ortólogos, quando a ancestralidade comum entre duas espécies pode ser traçada até um evento inicial de especiação e (ii) parálogos, quando a ancestralidade comum de duas espécies pode ser traçada até um evento de duplicação gênica.

Em 1997 foi criado o UniGene (http://www.ncbi.nlm.nih. gov/ UniGene/), um sistema computacional que tem o objetivo de organizar as seqüências depositadas no GenBank em grupos de seqüências não redundantes que representam genes (*Clusters*). Ou seja, cada *Cluster* do UniGene contem seqüências (mRNAs completos ou ESTs) oriundas de um único gene, bem como informações sobre os tecidos que originaram as seqüências, a localização cromossômica e quantas seqüências foram geradas a partir do gene em questão. Atualmente seqüências humanas e de camundongos estão sendo processadas dentro do UniGene. A princípio, estas espécies foram escolhidas porque possuem uma grande quantidade de ESTs depositadas (http://www.ncbi.nlm.nih.gov/Web/Newsltr/aug96.html).

1.6 Projetos para geração de ESTs humanas:

1.6.1 The Institute for Genomic Research (TIGR):

A *TIGR* é uma empresa do setor privado, criada em 1992 pelo Dr. Craig Venter. No Projeto Genoma Humano, a construção de bibliotecas de cDNA associada às estratégias de geração de ESTs, permitiu a publicação de dados de milhares de clones. Usando esta estratégia, foi publicado um dos mais extensos artigos da revista Nature, tratando do diretório do projeto genoma humano (ADAMS et al. 1995). Neste fascículo, o primeiro artigo, de 174 páginas, trata unicamente da análise e descrição de dados de 290.000 ESTs geradas de diversas bibliotecas de cDNA humano. Esta foi a primeira oportunidade de classificar sistematicamente a maioria dos genes altamente expressos no homem. Por ser uma empresa do setor privado os dados obtidos não estão, necessariamente, disponíveis nos bancos de dados públicos de acesso gratuito.

1.6.2 The Merck Gene Index:

O projeto *The Merck Gene Index* foi um projeto feito em colaboração entre as seguintes instituições: *Washington University*, *IMAGE Consortium* (*Integrated Molecular Analysis of Gene Expression*), *University of Pennsylvania*, *Merck Research Laboratories Department of Bioinformatics*, *NCBI* (*National Center for Biothecnology Information (NIH)*) e *GDB* (*Genome Sequence Database*). O objetivo deste projeto foi a geração de ESTs e a produção de um clone de cDNA para cada gene identificado. Este projeto, coordenado e mantido pela *Merck & Co. Inc.*, contribuiu com cerca de 600,000 seqüências para o banco de dados público GenBank (dbEST), (http://www.merck.com/mrl/merck_gene_index.2.html).

1.6.3 Projeto Anatomia Genômica do Câncer (CGAP):

Dentro do contexto do genoma humano foi iniciado em 1997, pelo Instituto Nacional do Câncer (*National Cancer Institute,* NCI), um projeto denominado *Cancer Genome Anatomy Project* (CGAP) que tem como objetivo determinar a seqüência de todos os genes envolvidos no processo de transformação maligna de uma célula, eventualmente aplicando estes conhecimentos na prevenção, detecção e tratamento do câncer. A princípio, foram determinados cinco tipos de tumores a serem estudados: ovário, mama, próstata, pulmão e cólon. Atualmente, inúmeros tipos de tumores provenientes de vários tecidos estão sendo incluídos. Os dados são depositados em um banco de dados do projeto (http://www.ncbi.nlm.nih.gov/cgap), criando uma valiosa fonte de informações sobre os genes expressos em células neoplásicas.

O projeto CGAP oferece um novo avanço na identificação de genes humanos (fig 1), sendo atualmente o principal projeto de geração de ESTs em curso, tendo disponibilizado nos bancos de dados cerca de 1.067.938 seqüências humanas (28/11/2000) (http://www.ncbi. nlm.nih.gov/ CGAP/hTGI/sumtab/cgapba.cgi). Quando este projeto foi iniciado, o UniGene possuía apenas 45.000 *clusters*, hoje este número cresceu para cerca de 95.000 *clusters* que, a princípio representam diferentes genes,

identificados através da geração de ESTs derivadas de bibliotecas de cDNA normalizadas (20%) e não normalizadas (80%) de tecidos normais (46%), pré-malignos (6%) e malignos (48%) (http://www.ncbi.nlm.nih.gov/CGAP /hTGI/sumtab/cgapba.cgi). Estes dados indicam o importante papel deste projeto no que diz respeito à descoberta de genes humanos e caracterização dos principais transcritos diferencialmente expressos na tumorigênese (STRAUSBERG et al. 2000).

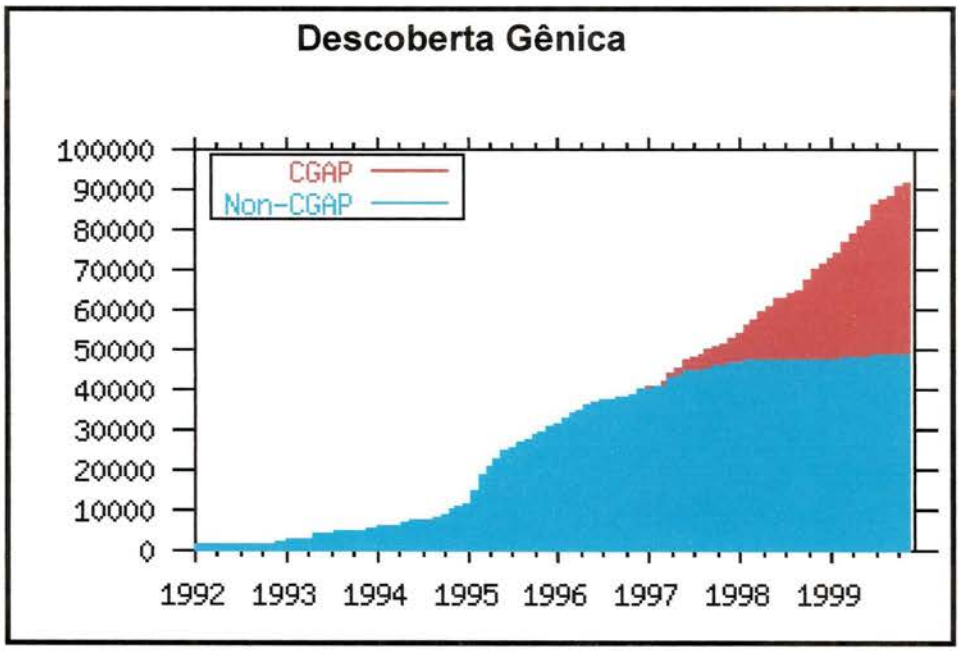


Figura 01: Identificação de genes humanos segundo o CGAP:

Gráfico ilustrativo mostrando o acúmulo de clusters no UniGene (eixo y) ao longo do tempo (eixo x). Em azul, porção que indica a contribuição de projetos não ligados ao CGAP na geração de novos *clusters* (Non-CGAP). Em vermelho (delineada a partir de 1997), porção que indica a contribuição do CGAP na obtenção de seqüências de novos genes humanos (http://www.ncbi.nlm.nih.gov/ncicgap/)

1.6.4 Projeto Genoma do Câncer Humano ILPC / FAPESP:

Este projeto, lançado em março de 1999, é resultante de uma colaboração entre o Instituto Ludwig de Pesquisa Sobre o Câncer e a

FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). O objetivo inicial foi produzir seqüências derivadas dos tumores humanos, com seu foco principal voltado para tumores mais relevantes para o Brasil e que não foram definidos como prioridade no CGAP. Uma síntese da análise dos tumores estudados está descrita na tabela 2. A estratégia adotada neste projeto, denominada ORESTES (*Open Reading Frame ESTs*) (DIAS NETO et al. 2000a), permite a obtenção preferencial de seqüências derivadas de regiões codificadoras (centrais) dos genes humanos e, além disto, favorece uma normalização dos genes etiquetados. A obtenção de ESTs derivadas da porção interna dos genes é muito útil, uma vez que do ponto de vista de distribuição posicional estas são complementares às demais ESTs humanas geradas por outros projetos. Esta complementaridade existe, pois as seqüências geradas pelos grupos, internacionais são derivadas das extremidades 3' e 5' dos transcritos, fazendo com que ORESTES seja complementar na formação de *contigs* (transcrito montado por consensos de ESTs) com a eventual obtenção da seqüência completa dos transcritos humanos. O primeiro objetivo deste projeto é gerar 1.000.000 seqüências da porção central de transcritos humanos, contribuindo de maneira significativa para a determinação de todos os genes humanos. Atualmente (28/11/2000), este projeto tem depositado em seu banco de dados cerca de 903.630 ESTs; destas 269.096 (29%) representam novas seqüências (http://www.ludwig.org.br /ORESTES/). Com os resultados gerados neste projeto a ciência brasileira está vivendo um momento histórico, visto que está contribuindo efetivamente para a conclusão de um projeto de âmbito mundial, o Projeto Genoma Humano.

Tabela 02: Dados gerados pelo Projeto Genoma do Câncer Humano ILPC/FAPESP:

Tecido Estudado	Seqüências produzidas	Seqüências novas
Aminion Normal	16,955	4,784 (28%)
Bexiga normal	-	-
Bexiga tumor	1,916	441 (23%)
Cabeça e Pescoço normal	5,910	1,668 (28%)
Cabeça e pescoço tumor	162,197	44,074 (27%)
Colon Normal	8,814	1,960 (22%)
Cólon tumor	76,498	23,187 (20%)
Cólon tumor estável p/ microssatelite (EST)	11,451	2,300 (20%)
Cólon tumor instável p/ microssatelite (INS)	26,710	6,054 (22%)
Epididimo tumor	4350	914 (21%)
Estômago Normal	13,805	4,160 (30%)
Estômago tumor	39,160	11,044 (28%)
Leiomiossarcoma	11,632	3,032 (26%)
Mama Normal	67,677	21,902 (32%)
Mama tumor	81,827	23,530 (28%)
Medula	28,448	9,756 (39%)
Ovário Tumor	19,649	5,183 (26)
Placenta normal	50,905	18,340 (36%)
PNET (Tumor Neuroectodermico Primitivo)	6,378	1,701 (26%)
Próstata normal	21,223	7,678 (36%)
Próstata tumor	22,315	6,492 (29%)
Pulmão normal	13,172	3,729 (28%)
Pulmão tumor	27,093	7,810 (28%)
Rim tumor	3,001	978 (32%)
Tecido Nervoso Normal	68,901	24,870 (36%)
Tecido Nervoso Tumor	29,796	8,050 (25%)
Testículo normal	27,484	8,364 (30%)
Testículo tumor	3,465	1,483 (42%)
Útero tumor	21,488	4,793 (22%)
Wilms tumor, associado à síndrome de Denis Drash	11,364	2,701 (23%)
Total	903,630	269,096 (29%)

28/11/2000 (http://www.ludwig.org.br/ORESTES/)

1.6 Limitações na geração de ESTs:

1.6.1 Bibliotecas de cDNA não normalizadas:

Apesar de vir sendo utilizado há mais de 15 anos (CONSTANZO et al. 1983), o seqüenciamento parcial de clones de bibliotecas direcionais de cDNA selecionados ao acaso, representa até os dias de hoje uma das principais estratégias empregadas no processo de descoberta de genes nos diversos Projetos Genoma. Esta metodologia já se mostrou bastante útil na obtenção de importantes informações tais como a determinação de qual o número e o tipo de genes mais abundantemente expressos em um determinado tecido, ou órgão em estágio evolutivo, ou ainda em determinadas fases do desenvolvimento de determinado órgão, estado patológico ou organismo. Uma característica marcante destas bibliotecas é que elas são fortemente influenciadas pelo alto nível de expressão de certos genes dentro da população global de mRNAs. Desta maneira, em uma biblioteca não normalizada, os clones mais abundantes correspondem proporcionalmente aos transcritos mais freqüentes no material biológico em estudo. Esta característica, sob o ponto de vista de descoberta gênica, é prejudicial uma vez que a quantidade de clones redundantes irá comprometer não só o objetivo principal como também irá tornar o projeto muito caro.

1.6.2 Bibliotecas de cDNA normalizadas:

A distribuição dos mRNAs em uma célula tem sido matéria de estudo a mais de 25 anos. Em 1974, BISHOP et al. usando uma abordagem baseada na hibridação de cDNA sintetizado com oligo dT (10) e RNA derivados de célula imortalizada HeLa, verificaram a distribuição da população de transcritos em três classes distintas: (I) transcritos abundantes, compreendendo 22% da massa total de RNA, correspondendo a cerca de 17 transcritos contendo 8,000 cópias por célula; (ii) transcritos intermediários, compreendendo 28% da massa total de RNA, correspondendo a cerca de

370 transcritos contendo 440 cópias por célula; e (iii) transcritos raros, compreendendo 50% da massa total de RNA, correspondendo a cerca de 33,000 transcritos contendo 8 cópias por célula. DAVIDSON e BRITTEN (1979) publicaram uma revisão sobre a distribuição dos mRNAs em diferentes células de diferentes organismos. Neste trabalho eles mostraram a heterogeneidade da expressão gênica em diferentes tecidos de diferentes organismos, também dividindo os transcritos em três classes distintas, a saber: (i) complexa, (ii) prevalência moderada e (iii) super-prevalente (maiores detalhes veja DAVIDSON e BRITTEN 1979).

Estudos mais recentes baseados na técnica de SAGE (*Serial Analysis of Gene Expression*) (VELCULESCU et al. 1995), sugerem a divisão dos genes em quatro classes de acordo com o número de transcritos encontrados em célula normal de cólon e célula de tumor de cólon (VELCULESCU et al. 1999) (Tabela 03). Embora as técnicas para o estudo da distribuição dos transcritos na década de 1970 fossem diferentes, os valores são proporcionais.

Tabela 03: Distribuição dos genes em classes de acordo com o nível de expressão:

Célula normal de colon			Célula de Tumor de cólon		
Cópias/Célula	N° Genes	Massa total de mRNA%	Cópias/Célula	N° Genes	Massa total de mRNA%
> 500	55 (0,04%)	18	> 500	61 (0,08%)	20
50 – 500	578 (0,43%)	27	50 – 500	562 (0,81%)	27
5 – 50	6160 (4,59%)	30	5 – 50	6358 (9,16%)	30
≤ 5	127342 (94,93%)	25	≤ 5	62400 (89,93%)	25
Total	134135 (100%)	100	Total	69381 (100%)	100

Adaptado de Velculescu et al. 1999

Diante da heterogeneidade do nível de expressão de genes em uma célula, surgiram técnicas alternativas que permitiram equilibrar parcialmente a representatividade dos transcritos. A idéia básica usada na

maioria destas técnicas, consiste na hibridação em fase líquida para reduzir a representatividade dos transcritos mais abundantes aumentando, em contrapartida, a freqüência dos transcritos mais raros. As populações dos diversos transcritos são então equilibradas e usadas para a construção de bibliotecas normalizadas com a hibridação e subtração dos clones mais abundantes. No final de 1996, BONALDO et al. descreveram alguns processos de normalização e subtração que permitiram acelerar os processos de descoberta de genes dos projetos genoma. As bibliotecas são construídas com cerca de 1μg de mRNA, que é usado como molde para a síntese da primeira fita de cDNA com iniciadores poli-T (18) contendo um sítio para enzima NotI na sua porção 5'. Para a construção das bibliotecas, as moléculas de cDNA são ligadas a adaptadores, digeridas e ligadas direcionalmente no vetor. Estas bibliotecas são então normalizadas, objetivando a diminuição da freqüência dos transcritos mais abundantes, e o aumento da representação dos clones mais raros. Para isto os plasmídeos são convertidos em fita simples e hibridados a oligonucleotídeos bloqueadores e RNA, em fase líquida. Este RNA é produzido pela transcrição *in vitro*, de alguns microgramas de plasmídeos da própria biblioteca, usando o sistema da T3 ou T7 RNA polimerase. Após a hibridação é feita uma cromatografia em coluna de hidroxiapatita, onde os ácidos nucléicos que permaneceram como fita simples são separados dos ácidos nucléicos que se tornaram fita dupla. A fase que contém apenas os plasmídeos fita simples pode ser convertida em fita dupla e usada para transformação de bactérias competentes, constituindo desta forma uma biblioteca normalizada.

A eficiência dos processos de normalização pode ser exemplificada nos resultados de LAMERDIN et al. (1995), onde os problemas normalmente encontrados com as bibliotecas usuais de cDNA foram praticamente resolvidos com o uso de uma biblioteca normalizada. Dentre 333 ESTs geradas neste trabalho, apenas dois clones continham rRNA, dois continham mtDNA, um representava contaminação com *Escherichia coli* e apenas um era vetor sem inserto. O seqüenciamento

repetido de clones aconteceu apenas duas vezes, com dois clones que foram seqüenciados duas vezes, ou seja, os 292 clones com seqüências úteis devem representar 290 genes diferentes, indicando uma redundância de apenas 0,7 %. Um dado impressionante foi a alta freqüência (79,5%) de ESTs que se mostraram distintas de outros genes humanos, demonstrando o universo de genes que podem ser analisados quando há a redução dos genes abundantes. Também utilizando bibliotecas de cDNA normalizadas para geração de ESTs, BONALDO et al. (1996) publicaram um trabalho onde foram produzidas 319,311 ESTs, da porção 3' e 5' do transcrito, partindo de 17 diferentes tecidos em 3 estágios evolutivos diferentes com o objetivo de produzir uma grande quantidade de seqüências de diferentes genes para a deposição em bancos de dados públicos. Infelizmente, a construção de bibliotecas de cDNA normalizadas requer a disponibilidade de alguns microgramas de mRNA, um fator limitante em situações onde a quantidade de material biológico é pequena. Este fator, associado à complexidade técnica, faz com que o uso desta técnica seja restrito.

1.7 Identificação de genes humanos em tecidos normais:

Apesar das limitações técnicas de diversos projetos, o acúmulo de centenas de milhares de ESTs humanas, fez com que a maioria dos genes ativos nos tecidos mais estudados fossem conhecidos. Em meados de 1997, o número de ESTs humanas derivadas de novos genes produzidas pelos diferentes projetos, começou a atingir um *plateau*. Isto foi um reflexo do crescente número de seqüências redundantes e portanto a baixa produção de novos *Clusters* na geração maciça de ESTs (Figura 01).

1.8 O uso de tecidos tumorais na descoberta de genes humanos:

A limitação na identificação de novos genes humanos estimulou o uso de novos tecidos em diferentes estados patológicos. De acordo com o gráfico (Fig 01) o projeto CGAP vem contribuindo de forma significativa no que diz respeito à descoberta de genes humanos. Também usando tecidos

neoplásicos, o projeto genoma humano do câncer (FAPESP/ILPC) está contribuindo significativamente neste sentido (Tabela 02). Possivelmente o uso de tecidos neoplásicos favorece a descoberta de genes devido à replicação desordenada de suas células, provocando desta forma a expressão também desordenada dos genes.

1.8.1 Tumores mais estudados nos projetos de geração de ESTs:

Carcinoma de cólon: o carcinoma de cólon e reto afeta cerca de 1 pessoa a cada 20 nos EUA, com mais de 155,000 novos casos diagnosticados por ano, representando cerca de 15% de todos os cânceres neste país. A despeito disso, o carcinoma de cólon é a segunda causa mais comum de morte por câncer, sendo assim, um dos grandes problemas públicos do país. As evidências até o momento sugerem que a natureza da dieta esteja diretamente relacionada com o desenvolvimento desta doença (COHEN et al. 1997). Tumores de cólon constituem um dos tecidos neoplásicos mais estudados nos projetos para geração de ESTs com mais de 55 mil seqüências geradas dentro do CGAP e mais de 120,000 seqüências geradas no Projeto Genoma do Câncer Humano.

Carcinoma de mama: este tumor tem origem nas estruturas glandulares e ductais da mama. O carcinoma mamário é a causa número um de mortes por câncer entre as mulheres, ocupando o primeiro lugar entre os cânceres em número de intervenções cirúrgicas, em tratamentos por radioterapia, administrações de hormônios e quimioterapia. No diagnóstico do câncer é o primeiro em número de biópsias. Este tipo de tumor também constitui um dos tumores mais estudados nos projetos para geração de ESTs, com cerca de 16,000 seqüências geradas dentro do CGAP e mais de 145,000 seqüências geradas no Projeto Genoma do Câncer Humano.

1.9 Tumores de baixa incidência como substrato para a descoberta de genes:

O avanço na identificação de novos genes depende de fatores tais como expressão gênica nos diferentes tecidos, diferenças de expressão

dos diversos genes e dos limites metodológicos que não favorecem a identificação de genes pouco expressos. Neste sentido um aumento na velocidade da descoberta de novos genes humanos poderia ser alcançado com a utilização de bibliotecas parcialmente normalizadas de cDNA construídas a partir de diferentes tecidos sob diferentes condições patológicas, ainda não avaliados no contexto do CGAP/PGH. Baseado na premissa de que diferentes tecidos neoplásicos possuem expressão gênica diferenciada, estes tumores também podem conter uma porcentagem de transcritos ainda não evidenciados nos tecidos atualmente em estudo.

1.10 A técnica de PCR para análise de genomas:

1.10.1 O uso da PCR em baixa estrigência para análise do DNA:

As abordagens denominadas RAPD (*Randomly Amplified polymorphic DNA*) (WELSH e MCCLELLAND 1990) e AP-PCR (*Arbitrarily Primed* - PCR) (WILLIAMS et al. 1990) são baseadas na amplificação de fragmentos de DNA genômico sob condições de baixa estringência. Ambas foram inicialmente usadas para evidenciar diferenças entre o DNA genômico de vários organismos.

1.10.2 A combinação das técnicas de síntese de cDNA (RT) e PCR de baixa estringência para análise de genes expressos:

O primeiro trabalho extendendo a técnica de AP-PCR para o estudo do RNA (*RNA Arbitrarily Primed* - RAP-PCR) também foi elaborado por WELSH et al. (1992). Eles conseguiram evidenciar através da técnica RAP-PCR, diferenças de expressão gênica em vários tecidos, usando o camundongo como modelo experimental. Também em 1992, foi publicada a técnica denominada *Differential Display* (DDRT-PCR) por LIANG e PARDEE (1992) onde foram usados iniciadores ancorados na porção 3' do mRNA para a transcrição reversa e iniciadores arbitrários para a síntese da segunda fita do cDNA. Esta técnica, representa uma poderosa ferramenta para a detecção de genes diferencialmente expressos, podendo também

evidenciar um complexo "fenótipo", através de bandas diferenciadas em gel de poliacrilamida não desnaturante, que reflete mudanças no nível de expressão gênica sob várias condições. Desde a sua introdução, uma grande quantidade de artigos foram publicados, visando melhoramentos ou diferentes aplicações (LIANG et al. 1993) (VEDOY et al. 1999). Desta forma, o método de RNA *fingerprinting* e seus variantes vêm sendo usados com sucesso na determinação de genes diferencialmente expressos e também na descoberta de genes (McCLELLAND et al. 1995). Um dos mais recentes trabalhos utilizando esta abordagem foi submetido a publicação em maio deste ano (DE VRIES et al. 2000), onde foram identificados 40 mRNAs em células musculares lisas formadoras das camadas média e adventícia de vasos sanguíneos, mostrando complexas alterações no espectro de expressão gênica sob diferentes estímulos, neste caso, foram usados citoquinas e fatores de crescimentos secretados por macrófagos ativados como estímulo diferencial.

1.11 A combinação da técnica de DDRT-PCR e tumores de baixa incidência para a identificação de genes humanos:

Na tentativa de identificar novos genes humanos, buscamos utilizar a técnica de DDRT-PCR em tumores de baixa incidência para geração de mini-bibliotecas de cDNA, parcialmente normalizadas, contendo clones que representem preferencialmente a porção 3' *UTR* de transcritos. A importância de gerar seqüências a partir da região 3' *UTR* é que a maioria das ESTs presentes nos bancos de dados é proveniente desta região, permitindo desta forma, a comparação das ESTs geradas com as ESTs dos bancos de dados. Estas características permitem avaliar o conteúdo de novos genes expressos em um dado tecido, além de permitir avaliar o nível de expressão dos genes encontrados, por comparação com o UniGene *cluster*. A técnica de *Difential Display* (DDRT-PCR), é usada para evidenciar a expressão diferencial dos genes se baseando na técnica de RT-PCR (SAMBROOK et al. 1989b), onde são usados iniciadores poli-T, os quais são denominados ancorados, do tipo T_{11VN} onde V = A, C ou G e N = A, C, G ou

T, dividindo a população de mRNA em 12 subpopulações de cDNA de fita simples. A amplificação de pequenos fragmentos de cDNA é feita por reação de PCR sob baixas condições de estringência, agregando um segundo iniciador de seqüência arbitrária contendo 10 nucleotídeos. Esses iniciadores permitem a vantagem de identificar os mRNAs independente da sua prevalência na amostra estudada (WAN et al. 1996). A combinação entre os iniciadores usados para síntese e os usados para a amplificação dos fragmentos de cDNA irá gerar substrato para a construção de mini-bibliotecas parcialmente normalizadas. Esta técnica normalmente é usada na identificação de genes diferencialmente expressos, onde é necessária a obtenção de perfis de bandamento que permitam a individualização de bandas. Neste trabalho objetivamos gerar um perfil mais complexo de bandas que representem um grande número de transcritos a partir de cada perfil, gerando desta maneira mini-bibliotecas de cDNA, parcialmente normalizadas e ancoradas nas porções mais ricas em A/T dos transcritos, prefencialmente na porção 3'.

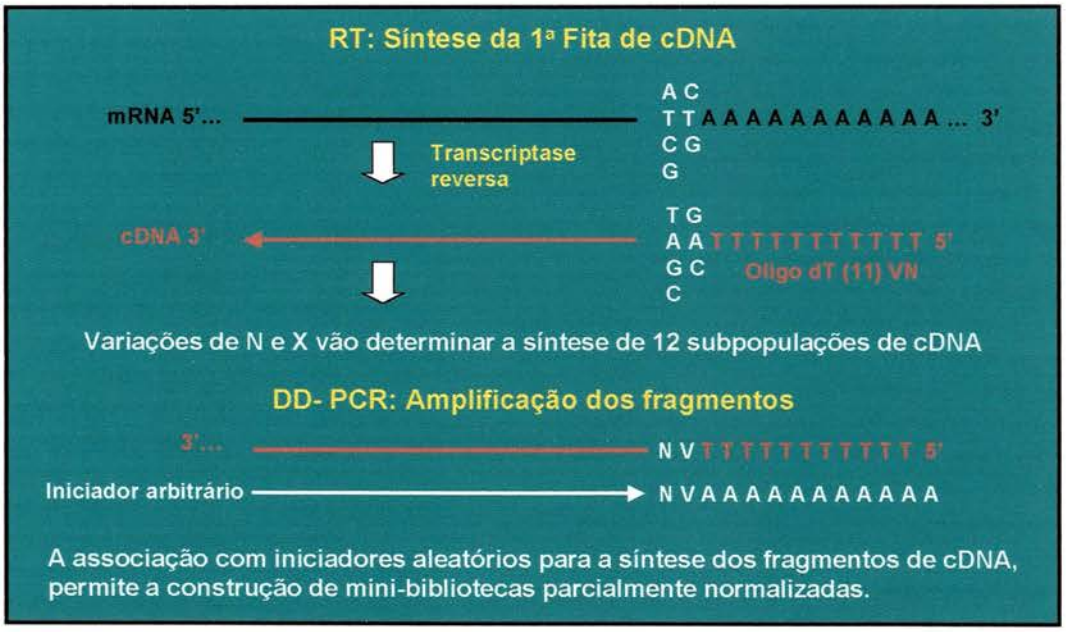


Figura 02: Representação esquemática para a geração de mini bibliotecas parcialmente normalizadas pela técnica de DDRT-PCR:

Os iniciadores usados, do tipo T_{11vn} onde V = A, C ou G e N = A, C, G ou T em associação aos iniciadores arbitrários, proporcionam a vantagem de identificar os transcritos independente da sua prevalência na amostra estudada, devido aos dois nucleotídeos degenerados (V e N) na porção 3' do iniciador do tipo $T(11)_{VN}$ no momento da síntese de cDNA e aos iniciadores arbitrários usados para a amplificação dos fragmentos de cDNA, permitindo assim, a normalização parcial dos transcritos.

