BIBLIOTECA
INSTITUTO DE QUÍMICA
Universidade de São Paulo
t: 19.629
UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

15211-102-1-28/10-1
15211-102-1-28/10-1
OF CIACS

**GERAÇÃO DE “ETIQUETAS DE SEQUÊNCIAS EXPRESSAS” DIRIGIDAS
PARA PORÇÕES CODIFICADORAS DOS GENES (ORESTES) :
IDENTIFICAÇÃO DE NOVOS GENES HUMANOS
EXPRESSOS EM CÂNCER DE MAMA.**

RICARDO GARCIA CORRÊA

TESE DE DOUTORADO

**ORIENTADOR: RICARDO R. BRENTANI
CO-ORIENTADOR: ANDREW J. G. SIMPSON**

SÃO PAULO

- 2001 -

DEDALUS - Acervo - CQ

30100003896

Ficha Catalográfica

Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

C824g Corrêa, Ricardo Garcia
Geração de "etiquetas de seqüências expressas" dirigidas
para porções codificadoras dos genes (ORESTES): identificação
de novos genes humanos expressos em câncer de mama /
Ricardo Garcia Corrêa. -- São Paulo, 2001.
142p

Tese (doutorado) - Instituto de Química da Universidade
de São Paulo. Departamento de Bioquímica.
Orientador : Brentani, Ricardo Renzo
Co-orientador : Simpson, Andrew John George

1. Biologia molecular 2. Genética 3. Bioquímica I. T.
II. Brentani, Ricardo Renzo, orientador III. Simpson,
Andrew John George, co-orientador.

574.88 CDD

“Geração de “Etiquetas de Sequências Expressas” Dirigidas Para
Porções Codificadoras dos Genes (Orestes): Identificação de Novos
Genes Humanos Expressos em Câncer de Mama”

RICARDO GARCIA CORRÊA

*Tese de Doutorado submetida ao Instituto de Química da
Universidade de São Paulo como parte dos requisitos necessários à obtenção do
grau de Doutor em Ciências - Área: Bioquímica.*

Aprovado por:

Prof. Dr. RICARDO RENZO BRENTANI
IQ - USP
(Orientador e Presidente)

Prof. Dr. SEGIO VERJOVSKI DE ALMEIDA
IQ - USP

Profª. Dra. MARI CLEIDE SOGAYAR
IQ - USP

Prof. Dr. LUIS EDUARDO SOARES NETO
IB - USP

Profª. Dra. ELIZABETH ANGÉLICA LEME MARTINS
IB

SÃO PAULO
16 DE FEVEREIRO DE 2001.

“Sonhar, apesar das decepções.
Caminhar, apesar dos obstáculos.
Lutar, apesar das barreiras.
Acreditar, acima de tudo.”

Autor desconhecido.

“Quando Deus quer
enlouquecer alguém, satisfaz todos
os seus desejos.”

Paulo Coelho, *As Valquírias*

“A ciência prometeu a felicidade?
Ela prometeu a verdade e a questão é saber
se algum dia faremos felicidade com verdade”

Émile Zola, *Discours aux étudiants*
Paris, 1893

“Somos uma terrível mistura de ácidos nucleicos
e lembranças, de desejos e de proteínas.”

François Jacob, *O rato, a mosca e o homem.*

Ao meu avô
Oswaldo Corrêa,
pelo exemplo de vida desenhado
em inesquecíveis lembranças.

À minha amada
Katia,
pela vida que me proporciona
e me impulsiona a viver.

Ao professor
Andrew J. G. Simpson,
pela amizade, confiança
e valorização da ciência brasileira.

AGRADECIMENTOS

Mais do que uma tese pessoal de Doutorado, o presente trabalho foi fruto da intensa colaboração de vários cientistas que, em suas respectivas áreas de atuação, deram o suporte mais do que devido à conclusão desta tese. A competência científica e política destes pesquisadores foi responsável pelo meu aprendizado.

Aos pais Oswaldo e Eliana, pela minha vida e pela oportunidade de estudar e desenvolver meus estudos no ramo científico.

Ao professor Ricardo R. Brentani, pela oportunidade única de trabalhar em uma das maiores instituições de Pesquisa no Brasil - o Instituto Ludwig de Pesquisa sobre o Câncer.

Ao amigo Emmanuel D. Neto, pelo suporte (prático e teórico) na técnica de RAP-PCR e identificação gênica.

Ao amigo Sandro J. de Souza e grupo de Bioinformática, pelas análises computacionais das ORESTES geradas neste trabalho e real colaboração na caracterização de novos genes humanos.

Ao amigo Ricardo P. de Moura, pelo apoio na produção de mini-bibliotecas e geração de ORESTES e, principalmente, pelas longas conversas (e risadas) dadas durante os anos “genômicos”.

Aos professores Bayardo B. Torres, Fernando F. Costa, Fernando Soares, Humberto Torloni, Marcelo R. S. Briones, Marco A. Zago, Mari C. Sogayar, Maria A. Nagai, Maria M. Brentani, Luiz F. L. Reis, Luiza L. Villa e Sergio V. Almeida, pelos ensinamentos e auxílio na realização deste trabalho.

Aos amigos do Instituto Ludwig: Adriana Abalen, Adriana Mercadante, Alex, Anamaria, Anna Christina, Ana Paula, Andrea, Aristóbolo, Beatriz, Carla, Carlinhos, Carol, Catarina, Daniel (Chuck), Edina, Eduardo Abrantes, Eliane, Elisangela, Eloisa, Emozila (Zilinha), Enrique, Sr. Eudes, Fabrício, Gaëlle, Jacqueline, Jane, João (Djoninho), José (Zezão), José Cláudio, Juliana, Kil, Laura, Lara, Marilza, Miyuki, Nídia, Lourdes, Otávia, Patrícia, Paulo Maciag (Paulinho), Paulo Oliveira (Paulão), Ricardo Gabriel, Rosa, Silvio Zanata, Sampaio, Sibeli, Valeska e Vilma, pela convivência diária, repleta de bons momentos e descontração.

Ao CNPq e FAPESP, pelo auxílio prestado.

Ao Deus que me acompanha, em cada passo dado pela estrada das provas e vicissitudes, meus mais sinceros agradecimentos.

ÍNDICE

ABREVIACÕES.....	01
SUMMARY.....	04
RESUMO.....	05
1. INTRODUÇÃO	
1.1. Visão global do genoma humano.....	06
1.2. O Projeto Genoma Humano (PGH).....	09
1.2.1. Origem do PGH.....	09
1.2.2. Organização e estabelecimento do PGH no setor público.....	11
1.2.3. O Instituto de Pesquisa Genômica (TIGR).....	13
1.2.4. A contribuição do setor privado no PGH.....	16
1.3. Estratégias utilizadas no sequenciamento completo do genoma humano.....	17
1.4. O Instituto Genômico Virtual de São Paulo / Brasil (ONSA).....	19
1.5. O Projeto Genoma da Anatomia do Câncer (CGAP).....	19
1.6. “Etiquetas de sequências expressas” como ferramenta para os projetos genoma.....	21
1.7. A técnica de RAP-PCR na geração de “etiquetas de sequências expressas”.....	24
1.8. O câncer de mama como modelo experimental.....	27
2. OBJETIVOS.....	31
3. MATERIAIS E MÉTODOS	
3.1. Material biológico.....	32
3.2. Purificação do mRNA.....	32