1.11.1 Tumores de baixa incidência candidatos à prospecção gênica:

A coleção de tumores selecionados para o presente trabalho teve um maior enfoque em tumores infantis, uma vez que mundialmente estes tumores representam apenas 2% de todas as neoplasias malignas.

1- Tumor Neuroectodérmico Primitivo (PNET): ou Tumor de Askin, é um tumor maligno de pequenas células que acomete as regiões toraco-pulmonar de crianças e adolescentes. Este tumor está inserido na família dos Sarcomas de Ewing, tumor de origem neural, que é composta pelo sarcoma de Ewing ósseo (87%), sarcoma de Ewing extraósseo (8%) e tumor neuroectodérmico primitivo (PNET) (5%). Em um estudo retrospectivo de 105 pacientes com sarcoma de Ewing no departamento de pediatria do Hospital do Câncer, a mediana de idade foi 11 anos, 51% eram do sexo masculino e 81% brancos. Este tumor representa 16% dos tumores malignos de osso, a maioria dos pacientes são crianças ou adultos jovens, ocorrendo mais comumente na segunda década de vida (64%) e sendo muito raro antes dos 5 e após os 30 anos de idade. Os locais mais comuns de ocorrência de metástases incluem ossos e estruturas torácicas. (COSTA e DE CAMARGO 2000).

2- Tumor do Seio Endodérmico: este tumor pertence à família dos tumores de células germinativas (TCG) que podem ser benignos ou malignos, derivados das células germinativas primordiais , podendo ocorrer em sítios gonadais e extragonadais. São caracterizados por distintos achados histológicos que influenciam o prognóstico. Correspondendo a um grupo heterogêneo, cujo comportamento é de dificil caracterização. Os tumores malígnos são divididos histologicamente em dois grandes grupos: seminomatosos e não seminomatosos. A ocorrência anual é da ordem de 0,1 caso por 500.000 crianças abaixo de 15 anos. Representam 3% dos tumores malignos que acometem crianças e adolecentes. Este tumor tem como principal característica um alto poder de metastatização e pode ser encontrado no testículo, ovário, região anal, entre outros (LOPES et al. 2000).

3- Rabdomiosarcoma Embrionário: os rabdomiossarcomas requerem um acurado método diagnóstico, pois, são subdivididos em vários subgrupos, cada um exigindo uma terapêutica específica. Além disso, estes tumores são altamente agressivos e por muitas vezes letais. Este tumor é originário da musculatura estriada sendo hoje considerado exclusivo de crianças. Histologicamente são subdivididos, basicamente, em 2 grupos: embrionário (de baixa incidência) e alveolar (associado a um prognóstico menos favorável), sendo que, na maioria dos casos, cada um apresenta distintas características clínicas. Uma característica molecular importante neste tumor é um polimorfismo envolvendo a perda de um alelo no cromossomo 11. A região de localização é a 11p15.5, região similar a encontrada em pacientes com síndrome de Beckwith-Wiedemann, seja em tumor de Wilms ou hepatoblastoma (DIAS NETO et al. 2000b).

4- Hepatoblastoma: Os tumores malignos do fígado correspondem a 5% das neoplasias pediátricas, dos quais a grande maioria é representada pelo hepatoblastoma seguido do hepatocarcinoma. O hepatoblastoma é o terceiro tumor intra-abdominal mais frequente em crianças, após o tumor de Wilms e o neuroblastoma. Geralmente são diagnosticados antes dos 3 anos de idade, principalmente por volta de 1 ano, com predomínio de incidência no sexo masculino (1,5:1) (COSTA et al. 2000). A incidência anual destes tumores nos E.U.A. por milhão de crianças é de 1.6, sendo 0.9 para hepatoblastoma e 0.7 para hepatocarcinoma. Embora a etiologia seja desconhecida, devido à baixa idade dos pacientes, dados epidemiológicos sugerem uma origem embrionária. Existem dois tipos morfológicos: o tipo epitelial puro, que contém células embrionárias ou fetais, e o tipo misto, contendo tecidos mesenquimais em adição aos elementos epiteliais (Greeberg M e Filler RM, 1997).

5- Neurofibrosarcoma: também denominado *Schwannoma* maligno, são tumores originalmente benignos da bainha neural. Estes tumores estão associados à doença de *Von Rechlinghansen*, uma síndrome dominante. Os sítios primários mais comuns são extremidades abdominal e torácica. As características genéticas mais comumente encontradas são

deleções do cromossomo 17q 11.2 e mutação pontual no gene p53 (GREENBERG e FILLER 1997).

6- Tumor de Wilms: A incidência do tumor de Wilms é baixa, cerca de 8,1 casos por milhão de crianças brancas menores de 15 anos de idade. Em 1991, este tumor representou 5% a 6% dos casos de cânceres em crianças nos EUA, onde a incidência total foi estimada em 460 casos por ano. O tumor de Wilms selecionado para este estudo está associado à síndrome de Denis Drash que tem como principais características o hermafroditismo e doenças renais. Esta síndrome apresenta alterações no mesmo segmento cromossomico da síndrome de *WAGR* (*Wilms tumor, Aniridia, Genitourinary malformation, mental Retardation*). O tumor de Wilms é composto, ou derivado do blastema primitivo metanéfrico. Vários tipos celulares são encontrados neste tipo de neoplasia. Agregados às células do desenvolvimento normal do rim, podem ser encontrados tecidos como, músculo esquelético, cartilagem e epitélio escamoso. Esta heterogeneidade de tecidos pode ser atribuída ao potencial de originar outros tecidos do blastema metanéfrico, cujo potencial não é encontrado na nefrogênese normal (GREEN et al. 1997).

7- Leiomiossarcoma: Os leiomiossarcomas atingem cerca de 7% dos sarcomas de partes moles em adultos, porém estes tumores são raros em crianças, respondendo por menos de 2% dos casos. Os leiomiossarcomas são tumores agressivos com grande incidência de recorrência local, dando metástases por via sanguínea para o pulmão e fígado, tendo também envolvimento dos nódulos linfáticos regionais (MISER et al. 1997).

8- Hibernoma: A incidência deste tumor em crianças e adolescentes é rara. Termos como lipoma do tecido adiposo imaturo, lipoma da gordura embrionária e lipoma fetal são propostos por alguns autores, porém, o termo hibernoma, é devido à gordura marrom, que está presente nos estágios iniciais do desenvolvimento da gordura branca e tem a aparência do tecido adiposo presente nos ursos no período de hibernação. Clinicamente, os hibernomas são tumores indolores de crescimento lento

que ocorrem na derme (subcutis) ou raramente em músculos e normalmente são notados por vários anos antes de serem excisados. Eles são bem definidos, moles, móveis e normalmente variam de 5 a 10 cm de diâmetro, a tonalidade de coloração varia de marrom para vermelho amarronzado (ENNZIGER e WEISS 1995).

9- Leiomioma: são tumores benignos de músculo liso, podem ser encontrados em qualquer parte do corpo, como, por exemplo, nas túnicas musculares do intestino, na camada média dos vasos sanguíneos, no esôfago e no estômago, mas sua localização mais comum é no útero. No entanto, na maioria dos casos os leiomiomas são benignos (ROBBINS e COTRAN 1993).

O Centro de Tratamento e Pesquisas do Hospital do Câncer - A. C. Camargo - São Paulo/SP, é um centro de referência nacional, atendendo anualmente cerca de 300 novos casos de cânceres em crianças, oriundas de todas as partes do país. A tabela 04 informa a freqüência dos tumores, selecionados para o presente trabalho, com base nos dados coletados no hospital do câncer em dois períodos, 1988 e 1994.

Tabela 04: Freqüência dos tumores selecionados no Hospital do Câncer nos anos de 1988 e 1994:

Tumor	Nº de casos em 1988	Nº de casos em 1994
PNET	04	06
Tumor do seio endodermico	08	07
Rabdomiossarcoma embrionário	10	05
Hepatoblastoma	01	01
Neurofibrossarcoma	00	00
Tumor de wilms	22	19
Leiomiossarcoma	01	00
Hibernoma	00	00
Leiomioma	00	00

Dados extraídos do registro hospilar de câncer pediátrico do Centro de Tratamento e Pesquisas do Hospital do Câncer - A. C. Camargo - São Paulo/SP, períodos 1988 & 1994. (Ribeiro et al. 1999)

2 OBJETIVOS:

2.1 Objetivo Geral:

Avaliar o processo de descoberta de genes humanos através da geração de ESTs a partir de uma coleção de tumores de baixa incidência, ainda pouco ou não estudados no contexto do CGAP/HPG .

2.2 Objetivos Específicos:

Avaliar a estratégia de descoberta gênica utilizando PCR em condições de baixa estringência.

Verificar a especificidade desta técnica em termos de distribuição posicional das ESTs e capacidade de normalização

Gerar 500 ESTs como um controle da estratégia usada, partindo de tumor de cólon, um tumor já bastante estudado.

Gerar 5000 ESTs a partir de mini-bibliotecas de cinco tumores de baixa incidência, pouco estudados em termos de avaliação de genes expressos.

3 MATERIAL E MÉTODOS:

3.1. Material biológico:

As amostras de tumores de baixa incidência foram selecionadas, com o auxílio do patologista Dr Fernando Soares e do pediatra Dr. Luiz Fernando Lopes, dentre as existentes no Banco de Tumores do Centro de Tratamento e Pesquisas do Hospital do Câncer - A. C. Camargo, ligado ao Instituto Ludwig de Pesquisa sobre o Câncer - São Paulo/SP (ILPC). As amostras foram isoladas após a cirurgia e mantidas a -70°C. Apenas a primeira amostra, carcinoma de cólon (9002364), passou por uma etapa de microdissecção experimental. Onde o tecido não tumoral adjacente foi retirado pelo Dr. Fernando Soares (Tabela 05).

Tabela 05: Tumores estudados:

Número de Registro	Idade do Paciente	Tumor	Região anatômica	Extração de RNA total	Quantidade de RNA total
9002364	36	Carcinoma de cólon	Cólon	*RNAzol™ B Isolation of RNA*	57,04µg
922645	32	Carcinoma de cólon	Cólon	*QuickPrep™* (Pharmacia)	?
9804162 -2	32	Carcinoma de mama	Mama	*TRIzol®*	361,20 µg
9901194-8	65	Carcinoma de cólon	Cólon	*TRIzol®*	3,00 µg
97-4111	7 anos	PNET	Parede do tórax (costela)	*TRIzol®*	1,26 µg
97-4335	16 anos	Tumor do Seio Endodérmico	Ovário	*TRIzol®*	Degradado
97-04837	14 anos	Rabdomiossarcoma embrionário	Retroauricular	*TRIzol®*	Degradado
98-00149	3 anos	Hepatoblastoma	Fígado	*TRIzol®*	Degradado
97-04929	5 anos	Neurofibrossarcoma	Pescoço	*TRIzol®*	Degradado
98-03992	8 meses	Wilms/Denis Drash	Rim	*TRIzol®*	2,40µg
99-1455	5 anos	Leiomiossarcoma	Retroauricular	*TRIzol®*	2,40µg
9805203	16 anos	Hibernoma	Coxa direita	*TRIzol®*	0,52 µg
97-04082	17 anos	Leiomioma	Estômago	*TRIzol®*	1,40µg

Tumores selecionados do banco de tumores do Centro de Tratamento e Pesquisas do Hospital do Câncer - A. C. Camargo, ligado ao Instituto Ludwig de Pesquisa sobre o Câncer - São Paulo/SP (ILPC).

3.2 Purificação de RNA:

Antes da otimização da metodologia duas estratégias foram utilizadas para as extrações de RNA baseadas na extração pelo (i) reagente *RNAzol™ B Isolation of RNA* (Cinna/Tel-Test), seguido de tratamento com

DNase e (ii) utilização do *kit QuickPrep*^{TM} (Pharmacia) segundo protocolo do fabricante.

Posteriormente as preparações e os controles de mRNA foram feitas pelo grupo do Prof. Dr. Luis F. Lima Reis (Laboratório de Inflamação, Instituto Ludwig de Pesquisa sobre o Câncer, São Paulo), seguindo os protocolos desenvolvidos para o Projeto Genoma Humano do câncer (ILPC/FAPESP):

3.2.1 Extração de RNA total:

O RNA total foi isolado utilizando o reagente *TRIzol*® (*Life Technologies, USA*). O *TRIzol*® corresponde a uma solução monofásica de fenol e isotiocianato de guanidina. A extração de RNA utilizando este reagente é uma adaptação do método desenvolvido por CHOMCZINSKI e SACCHI (1987).

Para a extração do RNA total, os fragmentos de tecido foram colocados em tubos contendo 1ml de *TRIzol*® para cada 100mg de tecido, e homogeneizados com o auxílio de um triturador de tecidos (Polytron). A purificação do RNA foi feita segundo o protocolo do fabricante.

A quantificação do RNA obtido foi feita através de leitura de uma alíquota da amostra no espectrofotômetro em comprimento de onda equivalente a 260 nm, considerando que $1DO_{260nm}$ equivale a 40μg/ml de RNA (SAMBROOK et al. 1989a). A relação entre as leituras realizadas a 260 e 280nm foi utilizada como parâmetro na estimativa do grau de contaminação do RNA por proteínas. Invariavelmente, o valor desta relação estava na faixa de 1,6 a 1,8.

3.2.2 Tratamento com DNase I- livre de RNase:

Este passo tem como objetivo eliminar contaminações de DNA genômico nas amostras de RNA. Para o tratamento dos RNAs com DNase I (Promega), adiciona-se 10 unidades desta enzima para cerca de 50μg de RNA (50μl) em tampão 5,6 mM $MgCl_2$; 40 U de RNAsin (inibidor de RNase, Promega), 0,1 mM de DTT, 50mM de KCl e 20 mM de Tris-HCl pH 8,4. A

reação se processa por 60 minutos a 37°C sendo interrompida pela adição de 25μl de solução de neutralização (2,5μl EDTA 0,5M, 12,5μl NaOAC 3M, 1,25μl SDS 20% e água q.s.p. 25μl), em seguida, tratada com 1 volume de Fenol/ Clorofórmio/ álcool.-.isoamílico nas respectivas proporções: 25/24/1. A solução é agitada vigorosamente por 15 segundos e centrifugada a 12000Xg por 15 minutos a 4°C. O RNA, presente na fase aquosa, é coletado e precipitado com 10% de NaCl 5M e 3X o volume de etanol 80% em *freezer* a -70°C por 12 a 16 horas. A solução é novamente centrifugada a 12000Xg por 15 minutos a 4°C e, após descarte do sobrenadante, o precipitado é lavado com etanol a 70%. Após uma nova centrifugação, a solução de lavagem é descartada e o precipitado é seco a temperatura ambiente. O RNA é ressuspendido em 50μl de água deionizada tratada com DEPC (Di-Etil pirocarbonato, Sigma).

3.2.3 Purificação do mRNA:

O mRNA foi purificado através do sistema *mRNA Isolation kit* (*Miltenyi Biotec*). Este método se baseia na utilização de esferas magnéticas contendo oligo-dT (*MACS oligo (dt) MicroBeads, Miltenyi Biotec*) que promovem a adsorção da cauda poli-A das populações de mRNA a serem purificadas. A purificação foi feita segundo o protocolo do fornecedor e a concentração da solução de mRNA foi mensurada através da leitura espectrofotométrica de absorbância a 260 nm de uma diluição 1/100. O grau de pureza da preparação foi avaliado pela relação das medidas de absorbância a 260 e 280nm.

3.2.4 Análise da qualidade do mRNA extraído:

Para avaliar o grau de contaminação do RNA com DNA, foi feita a amplificação específica da região do D-Loop mitocondrial (região conservada, não expressa) e amplificação específica do fragmento do gene p53 entre os exons 02 e 04, usando como molde da reação o mRNA extraído. As reações foram feitas da seguinte forma:

- D-Loop DNA mitocondrial: 100ng de RNA tratado com DNase I foram adicionados ao tubo de reação contendo 20 pmoles de cada iniciador (Mitfw: 5´-AATAACAATTGAATGTCTGCAC-3´; MITrv: 5´-TTGAGGAGGTAAGCTACATA-3´); 200 μM de dNTPs, 1,5mM de $MgCl_2$ e 0,5 unidade de *Taq* DNA polimerase (Gibco-BRL) em tampão 50 mM KCl e 20mM Tris-HCl pH 8,4 (volume final de 20μl). A reação foi submetida a um ciclo inicial de desnaturação a 95°C por 5 minutos, seguidos por 35 ciclos de desnaturação a 95°C por 1 minuto, pareamento a 55°C por 1 minuto e extensão a 72°C por 1 minuto e um ciclo final de extensão a 72°C por 5 minutos. Cerca de 5 μl do produto da reação foram aplicados em gel de poliacrilamida a 8% e, após eletroforese, analisado por coloração com prata (SANGUINETTI et al. 1994). A banda referente ao fragmento do D-Loop amplificado é de aproximadamente 379 pares de bases, observada no controle positivo onde foi usado DNA.

- p53: 1μl de cDNA (e/ou diluição 1/10), sintetizado a partir de 10-100ng de RNA com a enzima transcritase reversa (*SuperScript II* Gibco-BRL) na presença de pd $(N)_6$ ou oligo dT (11-18) em um volume final de 20μl, foi adicionado ao tubo de reação contendo 20 pmoles de cada iniciador (P53FWE2: 5´-GGAGGAGCCGCAGTCAGA -3´; P53RVE4: 5´-CAAGAAGCCCAGACGGAAAC-3´); 200 μM de dNTPs, 1,5mM de $MgCl_2$ e 0,5 unidade de *Taq* DNA polimerase (Gibco-BRL) em tampão 50 mM KCl e 20mM Tris-HCl pH 8,4 (volume final de 20μl). A reação foi submetida a um ciclo inicial de desnaturação a 95°C por 5 minutos, seguidos por 35 ciclos de desnaturação a 95°C por 1 minuto, pareamento a 55°C por 1 minuto e extensão a 72°C por 1 minuto, e um ciclo final de extensão a 72°C por 5 minutos. Cerca de 5 μl do produto da reação foram aplicados em gel de poliacrilamida a 8% e, após eletroforese, analisado por coloração com prata. A detecção de DNA dependerá do tamanho do produto amplificado, sendo 340 pb quando não existe contaminação por DNA genômico e 548 pb quando existe contaminação, o que pode ser observado no controle positivo

onde foi usado DNA humano. A diferença no tamanho do produto amplificado se deve à região intrônica que será parte do produto amplificado.

Para avaliar o grau de degradação do RNA extraído, foi feito *Northern Blot* utilizando 5μg de RNA total hibridando com sonda de cDNA do gene Gliceraldeído 3-fosfato desidrogenase humano (GAPDH, gene constitutivo), fragmento contendo cerca de 600 pb marcado com [α^{32} P]dCTP. Para isto, os RNAs extraídos conforme descrito anteriormente foram analisados em gel de agarose 1% contendo: 0,02 M de ácido bórico (pH 8,3), 0,5 mM de EDTA (pH 8,0) e 6% de de formaldeído.

As amostras contendo 5 μg de cada RNA em um volume final de 20 μl foram adicionadas a 20 μl de tampão de amostra (20% glicerol, borato 0,02 M, 0,2 mM EDTA, 50% formamida, 16% formaldeído e azul de bromofenol) e então incubadas em banho-maria a 65°C durante 10 minutos sendo imediatamente transferidas para gelo (5 minutos). As amostras foram aplicadas no gel e submetidas a eletroforese a 70 volts em tampão de corrida (borato 0,02 M, 0,5 mM EDTA e 5% formaldeído) durante 4 a 5 horas.

O gel foi lavado 2x em água deionizada durante 20 minutos a fim de eliminar o excesso de formaldeído. As amostras foram transferidas por capilaridade para uma membrana de *nylon* (Hybond-N) com 10x SSC (3M NaCl. 0.3M citrato de sódio pH 6.45) durante 12 a 16 horas. A eficiência de transferência foi conferida mediante visualização do RNA utilizando uma lâmpada UVG-11 (*Mineralight*) de baixo comprimento de onda (254 nm) e os RNAs foram fixados na membrana, quantidade de energia constante 120,000 μj/cm^2, utilizando o equipamento *UV Crosslinker FB-UVXL-1000* (FisherBiotech).

As sondas para GAPDH foram marcadas com [α^{32} P]dCTP pelo método de *random priming* utilizando o kit *redi*prime™ *II* (Pharmacia Biotech). A hibridação foi realizada a 65°C durante 16 horas em solução de hibridação (0,25 M Na_2HPO_4, 7% SDS, 1% BSA, 1mM EDTA). Posteriormente, a membrana foi lavada com uma solução de lavagem (0.25 M Na_2HPO_4, 1% SDS e 1 mM EDTA) durante 30 minutos, a 65°C a fim de

retirar o excesso de sonda. Finalmente, filmes de raios X foram expostos às membranas a temperatura de -70°C, durante 36 a 48 horas, onde o resultado foi visualizado.

3.3 Construção de mini-bibliotecas de cDNA:

3.3.1 Síntese de cDNA:

A princípio, a síntese de cDNA foi feita utilizando a enzima *SuperScript II* (Gibco-BRL), de acordo com o protocolo do fabricante. Os iniciadores utilizados foram (i) do tipo $T(11)_{VN}$, descritos para a técnica de *Differential Display*, ou (ii) do tipo $T(25)_{CG}$, desenhado e sintetizado exclusivamente para este trabalho.

Após a padronização, a síntese de cDNA foi desenvolvida da seguinte forma: uma alíquota de 10-100 ng de mRNA foi misturada com 50 pmoles de iniciador oligo-dt (20) e aquecida a 65°C por 10 minutos. Em seguida foi adicionada solução contendo: 0,5 mM de dNTPs, 40 unidades de *RNAseOUT* ™ e 15 unidades de *ThermoScript* RT (Gibco-BRL), 5mM de DTT e tampão contendo 250 mM de Tris-Acetato (pH 8.4) 375 mM Acetato de Potássio, 40mM Acetato de Magnésio e estabilizador (volume final 20μl). A mistura é incubada a 50°C por 1 hora e, posteriormente, as enzimas foram desnaturadas a 85°C por 05 minutos. O cDNA total obtido foi estocado a -70 °C, para posterior amplificação dos fragmentos.

3.3.2 Amplificação dos fragmentos de cDNA:

A princípio a amplificação dos fragmentos de cDNA foi feita de acordo com o protocolo descrito para a técnica de *Differential display*, usando os iniciadores do tipo $T(11)_{VN}$. Durante a padronização, a amplificação dos fragmentos de cDNA foi desenvolvida utilizando os iniciadores do tipo $T(25)_{CG}$. As seqüências dos iniciadores usados para esta etapa estão listadas na tabela 6.

Tabela 06: Iniciadores usados para amplificação dos fragmentos de cDNA.

Nome	Seqüência 5' 3'	Observações
$T(11)_{VN}$	TTTTTTTTTTVN	V= A, C ou G N= A, C, G ou T
$T(11)_{VTC}$	TTTTTTTTTTVTC	V= A, C ou G
$T(11)_{VVV}$	TTTTTTTTTTVVV	V= A, C ou G
$T(25)_{CG}$	TTTTTTTTTTTTTTTTTTTTTTTTTCG	Usado durante a padronização
Ldd1 (13)	CTGATCCATG	Usado durante a padronização
Ldd2 (14)	CTGCTCTCAA	Usado durante a padronização
Ltk3 (15)	CTTGATTGCC	Usado durante a padronização

Após padronização, a amplificação dos fragmentos de cDNA para a geração das mini-bibliotecas foi feita com oligonucleotídeos do tipo OligodT$(11)_{VN}$. Procedemos esta reação da seguinte forma: 1μl de cDNA sintetizado com o kit *ThermoScript* RT (Gibco-BRL) (diluição 1/10 e 1/20) foi adicionado ao tubo de reação contendo 10 pmoles do iniciador $T(11)_{VN}$; 200 μM de dNTPs, 5mM de $MgCl_2$ e 1,5 unidade de *Taq* DNA Polimerase (*Sequencing Grade-Promega*) em tampão 50 mM KCl e 20mM Tris-HCl pH 8,4 (volume final de 20μl). A reação foi submetida a um primeiro passo de desnaturação a 65°C por 3 minutos, pareamento a 38°C por 1 minuto e extensão a 72°C por 1 minuto, seguidos por 45 ciclos de desnaturação a 95°C por 1 minuto, pareamento a 38°C por 1 minuto e extensão a 72°C por 1 minuto, e um ciclo final de extensão a 72°C por 5 minutos. Cerca de 3 μl do produto da reação foram aplicados em gel de poliacrilamida a 8% e, após eletroforese, corado pela prata para a análise do perfil de bandamento.

3.3.3 Ligação dos produtos amplificados em vetores:

A clonagem do produto da reação anterior foi feita em pUC 18 utilizando o *SureClone Ligation Kit* (*Pharmacia Biotech*) de acordo com o protocolo do fabricante. Os produtos da reação de PCR são preparados para a reação de ligação utilizando a atividade exonucleásica 3' → 5' do fragmento *Klenow* da DNA polimerase I visando remover bases únicas não-pareadas das extremidades 3' do fragmento. Ao mesmo tempo estes fragmentos são fosforilados pela T4 polinucleotídeo quinase na presença de ATP. Após extração orgânica com fenol/clorofórmio e purificação em *MicroSpin Column (Pharmacia Biotech),* o produto de PCR foi ligado ao vetor pUC18. Este vetor é pré-digerido com enzima *Sma*I para formação de sítios de extremidades abruptas (*blunt ends*) e tratado com fosfatase alcalina para evitar a re-ligação do vetor sem inserto.

A eficiência da ligação dos perfis de cDNA em vetor pUC 18 foi avaliada por PCR, utilizando *primers* específicos ao sítio de clonagem do vetor (M-13F: 5'-TCCCAGTCACGA CGT-3'; M-13R: 5'-AACAGCTATGACCATG-3') (DIAS NETO et al. 1996). A reação foi feita utilizando 1μl de produto de ligação, 3 pmoles de cada *primer*, 200 μM de dNTPs e 0,5 unidade de *Taq* DNA polimerase (Gibco-BRL) em tampão 1,5 mM $MgCl_2$; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 10 μl). A reação se processou por 35 ciclos de desnaturação a 95°C por 45 segundos, anelamento a 55°C por 45 segundos e extensão a 72°C por 1 minuto, com uma extensão final de 72°C por 5 minutos. Após amplificação, um volume de 3 a 5 μl de cada amostra foi analisado em gel de poliacrilamida a 8% corado com prata. Após a verificação da eficiência de ligação, uma alíquota deste produto foi usada para transformar bactérias competentes *E.coli* JM109.

3.3.4 Preparação de bactérias competentes:

Para a preparação de bactérias competentes foram utilizadas linhagens bacterianas de *E. coli* JM109. Uma alíquota do estoque (-70°C) destas células foi plaqueada em LB-agar (Peptona de caseína 1%, extrato de levedura 0,5%, NaCl 0,5% e Agar 1,5%) e incubada por 16 horas a 37°C.

Em seguida, foram inoculadas dez colônias isoladas em 260 ml de meio LB (Peptona de caseína 1%, extrato de levedura 0,5%, NaCl 0,5%) contido em um frasco de 2 litros. Esta suspensão de células foi incubada a 18 - 22°C sob agitação vigorosa (200rpm) até atingir um crescimento correspondente a uma D.O. de 0,6 a 600nm. Ao atingir esta densidade óptica, o frasco foi imediatamente colocado em gelo por 10 minutos e depois distribuído em seis alíquotas de 42 ml em tubos de centrifuga de 50 ml. Os tubos foram centrifugados a 3500 rpm (centrifuga sorvall RC5B) por 15 minutos a 4°C. Cada precipitado foi ressuspendido em 12 ml de TB (*Transformation Buffer:* 10mM Pipes, 55mM $CaCl_2$, 250mM KCl) gelado, após a ressuspensão o conteúdo final foi reunido em apenas dois tubos com 36 ml cada um. Os tubos foram incubados no gelo por 10 minutos e centrifugados novamente como descrito acima. O precipitado foi ressuspendido em 18,6 ml de TB contendo 1,4 ml de DMSO (TB a 7% de dimetilsulfóxido). A suspensão de células foi incubada em gelo por 10 minutos, em seguida foram aliqüotadas em tubos para congelamento e imediatamente imersas em nitrogênio, onde foram mantidas até o momento de sua utilização (INOUE et al. 1990).

3.3.5 Transformação:

A transformação das bactérias competentes foi feita através de choque térmico por cerca de 50 segundos a 42°C utilizando 5μl do produto ligado através do kit *SureClone* para 100μl de bactérias competentes, preparadas conforme etapa anterior. Após a transformação, os clones recombinates são selecionados pela resistência à amplicilina em placas de 14cm de diâmetro contendo meio LB-agar (Luria Bertani) suplementado com 100μg/ml de ampicilina, 40 μg/ml de 5-bromo-4-cloro-3-indoil-β-D-galactosídeo (X-gal) e 0,5 mM de IPTG (iso – propyl-β-D-tiogalactopiranosídeo). Os vetores de clonagem da série pUC proporcionam uma seleção prévia de clones por α-complementação (Sambrook *et al.*, 1989), onde as colônias contedo inserto podem ser selecionadas pela ausência de coloração azul (devido a incapacidade destas colônias de

produzirem a enzima β-galactosidase, que degrada seu substrato cromogênico X-gal, originando colônias brancas).

3.4 Preparação das amostras para seqüenciamento:

A princípio, o DNA molde para o seqüenciamento foi gerado através de purificação de plasmídeos utilizando o kit *Wizard Plasmid Minipreps* (Promega, EUA), que é um método modificado para lise alcalina, originalmente descrito por OZAKI e CSEKO (1984). Para tanto, as colônias contendo os plasmídeos a serem seqüênciados são crescidas a 37°C por 12-16 horas, sob agitação, em 5 ml de meio LB suplementado com 50-100 μg/ml de ampicilina. Cada preparação resulta em aproximadamente 20 μg de plasmídeos ressuspendidos em 50 μl de água.

Após a padronização da metodologia, os clones recombinantes foram crescidos em placas de 96 poços e a obtenção de DNA molde para o seqüenciamento foi feita através da amplificação dos fragmentos clonados por PCR, com uso dos iniciadores específicos para o sítio de poli-clonagem do vetor pUC18. Procedemos esta reação da seguinte forma: 1μl de cada um dos clones recombinantes crescidos em meio LB líquido contendo 50 μg/ml de ampicilina de 12 a 14 horas é adicionado em um poço de uma placa de 96 poços contendo: 1pmol dos iniciadores (M13 -40 5´-CGCCAGGGTTTTCCCAGTCACGAC-3´, M13 Rev 5´-TTTCACACAGGAAACAGCTATGAC-3´); 0,8 μM de dNTP, 1.1mM de $MgCl_2$ e 0,6 unidade de *rTth* DNA Polimerase (*kit GeneAmp XL PCR – Perkin-Elmer*) em tampão Tris, Acetato de Potássio, Glicerol e DMSO (concentração omitida pelo fabricante), volume final de 10μl. A reação foi submetida a um ciclo inicial de desnaturação a 95°C por 5 minutos, seguidos por 40 ciclos de desnaturação a 95°C por 1 minuto, pareamento a 55°C por 1 minuto e extensão a 72°C por 1 minuto, e um ciclo final de extensão a 72°C por 5 minutos.

Uma amostragem de 14 produtos de PCR de cada placa foi previamente analisada por gel de poliacrilamida da seguinte forma: cerca de

3 μl do produto da reação foram aplicados em gel de poliacrilamida a 8% e, após eletroforese, o resultado foi visualizado por coloração pela prata. Este procedimento permite o controle de qualidade de transformação e da produção de DNA molde para o seqüenciamento de cada mini biblioteca reduzindo o seqüenciamento de bibliotecas com alta porcentagem de plasmídeos sem inserto ou de placas cuja PCR não foi de boa qualidade. Desta forma, foram levadas para a etapa de seqüenciamento apenas as placas que apresentavam mais de 70% de aproveitamento.

3.5 Seqüenciamento:

O seqüenciamento foi feito pelo método de terminação da cadeia por 2',3'-didesoxirribonucleotídeos (ddNTPs) (SANGER et al. 1977) utilizando os iniciadores, M13 -40 5´-CGCCAGGGTTTTCCCAGTCACGAC-3´ ou M13 Rev 5´-TTTCACACAGGAAACAGCTATGAC-3´.

3.5.1 Seqüenciamento em ALFexpress™ DNA Sequencer (Amersham Pharmacia Biotech)

Inicialmente o seqüenciamento das amostras deste trabalho, partindo de plasmídeos purificados, foi desenvolvido no equipamento ALF*express*™. As reações de seqüenciamento foram feitas por ciclagem térmica, utilizando-se o kit *Thermosequenase™* (*Amersham*). O programa consiste de uma desnaturação inicial a 95°C por 2 minutos, seguida por 25 ciclos de pareamento a 55-60°C por 36 segundos, extensão a 72°C por 84 segundos e desnaturação a 95°C por 36 segundos. A extensão final foi prolongada por um período de 5 minutos. O produto final das reações foi desnaturado conforme protocolo do kit e submetido à eletroforese em gel de poliacrilamida 6% e uréia 7M para o seqüenciamento.

3.5.2 Seqüenciamento em ABI Prism 377™ DNA Sequencer (P. E. Applied Biosystems)

As reações de seqüenciamento foram feitas por ciclagem térmica, utilizando-se o kit *BigDye™ Terminator Cycle Sequencing Ready Reaction Kit* (Perkin Elmer). O programa consiste de uma desnaturação inicial a 96°C

por 5 minutos, seguida por 25 ciclos de desnaturação a 96°C por 30 segundos, pareamento a 50°C por 15 segundos e extensão a 60°C por 4 minutos. O produto amplificado foi precipitado conforme protocolo do kit e então, ressuspendido em 3 μl de tampão de amostra (1:3 de *blue dextran* e formamida). A amostra foi submetida à eletroforese em gel desnaturante (Perkin Elmer) por um período de 3 a 7 horas para o seqüenciamento.

3.5.3 Seqüenciamento em MegaBace™ DNA Sequencer (Amersham Pharmacia Biotech)

Este processo de seqüenciamento foi usado em apenas duas das mini bibliotecas geradas. As reações de seqüenciamento foram feitas por ciclagem térmica, utilizando o kit *Thermo Sequenase™ II Dye Terminator Cycle Sequencing Kit* (Amersham Pharmacia). O programa é composto por uma série de 40 ciclos de desnaturação a 95°C por 30 segundos, pareamento a 55°C por 15 segundos e extensão a 60°C por 1 minuto. O produto amplificado é processado conforme protocolo do kit. Para o seqüenciamento, a amostra é injetada a 3 kV por 50 segundos e submetida à eletroforese capilar por um período de 100 minutos a uma voltagem de 9 kV.

3.6 Análise dos resultados:

A princípio, as análises das seqüências produzidas foram feitas uma a uma através de programa Blast em buscas contra os bancos de dados por acesso remoto (dbEST, Blast (nr) e Blast (x)).

Posteriormente, devido à grande quantidade de dados gerados após a padronização da metodologia, as ESTs produzidas passaram a ser analisadas utilizando um protocolo automatizado, desenvolvido pelo setor de bioinformática do ILPC para busca contra bancos de dados públicos. As seqüências produzidas foram analisadas quanto à qualidade pelo programa *Phred* (EWING et al. 1998), utilizando parâmetros definidos pelo laboratório de bioinformática. As regiões de baixa qualidade são excluidas utilizando-se janelas. Em um primeiro passo uma janela contendo 29 nt é analisada,

destes 29 nt se 07 nt apresentarem *score* menor que 10 (dado gerado pelo programa *Phred*), a janela é eliminada e uma nova janela é avaliada com os mesmos parâmetros. Em um segundo passo os próximos 19 nt são "pulados" (ignorados) para então se fazer uma análise de uma janela contendo 59 nt, destes 59 nt se 29 nt apresentarem *score* menor que 10, todo este fragmento será excluido. Após a remoção de seqüências de baixa qualidade o restante é editado para remoção de seqüências do plasmídeo e analisado frente a um banco de dados específico para elementos repetitivos (*RepeatMasker*), onde as seqüências que contém tais elementos são "mascaradas". Posteriormente, foram feitas as submissões contra bancos de dados para cada uma das ESTs geradas seguindo a seguinte ordem: (i) bancos de dados contendo seqüências de bactérias, DNA mitocondrial e RNA ribossomal; (ii) bancos de dados de *contigs* humanos (UniGene); (iii) bancos de dados de ESTs humanas não agrupadas no UniGene (dbEST).

As ESTs que apresentaram identidade com genes humanos totalmente seqüenciados foram usadas para análise de normalização, onde foi avaliado o nível de expressão do gene que deu origem a EST, levando em consideração o número de entradas (seqüências depositadas) no seu respectivo *cluster* no UniGene (http://www.ncbi.nlm.nih.gov/unigene), assumindo que o número de entradas de ESTs nos *clusters* do UniGene é diretamente proporcional ao nível de expressão gênica.

O nível de redundância das bibliotecas geradas foi determinado com o auxílio do programa Cap3 (HUANG e MADAN 1999) utilizando parâmetros definidos pelo setor de bioinformática. Este programa normalmente é usado para a montagem de *contigs* a partir de um grupo de seqüências. Porém, pode ser usado também para a determinação do nível de redundância, uma vez que a montagem de *contigs* é baseada no nível de identidade entre as seqüências do grupo em estudo.

A categoria de ESTs, que apresentaram identidade com genes humanos totalmente seqüenciados também foi usada para análise da distribuição posicional, cujo objetivo é informar qual a posição média do alinhamento das ESTs com os genes. Para a obtenção deste dado foram

usados os números relativos ao início e final do alinhamento com o gene, onde o resultado da média aritmética destes dois números dividido pelo tamanho do gene informou a posição relativa das ESTs nos genes avaliados.

As análises estastíscas foram realizadas com o auxílio do programa Epi Info 6 versão 6,04b (1997).

4- Resultados:

4.1 Aplicações iniciais:

A metodologia inicial do trabalho descrito nesta tese, foi baseada na técnica de *Differential Display* DDRT-PCR para a construção de mini-bibliotecas de cDNA parcialmente normalizadas. Seguindo esta metodologia, foram geradas cerca de 200 ESTs (tab 07) partindo de 04 mini bibliotecas construídas a partir de mRNA derivado de carcinoma de cólon (fig 03). A busca de similaridades destas ESTs foi feita usando o programa Blast (Blast N e Blast X) contra os bancos de dados (dbEST, Blast nr e banco de proteínas) de acesso remoto, onde os iniciadores usados, o número da mini biblioteca e a quantidade de ESTs geradas estão listados na tabela 07.

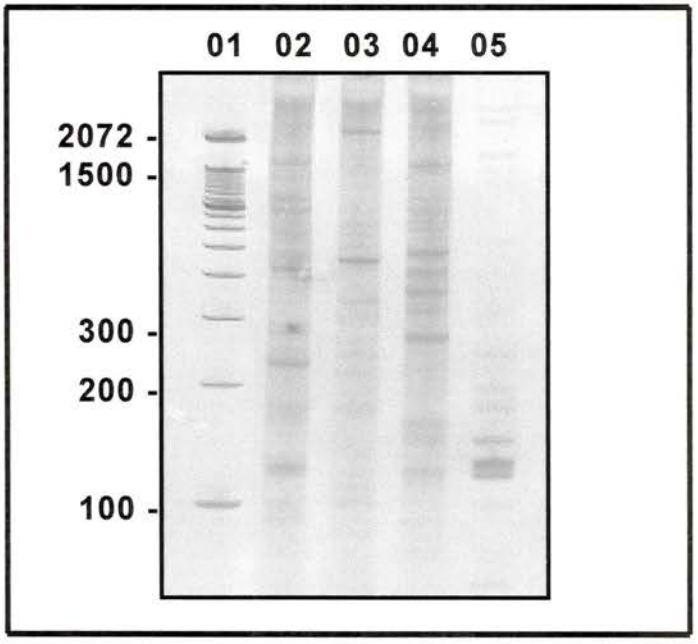


Figura 03: Primeiras mini-bibliotecas geradas usando a técnica de DDRT-PCR partindo de tumor de cólon.

Gel de poliacrilamida a 8% corado pela prata. Neste gel pode ser observado o perfil de amplificação de cDNAs para construção de mini bibliotecas geradas de acordo com a metodologia original. Canaletas: 1- 100bp ladder, 2- iniciadores $T11_{VT}$ /Ldd1, 3- iniciadores $T11_{VT}$ /Ldd2, 4- iniciadores $T11_{VT}$ /Ltk3 e 5- controle negativo da reação de PCR.

Tabela 07: ESTs geradas usando a técnica de DDRT-PCR, partindo de carcinoma de cólon:

Biblioteca	Iniciador 1	Iniciador 2	Nº de ESTs
01	T(11)VT	13 (Ldd1)	15
02	T(11)VT	14 (Ldd2)	35
03	T(11)VT	15 (Ltk3)	24
04	T(11)VC	Ldd1, Ldd2 e Ltk3	140
TOTAL			214

Iniciador 01: usado na síntese de cDNA; Iniciador 02 usado na amplificação dos fragmentos de cDNA (PCR). A biblioteca 04 foi construida com uma mistura dos produtos de PCR gerados com os iniciadores Ldd1, Ldd2 e Ltk3.

Estes primeiros resultados gerados a partir da estratégia originalmente pretendida foram importantes para a detecção de problemas técnicos, tais como: (i) baixo rendimento de ESTs provenientes da região *3' UTR* (*Untranslated Region*) e (ii) a grande maioria das seqüências produzidas foram de pequeno tamanho (<100pb).

Embora os resultados não fossem satisfatórios, procuramos fazer algumas análises. Mesmo sabendo que menos de 10% das ESTs apresentavam sinal de poliadenilação, selecionamos 19 ESTs não redundantes, que apresentaram identidade com genes totalmente seqüênciados, para fazer uma análise posicional (Tabela 08). O resultado desta análise pode ser observado na figura 04.

Tabela 08: Análise da posição média do alinhamento das primeiras ESTs geradas a partir de carcinoma de cólon:

Identidade da EST	Tamanho da EST	Tamanho do gene	Ponto médio	Posição (%) no gene	Gene
VC-pool-24	116	940	572	60	H.sapiens mRNA for olfactory receptor X87825.1
VC-pool-32	191	6455	955	14	RBP2 Retinoblastoma biding protein 2
VC-pool-38	61	323	291	90	Keratinocyte cDNA
VC-pool-58	164	672	472	70	mRNA for 23 KD highly base protein
VC-pool-66	67	7041	5328	75	Human acetyl-COA carboxylase mRNA
VC-pool-84	45	4782	2831	59	Human leucine-rich protein mRNA
VT 13/2	153	1331	1249	93	H sapiens Aflatoxin aldehyde redutase
VT 13/8	96	2228	2128	98	HS JM1 protein
VT 13/11	213	1466	658	44	H sapiens XPGC gene
VT 14/3	116	1922	1056	86	Metallothionein-IG (MT 1G) gene
VT 14/22	161	1552	1473	94	PMIT Isozyme II
VT 14/33	146	3971	3909	98	Laminin-5beta3-chain
VT 15/2	70	3110	3084	99	Human transcription factor ETR103
VT 15/4	42	1757	1365	77	3-hydroxy-3-methylglutaryl-coenzyme A
VC-pool-111	41	10942	10580	96	Human nucleolin gene
VC-pool-113	26	1349	417	30	H. sapiens mRNA for uridine phosphorylase
VC-pool-116	23	5102	2758	54	Human dioxin-inducible cytochrome β 450
VC-pool-139	69	4782	2837	59	Human leucine-rich protein mRNA
VC-pool-140	41	10942	10579	96	Human nucleolin gene

Nesta tabela estão listadas 19 ESTs não redundantes geradas a partir de carcinoma de cólon, tamanho médio 96 nt. Estas ESTs apresentaram pelo menos de 80% de identidade com genes humanos totalmente sequenciados com o valor de $e \leq 10^{-15}$. Na coluna "Ponto médio" foi colocado o resultado da média aritmética do ponto inicial e final do alinhamento da EST, este resultado dividido pelo tamanho do gene informa a posição relativa da EST no gene. Desta forma, se o tamanho do gene é 100%, podemos considerar que quanto maior o valor descrito nesta coluna mais no final (3') do gene a EST estará posicionada.

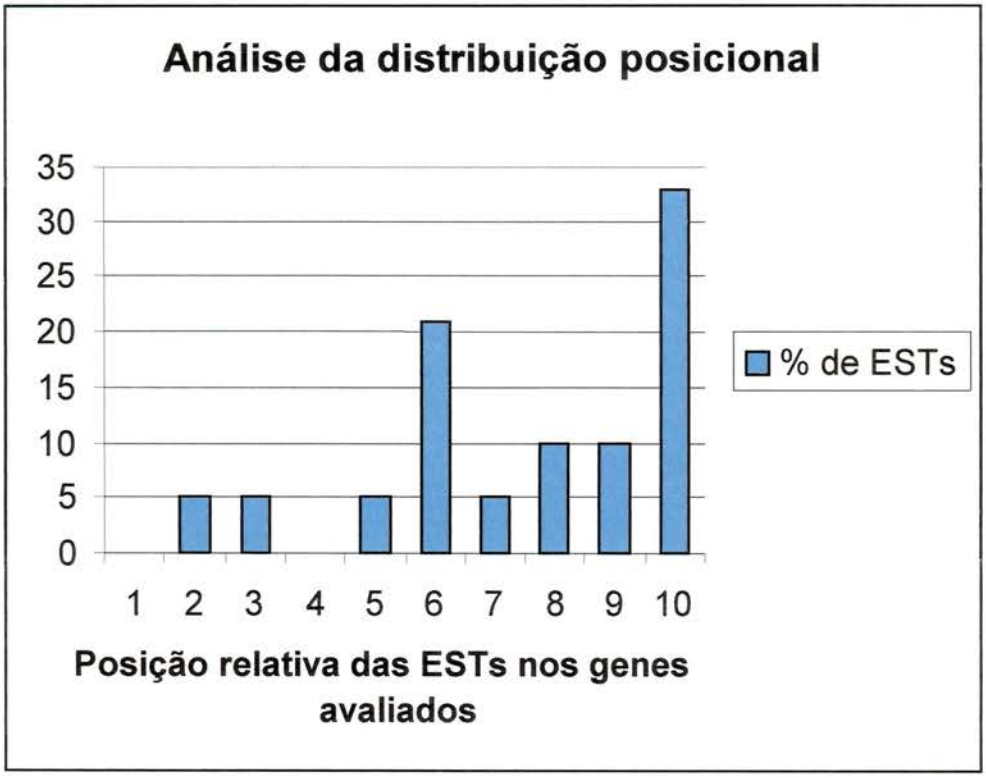


Figura 04: Análise da distribuição posicional das primeiras ESTs geradas a partir de carcinoma de cólon:

Análise posicional de 19 ESTs geradas a partir de carcinoma de cólon. No gráfico está representada a posição relativa das ESTs de acordo com o alinhamento ao longo dos genes (eixo do X). Para os cálculos desta análise, consideramos o tamanho total do gene equivalendo a 100%. Desta forma o gene foi fracionado em 10 segmentos, cada um equivalendo a 10%, sendo o primeiro segmento relativo a 10% da porção 5' e o último segmento relativo a 10% da porção 3'.

4.2 Padronização:

Na tentativa de elevar a estringência para evitar sítios internos de pareamento no momento da síntese do cDNA, foram desenhados novos iniciadores para substituir os sugeridos por LIANG e PARDEE (1992), os quais passaram a ter a seguinte forma: $T(25)_{CG}$. Com o aumento do número de T (timinas) para 25, a temperatura de pareamento podia ser elevada para no mínimo 50°C, favorecendo o pareamento específico do iniciador na cauda poli-A dos mRNAs, no momento da síntese de cDNA. Também com o objetivo de melhorar o rendimento das ESTs geradas a partir da região *3´ UTR*, o RNA foi extraído através do *Quick Prep Micro mRNA Purification kit* (Pharmacia). Este kit utiliza uma resina contendo poli-T e purifica apenas mRNA contendo a cauda poli A, o que facilita o pareamento dos iniciadores do tipo $T(25)_{CG}$ na região *3´UTR* do transcrito. Assim, foi suprimido o uso do iniciador arbitrário na PCR.

Através destas modificações, bem como a utilização do iniciador do tipo $T(25)_{CG}$ também na etapa de amplificação dos fragmentos de cDNA, foram produzidas 02 mini-bibliotecas também partindo de carcinoma de cólon (mini bibliotecas 05 e 06). As temperaturas de pareamento usadas na reação de PCR foram de: 50°C para a mini biblioteca 05 (figura 05) e 55°C para a mini biblioteca 06 (dados não mostrados). Os produtos de PCR usados para a construção destas mini bibliotecas partiram da diluição do cDNA na fração de 1/100, uma vez que o controle negativo da reação de síntese do cDNA (sem enzima transcriptase reversa), nesta mesma proporção, apresentava um índice menor de contaminação (figura 05). Apesar dos controles negativos apresentarem provável contaminação por DNA genômico, foram geradas e analisadas 314 ESTs (tab 09).

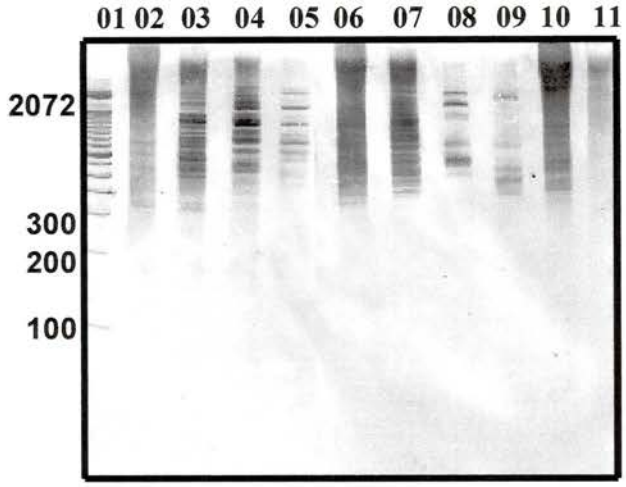


Figura 05: Perfil de amplificação de cDNAs de tumor de cólon com iniciador $T25_{GC}$:

Gel de poliacrilamida a 8% corado pela prata. Nesta figura pode ser observado o perfil de amplificação de cDNAs de carcinoma de cólon, sintetizados com o mRNA extraído pelo *Quick Prep Micro mRNA Purification kit* (Pharmacia), para a construção de mini-bibliotecas. Canaleta 1, padrão de peso molecular (100bp *Ladder*). Nas canaletas 2, 3, 4 e 5, perfil de amplificação partindo de diluições de cDNA nas seguintes proporções: 1µl, 1/10, 1/100 e 1/1000, para a construção da mini biblioteca 05. Nas canaletas 06, 07 08 e 09 estão os controles negativos da reação de síntese de cDNA (sem transcriptase reversa), usando as mesmas proporções de diluição que foram usadas para o cDNA. Canaleta 10 contém o produto de PCR gerado com 50ng de DNA humano e na canaleta 11 está o controle negativo da reação de PCR.

Tabela 09: ESTs geradas após as primeiras modificações na metodologia, partindo de carcinoma de cólon:

Biblioteca	N° de ESTs	Sinal de poli-A	*3´UTR*	Identidade ≥ 95%	Elementos repetitivos
05 (A)	95	34 (35%)	64 (67%)	44 (46%)	40.43%
05 (B)*	84	22 (26%)	63 (75%)	30 (35%)	35.42%
06 (A)	81	25 (30%)	53 (65%)	38 (46%)	26.28%
06 (B)*	54	30 (37%)	39 (72%)	49 (90%)	10.16%
total	314	111 (35%)	219 (69%)	161 (51%)	-

*As mini-bibliotecas foram divididas para fazer a análise usando o programa *RepeatMasker* (http://ftp.genome.washington. edu/RM/webhelp.html). Uma vez que este programa não consegue analizar, de uma só vez, arquivos contendo grandes volumes de dados.

Biblioteca: denominação da mini-biblioteca gerada; **N° de ESTs:** quantidade de ESTs analisadas; sendo o tamanho da maioria das ESTs geradas de cerca de 200nt; **Sinal de poli A:** quantidade de ESTs apresentando sinal de poliadenilação; ***3´UTR*:** quantidade de ESTs apresentando identidade, no início ou com uma diferença de até 50nt do início, com ESTs depositadas no banco de dados dbEST geradas a partir da região *3´UTR*.; **Identidade:** quantidade de ESTs apresentando identidade igual ou superior a 95% com ESTs depositadas no banco de dados dbEST, **Elementos repetitivos:** porcentagem de elementos repetitivos segundo análise usando o programa *RepeatMasker*.

Visando o monitoramento da contaminação proveniente de DNA genômico, foi avaliada a presença de elementos repetitivos utilizando o programa *RepeatMasker* (http://ftp.genome. washington. Edu/RM/webhelp. html). Como estes elementos repetitivos são encontrados com uma freqüência maior no DNA genômico que no cDNA, uma porcentagem acima de 10% provavelmente significa contaminação. Para executar esta análise, foram excluidas as seqüências dos iniciadores para evitar que estes fossem somados ao total de nucleotídeos mascarados.

Durante o seqüenciamento da biblioteca 06, outra modificação foi introduzida com o objetivo de automatizar e minimizar o tempo e o custo no processo de seqüenciamento dos clones. Esta modificação consistiu em gerar seqüências a partir de produto de PCR das colônias individualizadas, eliminando a etapa de crescimento e purificação dos plasmídeos.

Para verificar a complementaridade ao longo dos oligonucleotídeos $T25_{GC}$ nas reações de síntese de cDNA (RT) e de amplificação dos fragmentos (PCR), foi feito um estudo em um lote de 58 ESTs selecionadas das mini-bibliotecas 05 e 06 (fig 07), tendo como referência a EST depositada no dbEST. O critério para a seleção dos clones informativos para este tipo de análise foi baseado nas nossas ESTs que apresentaram: (i) a seqüência dos iniciadores nas duas extremidades e (ii) identidade maior ou igual a 95% com a EST depositada no dbEST; conforme "Ficha de análise da submissão da seqüência" (figura 06).