3.3. Validação do mRNA.....	34
3.3.1. PCR específica para gene da β -globina.....	35
3.3.2. PCR específica para região D-loop de DNA mitocondrial.....	35
3.4. Síntese da primeira fita de cDNA.....	36
3.5. Obtenção de perfis de expressão por RAP-PCR.....	36
3.5.1. Obtenção de perfis X-X.....	38
3.5.2. Obtenção de perfis X-Y.....	38
3.5.3. Obtenção de perfis X-Xn.....	39
3.5.4. Obtenção de perfis radioativos para controle de qualidade.....	39
3.6. Ligação dos perfis de cDNA a vetores plasmidiais.....	40
3.7. Preparação de células bacterianas competentes.....	41
3.8. Produção de mini-bibliotecas de cDNA.....	41
3.9. Seleção das colônias de interesse.....	42
3.10. Mini-preparação de plasmídeos.....	43
3.11. Geração de ESTs.....	43
3.11.1. Sequenciamento em <i>ALFexpressTM DNA Sequencer</i> (Amersham Pharmacia Biotech).....	43
3.11.2. Sequenciamento em <i>ABI Prism 377TM DNA Sequencer</i> (Perkin Elmer Applied Biosystems).....	44
3.11.3. Sequenciamento em <i>MegaBaceTM DNA Sequencer</i> (Amersham Pharmacia Biotech).....	44
3.12. Validação de mini-bibliotecas de cDNA.....	44
3.13. Análise de similaridade das ESTs frente a bancos de dados de acesso remoto.....	45
3.14. Obtenção da sequência completa de novos genes humanos.....	46
3.14.1. Amplificação por RACE-PCR (<i>Rapid Amplification of Complementary DNA Ends</i>).....	46
3.14.2. Amplificação por <i>primers</i> desenhados a partir da análise racional dos bancos de dados.....	47
3.14.3. Sequenciamento e montagem de cDNAs.....	49

3.15. Caracterização preliminar de novos genes humanos.....	49
3.15.1. Ensaios quantitativos de RT-PCR.....	49
3.15.2. Mapeamento genético.....	50
3.15.3. Ensaios de <i>Northern blot</i>	51
 4. RESULTADOS	
4.1. Critérios de qualidade para preparação de mRNA.....	53
4.2. Produção de perfis de cDNA por RAP-PCR.....	56
4.3. Critérios de qualidade para produção de perfis por RAP-PCR.....	61
4.4. Geração de mini-bibliotecas de cDNA.....	63
4.5. Critérios de qualidade para a utilização de mini-bibliotecas de cDNA na geração de ESTs (validação de mini-bibliotecas).....	64
4.6. Geração de 6000 "etiquetas de sequências expressas" em câncer de mama.....	68
4.7. Análise computacional de normalização de bibliotecas de cDNA.....	70
4.8. Análise posicional das ESTs geradas ao longo dos respectivos genes humanos.....	71
4.9. Avaliação global de redundância das sequências expressas.....	73
4.10. Avaliação do percentual de singletons e potencial codificador das sequências " <i>no matches</i> ".....	74
4.11. Distribuição parcial de sequências expressas similares a genes humanos conhecidos em distintas classes gênicas.....	75
4.12. Identificação de novos genes humanos (ortólogos e parálogos) por análise de similaridade.....	79
4.13. Caracterização preliminar de novos genes humanos identificados por análise de similaridade.....	81
4.13.1. <i>NABC1_5B</i> - uma nova isoforma do gene <i>NABC1</i>	81
4.13.2. Gene humano semaforina 6B (<i>Sema 6B</i>).....	86
4.13.3. Gene humano <i>Notch 2</i>	97

5. DISCUSSÃO

5.1. Estabelecimento do número total de ESTs geradas e modelo experimental de estudo.....	106
5.2. Padronização metodológica e análise comparativa da qualidade de preparações de mRNA.....	107
5.3. Eficiência dos diferentes tipos de perfis de cDNA na geração de sequências expressas informativas e não-redundantes.....	110
5.4. Análise de similaridade das sequências expressas geradas.....	112
5.5. Normalização parcial das mini-bibliotecas de cDNA geradas.....	113
5.6. Análise posicional e redundância global das sequências expressas geradas.....	115
5.7. Utilização prática das sequências expressas geradas na caracterização de novos genes humanos.....	117
5.7.1. <i>NABC1_5B</i> - uma nova isoforma do marcador tumoral <i>NABC1</i>	117
5.7.2. Semaforina 6B (<i>Sema 6B</i>) - um novo gene humano da família de semaforinas.....	118
5.7.3. <i>Notch 2</i> - um novo gene humano com provável envolvimento na diferenciação e destino celular.....	121
5.8. Conclusões finais.....	124
6. REFERÊNCIAS BIBLIOGRÁFICAS.....	125
7. CURRICULUM VITAE.....	140
8. ARTIGOS CIENTÍFICOS.....	142

ABREVIACES

µg - micrograma

µl - microlitro

µM - micromolar

3'UTR - *3'untranslated region*

5'UTR - *5'untranslated region*

ATP - adenosina trifosfato

ATRA - *all-trans-retinoic acid*

BAC - *bacterial artificial chromosome* (Cromossomo artificial de bactria)

BLAST - *Basic Local Alignment Search Tool*

BSA - albumina de soro bovino

cDNA - DNA complementar

CGAP - *Cancer Genome Anatomy Project* (Projeto de Anatomia Genmica do Cncer)

DDBJ - *DNA Data Bank of Japan*

ddNTPs - didesoxirribonucleotdeos

DEPC - dietilpirocarbonato

DMSO - dimetilsulfxido

dNTP - desoxirribonucleotdeo

DNA - cido desoxirribonuclico

DNase - desoxirribonuclease

DOE - *Department of Energy / USA*

EDTA - cido etileno-diamino-tetractico

EMBL - *European Molecular Biology Laboratory*

EST - *expressed sequence tag*

FAPESP - Fundao de Amparo  Pesquisa do Estado de So Paulo

GAPDH - gliceraldedo-3 fosfato desidrogenase

GSS - *Genomic Survey Sequence*

gb - gigabases

HTGS - *High-Throughput Genome Sequence*

HUGO - *Human Genome Organization*

ILPC - Instituto Ludwig de Pesquisa sobre o Câncer
IPTG - isopropil- β -D-tiogalactopiranosídeo
kV - quilovolts
LINEs - *long interspersed repeats*
M - molar
Mb - megabases
MGC - Mammalian Gene Collection
ml - mililitros
mRNA - RNA mensageiro
mtDNA - DNA mitocondrial
NCBI - *National Center for Biotechnology information*
ng - nanogramas
nm - nanômetros
ONSA - *Organization for Nucleotide Sequencing and Analysis*
ORESTES - *Open Reading Frame ESTs*
PAC - *P1 artificial chromosome* (cromossomo artificial derivado do fator P1)
pb - pares de bases
PCR - Reação em cadeia da polimerase
PGH - Projeto Genoma Humano
pH - potencial hidrogeniônico
Poli-A⁺ - cauda poli(A) do RNA mensageiro
pUC - *plasmid for Universal Cloning*
RAP-PCR - *RNA arbitrarily primed polymerase chain reaction*
RNA - ácido ribonucleico
RNase - ribonuclease
rpm - rotações por minuto
rRNA - RNA ribossomal
RT-PCR - Reação em cadeia da polimerase precedida da reação de transcrição reversa
SDS - dodecil-sulfato de sódio
SINEs - *small interspersed repeats*
SNPs - *single nucleotide polymorphisms*

STS - *sequence tagged site*

THCs - *tentative human consensus*

TIGR - *The Institute for Genomic Research*

UV - Ultravioleta

X-gal - 5-bromo-4-cloro-3-indolil- β -D-galactopiranosídeo

YAC - *yeast artificial chromosome* (Cromossomo artificial de levedura)

SUMMARY

Expressed sequence tags (ESTs) are of fundamental importance for the identification of genes within the human genome and defining gene expression characteristics.