Ficha de análise da submissão da seqüência

CLONE - CO21
Biblioteca – 05; colônia 39
Construção - pUC-18, JM109
Data do seqüenciamento - 1998-09
Seqüenciado com o iniciador - M-13F
Seqüenciado no equipamento - ALF*express*-ILPC
Sequencia de nucleotídeos -

```
>CO39
TTTTTTTTTTTTTTTTTTTTTTTTGCCCTTGCACCAGAGTATGAAATAGCTTCCAGGAGCTCCAGCTATAAGCTTGGAAGTGTC
TGTGTGATTGTAATCACATGGTGACAACACTCAGAATCTAAATTGGACTTCTGTTGTATTCTCACCACTCAATTTGTTTTTTAGC
AGTTTAATGGGTACATTTTAGAGTCTTCCATTTTGTTGGRAATTAGATCCTCCCCTTCAAATGCTGTAATTAACAACACTTAAAA
AACTTGAATAAAATATTGAAACCGCAAAAAAAAAAAAAAAAAAAAAAAA
```

Tamanho - 305nt
Ambiguidades - 0
Data da análise - Setembro/1998.
Resultados das submissões:

Blast (dbEST) -
gb|AA828506|AA828506 oc46g01.s1 NCI_CGAP_GCB1 Homo sapiens cDNA clone IMAGE:1352784 3' similar to gb:L06132 OUTER MITOCHONDRIAL MEMBRANE PROTEIN PORIN (HUMAN); Length = 513 Score = 524 bits (264), Expect = e-147 **Identities = 274/276 (99%)**, Positives = 274/276 (99%), Gaps = 1/276 (0%)

```
Query: 16  ttttttttttgcccttgcaccagagtatgaaatagcttccaggagctccagctataagct 75
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 275 ttttttttttgcccttgcaccagagtatgaaatagcttccaggagctccagctataagct 216

Query: 76  tggaagtgtctgtgtgattgtaatcacatggtgacaacactcagaatctaaattggactt 135
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 215 tggaagtgtctgtgtgattgtaatcacatggtgacaacactcagaatctaaattggactt 156

Query: 136 ctgttgtattctcaccactcaatttgtttttttagcagtttaatgggtacattttagagtc 195
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 155 ctgttgtattctcaccactcaatttgtttttttagcagtttaatgggtacattttagagtc 96

Query: 196 ttccattttgttggraattagatcctcccccttcaaatgctgtaattaacaacacttaaaa 255
           |||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 95  ttccattttgttgg-aattagatcctcccccttcaaatgctgtaattaacaacacttaaaa 37

Query: 256 aacttgaataaaatattgaaaccgcaaaaaaaaaaa 291
           |||||||||||||||||||||| ||||||||||||
Sbjct: 36  aacttgaataaaatattgaaacctcaaaaaaaaaaa 1
```

EST original, depositada no dbEST:

```
LOCUS       AA828506      513 bp    mRNA            EST       07-APR-1998
DEFINITION  oc46g01.s1 NCI_CGAP_GCB1 Homo sapiens cDNA clone IMAGE:1352784 3'
            similar to gb:L06132 OUTER MITOCHONDRIAL MEMBRANE PROTEIN PORIN
            (HUMAN);, mRNA sequence.
ACCESSION   AA828506
FEATURES    cDNA synthesis was primed with a Not I - oligo(dT) primer
            [5'-TGTTACCAATCTGAAGTGGGAGCGGCCGCCTCATTTTTTTTTTTTTTTTT-3'].
            Double-stranded cDNA was ligated to Eco RI adaptors (Pharmacia), digested
            with Not I and cloned into the Not I and Eco RI sites of the modified
pT7T3
            vector. Library went through one round of normalization, and was
            constructed  by Bento Soares and M. Fatima Bonaldo."/db_xref="taxon:9606"
            /clone="IMAGE:1352784"/clone_lib="NCI_CGAP_GCB1"/tissue_type="germinal
            center B cell"/lab_host="DH10B"

            BASE COUNT      176 a    100 c    100 g    137 t  ORIGIN
          1 ttttttttt tgaggtttca atattttatt caagtttttt aagtgttgtt aattacagca
         61 tttgaagggg aggatctaat tccaacaaaa tggaagactc taaaatgtac ccattaaact
        121 gctaaaaaac aaattgagtg gtgagaatac aacagaagtc caatttagat tctgagtgtt
        181 gtcaccatgt gattacaatc acacagacac ttccaagctt atagctggag ctcctggaag
        241 ctatttcata ctctggtgca agggcaaaaa aaaaacacaa cacaagaagg aataagtcct
```

```
      301 gaattattgg cttcatcaca tccaccttct ccaccccaaa atggcacaaa agaaacagtt
      361 accacaccct gcagaccttt tggtgtaaaa gagatgatga tgaactgggg tgggaacagg
      421 tcatgaagat ctgtccaaaa aagtcccatt caggtgagtt tgtacacacc atcaagcagc
      481 gagcctctca tcaattaggg ttagggaacc aag

Blast (nr) -
gb|L06132|HUMVDAC1X Human voltage-dependent anion channel isoform 1 (VDAC) mRNA,
complete cds.Length = 1806 Score =  499 bits (251), Expect = e-139 Identities =
261/263 (99%), Positives = 261/263 (99%), Gaps = 1/263 (0%)
Query: 16   ttttttttttgcccttgcaccagagtatgaaatagcttccaggagctccagctataagct 75
            ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1543 ttttttttggcccttgcaccagagtatgaaatagcttccaggagctccagctataagct 1602

Query: 76   tggaagtgtctgtgtgattgtaatcacatggtgacaacactcagaatctaaattggactt 135
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1603 tggaagtgtctgtgtgattgtaatcacatggtgacaacactcagaatctaaattggactt 1662

Query: 136  ctgttgtattctcaccactcaatttgtttttagcagtttaatgggtacattttagagtc 195
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1663 ctgttgtattctcaccactcaatttgtttttagcagtttaatgggtacattttagagtc 1722

Query: 196  ttccattttgttggraattagatcctccccttcaaatgctgtaattaacaacacttaaaa 255
            |||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1723 ttccattttgttgg-aattagatcctccccttcaaatgctgtaattaacaacacttaaaa 1781

Query: 256  aacttgaataaaatattgaaacc 278
            ||||||||||||||||||||||
Sbjct: 1782 aacttgaataaaatattgaaacc 1804

BlastX -
pir||S69466 hypothetical protein YPR142c - yeast (Saccharomyces cerevisiae)
>gi|2347172 (U40829) Ypr142cp [Saccharomyces cerevisiae] Length = 187 Score = 30.1
bits (66), Expect = 4.1 Identities = 23/77 (29%), Positives = 39/77 (49%), Gaps =
3/77 (3%)

Query: 5    FFFFFFFALAPEYEIASRSSSYKLGSVCVIVITW*QHSESKLDFCC---ILTTQFVF*QF 175
             FF FFF   +     +S SSS  L S+ +I+       +S L+  C     ++  +   F
Sbjct: 33   FFDFFFLGKSSSSSSSSSSSSASLCSLSIIL------DDSSLELFCSSSSASSPSIIVSF 86

Query: 176  NGYILESSILLXIRSSPSNA 235
             +G +L S + L + S P++A
Sbjct: 87   SGSLLNSWLPLFLFSRPNSA 106
```

Figura 06: Exemplo de EST selecionada para estudo da complementaridade ao longo dos oligonucleotídeos $T25_{GC}$ nas reações de síntese de cDNA (RT) e amplificação dos fragmentos (PCR), tendo como referência a EST depositada no dbEST.

Os iniciadores (grifados) podem ser observados nas duas extremidades da seqüência usada para submissão e a porcentagem de identidade ≥ a 95%, pode ser observada na análise contra o banco de dados dbEST (negrito). O provável sinal de poliadenilação (grifado TTTATT) pode indicar que esta seqüência foi gerada a partir da região 3' do transcrito. Outros prováveis indicativos disso são, a seqüência ter apresentado identidade com o início da EST depositada no dbEST, uma vez que a biblioteca que gerou a EST original foi construida com iniciadores do tipo Oligo dT e a baixa identidade apresentada contra o banco de dados de proteínas (Blast (X))

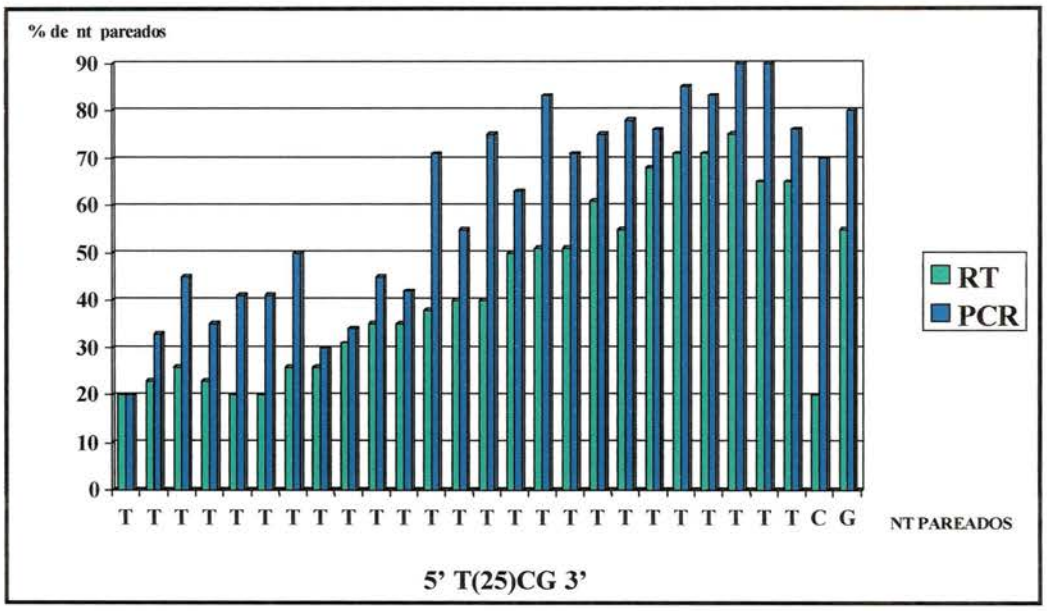


Figura 07: Gráfico representando a freqüência do pareamento dos nucleotídeos ao longo do iniciador $T25_{GC}$, na síntese de cDNA (RT) e na amplificação dos fragmentos (PCR).

No eixo das abscissas estão representados os 27 nucleotídeos do iniciador $T25_{GC}$ e no eixo das ordenadas a porcentagem de nucleotídeos pareados em cada uma das posições do iniciador. A predição do pareamento (*match*) foi feita com base no alinhamento de 60 ESTs geradas contra seqüências depositadas no dbEST.

Como era de se esperar, há um melhor nível de complementaridade entre os iniciadores e o DNA molde amplificados na porção 3', dado indicado pela quantidade de ESTs com sinal de poli A ou apresentando identidade com ESTs derivadas da porção 3' depositadas no dbEST. No entanto, foi possível observar uma alta taxa de *mismatches* nos dois últimos nucleotídeos da posição 3' do iniciador usado ($T25_{CG}$), durante a síntese de cDNA. Uma vez que, em um total de 60 ESTs analisadas, apenas 55% apresentaram *match* na última posição do iniciador durante a síntese de cDNA, contra 80% que apresentaram este mesmo *match* durante a PCR.

Para avaliar o grau de normalização das mini bibliotecas produzidas nesta etapa, foram selecionadas ESTs que apresentaram: (i) mais de 90% de identidade contra o banco de dados dbEST, (ii) pelo menos

um dos iniciadores presentes na seqüência obtida, (iii) sítio de poliadenilação, dos tipos TTTATT ou TTTAAT, presentes nos primeiros 40 nucleotídeos a partir da cauda poli A. Desta forma foram selecionadas 73 ESTs (não redundantes) provenientes da região 3' do transcrito. Com estas ESTs selecionadas foi feito um estudo buscando *clusters* de ESTs no banco UniGene (fig 08), onde o valor médio de entradas foi 355. Para comparar os resultados do estudo de normalização, foram colocados resultados de bibliotecas construídas a partir de tumor de mama utilizando outras estratégias, as quais são (i) Não Normalizada e (ii), Normalizada onde o número médio de entradas foi respectivamente 184 e 119.

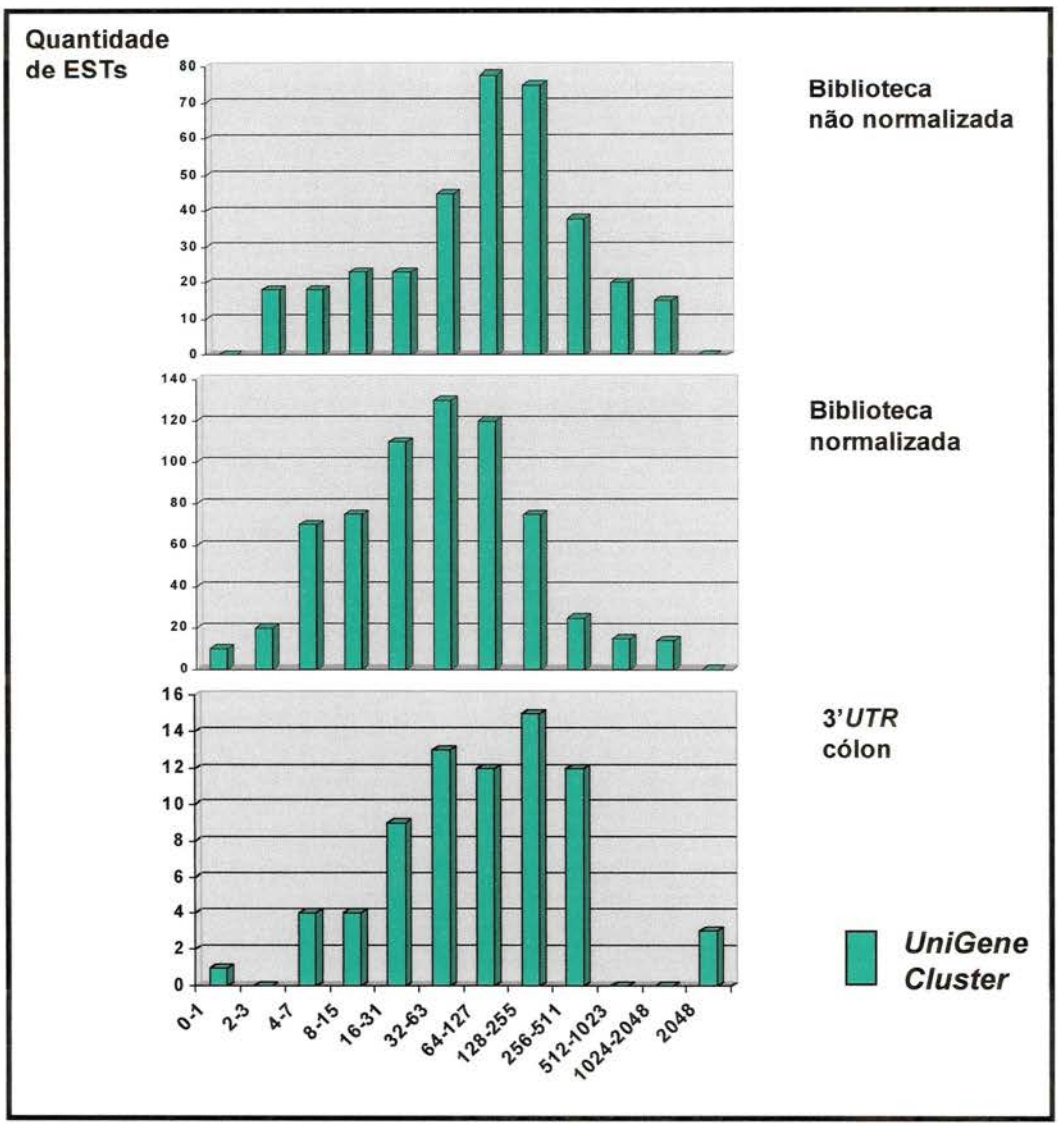


Figura 08: Gráfico representando o estudo comparativo da normalização em diferentes bibliotecas de cDNA.

Análise de abundância das ESTs por comparação com *clusters* do UniGene. Este gráfico representa o estudo da normalização de diferentes bibliotecas de cDNA, a saber: (i) não normalizada NCI CGAP Br1.1, (ii) normalizada NCI CGAP Br2 e (iii) 3' *UTR* cólon, o número médio de entradas nos clusters do UniGene foi respectivamente de 184, 119 e 355. No eixo das ordenadas esta representada a quantidade de ESTs estudada e no eixo das abscissas estão representados os diferentes *clusters* do UniGene.

4.3 Otimização dos protocolos de produção de mini bibliotecas e protocolos de seqüenciamento:

Nossos dados experimentais sugeriam problemas, tais como: (i) persistente contaminação por DNA genômico, (ii) promiscuidade no pareamento dos iniciadores do tipo $T(25)_{CG}$ tanto na síntese quanto na amplificação dos fragmentos de cDNA, além de baixa eficiência na reprodutibilidade das reações e (iii) dificuldades na etapa de seqüenciamento dos clones, devido aos iniciadores conterem uma região repetitiva do nucleotídeo timina. Estes iniciadores provocam um erro denominado *"slippage"*, que é um erro de amplificação em regiões repetitivas do DNA, devido à dificuldade da enzima *Taq* polimerase em amplificar estas regiões com fidelidade. Por causa destes problemas, novas alterações foram feitas:

1- Como o rendimento do kit usado para extração de mRNA *QuickPrep™* (Pharmacia) é baixo, o tratamento com DNase neste material se torna inviável. Desta forma foi feita a extração do RNA com o reagente *Trizol*, seguido de tratamento com DNase por 60 minutos e posterior purificação do mRNA poli-A através do *kit Macs Oligo(dt) Microbeads*.

2- Devido à falha na normalização parcial das mini-bibliotecas (figura 06:), provavelmente causada pela promiscuidade no pareamento dos iniciadores na síntese de cDNA e a baixa quantidade de ESTs geradas na porção *3' UTR*, foi feita a síntese do cDNA utilizando Oligo dT (20) e uma nova enzima transcritase reversa, a *ThermoScript* (Gibco-BRL), que é capaz de sintetizar o cDNA usando temperaturas mais elevadas (até 70°C).

3- Para fazer a amplificação dos fragmentos de cDNA através da reação de PCR, foi sugerida a volta dos iniciadores do tipo $T11_{VN}$ (LIANG e PARDEE 1992). O uso destes iniciadores menores tem o objetivo de melhorar a eficiência no funcionamento desta etapa, além de minimizar o problema de *"slippage"* que ocorre durante a geração de DNA molde para a etapa de seqüenciamento.

De acordo com as novas modificações, foram geradas 03 mini-bibliotecas a partir de carcinoma de mama, mini biblotecas 07, 08 e 09 (figura 09, tabela 10). O estudo deste tumor não foi previsto no projeto, porém, neste momento o mRNA de mama era o único aprovado por todos os controles de qualidade, não apresentando contaminação por DNA genômico detectável. O uso desse mRNA também tornava mais adequada a análise comparativa de normalização, uma vez que as bibliotecas usadas para esta análise são derivadas de tumor de mama.

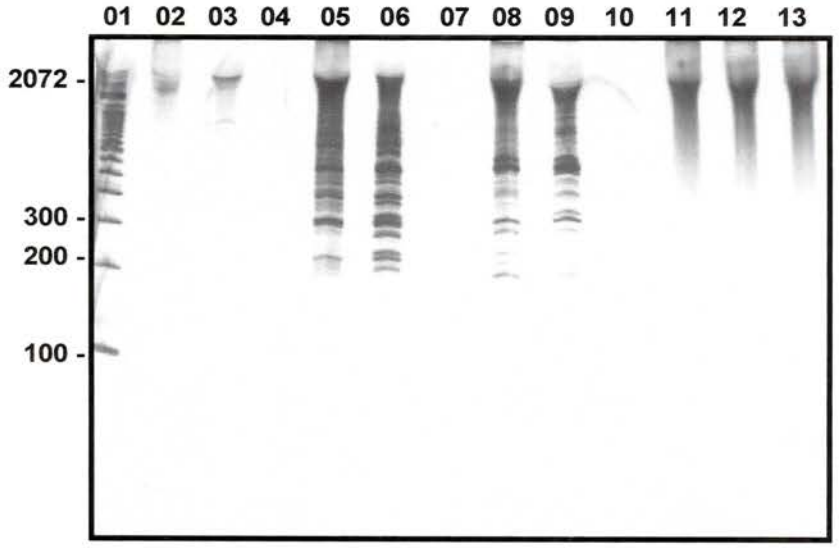


Figura 09: Perfil de amplificação de cDNAs para a geração de mini bibliotecas de carcinoma de mama.

Gel de poliacrilamida a 8% corado pela prata, onde estão representados os perfis de amplificação de cDNAs de carcinoma de mama. Canaleta 1 padrão de peso molecular (100bp *ladder*). Canaletas 02 e 03, perfil de amplificação de cDNAs a partir do iniciador $T11_{VA}$, na diluição de cDNA nas proporções: 1/10, 1/100, respectivamente, canaleta 04 controle negativo da reação de PCR. Seguindo o mesmo padrão de organização, pode ser observado o perfil de amplificação de cDNAs com o iniciador $T11_{VG}$, canaletas 05, 06 e 07; com o iniciador $T11_{VC}$, canaletas 08, 09 e 10; e com o iniciador $T11_{VT}$, canaletas 11, 12 e 13.

As mini-bibliotecas, produzidas a partir de mRNA de carcinoma de mama, geraram os seguintes resultados (Tabela 10):

Tabela 10: ESTs geradas a partir de tumor de mama:

Biblio	Iniciador	CLASS 01	CLASS 02	CLASS 03	CLASS 04	CLASS 05	CLASS 06	CLASS 07	CLASS 08	CLASS 09	CLASS 10
07	T(11)VA	50	50	16	31	33	34	28	06	01	00
08	T(11)VC	30	19	01	10	13	16	15	01	01	00
09	T(11)VG	30	11	0	06	06	06	06	0	0	01
	TOTAL	110	80 (72,7%)	17 (21,25)	47 (74%)	52 (82,5%)	56 (88%)	49 (87%)	07 (14%)	02 (3,57)	01 (1,7%)

BIBLIO: denominação da mini-biblioteca gerada; **Iniciador**: iniciador usado para a amplificação dos fragmentos de cDNA; **CLASS 01:** quantidade de clones seqüenciados; **CLASS 02:** quantidade de seqüências de boa qualidade para a análise; **CLASS 03**: quantidade de ESTs redundantes; **CLASS 04:** quantidade de ESTs apresentando sinal de poliadenilação; **CLASS 05:** quantidade de ESTs apresentando identidade, no início ou com uma diferença de até 50nt do início, contra ESTs depositadas no banco de dados dbEST geradas a partir da região *3'UTR*; **CLASS 06:** ESTs analisadas contra o UniGene; **CLASS 07:** quantidade de genes representados; **CLASS 08:** genes representados mais de uma vez na referida mini-biblioteca; **CLASS 09:** quantidade de ESTs apresentando identidade contra ESTs ainda não agrupadas em *clusters* do UniGene; **CLASS 10:** ESTs que não apresentaram identidade com seqüências nos bancos de dados avaliados.

Devido à baixa qualidade das seqüências, resultando em um baixo aproveitamento dos clones seqüenciados, avaliamos o uso da enzima *Tth* (*rTth DNA Polymerase kit GeneAmp XL PCR* – Perkin-Elmer) para reação de PCR dos clones, evitando assim o problema de *"slippage"* nos produtos usados como molde para o seqüenciamento, proporcionando um melhor rendimento e confiabilidade nesta etapa, que pode ser observado nos cromatogramas das seqüências obtidas (figuras 10A e 10B). Desta forma, foi feito o seqüenciamento dos clones partindo de mini bibliotecas de carcinoma de cólon (Tabela 11).

Figura 10A: Cromatograma obtido a partir de DNA molde produzido pela *Taq* DNA polimerase:

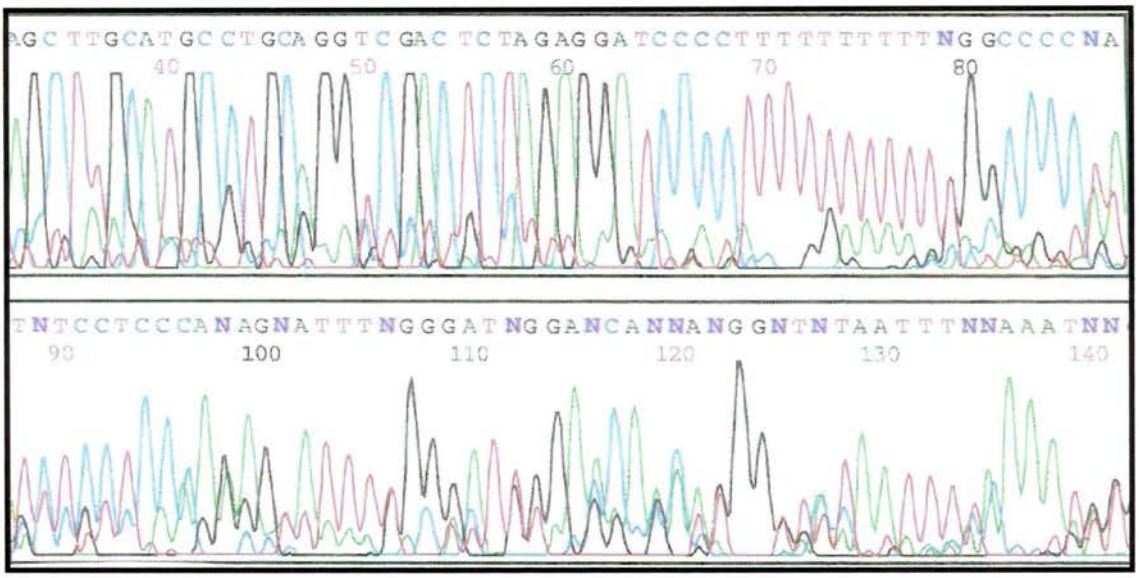
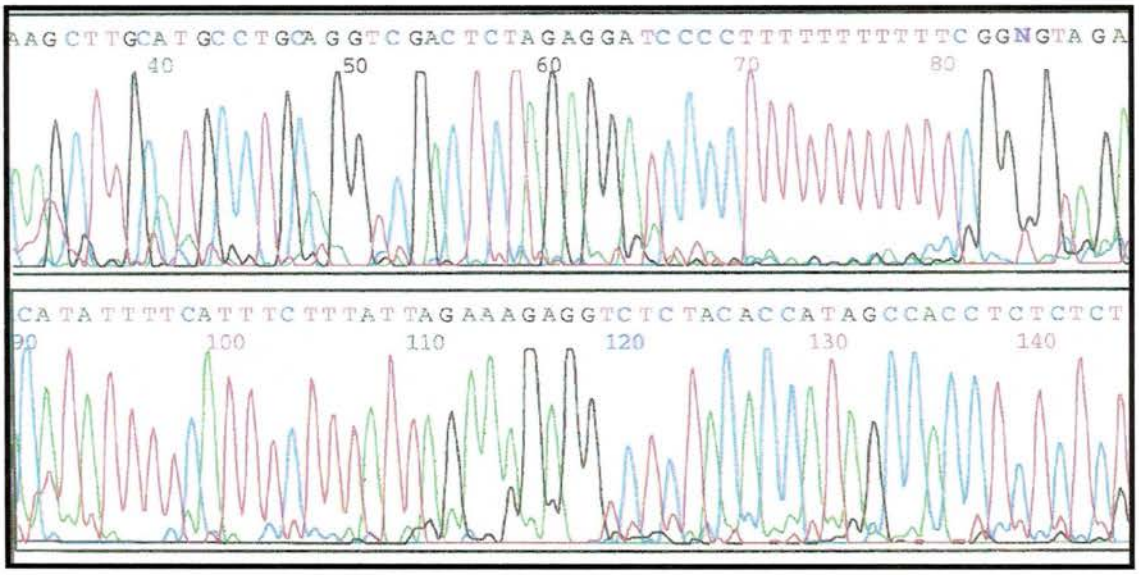


Figura 10B: Cromatograma obtido a partir de DNA molde produzido pela *Tth* DNA polimerase:



Nas figuras 10A e 10B, estão representados respectivamente, os cromatogramas de seqüências obtidas a partir de DNA molde produzido por PCR usando a *Taq* polimerase comun e a *Tth* polimerase, onde a diferença de qualidade de seqüência pode ser notada pela quantidade de ambigüidades (N). Na figura 10A o resultado do seqüenciamento é comprometido após o término da seqüência do iniciador, o que não ocorre quando o DNA molde é gerado pela enzima *Tth*, figura 10B.