In this work, we describe a new approach for generating cDNA libraries using essentially arbitrary primers to construct PCR-based minilibraries from breast tumor mRNA. Clones from these libraries were sequenced to generate 6,029 ESTs. Using this approach, we were able to observe a significant normalization of the different mRNA subpopulations and a preferential amplification of the central portions of the genes.

Bioinformatic analysis of these sequences shows that 3,350 ESTs (56%) have significant similarity to known DNA and/or cDNA sequences (annotated sequences) from different organisms and 1,509 ESTs (25%) show no similarity to any sequences on different databases. From the annotated sequences, we have identified some sequences with high similarity to known genes from different organisms, indicating the discovery of some homologous genes possibly correlated with carcinogenic processes. For instance, we have isolated and partially characterized (i) a new *NABC1* (*novel amplified sequence in breast carcinoma 1*) isoform which is down-regulated in colorectal tumors, (ii) a novel semaphorin member of axon guidance molecules that is down-regulated in glioblastoma cell lines treated with all-trans-retinoic acid, an anti-tumor agent and (iii) the ortholog *Notch 2* human gene, apparently overexpressed in breast tumors with higher malignancy.

RESUMO

Etiquetas de sequências expressas (ESTs) são fundamentais para a identificação de genes no genoma humano e para definir características de expressão gênica.

Neste trabalho, descrevemos uma nova abordagem para a geração de bibliotecas de cDNA, utilizando iniciadores arbitrários para a produção, por PCR, de mini-bibliotecas a partir de mRNA derivado de câncer de mama. Clones destas bibliotecas foram sequenciadas para gerar 6029 ESTs. Utilizando esta abordagem, foi possível observar uma significativa normalização das diferentes sub-populações de mRNA e amplificação preferencial de porções centrais dos genes.

Análise bioinformática destas sequências mostra que 3.350 ESTs (56%) tem similaridade significativa a sequências de DNA e/ou cDNA já conhecidas (sequências anotadas) descritas em diferentes organismos, e 1509 ESTs (25%) não possuem qualquer similaridade a diferentes bancos de dados. Dentre as sequências anotadas, identificamos algumas sequências com alta similaridade a genes conhecidos em diferentes organismos, indicando a descoberta de alguns genes homólogos possivelmente envolvidos com processos carcinogênicos. Como exemplo, isolamos e caracterizamos parcialmente (i) uma nova isoforma do gene *NABC1* (*novel amplified sequence in breast carcinoma 1*), o qual é pouco expresso em tumores coloretais, (ii) um novo gene da família de semaforinas (moléculas de motilidade axonal) que apresenta uma baixa expressão em linhagens celulares de glioblastoma tratadas com ácido retinóico, um agente anti-tumoral e (iii) o gene ortólogo humano *Notch 2*, aparentemente superexpresso em tumores mamários com maior malignidade.

1. INTRODUÇÃO

1.1. Visão global do genoma humano

As informações da hereditariedade de uma espécie animal ou vegetal, assim como o código informativo para a geração de moléculas orgânicas - proteínas e peptídeos - envolvidas com o crescimento, desenvolvimento e funcionamento celular, estão contidas na molécula de ácido desoxirribonucleico (DNA). O DNA é constituído por duas longas fitas nucleotídicas, estabilizadas por pontes de hidrogênio entre bases nitrogenadas complementares, com a peculiar característica de sofrer uma replicação semi-conservativa durante o processo de divisão celular, de modo a perpetuar a informação genética na espécie.

O DNA humano se encontra distribuído em 23 pares de cromossomos, enriquecidos por enzimas nucleares, proteínas reguladoras e histonas que favorecem a estabilidade, o empacotamento e a organização do material genético (cromatina) no núcleo celular. Cada cromossomo contém uma única molécula bifilamentar e linear de DNA com um comprimento médio entre 34.000 (cromossomo 21) e 250.000 kb (cromossomo 1). O conteúdo nucleotídico do DNA humano presente nos 24 diferentes cromossomos humanos (22 autossomos e 2 cromossomos sexuais, X e Y) é denominado genoma humano e corresponde a um número estimado de 3.12 bilhões de nucleotídeos (em torno de 3 Gb).

De maneira geral, 70 a 80% do genoma humano se refere a DNA extragênico (97% incluindo sequências intrônicas), ou seja, sequências de DNA não relacionadas a genes humanos (Brown, 1999). Dentro desta classe, estão presentes sequências cópia única e repetitivas de DNA, que podem estar distribuídas, principalmente, na forma de satélites de DNA, SINEs (*small interspersed repeats*) e LINEs (*long interspersed repeats*) (Brown, 1999).

Satélites de DNA correspondem a repetições em tandem (agrupadas) de unidades nucleotídicas, constituídas geralmente por arranjos de 5 a 200 nucleotídeos, com um possível envolvimento na manutenção da estrutura cromossômica (Cooper, 1997). Esta classe de sequências repetitivas pode ser diferentemente classificada de acordo com a sua extensão no genoma humano, podendo abranger (i) 100 a 5000 kb (satélite de DNA clássico), (ii) 100 pb a 20 kb (minissatélite de DNA) e (iii) extensões inferiores a 100 pb (microsatélite de DNA). Regiões

de microssatélites equivalentes a repetições de mono e dinucleotídeos estão presentes, respectivamente, em cerca de 0,3% e 0,5% do genoma humano total.

Sequências repetitivas de DNA dispersas ao longo do genoma são classificadas como SINEs (repetições pequenas inter-espaçadas) e LINEs (repetições longas inter-espaçadas). Os maiores representantes de SINEs no genoma humano são as sequências Alu, cuja denominação é derivada da presença de um usual sítio de restrição *AluI* na sequência. Cada sequência Alu possui um número estimado de 300 pb e aparece em torno de um milhão de vezes no genoma humano, correspondendo a 10% do DNA celular total. Embora as sequências Alu sejam transcritas, não codificam proteínas e sua função não é muito clara.

Os principais representantes de LINEs se referem às sequências L1 (família LINE 1). Cada sequência L1 é constituída por cerca de 6000 pb e aparece cerca de 50.000 vezes no DNA humano, correspondendo igualmente a 10% do genoma total. Sequências L1 são transcritas e codificam proteínas essenciais para a sua eventual transposição, por um processo que aparentemente envolve transcrição reversa (Lewin, 1997; Cooper, 1997; Brown, 1999).

Tanto as sequências Alu quanto L1 são elementos repetitivos tipicamente transponíveis, com a capacidade de se mover para diferentes regiões genômicas e podendo, com isso, exercer um papel de regulação da expressão gênica (Britten, 1996; Mazzarella & Schlessinger, 1997). Possivelmente, a presença destes elementos no genoma humano teve um importante papel evolucionário, proporcionando uma diversidade genética dentro da espécie humana (Cooper, 1997).

Os genes humanos, porções de DNA capazes de codificar proteínas, enzimas ou peptídeos, equivalem a cerca de 30% do genoma. Deste percentual, apenas 10% efetivamente possuem um papel codificador (ou seja, cerca de 3% do genoma global), sendo equivalentes a sequências denominadas exons, enquanto que 90% compõem estruturas gênicas não-codificantes (íntrons) e pseudogenes (Brown, 1999).

Pseudogenes são sequências de DNA semelhantes a genes funcionais porém, incapazes de serem traduzidos em proteínas. Esta inativação é derivada de eventos evolucionários que ocasionaram mutações gênicas, abolindo os sinais para início de transcrição e/ou *splicing* do RNA primário ou ainda promovendo terminação prematura do polipeptídeo a ser traduzido. Pseudogenes que representam versões inativas de genes normais tem sido identificados em

muitos sistemas, incluindo globina, imunoglobulinas e antígenos de histocompatibilidade (Lewin, 1997).

Os genes humanos variam grandemente em tamanho, desde menos de 100 pb (genes de tRNA e snRNA) até mais de 2.000 Kb (gene da distrofina). As regiões genômicas de maior densidade gênica estão relacionadas a porções de DNA ricas em bases G-C (53% do total de bases, tabela 01), que correspondem a cerca de 3% do genoma total (Gardiner, 1996).

Tabela 01. Relação entre conteúdo de G-C e densidade gênica no genoma humano

Porções genômicas (<i>isochore class</i>)	Conteúdo de GC (%)	Representação no genoma (%)	Estimativa gênica (%)*	Tamanho (Mb)*
L1	38	30	34	1860
L2	41	32		
H1	44	21	38	930
H2	49	10		
H3	53	3	39	90

* estimativa gênica e tamanho genômico referente a porções L1+L2, H1+H2 e H3, respectivamente.

Apesar da estimativa usual do número total de genes humanos ser de 50.000 a 100.000 genes (Lewin, 1997; Cooper, 1997; Brown, 1999), existem algumas controvérsias a este respeito. Para estimar o número de genes no genoma humano, Weissenbach e colaboradores (2000) analisaram 1.272 Mb de DNA humano não-redundante (*working draft sequence*, EMBL database, release 61) frente a algoritmos que detectam sequências ortólogas codificadoras (programa Exofish, *exon finding by sequence homology*), gerados a partir do genoma do vertebrado *Tetraodon nigroviridis*, e chegaram a um valor aproximado de 34.000 genes humanos. A comparação entre um milhão de sequências expressas (na forma de ESTs, geradas pelo *Washington University Genome Sequencing Center*) com dados da sequência completa do cromossomo 22 e sequências de mRNA completas não-redundantes, indicaram um número aproximado de 35.000 genes no genoma humano (Ewing & Green, 2000). Por outro lado, o estabelecimento de um índice gênico humano a partir da sobreposição de sequências de cDNA não-redundantes disponibilizadas no banco de dados dbEST (*GenBank*) evidenciou que o genoma humano contém cerca de 120.000 genes (Liang *et al.*, 2000). A finalização do

sequenciamento e montagem do DNA humano, através do projeto genoma humano, auxiliará na enumeração global correta dos genes humanos.

1.2. O Projeto Genoma Humano (PGH)

O Projeto Genoma Humano é um programa internacional de pesquisa científica multidisciplinar com a finalidade de gerar detalhados mapas genético e físico do genoma humano, assim como determinar a sua sequência nucleotídica completa, de modo a identificar os 50.000-100.000 genes estimados no genoma. O mapa genético é uma representação da ordem dos genes ao longo dos cromossomos, enquanto que o mapa físico é uma coleção de fragmentos de DNA sobrepostos uns aos outros, de acordo com a especificação de seu arranjo ao longo dos cromossomos.

Análise semelhante é feita com o genoma de outros organismos menos complexos, utilizados como modelos experimentais, com o intuito de relacionar certas funções biológicas e estabelecer um padrão evolucionário entre as espécies. Deste modo, o PGH visa fornecer informações a respeito da estrutura, organização e função do DNA humano, possibilitando um entendimento mais acurado do desenvolvimento e funcionamento do corpo humano. Em paralelo, avanços tecnológicos na análise genética de populações, identificação de mutações gênicas de importância clínica e desenvolvimento de novos fármacos a partir das informações geradas pelo PGH (Farmacogenômica) tem sido algumas das consequências do avanço constante deste projeto.

1.2.1. Origem do PGH

Conforme dito anteriormente, três motivos principais levaram a comunidade científica à implementação do projeto genoma: (i) a construção de mapas genéticos de *linkage* para o rastreamento de regiões cromossômicas que contêm características hereditárias, (ii) a produção de mapas físicos de grandes regiões cromossômicas para promover o processo de descoberta gênica a partir de informações estruturais do DNA e (iii) o sequenciamento total do DNA cromossomal, com a intenção de identificar genes e correlacionar mutações e polimorfismos com alterações biológicas evidentes no ser humano.

O advento tecnológico do sequenciamento de DNA (Maxam & Gilbert, 1977; Sanger *et al.*, 1977) foi a principal ferramenta que motivou a criação do Projeto Genoma Humano. Em questão de um ano após sua descoberta, o procedimento de sequenciamento de DNA se espalhou pela comunidade científica e, como consequência, tornou clara a necessidade de organizar e “catalogar” todos os dados de sequências geradas. Após alguns anos de discussão a respeito da organização de sequências de DNA em caráter internacional, foram criados dois bancos de dados para o armazenamento das sequências geradas, onde um banco de dados ficou sob o encargo do *European Molecular Biology Laboratory* (EMBL) em Heidelberg e o outro, nomeado como *GenBank*, foi inicialmente centralizado no *Los Alamos National Laboratory*. O grande avanço computacional ocorrido na década de 80 proporcionou o rápido crescimento destes bancos de dados e sua maior acessibilidade pela comunidade científica.

O *GenBank*, sob a atual responsabilidade do *National Center for Biotechnology Information* (NCBI), uma divisão da *National Library of Medicine* (situada no campus do *US National Institutes of Health*), comporta um número superior a 3.4 bilhões de bases, provenientes de 4.6 milhões de sequências diferentes e 55.000 espécies de organismos. Deste total de sequências, 56% correspondem a sequências humanas, sendo que 34% equivalem a ESTs humanas (Benson *et al.*, 2000).

De modo independente, três diferentes personagens foram responsáveis pela idéia inicial de sequenciamento do genoma humano. O primeiro destes, Robert Sinsheimer, então diretor da Universidade da Califórnia, Santa Cruz (UCSC) propôs, durante um *workshop* em maio de 1985, a elaboração deste projeto visando uma oportunidade de aquisição de fundos para o campus da UCSC e a construção de um Instituto de Pesquisa Genômica. Sua proposta acabou sendo infrutífera. O segundo personagem envolvido, Renato Dulbecco, pesquisador laureado Nobel do Instituto Salk (La Jolla, Califórnia), propôs o estabelecimento de um projeto de sequenciamento do genoma humano como ferramenta para estudos da origem epigenética do câncer. Seus comentários foram veiculados internacionalmente pela revista *Science* (1986) e, de certa maneira, serviram de informe e preparação da comunidade científica para este novo ideal (para revisão veja Bodmer & McKie, 1994).