Tabela 11: ESTs geradas a partir de tumor de cólon:

BIBLIO	Iniciador	CLASS 01	CLASS 02	CLASS 03	CLASS 04	CLASS 05	CLASS 06	CLASS 07	CLASS 08	CLASS 09	CLASS 10
10	T(11)VA	100	93	55	19	26	25	19	06	04	00
11	T(11)VT	65	23	07	11	11	12	07	05	02	00
12	T(11)VC	111	106	28	38	48	57	49	08	08	02
13	T(11)VG	100	51	04	17	33	36	17	19	02	00
	TOTAL	376	273 (72,6%)	94 (34,4%)	85 (47%)	118 (65,9%)	130 (72%)	92 (70,7%)	38 (29,2%)	16 (12,3%)	02 (1,5%)

BIBLIO: denominação da mini-biblioteca gerada; **Iniciador**: iniciador usado para a amplificação dos fragmentos de cDNA; **CLASS 01:** quantidade de clones seqüenciados; **CLASS 02:** quantidade de seqüências de boa qualidade para a análise; **CLASS 03**: quantidade de ESTs redundantes; **CLASS 04:** quantidade de ESTs apresentando sinal de poliadenilação; **CLASS 05:** quantidade de ESTs apresentando identidade, no início ou com uma diferença de até 50nt do início, contra ESTs depositadas no banco de dados dbEST geradas a partir da região *3'UTR*; **CLASS 06:** ESTs analisadas contra o UniGene; **CLASS 07:** quantidade de genes representados; **CLASS 08:** genes representados mais de uma vez na referida mini-biblioteca; **CLASS 09:** quantidade de ESTs apresentando identidade contra ESTs ainda não agrupadas em *clusters* do UniGene; **CLASS 10:** ESTs que não apresentaram identidade com seqüências nos bancos de dados avaliados.

De acordo com os resultados obtidos com mRNA derivado de tumores de cólon e mama, as alterações na metodologia foram positivas, devido: (i) a eliminação da contaminação do mRNA com DNA, segundo os controles de qualidade desenvolvidos (dados não mostrados); (ii) a combinação da enzima transcriptase reversa *ThermoScript* (Gibco-BRL) e os iniciadores Oligo dT (20) proporcionando um aumento da estringência na reação de síntese de cDNA fazendo com que o alvo do iniciador fosse a cauda poli A e não contínuos de adenina ao longo do transcrito, aumentando a porcentagem de ESTs com sinal de poliadenilação (54%); (iii) ao uso dos iniciadores do tipo $T11_{VN}$ na reação de amplificação dos fragmentos de cDNA para a construção das mini bibliotecas proporcionando uma melhora na eficiência do funcionamento desta etapa; e (iv) a combinação dos iniciadores $T11_{VN}$ na construção das bibliotecas e a enzima *Tth* (*rTth DNA Polymerase kit GeneAmp XL PCR* – Perkin-Elmer) para geração de DNA molde para o seqüenciamento, melhorando a qualidade das seqüências geradas (figuras 10A e 10B).

4.4 Iniciando o trabalho com a coleção de tumores de baixa incidência:

Baseado nos bons resultados obtidos a partir das bibliotecas de carcinoma de cólon e carcinoma de mama, foi feita a extração de mRNA dos tumores de baixa incidência. Segundo o controle de GAPDH, dentre os tumores listados no projeto inicial, apenas o mRNA derivado de PNET apresentou um baixo grau de degradação (fig. 11). Os outros controles do mRNA do tumor PNET (D-Loop e p53) não indicaram a presença de DNA (dados não mostrados). Sendo o mRNA extraído deste tumor de boa qualidade, foram construídas cinco mini bibliotecas (14,15,16,17 e 18), dois dos perfis de amplificação do cDNA usado para a construção de duas destas mini bibliotecas (15 e 17) podem ser visualizados na figura 12; as quais geraram resultados descritos na tabela 12.

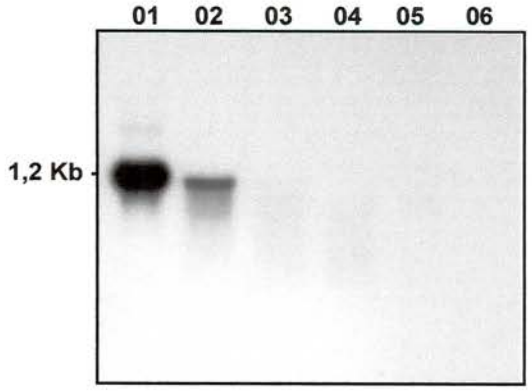


Figura 11: Ensaio de *"Northern Blot"* para controle de qualidade do RNA.

Autoradiograma de *"Northern-Blot"* onde 5μg de RNA total foram hibridados com sonda de cDNA de Gliceraldeído-3 fosfato desidrogenase humano marcado com αdCTP ^{32} (fragmento de 600pb clonado em pUC18). Nas canaletas 01, 02, 03, 04, 05 e 06 estão representados, respectivamente: controle positivo derivado de mRNA extraído de célula imortalizada cultivada (Fibroblasto humano GM 637), tumor neuroectodérmico primitivo (PNET), rabdomiossarcoma embrionário, tumor do seio endodérmico, neurofibrossarcoma e hepatoblastoma. Onde apenas o PNET apresentou baixo nível de degradação do mRNA.

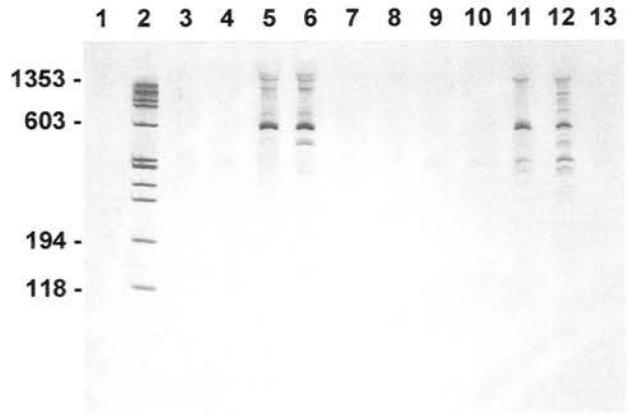


Figura 12: Perfil de amplificação de cDNAs do tumor PNET para a construção de mini bibliotecas.

Gel de poliacrilamida a 8% corado pela prata. Mini-bibliotecas geradas a partir de tumor PNET: Canaleta 02 padrão de peso molecular (100bp *ladder*). Nas canaletas 01 e 03 pode ser observado o fraco perfil de amplificação das mini-bibliotecas geradas partindo dos iniciadores $T11_{VA}$ respectivamente nas seguintes proporções de diluição do cDNA: 1/10, 1/100, na canaleta 04 controle negativo da reação de PCR. O mesmo ocorre com a mini-biblioteca $T11_{VT}$ canaletas 08, 09 e 10. Seguindo o mesmo padrão de organização, pode ser observado o perfil de amplificação do cDNA com os iniciadores $T11_{VG}$, canaletas 05, 06 e 07; e $T11_{VC}$, canaletas 11, 12 e 13.

Tabela 12: ESTs geradas a partir de PNET:

Biblio	Iniciador	CLASS 01	CLASS 02	CLASS 03	CLASS 04	CLASS 05	CLASS 06	CLASS 07	CLASS 08	CLASS 09	CLASS 10
14	T(11)VA	10	05	01	02	0	02	01	01	00	00
15	T(11)VG	96	84	25	19	26	30	21	09	00	09
16	T(11)VTC	30	20	04	08	04	10	07	03	01	00
17	T(11)VC	96	49	13	18	16	24	19	05	05	03
18	T(11)VTC /VC	106	90	50	26	29	32	25	07	04	01
	TOTAL	338	248 (70%)	93 (27%)	73 (47%)	75 (48%)	98 (63,2%)	73 (47%)	25 (16%)	10 (10,2%)	13 (13,2%)

BIBLIO: denominação da mini-biblioteca gerada; **Iniciador**: iniciador usado para a amplificação dos fragmentos de cDNA; **CLASS 01:** n° de clones seqüenciados; **CLASS 02:** quantidade de seqüências de boa qualidade para a análise; **CLASS 03**: n° de ESTs redundantes; **CLASS 04:** quantidade de ESTs apresentando sinal de poliadenilação; **CLASS 05:** quantidade de ESTs apresentando identidade, no início ou com uma diferença de até 50nt do início, contra ESTs depositadas no banco de dados dbEST geradas a partir da região *3'UTR*; **CLASS 06:** n° de ESTs analisadas contra o UniGene; **CLASS 07:** n° de genes representados; **CLASS 08:** genes representados mais de uma vez na referida mini-biblioteca; **CLASS 09:** quantidade de ESTs apresentando identidade contra ESTs ainda não agrupadas em *clusters* do UniGene; **CLASS 10:** ESTs que não apresentaram identidade com seqüências nos bancos de dados avaliados.

Tabela 13: ESTs geradas a partir de mini bibliotecas dos tumores de mama, cólon e PNET, construídas com Oligo dT (20) e analisadas manualmente:

Biblio	Iniciador	CLASS 01	CLASS 02	CLASS 03	CLASS 04	CLASS 05	CLASS 06	CLASS 07	CLASS 08	CLASS 09	CLASS 10
07	T(11)VA	50	50	16	31	33	34	28	06	01	00
08	T(11)VC	30	19	01	10	13	16	15	01	01	00
09	T(11)VG	30	11	0	06	06	06	06	0	0	01
Subtotal Mama	03	110	80 (72,7%)	17 (21,25)	47 (74%)	52 (82,5%)	56 (88%)	49 (87%)	07 (14%)	02 (3,57)	01 (1,7%)
10	T(11)VA	100	93	55	19	26	25	19	06	04	00
11	T(11)VT	65	23	07	11	11	12	07	05	02	00
12	T(11)VC	111	106	28	38	48	57	49	08	08	02
13	T(11)VG	100	51	04	17	33	36	17	19	02	00
Subtotal cólon	04	376	273 (72,6%)	94 (34,4%)	85 (47%)	118 (65,9%)	130 (72%)	92 (70,7%)	38 (29,2%)	16 (12,3%)	02 (1,5%)
14	T(11)VA	10	05	01	02	0	02	01	01	00	00
15	T(11)VG	96	84	25	19	26	30	21	09	00	09
16	T(11)VTC	30	20	04	08	04	10	07	03	01	00
17	T(11)VC	96	49	13	18	16	24	19	05	05	03
18	T(11)VTC /VC	106	90	50	26	29	32	25	07	04	01
Subtotal PNET	05	338	248 (70%)	93 (27%)	73 (47%)	75 (48%)	98 (63,2%)	73 (47%)	25 (16%)	10 (10,2%)	12 (13,2%)
TOTAL	12	824	601 (72,9%)	204 (33%)	205 (51%)	245 (61,7%)	284 (71,5%)	214 (75,3%)	70 (24,6)	28 (9,8%)	15 (5,2%)

CLASS 01: n° de clones seqüenciados; **Iniciador**: iniciador usado para a amplificação dos fragmentos de cDNA; **CLASS 02:** quantidade de seqüências de boa qualidade para a análise; **CLASS 03**: n° de ESTs redundantes; **CLASS 04:** quantidade de ESTs apresentando sinal de poliadenilação; **CLASS 05:** quantidade de ESTs apresentando identidade, no início ou com uma diferença de até 50nt do início, contra ESTs depositadas no banco de dados dbEST geradas a partir da região *3'UTR*; **CLASS 06:** n° de ESTs analisadas contra o UniGene; **CLASS 07:** n° de genes representados; **CLASS 08:** genes representados mais de uma vez na referida mini-biblioteca; **CLASS 09:** quantidade de ESTs apresentando identidade contra ESTs ainda não agrupadas em *clusters* do UniGene; **CLASS 10:** ESTs que não apresentaram identidade com seqüências nos bancos de dados avaliados.

A análise das ESTs geradas a partir das mini bibliotecas de carcinoma de mama, carcinoma de cólon e PNET, sugeriu que a nova estratégia para construção de mini bibliotecas parcialmente normalizadas para geração de ESTs provenientes preferencialmente da porção 3' do transcrito estava funcionando adequadamente. Uma nova análise do pareamento dos iniciadores, feita com os mesmos critérios usados anteriormente, demostrou um aumento na taxa de *matches* nos dois últimos nt da posição 3' do iniciador T (11) $_{VN}$ durante a PCR (Figura 13). Dentre 52 ESTs analisadas, quase 80% apresentaram *match* na última posição do iniciador no sentido sense e cerca de 98% apresentaram este mesmo *match* no sentido antisense, durante a PCR.

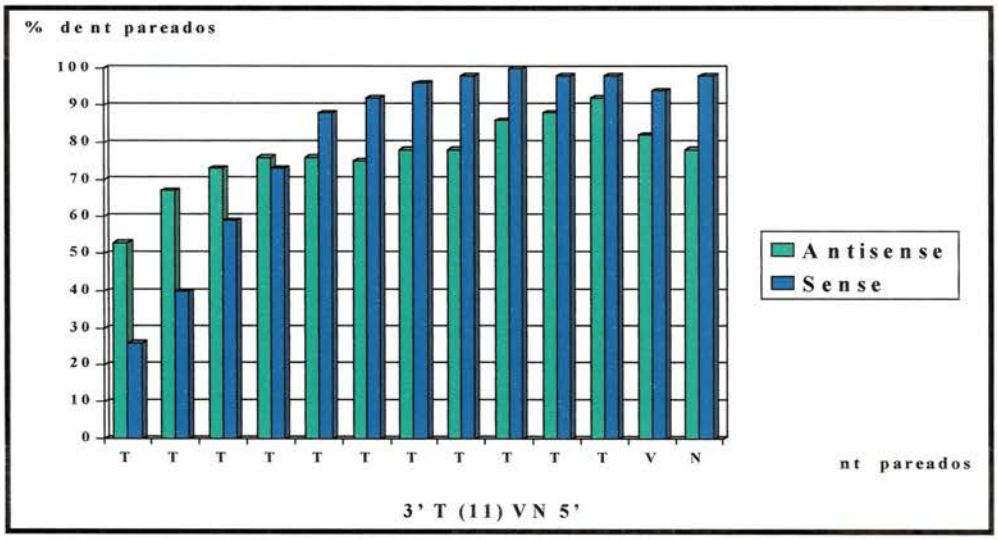


Figura 13: Gráfico representando a freqüência do pareamento dos nucleotídeos ao longo do iniciador T(11)$_{VN}$, na amplificação dos fragmentos (PCR).

No eixo das abscissas estão representados os 13 nt do iniciador T(11)$_{VN}$, e no eixo das ordenadas o percentual de nt pareados em cada uma das posições do iniciador. A predição do pareamento foi feita com base no alinhamento de 52 ESTs selecionadas contra seqüências depositadas no dbEST.

Provavelmente a diminuição da promiscuidade no pareamento dos nucleotídeos permitiu uma melhora na normalização das bibliotecas geradas, uma vez que grande parte das ESTs selecionadas se encaixavam em *clusters* de genes que apresentavam baixo ou médio nível de expressão (Tabela 14), sendo o valor médio destes *clusters* de 135 entradas, contra 240 (biblioteca não normalizada) e 194 (biblioteca normalizada) (Figura 14).

Tabela 14: Análise de normalização em função do número de entradas (ESTs) nos *clusters* do UniGene:

UniGene *Cluster*	Carcinoma de mama				Carcinoma de cólon				PNET					Total
mini bibliotecas	07 T(11)VA	08 T(11)VC	09 T(11)VG	T(11)VT	10 T(11)VA	12 T(11)VC	13 T(11)VG	11 T(11)VT	14 T(11)VA	16 T(11)VTC	17 T(11)VC	15 T(11)VG	18 T(11)VTC/VC	
01	01	00	00	-	01	01	00	01	00	01	01	00	00	06
2-3	01	00	01	-	00	00	01	00	01	00	01	00	00	05
4-7	02	01	00	-	01	06	01	00	01	02	01	02	00	17
8-15	00	06	00	-	02	04	02	02	00	05	00	01	02	24
16-31	03	01	01	-	03	04	03	00	00	01	02	03	00	21
32-63	08	04	02	-	03	08	01	01	01	03	04	05	00	40
64-127	04	01	00	-	02	02	03	00	00	03	01	04	01	21
128-255	01	01	02	-	02	10	02	01	01	02	03	02	01	28
256-511	06	00	01	-	01	05	02	00	01	01	00	01	00	18
512-1023	01	00	00	-	00	01	00	00	00	01	01	00	01	05
1024-2048	00	00	00	-	00	00	00	00	00	00	00	00	00	00
2048	00	00	00	-	00	00	01	00	00	01	00	00	00	02
Total	27	14	07		15	41	16	05	05	20	14	18	05	187

Para fazer a análise de normalização em função do número de entradas nos *clusters* do UniGene, foram selecionadas ESTs não redundantes, geradas com diferentes iniciadores dos diferentes tecidos estudados. A seleção levou em consideração a presença de sinal de poliadenilação e/ou identidade contra ESTs depositadas no banco de dados dbEST geradas a partir da região *3'UTR*. O valor médio de entradas dos diferentes tecidos foi: carcinoma de mama 115, carcinoma de cólon 135 e PNET 144.

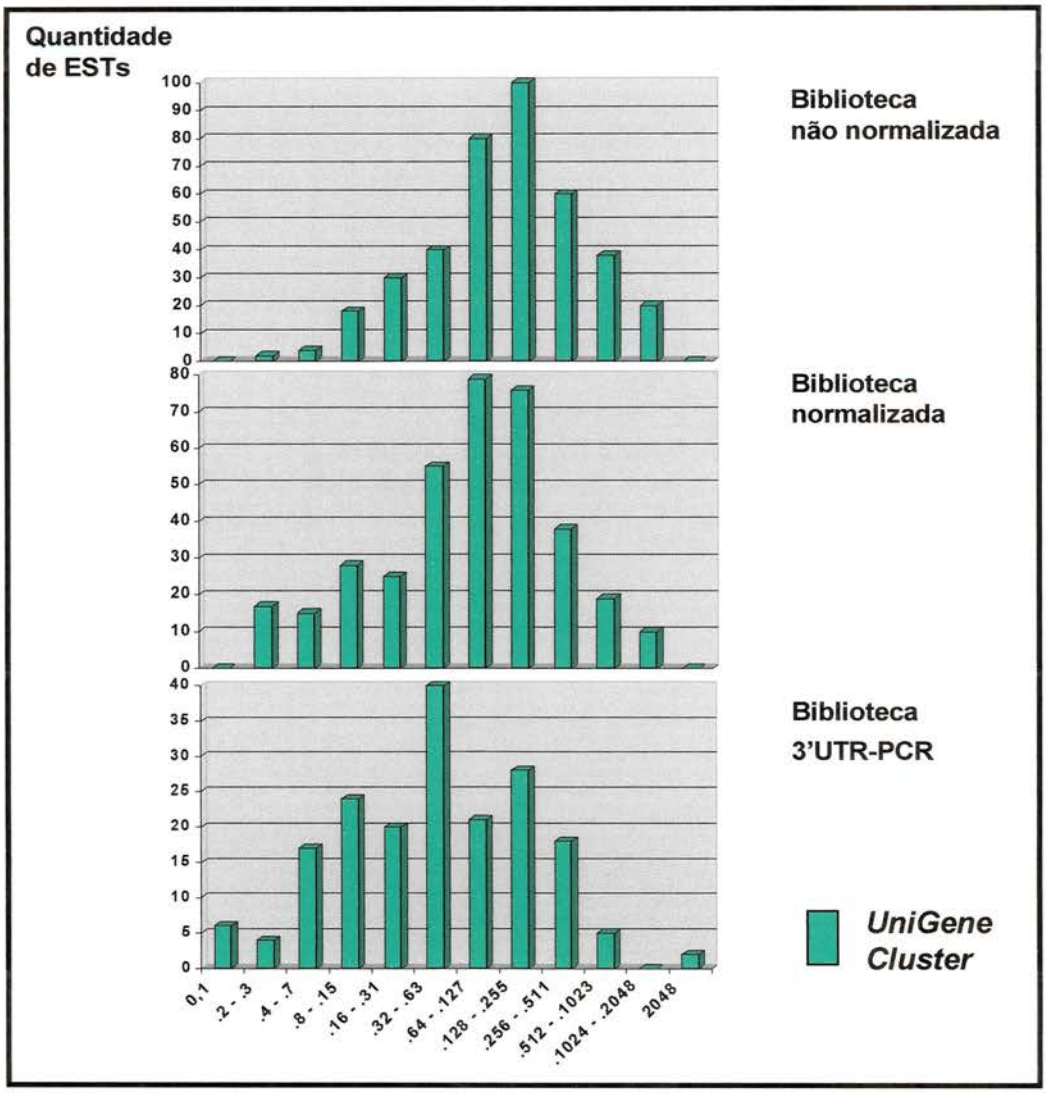


Figura 14: Gráfico representando o estudo comparativo da normalização usando a metodologia padronizada:

Análise de abundância das ESTs por comparação com *clusters* do UniGene. Este gráfico representa o estudo da normalização das ESTs geradas a partir de carcinoma de cólon, carcinoma de mama e PNET (mini bibliotecas construídas com Oligo dT (20)) onde o número médio de entradas foi de 135, em comparação com bibliotecas Não Normalizada NCI CGAP Br1.1 e Normalizada NCI CGAP Br2, onde os valores médios de entradas foram respectivamente 240 e 194. No eixo das ordenadas está representado a quantidade de ESTs e no eixo das abscissas estão representados os diferentes *clusters* do UniGene.

Devido ao fraco perfil de amplificação gerado pela reação contendo apenas um iniciador, foi testado o uso de combinações de iniciadores tais como $T11_{VG}$ / $T11_{VC}$, onde foi observado um aumento na complexidade dos perfis gerados (figura 15).

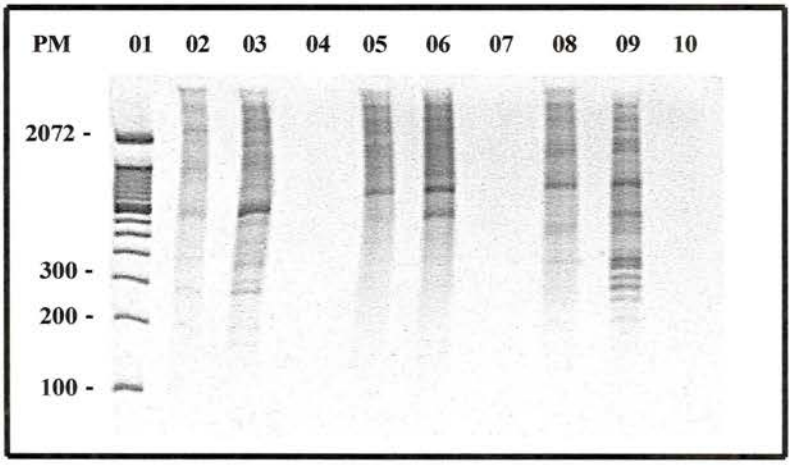


Figura 15: Perfil de amplificação de cDNAs de diferentes tumores, usando a combinação de iniciadores $T11_{VG}$ / $T11_{VC}$ para a produção de mini bibliotecas.

Gel de poliacrilamida a 8% corado pela prata. Canaleta 01 padrão de peso molecular (100bp *ladder*). Nas canaletas 02 e 03 pode ser observado o perfil de amplificação do cDNA para a construção das mini-bibliotecas partindo de carcinoma de mama nas seguintes proporções de diluição do cDNA: 1/10, 1/100. Seguindo o mesmo padrão de organização, pode ser observado o perfil de amplificação de cDNAs para a construção das mini-bibliotecas partindo de tumor de cólon, canaletas 05, 06; e para a construção das mini-bibliotecas partindo de PNET, canaletas 08, 09. Canaletas 04, 07 e 10 representam os controles negativos da síntese de cDNA, onde não foi usada a enzima transcritase reversa.

Tabela 15: Síntese dos dados gerados durante a padronização da metodologia:

Biblio Nº	Tumor	Iniciadores	Tempertura de pareamento	Quantidade de ESTs	Detalhes Técnicos	Observações
01	cólon	$T(11)_{VN}$ / Ldd1	38°C	15	DDRT-PCR	Baixo percentual de ESTs ancoradas na porção 3'
02	cólon	$T(11)_{VN}$ / Ldd2	38°C	35	DDRT-PCR	Baixo percentual de ESTs ancoradas na porção 3'
03	cólon	$T(11)_{VN}$ / LTK3	38°C	24	DDRT-PCR	Baixo percentual de ESTs ancoradas na porção 3'
04	cólon	$T(11)_{VN}$ / Ldd1/Ldd2/Ltk3	38°C	140	DDRT-PCR	Baixo percentual de ESTs ancoradas na porção 3'
05	cólon	$T(25)_{CG}$	50°C	179	DDRT-PCR com iniciadores maiores	Dificuldades na etapa de seqüenciamento
06	cólon	$T(25)_{CG}$	55°C	135	DDRT-PCR com iniciadores maiores	Dificuldades na etapa de seqüenciamento
07	mama	$T(11)_{VA}$	38°C	50	Oligo dT (20)	Baixa qualidade de seqüência
08	mama	$T(11)_{VC}$	38°C	30	Oligo dT (20)	Baixa qualidade de seqüência
09	mama	$T(11)_{VG}$	38°C	30	Oligo dT (20)	Baixa qualidade de seqüência
10	cólon	$T(11)_{VA}$	38°C	100	Oligo dT (20)	Metodologia padronizada
11	cólon	$T(11)_{VT}$	38°C	65	Oligo dT (20)	Metodologia padronizada
12	cólon	$T(11)_{VC}$	38°C	111	Oligo dT (20)	Metodologia padronizada
13	cólon	$T(11)_{VG}$	38°C	100	Oligo dT (20)	Metodologia padronizada
14	PNET	$T(11)_{VA}$	38°C	10	Oligo dT (20)	Metodologia padronizada
15	PNET	$T(11)_{VG}$	38°C	96	Oligo dT (20)	Metodologia padronizada
16	PNET	$T(11)_{VTC}$	38°C	30	Oligo dT (20)	Metodologia padronizada
17	PNET	$T(11)_{VC}$	38°C	96	Oligo dT (20)	Metodologia padronizada
18	PNET	$T(11)_{VTC/VC}$	38°C	106	Oligo dT (20)	Metodologia padronizada

Nesta tabela está síntetisado o total de mini bibliotecas produzidas até o momento. Todas as ESTs (824) geradas por estas mini bibliotecas foram analisadas manualmente.