O último personagem envolvido foi Charles DeLisi, pesquisador de Biologia matemática do *National Cancer Institute* (a maior instituição do *National Institutes of Health*, NIH/USA), que propôs, em 1985, o sequenciamento do genoma humano sob o apoio do Departamento de

Energia (DOE) Norte-Americano. Tal sugestão de apoio governamental tinha seu devido fundamento visto que o DOE já possuía um programa de estudo biológico em alta escala, denominado Projeto Manhattan. Este era um projeto científico pós-2ª Guerra mundial com a finalidade de estudar os diversos danos genéticos causados pela radiação emitida por bombas atômicas. Na opinião de DeLisi, o sequenciamento do genoma humano era essencial para a identificação gênica, servindo de suporte para a elucidação (e explicação biológica) de mutações hereditárias em crianças expostas a radiação semelhante àquela emitida pelas bombas de Hiroshima e Nagasaki. Graças a seu forte embasamento e posição dentro da pesquisa científica do governo norte-americano, Charles DeLisi foi o responsável pelo início do Projeto Genoma Humano (<http://www.fplc.edu/risk/vol5/spring/cookdeeg.htm>).

1.2.2. Organização e estabelecimento do PGH no setor público

Críticas com relação ao custo, organização e prioridades do PGH, assim como a evidência de que um grande percentual do genoma não corresponde a genes (sugerindo uma perda desnecessária de capital na geração de sequências não-informativas), criaram um conflito de idéias na comunidade científica internacional que retardou a realização do projeto por cerca de 3 anos. As prioridades do projeto foram redefinidas *pela National Research Council (NCR)*, o maior setor operacional da *National Academy of Sciences (USA)*, e em 1988, o PGH foi estabelecido com a favorável ação governamental e consequente liberação de fundos para o NIH-DOE. Em 1989, foi estabelecido *The National Center for Human Genome Research (NCHGR)*, posteriormente conhecido como *The National Human Genome Research Institute (NHGRI)*, com o objetivo de comandar o PGH do NIH-DOE.

Na mesma época, países como a Itália, União Soviética, Reino Unido, Japão, França e Canadá haviam igualmente desenvolvido seus programas nacionais de Genoma Humano e, em 1989, foi estabelecida a Organização do Genoma Humano (*Human Genome Organization, HUGO*), com a finalidade de coordenar e centralizar todos os programas envolvidos com o PGH, em âmbito mundial. As instituições *Howard Hughes Medical Institute, The Wellcome Trust e Imperial Cancer Research Fund* proporcionaram o apoio financeiro necessário para esta organização, enquanto que a Comunidade Européia auxiliou a HUGO na coordenação das atividades de pesquisa.

Para atingir os propósitos iniciais de mapeamento genético e sequenciamento global do genoma humano, tornou-se necessário desenvolver novas metodologias como (i) sistemas de vetores para clonagem de grandes fragmentos de DNA (YACs, BACs), (ii) métodos computacionais de montagem de sequências para devido mapeamento físico, (iii) automação do sequenciamento de DNA, entre outros. O estabelecimento de normas éticas e legais para a manipulação da informação genética (*Ethical, legal and social implications, ELSI*) foram igualmente definidas durante a etapa inicial do PGH. Em 1993, foi editado um plano de metas para os 5 anos subsequentes do PGH desenvolvido nos Estados Unidos. As principais metas discriminadas foram (i) completar um mapa de STSs (*Sequence tagged sites*) do genoma humano com uma resolução de 100 Kb, (ii) gerar uma quantidade de sequência genômica mínima equivalente a 50 Mb ao ano (assumindo um crescimento exponencial da capacidade de sequenciamento com o tempo) e (iii) completar o sequenciamento dos organismos modelos *Escherichia coli*, *Saccharomyces cerevisiae* e *Caenorhabditis elegans*. A estimativa inicial de conclusão do PGH feita pelo NIH-DOE foi para o ano 2005 (Collins & Galas, 1993).

Em 1998, um novo plano de metas do NIH-DOE para o período de 1998-2003 foi publicado (Collins *et al.*, 1998). No final de 1998, o PGH havia atingido as metas anteriormente sugeridas, com (i) uma produção de 90 Mb de sequência ao ano e a implementação de eletroforese em sistema capilar como nova tecnologia de sequenciamento, (ii) o mapeamento de 52000 STSs, (iii) a geração de 180 Mb de sequência humana e 111 Mb não-humana e (iv) o sequenciamento completo dos organismos *E. coli*, *S. cerevisiae* e *C. elegans*. Neste novo plano, o tempo estimado de conclusão do sequenciamento do genoma humano foi antecipado em 2 anos (de 2005 para 2003), com a esperada contribuição do NIH-DOE de 60-70% das sequências e o restante sob o encargo do *Sanger Centre* (Cambridge, Inglaterra), apoiado pelo *The Wellcome Trust*, e outros parceiros internacionais. Para o final de 2001, foi estabelecida a meta de concluir 1 Gb de sequência humana finalizada (erro inferior a 1 em cada 10.000 bases e alta contiguidade), equivalente a 1/3 do genoma total.

No entanto, a possibilidade do PGH fornecer informações sobre novos genes humanos com provável aplicação diagnóstica e desenvolvimento de novos medicamentos se tornou cada vez mais clara e atraiu a participação do setor privado na conclusão do PGH desde o ano de 1992. O anúncio da empresa *Celera Genomics*, em 1998, de que o sequenciamento do genoma humano seria finalizado até o ano 2001, provocou uma maior aceleração do PGH no setor

público, em virtude do risco de patenteamento de sequências pelo setor privado. Em março de 1999, o NHGRI promoveu maior suporte para o PGH, gerando verbas para três grandes centros de sequenciamento genômico, localizados na *Washington University*, *Baylor College of Medicine* e *Whitehead Institute for Biomedical Research*. Ao mesmo tempo, *The Wellcome Trust* gerou mais verbas para o avanço do processo de sequenciamento no *Sanger Centre*. Estes quatro grupos, aliados ao DOE, se comprometeram a sequenciar, ao menos, 90% do genoma humano, com uma redundância de cinco vezes até março de 2000. Até o final do ano 2000, o consórcio público do PGH concluiu a montagem de 30,4% do genoma humano (<http://www.ncbi.nlm.nih.gov/genome/seq>).

1.2.3. O Instituto de Pesquisa Genômica (TIGR)

Até o ano de 1991, um número inferior a 2000 genes humanos tinham sido completamente sequenciados e depositados nos bancos de dados públicos. Visando um aumento na eficiência de descoberta gênica aliada ao PGH, J. Craig Venter, chefe do *National Institute of Neurological Disorders and Stroke* (NINDS/NIH), desenvolveu uma técnica inovadora de geração de “etiquetas de sequência expressas” (*expressed sequence tags*, ESTs) (Adams *et al.*, 1991) com o intuito de identificar novos genes humanos e, utilizando programas computacionais de sobreposição de ESTs, determinar total ou parcialmente a sequência completa destes genes expressos. Utilizando esta técnica, foi possível identificar cerca de 2300 novos genes expressos em células de cérebro humano (Adams *et al.*, 1992), duplicando o número de genes humanos até então conhecidos.

Com o intuito de aplicar, em larga escala, esta nova tecnologia de identificação gênica, J. Craig Venter se desligou do NIH e estabeleceu, em 1992, *The Institute for Genomic Research* (TIGR), uma instituição privada sem fins lucrativos sob o suporte financeiro das incorporações *Human Genome Sciences* e *Smith-Kline-Beecham*. Em 1997, o TIGR dissolveu suas relações com o setor privado e começou a ter suporte financeiro do DOE-USA.

Dentre os benefícios científicos fornecidos pelo TIGR, desde a sua criação, podemos citar (i) a anotação parcial de, ao menos, 50% dos genes humanos expressos (Adams *et al.*, 1995), (ii) a acessibilidade da diversidade genética humana através da análise de SNPs (*single nucleotide polymorphisms*) e (iii) o sequenciamento do genoma completo de vários microorganismos

patogênicos como, por exemplo, *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Helicobacter pylori* (Tomb *et al.*, 1997) e *Vibrio cholerae* (Heidelberg *et al.*, 2000). Até o presente momento, o TIGR foi a instituição de maior índice de conclusão de genomas microbianos, ou seja, 14 em 38 organismos (36,8%) foram totalmente sequenciados pelo TIGR.

Tabela 02. Microorganismos com genoma totalmente sequenciado (até 28.10.2000). Tamanho estimado do genoma (em megabases, Mb), instituição responsável pelo trabalho e referências bibliográficas estão igualmente tabuladas.

Organismo (linhagem)	Tamanho (Mb)	Instituição	Referência bibliográfica
<i>Aeropyrum pernix</i> (K1)	1.67	Biotechnology Center / NITE	Kawarabayasi <i>et al.</i> (1999). <i>DNA Res.</i> 6, 83-101.
<i>Aquifex aeolicus</i> (VF5)	1.50	Diversa	Deckert <i>et al.</i> (1998). <i>Nature</i> 392, 353.
<i>Archaeoglobus fulgidus</i> (DSM4304)	2.18	TIGR	Klenk <i>et al.</i> (1997). <i>Nature</i> 390, 364-370.
<i>Bacillus subtilis</i> (168)	4.20	International Consortium / EC	Kunst <i>et al.</i> (1997). <i>Nature</i> 390, 249-256.
<i>Borrelia burgdorferi</i> (B31)	1.44	TIGR	Fraser <i>et al.</i> (1997). <i>Nature</i> 390, 580-586.
<i>Buchnera</i> sp. (APS)	0.64	Univ. Tokyo / RIKEN	Shigenobu <i>et al.</i> (2000). <i>Nature</i> 407, 81-86.
<i>Campylobacter jejuni</i> (NCTC 11168)	1.64	Sanger Centre	Parkhill <i>et al.</i> (2000). <i>Nature</i> 403, 665-668.
<i>Chlamydia pneumoniae</i> (CWL029)	1.23	UC Berkeley & Stanford	Kalman <i>et al.</i> (1999). <i>Nat Genet</i> 21, 385-389.
<i>Chlamydia pneumoniae</i> (AR39)	1.23	TIGR	Read <i>et al.</i> (2000). <i>Nucleic Acids Res.</i> 28, 1397-1406.
<i>Chlamydia trachomatis</i> (serovar D)	1.05	UC Berkeley & Stanford	Stephens <i>et al.</i> (1998). <i>Science</i> 282, 754-759.
<i>Chlamydia trachomatis</i> (MoPn)	1.07	TIGR	Read <i>et al.</i> (2000). <i>Nucleic Acids Res.</i> 28, 1397-1406.
<i>Deinococcus radiodurans</i> (R1)	3.28	TIGR	White <i>et al.</i> (1999). <i>Science</i> 286, 1571-1577.
<i>Escherichia coli</i> (K12)	4.60	University of Wisconsin	Blattner <i>et al.</i> (1997). <i>Science</i> 277, 1453-1474.
<i>Haemophilus influenzae</i> Rd (KW20)	1.83	TIGR	Fleischmann <i>et al.</i> (1995). <i>Science</i> 269, 496-512.
<i>Halobacterium</i> sp. (NRC-1)	2.57	Halobacterium genome consortium (NSF)	Ng <i>et al.</i> (2000). <i>Proc Natl Acad Sci USA</i> 97, 12176-12181.
<i>Helicobacter pylori</i> (26695)	1.66	TIGR	Tomb <i>et al.</i> (1997). <i>Nature</i> 388,

			539-547.
<i>Helicobacter pylori</i> (J99)	1.64	Astra Research Center/ Genome Therapeutics	Alm <i>et al.</i> (1999). <i>Nature</i> 397 , 176-180.
<i>Methanobacterium thermoautotrophicum</i> (delta H)	1.75	Genome Therapeutics & Ohio State Univ.	Smith <i>et al.</i> (1997). <i>J. Bacteriology</i> 179 , 7135-7155.
<i>Methanococcus jannaschii</i> (DSM2661)	1.66	TIGR	Bult <i>et al.</i> (1996). <i>Science</i> 273 , 1058-1073.
<i>Mycobacterium tuberculosis</i> (H37Rv)	4.40	Sanger Centre	Cole <i>et al.</i> (1998). <i>Nature</i> 393 , 537.
<i>Mycoplasma genitalium</i> (G-37)	0.58	TIGR	Fraser <i>et al.</i> (1995). <i>Science</i> 270 , 397-403.
<i>Mycoplasma pneumoniae</i> (M129)	0.81	Univ. of Heidelberg	Himmelreich <i>et al.</i> (1996). <i>Nucleic Acids Res</i> 24 , 4420-4449.
<i>Neisseria meningitidis</i> (MC58)	2.27	TIGR	Tettelin <i>et al.</i> (2000). <i>Science</i> 287 , 1809-1815.
<i>Neisseria meningitidis</i> (Z2491)	2.18	Sanger Centre	Parkhill <i>et al.</i> (2000). <i>Nature</i> 404 , 502-506.
<i>Pseudomonas aeruginosa</i> (PAO1)	6.30	Univ. of Washington / PathoGenesis	Stover <i>et al.</i> (2000). <i>Nature</i> 406 , 959-964.
<i>Pyrococcus horikoshii</i> (OT3)	1.80	Biotechnology Center / NITE	Kawarabayasi <i>et al.</i> (1998). <i>DNA Res.</i> 5 , 55-76.
<i>Rickettsia prowazekii</i> (Madrid E)	1.10	Univ. of Uppsala	Andersson <i>et al.</i> (1998). <i>Nature</i> 396 , 133-140.
<i>Saccharomyces cerevisiae</i> (S288C)	13.0	International Consortium / EC	Goffeau <i>et al.</i> (1996). <i>Science</i> 274 , 563-567.
<i>Synechocystis</i> sp. (PCC6803)	3.57	Kazusa DNA Research Inst.	Kaneko <i>et al.</i> (1996). <i>DNA Res.</i> 3 , 109-136.
<i>Thermoplasma acidophilum</i>	1.56	Max-Planck-Institute for Biochemistry	Ruepp <i>et al.</i> (2000). <i>Nature</i> 407 , 508-513.
<i>Thermotoga maritima</i> (MSB8)	1.80	TIGR	Nelson <i>et al.</i> (1999). <i>Nature</i> 399 , 323-329.
<i>Treponema pallidum</i> (Nichols)	1.14	TIGR / Univ. Texas	Fraser <i>et al.</i> (1998). <i>Science</i> 281 , 375-388.
<i>Ureaplasma urealyticum</i> (serovar 3)	0.75	Univ. Alabama / PEB-ACGT	Glass <i>et al.</i> (2000). <i>Nature</i> 407 , 757-762.
<i>Vibrio cholerae</i> (N16961)	4.0	TIGR	Heidelberg <i>et al.</i> (2000). <i>Nature</i> 406 , 477-483.
<i>Xylella fastidiosa</i> (9a5c)	2.68	ONSA Consortium	Simpson <i>et al.</i> (2000). <i>Nature</i> 406 , 151-157.