4.5 Dinâmica da produção de ESTs e análise dos resultados:

Diante dos bons resultados obtidos com as modificações feitas na metodologia, demos continuidade à extração de mRNA, construção de mini bibliotecas e geração de ESTs a partir de uma nova coleção de tumores de baixa incidência, a saber: (i) leiomioma, (ii) Wilms (associado à síndrome de Denis Drash), (iii) hibernoma e (iiii) leiomiossarcoma. Uma média de 08 mini-bibliotecas (com exceção do tumor leiomiossarcoma), diferenciadas pelo tipo de iniciador usado, foram construídas partindo de cada tumor, 96 clones de cada biblioteca foram seqüenciados, gerando cerca de 800 ESTs para cada tumor. Tanto as ESTs iniciais geradas a partir de carcinoma de cólon, carcinoma de mama e PNET, quanto aquelas ESTs geradas a partir dos novos tumores selecionados, foram analisadas no setor de bioinformática do ILPC, coordenado pelo Dr. Sandro de Souza. Os resultados destas análises estão listados na tabela 16:

Tabela 16: Total de ESTs* analisadas com o auxílio do laboratório de bioinformática do ILPC:

Biblio N°	Tumor	Iniciador T(11)	Biblio ID	Clones Sequen	Clones Analisados	% nt Masc	Outras	Genes humanos	Contigs	dbEST	*No matches*
07	Mama	VA	BT0001-003	62-ABI	50	12.9	00	14	28	00	06
08	Mama	VC	BT0001-004	17-ABI	05	00	01	01	03	00	00
09	Mama	VG	BT0001-005	56-ABI	30	7.0	01	03	06	04	13
19	Mama	VC/VG	BT0001-001	ABI-96	88	4,4	03	11	59	03	07
20	MAMA	VVV	BT0001-002	ABI-96	69	3.1	01	12	42	05	05
10	COLON	VA	CT0001-002	78-ABI	63	8.3	00	18	36	04	03
11	COLON	VT	CT0001-005	46-ABI	12	2.5	00	01	02	01	05
12	COLON	VC	CT0001-003	95-ABI	71	5.0	00	09	45	01	12
13	COLON	VG	CT0001-004	93-ABI	46	3.0	08	07	16	04	09
21	COLON	VC/VG	CT0001-001	ABI-47 Mega-49	33	5,0	00	04	20	01	05
				735	**467**		**14**	**80**	**257**	**23**	**65**
14	PNET	VA	HT0001-003	10-ABI	-		-	-	-	-	
15	PNET	VG	HT0001-005	96-ABI	83	26.0	13	09	34	00	17
16	PNET	VTC	HT0001-007	95-ABI	82	26.0	13	09	34	00	16
17	PNET	VC	HT0001-004	61-ABI	35	18.0	00	02	17	01	01
18	PNET	VTC/VC	HT0001-006	96-ABI	91	4.3	00	37	39	03	08
22	PNET	VC/VG	ST0999-304	36-ABI	32	22,5	00	03	17	02	06
23	PNET	VC/VG	HT0001-001	96-ABI	82	00	02	11	49	04	15
24	PNET	VVV	HT0001-002	96-ABI	72	0.82	00	09	53	04	06
				586	**477**		**28**	**80**	**243**	**14**	**69**
24	Wilms	VC	ST0001-001	ABI-96	89	10.4	00	13	46	07	15
25	Wilms	VG	ST0001-002	ABI-96	90	8.7	19	11	34	10	10
26	Wilms	VT	ST0001-003	Mega-96	17	10.8	00	00	12	00	04
27	Wilms	VVV	ST0001-004	ABI-96	82	1.6	01	16	51	04	04
28	Wilms	VTC	ST0001-005	ABI-96	80	4.5	00	08	49	04	14
29	Wilms	VA	ST0001-006	ABI-96	91	7.8	00	05	63	11	09
30	Wilms	VC/VG	ST0001-007	ABI-96	59	2.2	00	06	43	02	06
31	Wilms	VC/VT	ST0001-008	ABI-96	93	3.4	00	18	53	04	12
32	Wilms	VC/VA	ST0001-009	ABI96	57	0.98	00	05	39	03	06
				864	**658**		**20**	**82**	**390**	**46**	**80**
33	Leios	VG	ST0002-001	ABI-96	74	7.1	00	11	40	00	18
34	Leios	VVV	ST0002-002	ABI-96	83	2.5	01	21	46	01	08
35	Leios	VTC	ST0002-003	ABI-96	83	1.8	02	17	40	06	16
36	Leios	VC/VT	ST0002-004	ABI-96	82	6.6	00	18	47	04	09
37	Leios	VC/VA	ST0002-005	ABI-96	63	10.2	00	04	40	05	08
38	Leios	VC/VG	ST0002-006	ABI-96	93	7.1	00	24	54	02	11
39	Leios	VC	ST0002-007	ABI-96	83	4.8	00	18	48	02	11
				672	**561**		**03**	**113**	**315**	**20**	**81**
40	Leiom	VC/VG	HT0002-001	ABI-96	87	1,8	01	07	58	01	16
41	Leiom	VG	HT0002-002	ABI-96	86	0,6	00	08	58	09	07
				192	**173**		**01**	**15**	**116**	**10**	**23**
42	Hiber	VVV	ST0003-001	ABI-96	73	2.3	00	14	37	08	06
43	Hiber	VC/VA	ST0003-002	ABI-96	87	3.7	00	12	52	07	09
44	Hiber	VC/VG	ST0003-003	ABI-96	85	6.7	00	24	43	03	11
45	Hiber	VC/VT	ST0003-004	ABI-96	-	-	-	-	-	-	-
46	Hiber	VA/VT	ST0003-005	ABI-96	-	-	-	-	-	-	-
47	Hiber	VG	ST0003-006	ABI-96	-	-	-	-	-	-	-
48	Hiber	VC	ST0003-007	ABI-96	-	-	-	-	-	-	-
49	Hiber	VTC	ST0003-008	ABI-96	-	-	-	-	-	-	-
50	Hiber	VA	ST0003-009	ABI-96	-	-	-	-	-	-	-
				864	**245**		**00**	**50**	**132**	**18**	**26**
	TOTAL			**3913**	**2581** 65,9%		**66** 2,6%	**420** 16,2%	**1453** 56,2	**131** 5,0%	**344** 13,3%

* Todas as ESTs produzidas e analisadas estão em processo de deposição no banco de dados dbEST.

Tumor: Tumores estudados: carcinoma de mama, carcinoma de cólon, PNET, Wilms/DenisDrash, leiomiossarcoma (Leios), leiomioma (Leiom) e hibernoma (Hiber).

Iniciador T(11): iniciador usado para construção da mini-biblioteca de cDNA

Biblio ID: Código da mini-biblioteca para análise na bioinformática

Clones Sequen: número de clones seqüenciados

Clones Analisados: número de seqüências com tamanho superior a 100 nt e qualidade suficiênte para atender os parâmetros exigidos pelo programa Phred-Analysis

% nt Masc: percentual de nucleotídeos mascarados de acordo com programa *ReapetMasker*

Outras: ESTs que apresentaram identidade com DNA de bactérias, DNA mitocondrial e RNA ribossomal

Genes Humanos: ESTs que apresentaram identidade com genes humanos totalmente seqüenciados

Contigs: ESTs que apresentaram identidade com genes humanos parcialmente seqüenciados (contigs) agrupadas no UniGene.

dbEST: ESTs que apresentaram identidade com ESTs humanas que ainda não formaram contigs (dbEST)

No matches: ESTs que não apresentaram identidade contra seqüências depositadas nos bancos de dados.

A análise de normalização das mini bibliotecas geradas, foi baseada no número total de ESTs que apresentavam identidade contra genes humanos totalmente sequenciados (tabela 15), novamente levando em consideração a quantidade de seqüências depositadas (entradas) nos respectivos *clusters* do UniGene. O resultado desta análise pode ser observado na figura 16a (3'UTR-PCR). Com o objetivo de comparar o nível de normalização da estratégia usada, foi feita uma análise comparativa com outras duas bibliotecas, uma não normalizada e outra normalizada, ambas produzidas a partir de tumor de mama (Figuras 16a e 16b).

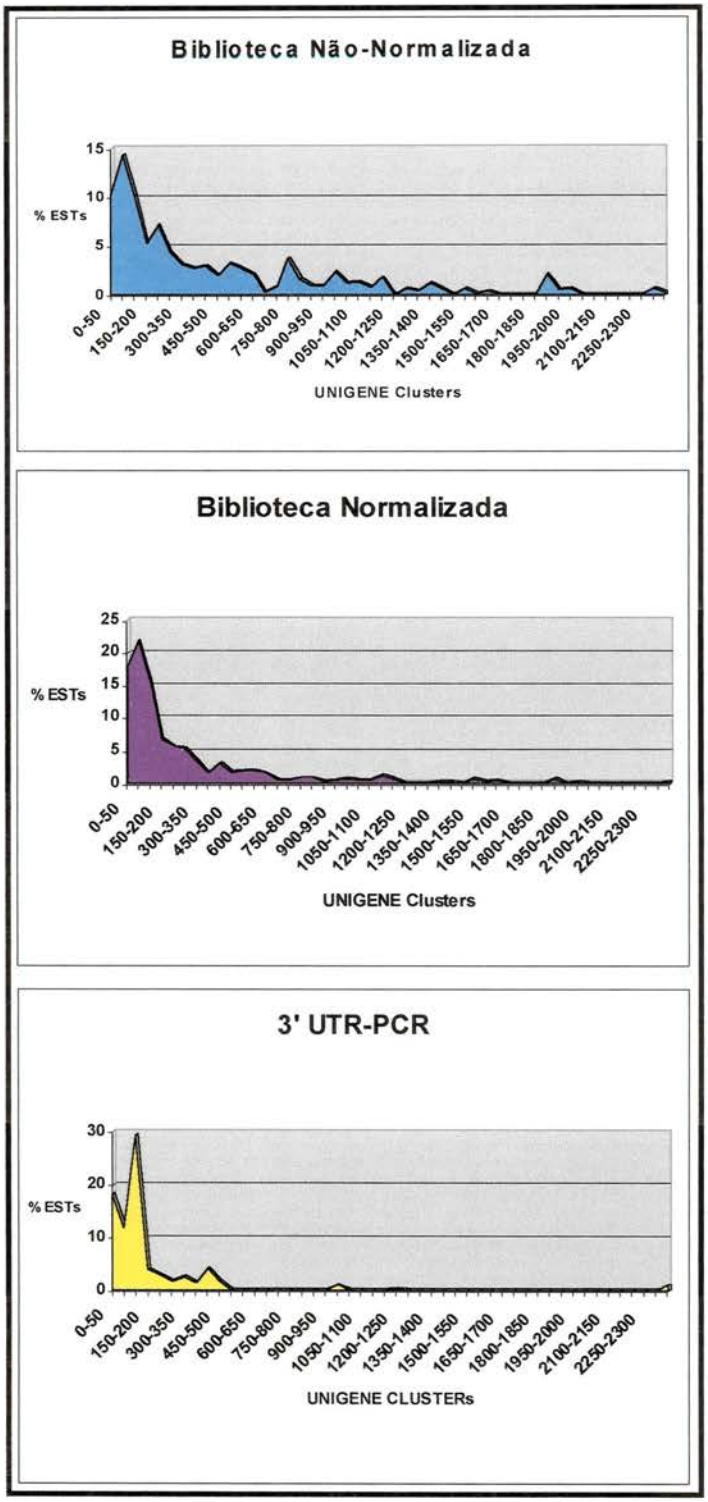


Figura 16a: Análise da normalização de bibliotecas de cDNA geradas por diferentes estratégias:

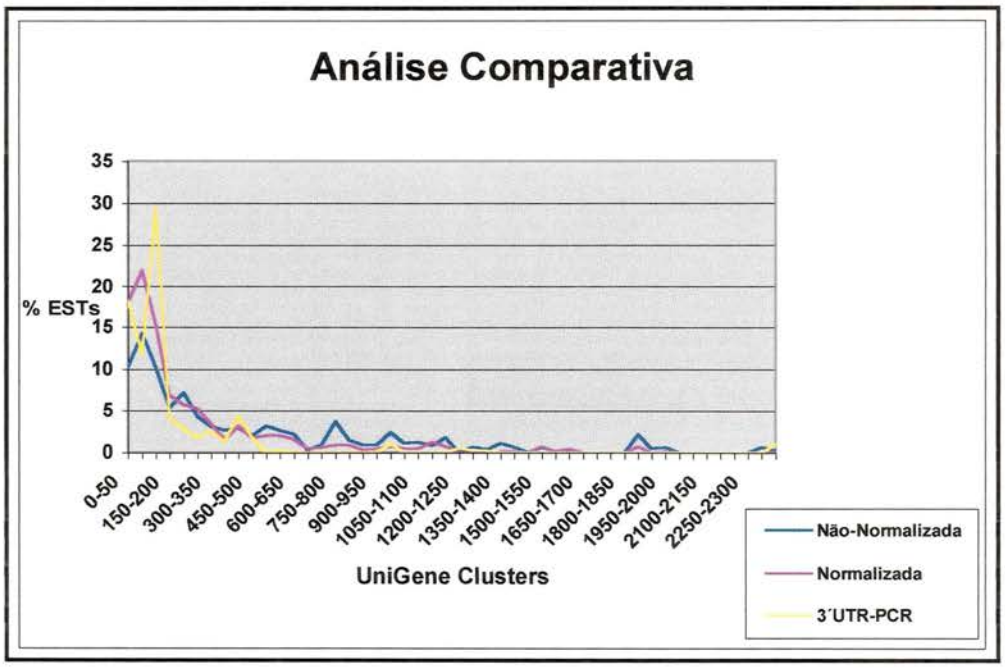


Figura 16b: Análise comparativa da normalização de bibliotecas de cDNA geradas por diferentes estratégias:

Na figura 16a estão representados os graficos de análise de normalização de bibliotecas de cDNA geradas por diferentes estratégias as quais são: bibliotecas não normalizadas NCI CGAP Br1.1, bibliotecas normalizadas NCI CGAP Br2, ambas partindo de tumor de mama geradas pelo grupo do Dr. Bento Soares e bibliotecas geradas pela estratégia desenvolvida neste trabalho (3' UTR-PCR), onde o valor médio de entradas nos *clusters* do UniGene foi respectivamente de 444, 255 e 194. Na figura 16b está representado o gráfico agrupando as análises de normalização das três metodologias, permitindo uma análise comparativa.

A redundância foi determinada com o auxílio do programa Cap3 (HUANG e MADAN 1999), com base no total de ESTs geradas por tecido estudado (Tabela 17). Esta análise tem o objetivo de informar a diversidade de genes que a técnica é capaz de detectar baseada no número de vezes que o fragmento de um mesmo gene foi sequenciado.

Tabela 17: índice de redundância das mini bibliotecas nos diferentes tecidos:

Tecido	Total de ESTs analisadas	*Contigs*	ESTs (*Singletons*)	Redundância
Carcinoma de mama	283	31 (120)	163	0,32
Carcinoma de cólon	186	30 (101)	85	0,39
PNET	427	72 (336)	91	0,62
Tumor Wilms	639	89 (388)	251	0,47
Leiomiossarcoma	558	59 (352)	206	0,53
Leiomioma	170	18 (118)	52	0,59
Hibernoma	235	35 (139)	96	0,45
Total	2498	334 (1554)	944	0,49

O programa Cap3 foi utilizado para a obtenção dos valores que estão listados na tabela 15. Os números colocados entre parênteses na coluna *Contigs* indica a quantidade de ESTs usadas para a montagem dos *Contigs* (gene). O percentual de redundância foi obtido através da soma dos *Contigs* com as ESTs que não formaram *contigs (singletons)*, dividido pelo total de ESTs analisadas.

Para saber se as seqüências estavam sendo geradas partindo da porção 3' *UTR*, foram selecionadas 100 ESTs, não redundantes, que apresentaram identidade com genes humanos totalmente seqüênciados, para fazer uma análise posicional (Tabela 18). O resultado desta análise pode ser observado na figura 17.

Tabela 18: Dados usados para análise da posição média do alinhamento das ESTs contra genes humanos totalmente seqüenciados:

Identidade da EST	Tamanho da EST	Tamanho do gene	Ponto médio	Posição (%) no gene
260899-002-h08	226	447	260	58
200999-004-c12	167	517	87	16
260899-002-b08	220	542	97	17
260899-001-a08	308	555	412	74
251099-003-c10	221	811	698	86
161199-003-f02	280	975	880	90
260899-001-b02	415	1013	743	73
251099-002-g06	346	1093	811	74
200999-005-h09	353	1194	962	80
250899-004-d01	279	1248	1105	88
121199-005-b09	154	1291	482	37
200999-006-d07	212	1305	979	75
260899-001-h08	164	1332	1250	93
251099-002-g11	279	1339	1205	89
161199-004-d09	263	1347	571	42
250899-004-f11	185	1402	991	70
181099-001-g01	195	1409	1279	90
260899-005-f02	235	1413	1299	91
030999-001-b06	351	1432	1017	71
260899-001-h11	139	1487	1449	97
161199-003-b04	395	1495	1290	86
161199-003-d05	336	1532	1368	89
260899-001-d07	190	1537	1455	94
200999-006-h04	209	1577	783	49
161199-003-b06	237	1582	1451	91
161199-002-d01	209	1591	1500	94
251099-003-e08	161	1628	1417	87
260899-001-f07	402	1653	1516	91
251099-007-b11	320	1661	1568	94
181099-001-f04	133	1707	1557	91
161199-006-g02	433	1718	1523	88
260899-001-f07	220	1778	901	50
030999-001-c04	355	1814	1505	82
260899-005-c08	220	1907	1790	93
251099-002-g01	195	1918	1736	90
250899-004-d11	306	1960	1802	91
030999-001-a03	404	2002	1838	91
030999-001-h01	263	2051	1917	93
030999-001-a03	404	2062	1838	89
260899-001-e04	368	2078	2005	96
171199-003-d04	283	2079	1580	75
251099-002-g12	348	2123	2061	97
200999-005-a11	312	2124	2003	92
070999-006-c08	351	2144	1973	92
250899-004-g12	165	2180	2117	97
161199-003-b07	231	2194	2064	94
260899-001-e03	199	2197	2109	95
260899-001-e08	203	2226	1165	52
161199-006-e08	392	2269	1385	61
200999-004-a09	147	2364	1518	64
200999-005-a01	385	2369	1041	43
161199-003-a01	471	2371	1788	75
161199-003-e09	332	2427	1797	74
121199-007-g05	326	2460	2296	93
161199-003-d09	332	2472	2282	92
200999-004-d08	208	2591	1213	46
251099-002-b12	190	2606	2503	96

200999-008-b04	293	2615	1072	41
171199-003-d10	443	2636	2066	78
181099-009-g11	208	2654	1441	54
260899-001-h01	280	2739	2648	96
161199-004-c11	298	2780	2627	94
251099-002-e03	232	2786	2361	84
161199-003-d11	176	2809	1351	48
161199-003-e10	341	2834	1899	67
250899-004-c01	206	2881	2817	97
260899-001-b05	375	2940	2606	88
121199-005-d05	223	3019	2893	95
161199-006-c07	272	3065	2459	80
121199-005-d06	108	3099	2576	83
251099-007-g06	281	3143	1499	47
251099-007-g08	148	3182	1637	51
200999-004-c11	253	3209	2283	71
200999-006-f10	106	3210	3164	98
260899-001-e12	110	3228	3178	98
251099-007-d10	280	3480	3326	95
250899-004-f04	297	3614	3495	96
030999-001-d11	279	3647	3511	96
251099-002-f07	289	3872	3780	97
200999-006-f07	209	3913	3820	97
210999-007-a05	199	3914	3821	97
251099-002-g07	236	4114	3605	87
030999-001-d03	367	4137	2600	62
260899-002-b03	183	4312	3607	83
200999-008-f09	231	4318	4032	93
260899-001-d11	373	4405	1941	44
101299-002-d08	254	4930	4772	96
260899-002-d03	124	5010	4937	98
251099-003-g05	408	5024	4828	96
251099-003-f03	391	5043	4838	95
251099-002-c12	323	5130	4859	94
210999-007-d01	174	5483	5401	98
251099-007-d10	280	5504	5361	97
161199-002-a03	216	5661	3619	63
181099-001-c10	224	5966	5858	98
200999-004-c05	255	5999	5511	91
030999-001-e12	296	7389	7022	95
251099-003-f02	306	8694	8346	99
070999-006-f09	117	13865	12932	93
	266,94			80,75

Nesta tabela estão listadas 100 ESTs não redundantes, que apresentaram identidade com genes humanos totalmente sequenciados com o valor de $e \geq 10^{-15}$. Na coluna "Ponto médio" foi colocado o resultado da média aritmética do ponto inicial e final do alinhamento da EST, este resultado dividido pelo tamanho do gene informa a posição relativa da EST no gene. Desta forma, se o tamanho do gene é 100%, podemos considerar que quanto maior o valor descrito nesta coluna mais no final (3') do gene a EST estará posicionada. O tamanho médio das ESTs selecionadas para este estudo foi de 266 nt.

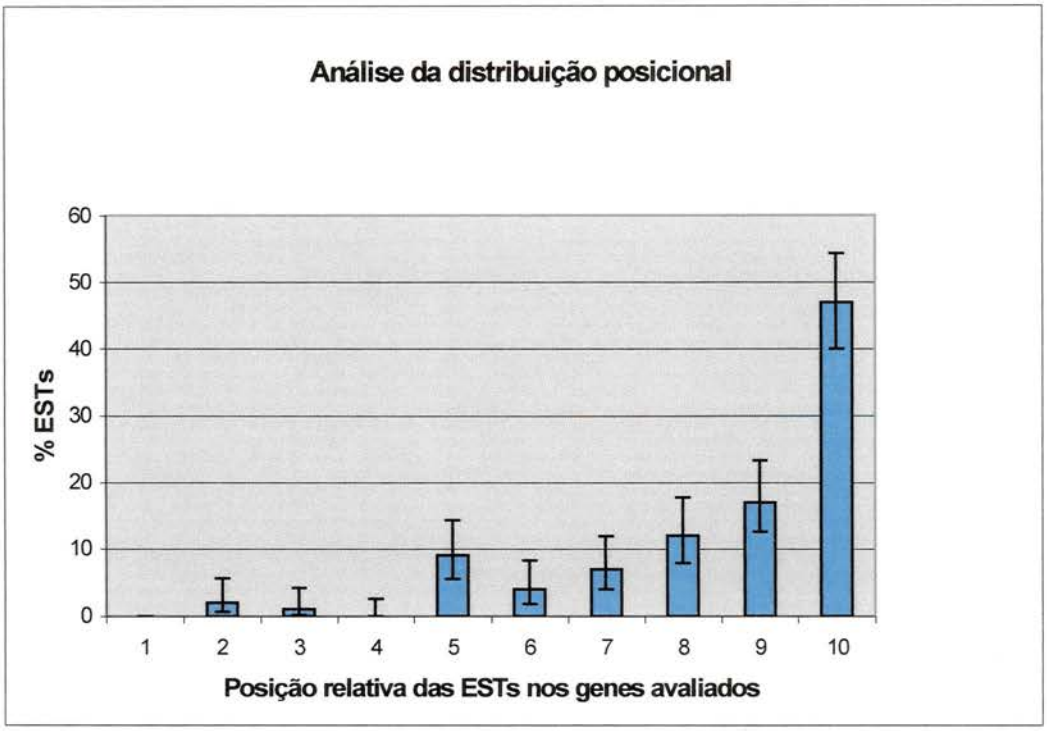


Figura 17: Análise da distribuição posicional:

No gráfico está representada a posição relativa de 100 ESTs geradas de acordo com o alinhamento ao longo dos genes (eixo do X). Para os cálculos desta análise, consideramos o tamanho total do gene equivalendo a 100%. Desta forma o gene foi fracionado em 10 segmentos, cada um equivalendo a 10%, sendo o primeiro segmento relativo a 10% da porção 5' e o último segmento relativo a 10% da porção 3'. As linhas verticais representam o intervalo de confiança de 95%, calculados a partir da distribuição binomial com correção para populações finitas (ZAR JH, 1996).

```
Name:   CM0-HT0002-101299-002-b09       View Chromatogram
Tissue: head_neck
Primer: P0001.P0003
Date:   101299
Sequence:
AATTTATTAGAAAATATTACACACTACATACATATTTATTTAAGAGTATA
ATTTCAGTAGACAGCACTCCAACTTCATAATACAAACAATATCTCATGGA
CCCTGGGGCAATCACTGTGCTTCTGAAGCTACATCTTTTGGATGGAACAT
TCCGGCAGTACACACTGTGGCCATTTAAAGGAAGGTAAGAGAAAAAAAAA
AA

Length: 202 nt
Six-frame translation and coding potential
High quality sequence : 7 to 202

Significant similarities:
ugc|Hs.181165.10  |Hs.181165 GENE=EEF1A1 PROTSIM=PID:g1072161 euk...   385  e-107
>ugc|Hs.181165.10|Hs.181165 GENE=EEF1A1 PROTSIM=PID:g1072161 eukaryotic translation
elongation factor 1 alpha 1 Length = 811 Score =  385 bits (194), Expect = e-107
 Identities = 197/198 (99%), Positives = 197/198 (99%)

Query: 1   aatttattagaaaatattacacactacatacatatttatttaagagtataatttcagtag 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 791 aatttattagaaaatattacacactacatacatatttatttaagagtataatttcagtag 732

Query: 61  acagcactccaacttcataatacaaacaatatctcatggaccctggggcaatcactgtgc 120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 731 acagcactccaacttcataatacaaacaatatctcatggaccctggggcaatcactgtgc 672

Query: 121 ttctgaagctacatcttttggatggaacattccggcagtacacactgtggccatttaaag 180
           |||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct: 671 ttctgaagctacatcttttggatggaacattccagcagtacacactgtggccatttaaag 612

Query: 181 gaaggtaagagaaaaaaa 198
           ||||||||||||||||||
Sbjct: 611 gaaggtaagagaaaaaaa 594

Bioinformática ILPC: Last update: Fri Dec 10 08:26:09 EST 1999
Blast (X): No significant similarity found
```

Figura 18: Análise de EST ancorada na região 3' do transcrito.