Visando o mapeamento e a anotação funcional de prováveis genes representados por ESTs, o TIGR desenvolveu um sistema de análise denominado Índices Gênicos do TIGR (*TIGR Gene Indices*), mantido para 10 dos 14 organismos mais representativos nos projetos públicos de geração de ESTs, com a intenção de agrupar sequências expressas em prováveis sequências consenso (*tentative consensus sequences*) de qualidade, representando transcritos únicos e de

maior extensão. Além dos objetivos citados, estas sequências consenso possibilitarão a identificação de genes ortólogos e parálogos na comparação entre diferentes índices gênicos (Quackenbush *et al.*, 2000).

1.2.4. A contribuição do setor privado no PGH

Várias empresas do setor farmacêutico e biotecnológico, como a *Merck* e a *Perkin-Elmer* (*Applied Biosystems*), iniciaram projetos particulares de estudo do genoma humano, devido a possibilidade do PGH gerar informações essenciais para a produção de novas metodologias de diagnóstico clínico e descoberta de novos medicamentos.

Como principal exemplo, em maio de 1998, a *Perkin-Elmer (PE) Corporation* juntamente com o pesquisador J. Craig Venter, fundador do TIGR, criaram a empresa *Celera Genomics* com o intuito de acelerar o processo de sequenciamento e anotação de novos genes humanos, comercializando estas informações genômicas para a comunidade em geral. A previsão de conclusão do sequenciamento e montagem completa do genoma humano foi estimada para o ano de 2001.

A visão comercial da *Celera* se choca diretamente com a proposta do PGH do NIH-DOE e do consórcio internacional organizado pela HUGO, os quais pretendem disponibilizar as informações geradas nos bancos de dados internacionais isento de monopólio e/ou patenteamento genômico.

Dentro da estrutura da *Celera Genomics*, foram incorporadas duas unidades específicas: (i) o *GenScope*, um provedor de serviços para descoberta gênica, análise funcional e farmacogenômica e (ii) o *AgGen*, um prestador imediato de serviços para aplicação de informações genômicas no melhoramento agrícola e pecuário (para maiores informações, veja a *homepage* da *Celera Genomics*, <http://www.celera.com>).

Dentro do setor privado, a *Celera* é considerada a maior empresa com recursos de sequenciamento de DNA, composta por um total aproximado de 300 sequenciadores capilares do tipo ABI PRISM® 3700 (PE Biosystems) e uma capacidade de decodificação aproximada de 140 milhões de nucleotídeos a cada 24 horas (este potencial possibilitaria o sequenciamento total do genoma humano em menos de 23 dias úteis!). O sistema computacional agregado à análise das

sequências geradas é o segundo mais poderoso do mundo, perdendo apenas para o Sistema de Defesa do DOE-USA.

No dia 26 de junho de 2000, foi anunciada a primeira montagem das sequências geradas para a finalização do genoma humano, através da colaboração entre o NIH-DOE e a empresa *Celera*. Esta data foi um marco da pesquisa genômica, que tornou mais próxima a esperança de utilização das informações gênicas para o benefício da humanidade. A partir de agora, uma maior ênfase para a análise computacional está sendo dada, de modo a proceder ao preenchimento de *gaps* presentes no genoma, organização de sequências repetitivas e diminuição de erros na sequência, para então vislumbrarmos a montagem final do genoma humano completo e suas aplicações.

1.3. Estratégias utilizadas no sequenciamento completo do genoma humano

O recurso mais utilizado pelo setor público no sequenciamento do genoma humano, denominado sequenciamento de “clone por clone” ou estratégia “divida e conquiste” (*divide-and-conquer strategy*), visa a construção inicial de bibliotecas utilizando vetores de clonagem de insertos grandes (em geral, YACs e BACs) e posterior subclonagem de fragmentos genômicos menores para um efetivo sequenciamento. Para tanto, o DNA genômico é clivado de maneira aleatória, gerando fragmentos de diferentes tamanhos. Após seleção de fragmentos de DNA com tamanho superior a 100 Kb, estes são clonados em vetores apropriados para devida propagação em bactérias ou fungos e posterior isolamento. Através da utilização de STSs para cada região cromossômica (número atual aproximado de 1 STS específico para cada 100 Kb), os clones são alinhados para a montagem de *clusters* (grupos) ordenados. Cada clone mapeado é clivado de maneira randômica (*shotgun phase*) para originar fragmentos menores de DNA, subclonado em vetores apropriados (em geral, plasmídeos ou fagos M13) e sequenciado. As várias sequências geradas são sobrepostas umas às outras por análise computacional, promovendo a montagem (formação de *contigs*) e finalização da sequência completa de cada clone (Venter *et al.*, 1996; Green, 1997). Este recurso foi amplamente utilizado para o sequenciamento genômico dos organismos *S. cerevisiae* (Goffeau *et al.*, 1996) e *C. elegans* (*The C. elegans Sequencing Consortium*, 1998), além dos cromossomos humanos 21 (Hattori *et al.*, 2000) e 22 (Dunham *et al.*, 1999).

Tabela 03. Tipos de bibliotecas utilizadas para mapeamento genético e sequenciamento.

Vetor	Tamanho médio do inserto (Kb)	Número de clones necessários para cobertura total do genoma humano
YAC (<i>yeast artificial chromosome</i>)	100-2000	3.000 (1.000 Kb / clone)
BAC (<i>bacterial artificial chromosome</i>)	80-350	20.000 (150 Kb / clone)
Cosmídeo	30-45	75.000 (40 Kb / clone)
Plasmídeo	3-10	600.000 (5 Kb / clone)
Fago M13	1	3.000.000 (1 Kb / clone)

Um recurso alternativo de sequenciamento foi utilizado inicialmente pelo TIGR para a conclusão do primeiro projeto genoma descrito, referente ao organismo *H. influenzae* (Fleischmann *et al.*, 1995) e, mais recentemente, para o segundo e maior genoma animal já sequenciado (180 Mb), correspondente a *Drosophila melanogaster* (Adams *et al.*, 2000). Este recurso, denominado *whole-genome shotgun*, visa a clivagem randômica de DNA cromossomal com a consequente geração de milhões de sequências divididas em dois grupos: (i) sequências de 5 a 20 Kb (insertos longos), a serem clonadas em plasmídeos, fagos ou cosmídeos e (ii) sequências de 0.4 a 1.2 Kb (insertos curtos), a serem clonadas em plasmídeos. O sequenciamento das extremidades de insertos longos permite uma maior eficiência no processo de montagem das sequências geradas, a ser realizado por programas computacionais específicos. Sequências já conhecidas, correspondentes a ESTs e STSs humanas, podem ser muito informativas para a montagem de sequências (análise sequencial de exons e íntrons), participação como sítios de nucleação para ligação entre *contigs* e consequente mapeamento físico (Weber & Myers, 1997).

O recurso de *whole-genome shotgun* foi utilizado pela empresa privada *Celera Genomics* para o sequenciamento do genoma humano. A montagem e ordenação das sequências geradas tem sido realizada por algoritmos específicos ao sistema computacional da empresa. Estas informações, aliadas a análise por acesso remoto dos bancos de dados públicos referentes aos clones já sequenciados, tem permitido um rápido avanço do sequenciamento e montagem do genoma humano.