Esta análise indica que o fragmento seqüenciado, possivelmente é proveniente da região 3' do gene. Os dados que podem evidenciar esta característica são: o tamanho do gene, o início e o final do alinhamento que estão marcados em negrito. *Query* (EST submentida a análise), *Subject* (Seqüência depositada no banco de dados).

```
Name:   CM0-CT0001-171199-003-d10       View Chromatogram
Tissue: colon
Primer: P0001.P0002
Date:   171199
Sequence:
TTTTTTTTTTTCCAATATTTAAATAACAAGCTGTTTAATATTAAAGCAGA
AGGTATTGCCACATTGTGACTGAAAAACCCTTTTATCCATAAGCCCTTCA
CACAATTATACATTAAATGCTATTTTTATTTAAGCAAGGCACCCCTACTT
GTTCTAAAATATGGGATGTACTACTCCATTTAAAAAGCAAATGAGGAGAC
TGATTTTTTTCCTATCTAGAGCTGGTTTAAATTCAAGTGAAGCCATTAAT
TTTCTTACCAGTCTGGACTGTTCTGACATGTCACTTCACAGTTTTGAGAC
TAACAAACACCCTTAGGTCTACCCCAAACCAAAGAAAGCCAGGGAGGGAC
AGTTTAGACAACTTTTAAGTTAAGACTAGAATGCCCCCTTAAGTAGATAG
GCACCACCGTTATACGCCTTTAGCACAAGGCTTAGAAAGGCAA

Length: 443 nt
Six-frame translation and coding potential
High quality sequence : 1 to 75
Significant similarities:
ugc|Hs.79295.3  |Hs.79295 GENE=GRSF1 PROTSIM=PID:g3880146 G-rich ...   767  0.0
>ugc|Hs.79295.3|Hs.79295 GENE=GRSF1 PROTSIM=PID:g3880146 G-rich RNA sequence binding
factor 1 Length = 2636 Score =  767 bits (387), Expect = 0.0 Identities = 427/439
(97%), Positives = 427/439 (97%), Gaps = 1/439 (0%)

Query: 6    ttttttccaatatttaaataacaagctgtttaatattaaagcagaaggtattgccacatt 65
             |||| ||||||||||||||||||||||||||||||||||||||| ||| |||||||||
Sbjct: 2285 ttttgtccaatatttaaataacaagctgtttaatattaaagcagaaagtactgccacatt 2226

Query: 66   gtgactgaaaaacccttttatccataagcccttcacacaattatacattaaatgctattt 125
             ||||| |||  ||   |||||||||| ||||||||||||||||||||||||||||||
Sbjct: 2225 gtgacagaagtacagctttatccataaacccttcacacaattatacattaaatgctattt 2166

Query: 126  ttatttaagcaaggcacccctacttgttctaaaatatgggatgtactactccatttaaaa 185
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2165 ttatttaagcaaggcacccctacttgttctaaaatatgggatgtactactccatttaaaa 2106

Query: 186  agcaaatgaggagactgatttttttcctatctagagctggtttaaattcaagtgaagcca 245
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2105 agcaaatgaggagactgatttttttcctatctagagctggtttaaattcaagtgaagcca 2046

Query: 246  ttaattttcttaccagtctggactgttctgacatgtcacttcacagttttgagactaaca 305
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2045 ttaattttcttaccagtctggactgttctgacatgtcacttcacagttttgagactaaca 1986

Query: 306  aacacccttaggtctaccccaaaccaaagaaagccagggagggacagtttagacaacttt 365
             |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct: 1985 aacacccttaggtctaccccaaaccaaaaaaagccagggagggacagtttagacaacttt 1926

Query: 366  taagttaagactagaatgccccctaagtagataggcacc-accgttatacgcctttagc 424
             |||||||||||||||||||||||||||||||||||| |||||||||||||||||||
Sbjct: 1925 taagttaagactagaatgccccctaagtagataggcaccaaccgttatacgcctttagc 1866

Query: 425  acaaggcttagaaaggcaa 443
             ||||||||||||||||||
Sbjct: 1865 acaaggcttagaaaggcaa 1847

Bioinformática ILPC: Last update: Wed Nov 17 09:30:06 EST 1999
BLAST (X): No significant similarity found
```

Figura 19: Análise de EST sugerindo possível sinal de poliadenilação alternativo.

Esta análise indica que o gene que gerou a seqüência possui sinal de poliadenilação alternativo. Os dados que podem evidenciar esta característica são: EST gerada contendo sinal de poliadenilação (grifado) e a diferença entre o tamanho do gene e o início do alinhamento da EST. O tamanho do gene o início e o final do alinhamento também estão marcados em negrito. *Query* (EST submentida a análise), *Subject* (Seqüência depositada no banco de dados).

Também visando o objetivo principal do projeto, foi feita uma análise mais detalhada das ESTs classificadas na categoria *"No match"*. Estas ESTs que não apresentaram identidade contra seqüências depositadas nos bancos de dados foram analisadas manualmente com o objetivo de checar se elas realmente identificaram novos genes humanos ou seria possível a identificação de algum tipo de artefato ou contaminação (tabela 19 e figura 20). Esta análise levou em consideração os seguintes aspectos: (i) percentual de nucleotídeos mascarados (*RepeatMaker*), (ii) qualidade da seqüência e (iii) a existência de sítio de poliadenilação.

Tabela 19: Prováveis novos genes:

Tumor	Clones analisados	*No Matches*	Prováveis novos genes	x^2 valor de p	Fisher valor de p
Mama e Cólon	467	65	10 (2,1%)	-	-
PNET	477	69	28 (5,8%)	0.003567	0.004354
Leiomiossarcoma	561	81	38 (6,7%)	0.000456	0.000507
Wilms/DanisDrash	658	80	18 (2,7%)	0.528426	0.566668
Leiomioma	173	23	07 (4,0%)	0.291768	0.265228
Hibernoma	245	26	03 (1,2%)	0.566317	0.558418
Total	2581	329	104 (3%)	-	-

Dados relativos aos prováveis novos genes identificados. O valor de p para o teste do x^2 e de Fisher, teve como categoria referêncial os valores obtidos para os tumores de mama e cólon (Epi Info 6 versão 6,04b 1997).

Name: CM0-HT0001-161199-005-b02 View Chromatogram
Tissue: PNET
Primer: P0001.P0002
Date: 161199
Sequence:
TTTTTTTTTTCGTAGGCAATTATAATTTATTTCGTGTATTATGTTTA
TTTCATTTTTGTAACTTTAAAATTTCTAAATGAGTATATGTTACTTTAG
TAGTAAGATATTCAGAGCTTCTTTATATAACCAAAAAAAAAAA
Length: 143 nt
Six-frame translation and coding potential
High quality sequence: 1 to 143

Significant similarities:
No more matches!

Figura 20: EST representando um provável novo gene humano:

Exemplo de EST representando a identificação de um provável novo gene, onde o sítio de poli-adenilação pode ser observado (em vermelho). A rotina de análises foi feita no setor de bioinformática do ILPC.

4.6 Re-análise das ESTs classificadas na categoria *"No match"*:

Uma vez que as ESTs geradas foram analisadas (tabela 15) contra os bancos de dados públicos antes do término do primeiro rascunho da decodificação do genoma humano, uma nova análise se faz necessária para atualização dos dados. Desta forma, foram feitas análises manuais utilizando o programa Blast contra o banco de dados nr, dbEST categoria *Human ESTs*, HTGS (*High-Throughput Genome Sequence*) e o banco de dados de proteínas Blast X, de todas as ESTs classificadas na categoria *"No match"* (Tabelas 20, 21 e 22). Em alguns casos a EST analisada não apresentou identidade com seqüências de cDNA, porém a seqüência já existe no banco de dados de DNA, sugerindo que possivelmente o transcrito deste gene ainda não foi seqüenciado, conforme figura 21 e tabela 20.

Tabela 20: Re-análise das ESTs da categoria *"No match"* :

TECIDO	cDNA	DNA	Sinal de poli A	*Seqüências novas*	Total de ESTs
Cólon/mama	14 (43%)	06 (18%)	09 (27%)	12 (37%)	32
PNET	19 (35%)	17 (31%)	18 (33%)	18 (33%)	54
Leiomioma	08 (36%)	01 (4%)	03 (13%)	13 (59%)	22
Wims/DenisDrash	10 (22%)	19 (42%)	15 (33%)	16 (35%)	45
Leiomiossarcoma	09 (27%)	09 (27%)	11 (33%)	15 (45%)	33
Total	60 (32%)	52 (27%)	56 (30%)	74 (39%)	186

Na tabela estão representados os resultados da reanálise de cerca de 2/3 das ESTs classificadas na categoria *"No match"*. A análise foi feita através do programa Blast contra os bancos de dados nr, dbEST, HTGS e banco de proteínas Blast X, além de ter sido verificado o percentual de ESTs contendo sinal de poliadenilação. O tamanho médio das ESTs que entraram para este estudo, foi de 198 nt. Nas colunas cDNA e DNA foram colocadas as ESTs que apresentaram identidade de pelo menos 80% com o valor de $e \geq 10^{-15}$, com seqüências expressas (cDNA ou mRNA) e com DNA respectivamente.

Tabela 21: Seqüências novas:

Tecido	Seqüências novas	Sinal de poli A	Redundantes	Seqüências diferentes
Colon/mama	12	03 (25%)	02 (16%)	10 (83%)
PNET	18	00	10 (55%)	08 (44%)
Leiomioma	13	01 (7%)	02 (15%)	11 (84%)
Wilms/DenisDrash	16	04 (25%)	00	16 (100%)
Leiomiossarcoma	15	06 (40%)	00	15 (100%)
Total	74	14 (18%)	14 (18%)	60 (81%)

Na tabela estão representados os resultados das análises mais detalhadas das ESTs classificadas na categoria "seqüências novas", que não apresentaram identidade nos bancos de dados.

Tabela 22: Resultados das ESTs re-analisadas:

EST gerada	Blast nr	dbEST Human	HTGS	Sinal de poli A	Identidade da EST
ST0001-181099-009-G06	AC007251.3*		AF257499.2	(-)	BAC Clone RP11-373G23
ST0001-181099-009-D06	AC000123.1*			(+)	Cosmídio g0771a106
ST0001-181099-009-H04			AL365504*	(+)	Crom 09
ST0001-181099-009-F02			AC012301.2*	(+)	Crom 18
ST0001-181099-009-D12		AC007055.3		(-)	?
ST0001-181099-009-C12	NM002107.1*	AU682482.1	AC018890.6	(-)	H3 histona H3F3a
ST0001-200999-008-F04			AC69514.10*	(-)	Crom 03
ST0001-200999-008-C09			AC16697.4*	(+)	Crom 02
ST0001-200999-008-F11			AC027187.2*	(-)	Crom 04
ST0001-200999-008-B12	AL132795.12*	BE467618.1		(-)	Crom 07
ST0001-200999-008-A01		BE175508.1	AC013394.4*	(-)	Crom 02
ST0001-210999-007-H02	AL136985.11*			(+)	AP1 Proto-oncogen C-jun
ST0001-210999-007-H01			AL160054.5*	(-)	Crom 09
ST0001-210999-007-G02			AC067942.3*	(-)	Crom 04
ST0001-210999-007-A06	AC020606.7		AC017021.4*	(-)	Crom 07
ST0001-070999-006-G09		AA927596.1	AC058794.2*	(-)	Crom 04
ST0001-070999-006-F11		AI624239.1		(-)	?
ST0001-070999-006-A11			AL356970.5*	(-)	Crm 06
ST0001-260899-005-E12	NM017754.1*	BE935636.1		(-)	Clone FLJ20302
ST0001-260899-005-E06			AC011332.5*	(+)	Crom 05
ST0001-260899-005-D06	AL1337000.6*			(+)	Crm 13 RP11-203116
ST0001-260899-005-A09			AC040910.2*	(-)	Crom 09
ST0001-030999-001-F11		AW751250.1		(-)	?
ST0001-030999-001-B05			AL357044.9*	(-)	Crom 01
ST0002-251099-002-B01	NM005463.1*	D12037.1	AC015456.4	(+)	Nuclear ribonucleoprotein HNRPDL
ST0002-251099-002-D08			AC023433.2*	(-)	RP11-114K22
ST0002-251099-002-H08		AI458816.1	AL365203.9*	(-)	Crom 10
ST0002-251099-003-A06		BE550475	AC012213.3*	(-)	CA54_HUMAN COLLAGEN ALPHA 5(IV)
ST0002-251099-003-A12			AC009806.3*	(-)	Crom 11
ST0002-251099-003-B06	NM004667.2*	AI811537.1	AC062028.2	(+)	RLD2 (Herc 2)
ST0002-251099-003-C02			AL356583.9*	(-)	Crom 01
ST0002-251099-003-C07		AI768843.1	AC020594.3*	(-)	Crom 02
ST0002-251099-003-F12	AC008572.6*	BE748354.1		(-)	Crom 05
ST0002-251099-007-B09			AC073364.7*	(+)	Crom 03
ST0002-251099-007-C02			AC016634.5*	(+)	Crom 05
ST0002-251099-007-C08			AC079319.7*	(+)	Crom 12q
ST0002-251099-007-E02		AW573079.1	AC073986.2*	(+)	Crom 05
ST0002-251099-007-G12	AC008554.7*			(-)	Crom Clone 19 CTC-513N18
ST0002-251099-007-H03	AL162457.11*			(-)	Crom 20 clone AP11-429E11
HT0002-101299-002-G06	AL355476.12*	BE673754.1	AC026085.4	(-)	Crom 10
HT0002-101299-002-E07	NM003670.1*	AI394355.1	AC019027.4	(-)	RP1144K16
HT0002-101299-002-A03		BE880550.1	AC069351.2*	(-)	Crom 08
HT0002-291199-001-H03	AK026308.1*	AW779544.1	AC073504.11	(-)	cDNA FLJ22655
HT0002-291199-001-F10			AC024257.10*	(-)	Crom 12
HT0002-291199-001-C01		R42880.1	AC020780.3*	(+)	Crom 08
HT0001-161199-005-B02			AC016770.6*	(+)	Crom 02
HT0001-161199-005-G08	NM001025.1*			(+)	Ribossomal protein S23
HT0001-161199-005-A05		BE739645.1		(-)	?
HT0001-161199-006-A10		AW594002.1		(+)	?
HT0001-121199-007-G01	NM003352.1*	BE738302.1	AC079354.1	(-)	Ubiquitin Like 1 (UBL 1)
HT0001-260899-001-B04		AW54648.1		(+)	?
HT0001-260899-001-B12			AC024973.2*	(-)	Clone RP11-542E4
HT0001-260899-001-C05		AW054648.1		(-)	?
HT0001-260899-001-C07	NM019597.1*	BE669449.1		(-)	Ribonucleoprot H2 (HNRPH2)
HT0001-260899-001-C10		AI870209.1		(-)	?
HT0001-260899-001-H12			AL139100.7*	(+)	Crom 06
HT0001-270899-002-B02		AW614117.1	AC016896.4*	(+)	Crom 11

HT0001-270899-002-D01	AL133551.13*	BE220897.1	AC022592.2	(+)	Crom 10 clone RP11-57G10
CT0001-161199-002-F03	AK025794.1*	BE542132.1	AC025289.5	(-)	cDNA FLJ22141
CT0001-161199-003-F12		AW003195.1		(-)	?
CT0001-161199-003-G02		AW189519.1		(-)	?
CT0001-161199-003-G04	AY008268.1*	AW191973.1	AC067749.3	(-)	GTP Binding protein 5AR1
CT0001-161199-003-C12			AC26294.3*	(+)	Crom 8
CT0001-161199-004-E11			AC022221.5*	(-)	Crom 15
CT0001-161199-004-F02		AW811705.1	AC068228.2*	(-)	Crom 03
CT0001-161199-005-B09		AW468444.1	AC046178.2*	(+)	Crom 13
BT0001-260899-003-A09			AL365355.3*	(+)	Crom 01
BT0001-260899-005-E03		AW811573.1		(-)	?

Na tabela estão representados os resultados da re-submissão de 67 ESTs (não redundantes) classificadas na categoria *"No match"* contra os bancos de dados. Nas colunas estão anotados os códigos das seqüências depositadas nos respectivos bancos de dados, cuja EST reanalisada apresentou identidade de pelo menos 80% com o valor de $e \geq 10^{-15}$. Na última coluna está anotado um identificador que indica a orígem da seqüência depositada nos bancos de dados. Sinal de poli A (+) encontrado e (-) não encontrado. * Banco de dados utilizado para informar a identidade da EST.

```
Name:   CM0-ST0002-251099-003-a12

Tissue: leiomiossarcoma
Primer: P0001.P0002
Date:   251099
Sequence:
TTTTTTTTTTTGCCTTAGGTTTATACCTAGGAGTTTACACTCAAGGGTA
TTTAAACATTTAACAGAGTTCAGGGCACATCTCCTCAGCCCTGAGCCTT
ATGAGTTATCTGACGGTAAACCCCGAAGGGAACACACTGGACCTCCTTC
ACTCATTTTTTAACCCTGACTGGGACACCAAAGACATGCTGCATCTTGT
ATAAGGGGGTCCATCTTGCAAAAGGGCTGGGCTCTTGAAAAATTCCCTG
TGAAGAAAATGGTTACAATCCCATTACATCACGGGCTTTTATAATTTAA
ATTGGAAATTGTTGGCTGGAAACAATTTTAACCCCAAATTGGGAGAAAA
AAAAAAA
Length: 352nt
High quality sequence : 1 to 16
Significant similarities:
dbEST:
>gb|BE799896.1|BE799896 601588012F1 NIH_MGC_7 Homo sapiens cDNA clone IMAGE:3942195
5'. Length = 999 Score = 38.2 bits (19), Expect = 0.97 Identities = 19/19 (100%)
Strand = Plus / Plus
Query: 124 aagggaacacactggacct 142
           ||||||||||||||||||
Sbjct: 76  aagggaacacactggacct 94

HTGS:
>gb|AC009806.3|AC009806 Homo sapiens chromosome 11 clone RP11-51B23 map 11, WORKING
DRAFT
SEQUENCE, 34 unordered pieces Length = 154177 Score =  424 bits (214), Expect = e-118
 Identities = 296/322 (91%), Gaps = 1/322 (0%) Strand = Plus / Plus
Query: 14    cttaggtttatacctaggagtttacactcaagggtatttaaacatttaacagagttcagg 73
              ||||||| |||||||||||||| |||||||||| ||||||||||||| ||||||||||
Sbjct: 28791 cttaggtatatacctaggagtatacactcaagtgtatttaaacattttacagagttcagt 28850

Query: 74    gcacatctcctcagccctgagccttatgagttatctgacggtaaaccccgaagggaacac 133
              |||||||||||||| ||||||||||||||||||||||| || ||||||||| || ||||
Sbjct: 28851 gcacatctcctcagtcctgagccttatgagttatctgactgttaaccccgaaaggtacac 28910

Query: 134   actggacctccttcactcatttttttaaccctgactgggacaccaaagacatgctgcatct 193
              |||||| ||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct: 28911 actggatctccttcactcatttttttaaccctgactgggacaccagagacatgctgcatct 28970

Query: 194   tgtataaggggtccatcttgcaaaagggctgggctcttgaaaaattccctgtgaagaaa 253
              ||||| ||| | | ||||||||| || ||||| |||| ||||| ||| ||||||||||||
Sbjct: 28971 tgtattaggtgtttcatcttgcagaatggctgtgctcctgaaatatttcctgtgaagaaa 29030

Query: 254   atggttacaatcccattacatcacgggctttttataatttaaattggaaattgttggctgg 313
              || |||||||||||||||||||| ||||||||| | ||||||||||||||||||||||
Sbjct: 29031 attgttacaatcccattacatcactggcttttatta-ttaaattggaaattgttggctgg 29089

Query: 314   aaacaattttaaccccaaattg 335
              |||||||||||||||||||||
Sbjct: 29090 aaacaattttaaccccaaattg 29111

BLAST (X): No significant similarity found
```

Figura 21: Re-análise de EST da categoria *"No match"*.

Análise da EST contra o banco de dados dbEST e o banco de dados HTGS (*High-Throughput Genome Sequence*), onde pode ser observada a identidade contra o banco de dados HTGS, o que não ocorre no banco de dados dbEST, sugerindo que o transcrito deste gene ainda não foi seqüenciado. *Query* (EST submetida a análise), *Subject* (Seqüência depositada no banco de dados).

5 - Discussão:

5.1 Padronização da construção de mini bibliotecas de cDNA parcialmente normalizadas:

5.1.1 Limitações da técnica de Differential Display:

A técnica de *Differential Display* é amplamente usada para identificar genes diferencialmente expressos. Duas etapas são usadas nesta técnica. A primeira etapa, compreende a síntese de cDNA em quatro reações distintas, usando 4 iniciadores diferentes do tipo T(11) $_{VN,}$ permitindo a divisão dos transcritos de uma amostra em 12 subpopulações diferentes. A segunda etapa envolve o uso de 3 iniciadores de seqüência arbitrária em uma PCR de baixa estringência para amplificar fragmentos de populações de cDNAs sintetizados. Esta combinação de reações, permite a construção de 12 mini bibliotecas de cDNA parcialmente normalizadas. A normalização parcial ocorre devido à degeneração presente na região 3' dos iniciadores do tipo T(11) $_{VN,}$ usado na síntese de cDNA e também aos iniciadores arbitrários usados em uma PCR de baixa estringência, onde os transcritos são amplificados de maneira relativamente independente de sua prevalência na amostra.

Os resultados obtidos com o uso desta técnica originalmente descrita por LIANG e PARDEE (1992) (Figura 03), foram importantes para a detecção de duas limitações dentro dos nossos objetivos: (i) a maioria das ESTs geradas foram de pequeno tamanho (< 100bp) e (ii) as baixas condições de rigor na PCR para a geração das mini bibliotecas resultaram em uma grande quantidade de ESTs que não apresentavam identidade com a região *3'UTR* próximo da cauda poli-A, porção usada para ancorar o oligo-dT. O primeiro problema pode ter sido causado pela maior facilidade de clonagem de pequenos fragmentos de produtos de PCR, nesse sentido, uma possível solução é a seleção, com relação ao tamanho, dos fragmentos gerados na reação de PCR, através de eletroforese em gel de agarose (*Size*

Selection). O segundo problema observado foi a pequena porcentagem (<10%) de ESTs que apresentavam os sinais de poliadenilação mais freqüentes presentes nesta região, que são: AAATAA, AAATTA e AATAAT. Segundo trabalho desenvolvido por GAUTHERET et al. (1998), onde foram agrupadas 164,000 ESTs em 15,000 *clusters* representando a região 3' do transcrito, cerca de 83% dos *clusters* gerados continham sinal de poliadenilação. O gráfico de distribuição posicional (figura 04) também confirma a baixa quantidade de ESTs provenientes da região 3' do gene.

5.1.2 Alterações na extração de RNA e no tamanho dos iniciadores:

Na tentativa de aumentar o rendimento de seqüências da região 3' *UTR* foi alterada a forma de extração de RNA, que passou a ser realizada através do *kit Quick Prep mRNA Purification* (Pharmacia) que utiliza resina contendo poli-T, purificando portanto apenas transcritos contendo a cauda poli A. Além disso, foi desenhado e sintetizado um iniciador do tipo $T(25)_{CG}$, que por ser maior, permite um aumento na temperatura de pareamento no momento da reação de amplificação dos fragmentos (PCR), permitindo assim, uma reação de maior estringência, o que resulta na produção de seqüências provenientes preferencialmente da porção *3' UTR*.

No entanto, o kit sugerido para extração de RNA não permite um rendimento suficiente para o tratamento com DNase, resultando em uma persistente contaminação com DNA genômico, que pode ser detectada tanto nos controles negativos das reações (figura 05) quanto na submissão das ESTs no programa *RepeatMasker* (tabela 07), cujos dados mostraram uma quantidade de elementos repetitivos maior que 20%. Segundo dados de MAKALOWSKI et al. (1998) a porcentagem de elementos repetitivos presentes na região *3'UTR* de transcritos humanos é cerca de 9,9%. Além disso, a eficiência de amplificação dos fragmentos de cDNA diminuiu consideravelmente com a introdução do iniciador do tipo $T(25)_{CG}$. Outra alteração proposta com o objetivo de automatizar e minimizar os custos na etapa de seqüenciamento foi a geração de seqüências a partir de produto de PCR dos clones, o que suprime as etapas de crescimento dos clones e

purificação dos plasmídeos. Porém os iniciadores do tipo $T(25)_{CG}$ possuem uma repetição muito grande do nucleotídeo T (timina), que provoca erro de extensão do fragmento, feito pela *Taq* DNA polimerase no momento da PCR dos clones. Este tipo de erro, que também ocorre em regiões repetitivas (microsatélites) do DNA genômico, é denominado *"slippage"*, que basicamente, se refere a um deslize que ocorre entre as cadeias de DNA durante a replicação celular, resultando na deleção ou inserção de um ou mais nucleotídeos preferencialmente nas regiões repetitivas. Este erro dificulta o seqüenciamento a partir do produto de PCR dos clones, uma vez que ele provoca a geração de fragmentos de diferentes tamanhos, e conseqüentemente, de seqüências diferentes. Visando resolver este problema, foi usado outro tipo de *Taq* DNA polimerase, a *rTth (rTth DNA Polymerase kit GeneAmp XL PCR – Perkin-Elmer)* originada do organismo *Thermus thermophilus* e obtida por engenharia genética. Esta enzima possui atividade exonucleásica 3' → 5' (*Proofreading*) que minimiza a geração de erros causados pelo *slippage* resultando em uma taxa de erro muito menor que a *Taq* DNA polimerase comum, porém esta modificação não foi eficaz para a solução deste problema, provavelmente devido ao grande tamanho da região repetitiva (25 nucleotídeos T).

5.1.3 Os resultados evidenciam promiscuidade no pareamento dos iniciadores:

Apesar dos problemas encontrados, foi feito um estudo em um lote de 58 ESTs para verificar a freqüência do pareamento dos nucleotídeos (*matchs*) ao longo do oligonucleotídeo $T(25)_{CG}$ nas reações de síntese de cDNA (RT) e de amplificação dos fragmentos (PCR) (fig 07). O estudo destas ESTs mostrou uma grande promiscuidade no pareamento dos iniciadores, evidenciada por uma elevada taxa de *"mismatchs"* nos dois últimos nucleotídeos da posição 3' no iniciador do tipo $T(25)_{CG}$. Esta alta taxa de *"mismatches"* pode ser causada pela baixa temperatura de extensão, (42°C) requerida pela enzima transcriptase reversa (*SuperScript II* Gibco-BRL), que eventualmente, também será a temperatura de pareamento dos iniciadores. Semelhante resultado foi obtido por FROST e GUGGENHENIM

(1999), que sintetizou cDNA partindo de mRNA de β-Actina usando 12 iniciadores ancorados do tipo $T(12)_{MN}$ (onde M = A, C, ou G e N = A, C, T ou G) combinados com um iniciador específico (5' TGGAGAAGAGCTACGAGCTGCCTG 3'). Usando esta estratégia, era esperado que apenas um dos iniciadores encontrasse a seqüência alvo e amplificasse um fragmento de 1067 pb, porém 11 dos 12 iniciadores amplificaram o fragmento. Após a confirmação por seqüenciamento, foi concluído que o erro que causa esta promiscuidade no pareamento do iniciador ocorre durante a síntese de cDNA.

Para avaliar o grau de normalização das ESTs geradas nesta etapa, foram selecionadas ESTs que apresentaram: (i) mais de 90% de identidade contra ESTs humanas, (ii) pelo menos um iniciador presente na seqüência gerada, (iii) sinal de poliadenilação, dos tipos AAATAA, AAATTA ou AATAAT, presentes nos primeiros 40 nucleotídeos a partir da cauda poli A. Desta forma, foram selecionadas 73 ESTs (não redundantes) provenientes da região *3' UTR* do transcrito. Com estas ESTs selecionadas foi feito um estudo buscando *clusters* de ESTs no UniGene. Teoricamente a quantidade de entradas em cada *cluster* deste banco é proporcional ao nível de expressão do gene, desta forma, um *cluster* com 1500 ESTs significa que fragmentos deste gene foram seqüenciados 1500 vezes. Os resultados gerados indicavam que a metodologia utilizada estava falhando na normalização parcial das mini bibliotecas, uma vez que grande parte das ESTs selecionadas se encaixavam em *clusters* de genes altamente expressos (curva do gráfico desviada para direita), sendo o valor médio de entradas nestes *clusters* de 355 (fig 08). Para comparar os resultados do estudo de normalização, foram colocados resultados de bibliotecas construídas utilizando outras estratégias, as quais foram (i) não normalizada e (ii) normalizada, derivadas de mama, onde os valores médios de entradas nos *clusters* do UniGene foram respectivamente de 184 e 119. A falha na normalização indicada por esta análise pode ser conseqüência da promiscuidade no pareamento dos iniciadores, fazendo com que a

prevalência do transcrito passe a ser um fator determinante de sua representação na biblioteca.

5.1.4 Maior rigor na avaliação da qualidade do mRNA e na síntese de cDNA:

Devido a problemas como: (i) persistente contaminação por DNA genômico, (ii) promiscuidade no pareamento dos iniciadores comprometendo a normalização parcial das mini-bibliotecas e (iii) dificuldades na etapa de seqüenciamento geradas pelo iniciador que contém seqüência repetitiva dos nucleotídeos timina (T); novas alterações foram propostas. A saber: (i) avaliação do mRNA utilizando métodos e controles de qualidade mais rigorosos, (ii) utilização de oligo dT (20) e uma nova enzima na síntese de cDNA, a *ThermoScript* (Gibco-BRL), cuja temperatura de extensão é de 50°C, (iii) amplificação dos fragmentos de cDNA pela reação de PCR voltando a utilizar os iniciadores do tipo $T(11)_{VN}$, sugeridos no projeto inicial. Com a temperatura de extensão a 50°C e o tamanho do Oligo dt (20) usado, a síntese de cDNA passou a ser mais específica. Nestas condições de rigor, o alvo do iniciador (Oligo dt(20)) será a cauda poli A, na região *3´UTR* do gene, e não contínuos de adeninas dispostos ao longo do gene. A volta do uso dos iniciadores $T(11)_{VN}$, tem o objetivo de permitir uma melhor eficiência na reação de amplificação dos fragmentos de cDNA, além de minimizar o problema de *Slippage* que ocorre no produto de PCR dos clones durante a geração de DNA molde para o seqüenciamento, proporcionando um melhor rendimento nesta etapa. Entretanto, esta alteração elimina a normalização parcial a partir da síntese de cDNA, uma vez que a porção 3' do iniciador usado não possui os dois últimos nucleotídeos degenerados. Desta forma, a normalização parcial dos transcritos agora é feita apenas na amplificação dos fragmentos de cDNA (PCR). Porém, a análise de pareamento dos iniciadores demostrou uma diminuição da promiscuidade no pareamento dos nt (Figura 13), o que pode ter proporcionado uma melhora na normalização das bibliotecas (Figura 14). A nova metodologia está esquematizada na figura 22.

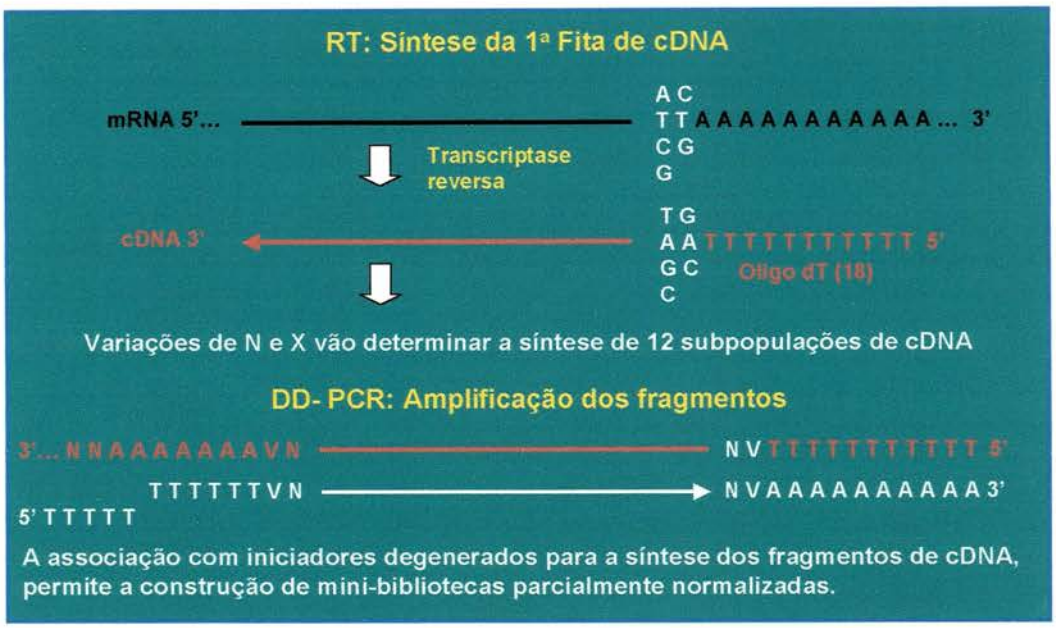


Figura 22: Representação esquemática da amplificação de fragmentos de cDNA visando a obtenção preferencial da porção 3' *UTR*.

Os iniciadores usados, do tipo $T(11)_{VN}$ onde V = A, C ou G e N = A, C, G ou T, proporcionam a vantagem de identificar os cDNAs de maneira parcialmente independente da sua prevalência na amostra estudada, devido aos dois nucleotídeos degenerados (V e N) na porção 3', permitindo assim, a normalização parcial dos transcritos. Devido à grande quantidade do nucleotídeo timina (T-11) nos iniciadores, teóricamente, a amplificação ocorre preferencialmente na região 3' *UTR* do gene, devido à grande quantidade de contínuos de adeninas (A) nesta região, o que gera uma grande quantidade de sítios de pareamento para os iniciadores usados (T_{11vn}).

Desta forma, foram construídas 12 mini bibliotecas a partir de carcinoma de mama, carcinoma de cólon e PNET que geraram 824 seqüências (tabela 13). Destas, 601 apresentaram qualidade (análise visual do cromatograma) suficiente para submissão contra os bancos de dados sendo que, 205 ESTs (51%) apresentaram sinal de poliadenilação e 245 ESTs (61,7%) apresentaram identidade contra ESTs depositadas no banco de dados dbEST, geradas a partir da região 3'*UTR*, resultando em 284 (71,5%) seqüências com alguma evidência de estarem ancoradas na região 3' dos transcritos. Estas 284 ESTs, que apresentaram evidências de serem provenientes da região 3' do transcrito, foram analisadas no UniGene (classe 6), destas, 70 ESTs representavam o mesmo gene mais de uma vez (classe 8) e 28 representavam ESTs ainda não agrupadas em clusters do UniGene (classe 10). A última classe (classe 11), representa a quantidade de ESTs

(15 ESTs) que não apresentaram identidade contra os bancos de dados, representando assim possíveis novos transcritos identificados.

Desta forma, um total de 186 ESTs não redundantes geradas a partir das mini-bibliotecas dos tumores de mama, cólon e PNET, foram selecionadas para submissão contra *clusters* do UniGene para fazer a análise de normalização. O resultado desta análise sugeriu que a atual metodologia para geração de ESTs estava funcionando adequadamente (figura 14 e tabela 14), uma vez que grande parte das ESTs selecionadas encaixavam em *clusters* de genes que apresentam baixo ou médio nível de expressão (curva do gráfico desviada para esquerda), sendo o valor médio de entradas nos *clusters* de 135. Para bibliotecas não normalizada e normalizada, derivadas de mama os valores médios de entradas foram respectivamente 239 e 194. No entanto, sabemos que esta não é a forma mais correta para fazer este estudo comparativo, uma vez que os tecidos usados são diferentes. Porém, não existem bibliotecas de cDNA destes tecidos com número de ESTs que permita uma análise deste tipo.

Estas análises sugerem que as mudanças foram positivas nesse sentido. Entretanto, de acordo com os resultados das análises descritos na classe 03 (tabela 13), a quantidade de seqüências redundantes estava alta, 33% (205 ESTs). Este fato pode ser explicado pela necessidade de reamplificação dos perfis gerados com os iniciadores $T(11)_{VA}$ e $T(11)_{VT}$ os quais, após tal procedimento, geram um perfil de bandamento menos complexo (dados não mostrados). Na tentativa de resolver este problema, foram produzidos perfis usando diferentes combinações de iniciadores o que produziu perfis mais complexos (Figura 15). Alternativamente, este problema poderia ter sido resolvido através da construção de iniciadores contendo algumas degenerações em diferentes posições, porém esta estratégia seria mais útil em um projeto semelhante que busca a produção de quantidades maiores de ESTs.

5.2 Análise dos dados com o auxílio do Laboratório de bioinformática:

Uma vez que a estratégia parecia funcionar adequadamente, foi enfatizada nesta parte do trabalho, a produção de ESTs dos tumores já em estudo e também de uma nova coleção de tumores de rara incidência. Neste sentido foi necessário analisar todas as ESTs geradas a partir da nova metodologia, inclusive as ESTs produzidas inicialmente e analisadas manualmente (tabela 13), com o auxílio do laboratório de bioinformática. A automatização desta etapa de análise é importante para, (i) controlar a qualidade das seqüências geradas (ii) excluir as seqüêcias dos vetores e (iii) automatizar a análise contra os bancos de dados, aumentando a velocidade e confiabilidade dos dados obtidos. Os resultados gerados estão descritos na tabela 16, onde foram submetidas 3,337 seqüências derivadas de 07 tumores, destas 2,581 foram aprovadas quanto à qualidade pelo programa *Phred*, podendo assim, entrar para uma rotina de submissão contra os bancos de dados. Em um primeiro passo, as ESTs foram submetidas ao programa *RepeatMasker* para mascarar regiões repetitivas. Após a análise pelo programa *RepeatMasker* as seqüências foram classificadas, de acordo com a identidade apresentada, em uma das seguintes categorias (i) seqüências de bactérias, DNA mitocondrial e RNA ribossomal, (ii) cDNAs humanos totalmente seqüenciados, (iii) cDNAs humanos parcialmente seqüenciados, (iv) ESTs ainda não agrupadas em *clusters* do UniGene (dbEST); as ESTs que não apresentaram identidade em nenhum destes bancos foram classificadas na categoria *"No match"*.

5.2.1 Análise de normalização:

A análise do grau de normalização das ESTs geradas, foi feita com base nas ESTs que apresentaram identidade contra *clusters* de transcritos totalmente seqüenciados (*Human Genes* Tabela 14), onde o valor médio do número de entradas foi de 194. Para comparar o nível de normalização da estratégia, foi feita uma análise comparativa com outras duas bibliotecas, uma não normalizada e outra normalizada, ambas produzidas pelo grupo do Dr. Bento Soares a partir de tumor de mama, o valor médio do número de

entradas nos *clusters* são respectivamente 444 e 255. De acordo com os gráficos (figuras 16a e 16b) é possível perceber que a estratégia desenvolvida neste projeto gera resultados bem próximos das bibliotecas normalizadas, uma vez que a quantidade de ESTs que apresentam identidade contra *clusters* com maior número de entradas é baixo. Outro fato que também corrobora isto é a diferença do percentual de ESTs no *cluster* de intervalo 0-50 (nº de entradas neste *cluster*) entre as 3 bibliotecas. Neste tamanho de *cluster,* podemos perceber que as bibliotecas normalizadas e 3' *UTR*-PCR possuem valores bem próximos, cerca de 18%; enquanto a biblioteca não normalizada possui este valor em torno de 10%.

5.2.2 Análise de redundância:

A análise de redundância foi feita separadamente para os diferentes tecidos estudados. Não foi observada nenhuma relação direta entre a quantidade de ESTs geradas e o nível de redundância. Apesar da variação nos diferentes tecidos, a média do nível de redundância é relativamente aceitável (0,5%) devido ao fato da estratégia ser dependente de iniciadores que têm como alvo preferencial a região 3' do gene.

5.2.3 Análise posicional:

A análise posicional foi feita baseada em 100 ESTs não redundantes que apresentaram identidade com genes humanos totalmente sequenciados (tabela 14 categoria *Human genes* e tabela 15). Esta análise tem como objetivo mostrar a posição das ESTs geradas em transcritos totalmente seqüenciados. Como era de se esperar a maioria das ESTs geradas estão apresentando identidade contra as regiões 3'*UTR* , o que pode ser explicado pela quantidade de adeninas e de timinas presentes nesta região, favorecendo desta forma o pareamento dos iniciadores usados para a construção das mini bibliotecas.

5.2.4 Prováveis novos genes identificados:

As ESTs que não apresentaram identidade com seqüências depositadas nos bancos de dados, aqui classificadas na categoria *No match*

(Tabela 16), foram analisadas manualmente para checar a probabilidade de representarem possíveis novos genes humanos. Esta análise levou em consideração os seguintes aspectos: (i) percentual de nucleotídeos mascarados menor que 50% (*RepeatMaked*), (ii) qualidade da seqüência (*high quality sequence*) e (iii) a existência de sinal de poliadenilação. Segundo esta análise (Tabela 19). Embora o percentual seja variado, todos os tumores de baixa incidência apresentaram uma quantidade maior de prováveis novos genes, em comparação com os tumores de cólon e mama, que foram usados como controle. Os testes estatíscos do x^2 e de Fisher demostraram que os valores foram significativos para os tecidos com maior número de ESTs analisadas.

5.2.4 Re-análise das ESTs classificadas na categoria "No match":

Devido ao grande volume de seqüências que são depositadas diariamente e também ao término do primeiro rascunho da decodificação do genoma humano, uma nova análise das ESTs anteriormente classificadas como *"No match"* se faz necessária. Esta análise foi feita manualmente em 186 ESTs (cerca de 2/3 das ESTs classificadas nesta categoria), utilizando o programa Blast contra os bancos de dados nr, dbEST categoria *Human ESTs* e HTGS (*High-Throughput Genome Sequence*), sendo útil para atualização dos dados gerados. Segundo esta atualização, 32% das ESTs (60) passaram a apresentar identidade contra cDNAs humanos recentemente identificados. Destas ESTs, grande parte foi proveniente dos tumores mais estudados (mama/cólon), sugerindo a prevalência de genes mais raros ou específicos em tumores de baixa incidência (Tabela 20). Em 109 casos que não apresentaram identidade com seqüências de cDNA, 52 apresentaram identidade com seqüências de DNA, confirmando a origem humana das seqüências geradas (Figura 20). Este resultado pode ser devido a uma contaminação por DNA genômico, ou ainda sugerir que a seqüência destes genes ainda não foi definida até o momento, reforçando a importância do seqüenciamento dos transcritos para a identificação e mapeamento de genes.

Cerca de 200 ESTs com um tamanho médio de 200 nt foram submetidas ao banco de dados de proteínas através do programa Blast, não apresentando identidade significativa em nenhuma delas. Segundo Pesole at al (1994), a região 3' UTR de transcritos humanos têm em média 520 pb. Estes dados corroboram a preferência da metodologia, desenvolvida neste projeto, em gerar seqüências da região 3' UTR.

6 - Conclusões:

A adaptação da técnica de DDRT-PCR para a construção de mini bibliotecas de cDNA e geração de ESTs permitiu a normalização parcial das mini bibliotecas e geração de seqüências preferencialmente da região 3' do gene. A otimização da metodologia ainda é necessária, uma vez que o índice de redundância é relativamente alto.

Os resultados sugerem que os tumores de baixa incidência permitem identificar um maior número de novos genes humanos.

A associação dos resultados gerados pela técnica, 3 'UTR – PCR, desenvolvida neste trabalho e a técnica ORESTES poderá ser de grande ajuda para a obtenção da seqüência completa de transcritos raros.

A associação dos dados gerados neste trabalho aos dados disponibilizados pelo sequenciamento do genoma humano poderão facilitar o mapeamento e o seqüenciamento completo de genes ainda desconhecidos.

7- REFERÊNCIAS BIBLIOGRÁFICAS

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H et al. Complementary DNA sequencing: Expressed Sequence Tags and human genome project. **Science** 1991; 252:1651-6.

Adams MD, Kerlavage AR, Fleischman RD, Fuldner RA, Bult CJ, Lee NH, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. **Nature** 1995; 377 Suppl:3-174.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE et al.The genome sequence of *Drosophila melanogaster*. **Science** 2000; 287:2185-95.

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD, editors. **Molecular Biology of the cell**. 3nd ed. New York: Garland Pu; 1994. p.394-5.

Aparicio SAJR. How to count...human genes. **Nature Genet** 2000; 25:129-30.

Bishop JO, Morton JG, Rosbash M, Richardson M. Three abundance classes in HeLa cell messenger RNA. **Nature** 1974; 250:199-204.
Boguski MS, Schuler GD. ESTablishing a human transcript map. **Nature Genet** 1995; 10:369-75.

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

Chomczynski P, Sacchi N. Single-step method of RNA isolation by Acid Guanidinum Thiocyanate-Phenol-Chloroform Extraction. **Anal Biochem** 1987; 162:156-9.

Cohen AM, Minsk BD, Schilsky RL. Cancer of the colon. In: De Vita Jr VT, Hellman S, Rosenberg AS, editors. **Cancer principles & practice of oncology**. 5th ed. Philadelphia: Lippincott-Raven; 1997. p.1144-84.

Collins F, Galas D. A new five-year plan for the U.S. human genome project. **Science** 1993; 262:43-7.

Constanzo F, Castagnoli L, Dente L, Arcari P, Smith M, Constanzo P et al. Cloning of several cDNA segments coding for human liver proteins. **EMBO J** 1983; 2:57-61.

Costa CML, de Camargo B. Tumores abnominais. In: de Camargo B, Lopes LF, editors. **Pediatria oncologica:** noções fundamentais para o pediatra. São Paulo: Lemar; 2000. p.119-34.

Costa CML, Mendes WL, Alves L, Penna V. Tumores osseos. In: de Camargo B, Lopes L F, editors. **Pediatria oncologica:** noções fundamentais para o pediatra. São Paulo: Lemar; 2000. p.161-73

Crollius HR, Jaillon O, Bernot A, Dasilva C, Borneau L, Fischer C et al. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. **Nature Genet** 2000; 25:235-8.

Davidson EH, Britten RJ. Regulation of gene expression: possible role of repetitive sequences. **Science** 1979; 204:1052-9.

De Vries CJM, Van Achterberg TAE, Horrevoets AJG, Ten Cate JW, Pannekoek H. Differential display identification of 40 genes with altered

expression in activated human smooth muscle cells: local expression in atherosclerotic lesions of smags, smooth muscle activation-specific genes. **J Biol Chem** 2000,275 (31): 23939-47.

Dias Neto E, Rahal P, Moura RP, Caballero O, Villa LL, Simpson AJG. DNA quality control by competitive PCR. **TIG** 1996; 12:341-2.

Dias Neto E, Correa GR, Verjovski-Almeida S, Briones MRS, Nagai MA, Silva Jr W et al. Shotgun sequencing of the human transcriptome with ORF Expressed Sequence Tags. **Proc Natl Acad Sci USA** 2000a; 97:3491-6.

Dias Neto E, Simpson A J G, Lopes L F. Biologia Molecular dos tumores infantis. In: Camargo B, Lopes L F, editors. **Pediatria oncologica:** noções fundamentais para o pediatra**.** São Paulo: Lemar; 2000b. p 29-45.

Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M et al. The DNA sequence Of human chromosome 22. **Nature** 1999; 402:489-95.

Ennziger FM, Weiss SW. **Soft tissue tumors**. 3rd ed. St. Louis: Mosby, 1995. p.381-430: Benign lipomatous tumors.

Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. **Science** 1999; 286:487-91.

Ewing B, Green P. Analysis of Expressed Sequence Tags indicates 35,000 human genes. **Nature Genet** 2000; 25:232-4.

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

Fields C, Adams MD, White O, Venter JC. How many genes in the human genome? **Nature Genet** 1994; 7:345-6.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science** 1995; 269:496-512.

Frost MR, Guggenhenim JA. Mammalian polyadenilation sites: implications for differential display. **Nucleic Acids Res** 1999; 27:1386-91.

Gautheret D, Poirot O, Lopez F, Audic S, Claverie JM. Alternative polyadenilation in human mRNA: a large-scale analyses by EST clustering. **Genome Res** 1998; 8:524-30.

Gilbert W. A vision of the grail. In: Kevles DJ, Hood AL, editors. **The code of codes**: scientific and social issues in the human genome project. Harvard: Harvard University; 1992. p.83-97.

Green DM, Coppes MJ, Breslow NE, Grundy PE, Ritchey ML, Beckwith JB, Thomas PRM, D'Ángio GJ. Wilms tumor. In: Pizzo PA, Poplack DG, editors. **Principles and practice of pediatric oncology**. 3rd ed. Philadelphia: Lippincott-Raven; 1997. p.733-60.

Greenberg M, Filler RM. Hepatic tumors. In: Pizzo PA, Poplack DG, editors. **Principles and practice of pediatric oncology**. 3rd ed. Philadelphia: Lippincott-Raven. 1997. p.717-32.

Harris J, Morrow M, Northon L. Malignant tumors of the breast, editors. In: De Vita Jr VT, Hellman S, Rosenberg AS, editors. **Cancer principles & practice of oncology**. 5th ed. Philadelphia: Lippincott-Raven. 1997. p.1557-602.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. **Nature** 2000; 405:311-9.

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

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

Lamerdin JE, Athwal RS, Kansara MS, Sandhu AK, Pantajali SR, Weissman SM, Carrano AV. Chromossomal localization and expressed sequence tag generation of clones from a normalized human adult thymus cDNA library. **Genome Res** 1995, 5:359-7

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

Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by menas of polymerase chain reaction. **Science** 1992; 267:967-71.

Liang P, Averboukh L, Pardee AB. Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. **Nucleic Acids Res** 1993; 21:3269-75.

Lopes LF, de Camargo B. Tumores ovarianos e testiculares. In: de Camargo B, Lopes LF, editors. **Pediatria oncologica:** noções fundamentais para o pediatra. São Paulo: Lemar; 2000. p.135-47.

Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. **Genome Res** 1998; 95:9407-12.

Marshall E. A high-stakes gamble on genome sequencing. **Science** 1999; 284:1906-9.

McClelland M, Mathieu-Daude F, Welsh J. RNA fingerprinting and differential display using arbitrarily primed PCR. **TIG** 1995; 11:242-6.

Miser JS, Triche TJ, Kinsella TJ, Pritchard DJ. Other soft tissue sarcomas of childhood. In: Pizzo PA, Poplack DG, editors. **Principles and practice of pediatric oncology**. 3rd ed. Philadelphia: Lippincott-Raven; 1997. p.865-88.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ et al. A whole-genome assembly of *Drosophila*. **Science** 2000; 287:2196-204.

Ozaki LS, Czeko YMT. Genomic DNA cloning related techniques. In: Morel CM, editor. **Genes antigens of parasites:** a laboratory manual. Rio de Janeiro: UNDP-Word Bank-Who and Fundação Oswaldo Cruz; 1984; 165-85.

Pesole G, Fiomarino G e Saccone C. Sequence analysis and compositional properties of untranslated regions of human mRNAs. Gene 1994; 140: 219-225.

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

Reeves RH. Recounting a genetic story. **Nature** 2000; 405:283-4.

Ribeiro KCB, de Camargo B, Torloni H, editores. **Registro hospitalar de câncer pediátrico 1988 & 1994**. São Paulo: Fundação Antônio Prudente; 1999.

Robbins SL, Cotran RS, editores. **Patologia estrutural e funcional**. 2 ed. Rio de Janeiro: Interamericana, 1993. p.157-77: Aspectos clínicos da neoplasia.

Roses AD. Pharmacogenetics and the practice of medicine. **Nature** 2000; 405:857-65.

Sambrook J, Fritsh EF, Maniats T. Appendix E: commonly used techniques in molecular cloning. In: Sambrook J, Fritsh EF, Maniats T, editors. **Molecular cloning: a laboratory manual**. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989a. p.8.3-8.82.

Sambrook J, Fritsh EF, Maniats T. Construction and analysis of cDNA cloning. In: Sambrook J, Fritsh EF, Maniats T, editors. **Molecular cloning: a laboratory manual**. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989b. p.8.3-8.82.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci USA** 1977; 74:5463-7.

Sanguinetti CL, Dias Neto E, Simpson AJG. Rapid silver staining and recovery of PCR products separated on polyacrylamid gels. **Biotechniques** 1994; 17:915-21.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K et al. A Gene map of human genome. **Science** 1996; 274:540-6.

Stracham T, Read AP, editors. **Human molecular genetics**. Oxoford: BIOS Scientific Pu; 1996a: 1-31: DNA estructure and function.

Stracham T, Read AP, editors. **Human molecular genetics**. Oxoford: BIOS Scientific Pu; 1996b: 241-75: Mutation and instability of Human DNA.

Stracham T, Read AP, editors. **Human molecular genetics**. Oxoford: BIOS Scientific Pu; 1996c: 335-67: The human genome project.

Stracham T, Read AP, editors. **Human molecular genetics**. Oxoford: BIOS Scientific Pu; 1996d:183-202: Human multigene families and repetitive DNA.

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

Vaughan D, editor. **To know ourselves**. California: The U.S. Departament of Energy and The Human Genome Project; 1996.

Vedoy CG, Bengtson MH, Sogayar MC. Hunting for differentially expressed genes. **Braz J Med Biol Res** 1999; 32:877-84.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. **Science** 1995; 270:484-6.

Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C et al. Analysis of human transcriptomes. **Nature Genet** 1999, 23:387-8.

Zar JH, editor. Biostatistical Analysis. 3rd edition. New Jersey: Printice Hal, 1996. p 521-7: Sampling a binomial population.

Wan JS, Sharp SJ, Poirier GM-C, Wagaman PC, Chambers J, Pyati J et al. Cloning differentially expressed mRNAs. **Nat Biotechnol** 1996; 14:1685-91.

Welsh J, McClelland M. Fingerprinting genomes using PCR with arbitrary primers. **Nucleic Acids Res** 1990; 18:7213-8.

Welsh J, Chada K, Dalal SS, Cheng R, Ralph D, McClelland M. Arbitrarily primed PCR fingerprinting of RNA. **Nucleic Acids Res** 1992; 20:4965-70.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Res** 1990; 18:6531-5.

8- OBRAS CONSULTADAS

Centers for Disease Control, World Helth Organization. Epi Info. Epidemiologia em microcomputadores: um sistema de processamento de texto, banco de dados e estatísticas [programa de computador]. versão 6,04b. Atlanta: OPAS/WHO; 1997.

http://www.celera.com/

http://www.ludwig.org.br /ORESTES/

http://www.merck.com/mrl/merck_gene_index.2.html

http://www.ncbi. nlm.nih.gov/ CGAP/hTGI/sumtab/cgapba.cgi

http://www.ncbi.nlm.nih.gov/ UniGene/

http://www.ncbi.nlm.nih.gov/

http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=1

http://www.ncbi.nlm.nih.gov/cgap

http://www.ncbi.nlm.nih.gov/CGAP/hTGI/sumtab/cgapba.cgi

http://www.ncbi.nlm.nih.gov/dbEST/

http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

http://www.ncbi.nlm.nih.gov/ncicgap/

http://www.ncbi.nlm.nih.gov/unigene)

http://www.ncbi.nlm.nih.gov/Web/Newsltr/aug96.html

http://www.ornl.gov/hgmis/

http://www.ornl.gov/TechResources/HumanGenome/hg5p/goal.html

http://www.sanger.ac.uk/

http://www.tigr.org/tdb/tgi.shtml