1.4. O Instituto Genômico Virtual de São Paulo / Brasil (ONSA)

Com a rápida expansão da pesquisa nas áreas de genética molecular e genômica ao redor do mundo, tornou-se claro o urgente desenvolvimento e estímulo desta linha de pesquisa dentro da perspectiva brasileira. Tal estímulo possibilitaria a geração de uma maior massa crítica e competitiva de pesquisadores e um maior reconhecimento da pesquisa científica brasileira no âmbito mundial. Para tanto, foi criada em 1997, no Estado de São Paulo, uma Instituição virtual de sequenciamento e análise genômica, constituída inicialmente por uma rede de 25 a 30 laboratórios, denominada ONSA (*Organization for Nucleotide Sequencing and Analysis*).

A estratégia utilizada pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) para a implementação da ONSA foi extremamente peculiar, pois não capacitou a montagem de qualquer infraestrutura física, utilizando apenas uma rede de conexão computacional dentro de uma comunidade científica pré-existente no Estado (Simpson & Perez, 1998).

No mesmo ano de criação da ONSA, iniciou-se, em São Paulo, o projeto genoma da *Xylella fastidiosa*, bactéria responsável pela clorose variegada cítrica (CVC) que atinge cerca de 30% do cultivo paulista de laranjas. Tal projeto recebeu estímulo do setor privado (Fundecitrus), visto que o estudo genômico de tal bactéria poderia sugerir novas estratégias efetivas no combate a CVC. Em julho de 2000, foi publicada a sequência genômica completa da *X. fastidiosa*, o primeiro projeto genoma criado e desenvolvido no Brasil. Seu genoma completo, com um tamanho aproximado de 2.7 Mb, composto por um cromossomo circular e dois plasmídeos, é constituído de 2.904 possíveis regiões codificadoras, das quais 47% possuem uma função biológica pré-estabelecida por análise computacional de homologia a outros organismos (Simpson *et al.*, 2000). A elucidação da sequência genômica da *X. fastidiosa* tem auxiliado na determinação das características metabólicas e replicativas desta agrobactéria, além de seus mecanismos particulares de patogenicidade (Lambais *et al.*, 2000).

1.5. O Projeto Genoma da Anatomia do Câncer (CGAP)

Em meados de 1997, o número de sequências novas produzidas pelos diferentes projetos de geração de ESTs, no modelo humano, começou a atingir um *plateau* (vide figura 01). Este

fato foi justificado pelo crescente número de sequências redundantes e baixa produção de novos *clusters* na geração maciça de ESTs (para maiores informações, veja a *homepage* do Unigene, <http://www.ncbi.nlm.nih.gov/UniGene/>).

Buscando um maior entendimento molecular sobre os diferentes tipos de câncer, o NCI (*National Cancer Institute, USA*) estabeleceu, em 1997, um projeto interdisciplinar denominado CGAP (*Cancer Genome Anatomy Project*). A meta deste projeto é realizar uma caracterização molecular de células normais, pré-cancerosas (transformadas) e malignas, proporcionando um maior entendimento da evolução carcinogênica. Os objetivos inicialmente propostos pelo CGAP foram (i) estabelecer um índice de genes expressos em tumores e (ii) apoiar o desenvolvimento e disseminação de novas tecnologias que permitam uma rápida e eficiente análise gênica e de expressão proteica, assim como a detecção de mutações (maiores detalhes vide *homepage* do CGAP). Este projeto ofereceu um novo avanço na identificação de novos genes humanos (fig. 01).

Figura 01: Gráfico ilustrativo mostrando o acúmulo de clusters no Unigene (eixo y) ao longo do tempo (eixo x). Porção em azul indica a contribuição de projetos não ligados ao CGAP na geração de novos clusters (Non-CGAP). Porção em vermelho (delineada a partir de 1997) indica a contribuição do CGAP na obtenção de sequências de novos clusters humanos (<http://www.ncbi.nlm.nih.gov/ncicgap/>). Um cluster é composto por sequências que correspondem a um único gene.

Mais recentemente, as informações geradas pelo CGAP deram origem a 5 novas iniciativas de implementação dos estudos moleculares do câncer, os quais seriam (i) o Índice de

genes humanos relacionados ao câncer (*The Human Tumor Gene Index, hTGI*), que visa a descoberta gênica e a identificação de genes envolvidos com a evolução tumoral; (ii) o Delineamento molecular (*Molecular Profiling, MP*), cuja intenção é demonstrar o conceito de análise molecular de espécies tumorais utilizando o câncer de próstata como modelo experimental; (iii) o Projeto de aberrações cromossômicas relacionadas ao câncer (*The Cancer Chromosome Aberration Project, CCAP*), que pretende integrar estudos citogenéticos e de mapeamento físico do genoma humano à análise de regiões cromossômicas aberrantes relacionadas ao câncer; (iv) o Índice de anotação genética (*The Genetic Annotation Index, GAI*), um programa de pesquisa para a identificação e caracterização de polimorfismos (*single nucleotide polymorphisms, SNPs*) em genes relacionados ao câncer e (v) o Índice de genes relacionados ao câncer murino (*The Mouse Tumor Gene Index, mTGI*) que pretende identificar e investigar genes envolvidos com a carcinogênese em murinos, de modo a comparar e auxiliar o entendimento de vias moleculares pouco estudadas na carcinogênese humana.

1.6. “Etiquetas de sequências expressas” como ferramenta para os projetos genoma

As sequências de DNA que estão sendo relacionadas no índice de genes expressos do CGAP são, na sua maioria, “etiquetas de sequências expressas” (Adams *et al.*, 1991). Estas sequências são obtidas, em geral, a partir do sequenciamento parcial de clones de bibliotecas de cDNA a partir de suas extremidades 5’ e/ou 3’, gerando milhares de sequências gene-específicas em um curto espaço de tempo. Esta estratégia vem a ser uma eficiente metodologia para a descoberta de novos genes (Marra *et al.*, 1998) e para o mapeamento genético (Bentley & Dunham, 1995; Schuler *et al.*, 1996; Touchman *et al.*, 1997).

A produção de “etiquetas de sequências expressas” (ESTs) tem sido uma ferramenta muito importante na identificação gênica dos diversos projetos genoma em andamento. Estas sequências estão sendo depositadas, em grande parte, em bancos de dados internacionais (dbEST, subdivisão do *GenBank*) com o intuito de disponibilizar as informações geradas para a comunidade científica (*IMAGE consortium*, <http://bbrp.llnl.gov/bbrp/image/>) e formar *clusters* (cada *cluster* é composto por sequências que correspondem a um único gene) com sequências já depositadas, permitindo uma ampliação dos dados em direção a sequência completa dos genes avaliados (*Unigene database*, <http://www.ncbi.nlm.nih.gov/Unigene/>).

Para a identificação de novos genes, a obtenção de ESTs a partir de bibliotecas de cDNA normalizadas é essencial pois, neste caso, há uma representatividade mais uniforme entre as populações de cDNAs altamente expressos e pouco expressos (Bonaldo *et al.*, 1996). No caso de bibliotecas não normalizadas, a população de cDNA está relacionada com o nível de expressão gênica, onde os clones mais abundantes correspondem proporcionalmente aos genes mais abundantemente expressos no material biológico em estudo. Isto causa uma aumento no nível de redundância dos genes expressos, ou seja, há um maior número de ESTs condizentes aos genes mais expressos, o que compromete o sequenciamento de genes mais raros. Os genes altamente expressos em uma célula somática típica totalizam 10 a 15 genes, que representam 10 a 20% da massa total de mRNA. Já os genes raros totalizam 15 mil a 20 mil genes, representando 40 a 45% do mRNA total da célula (Bishop *et al.*, 1974; Davidson & Britten, 1979).

O primeiro projeto piloto de geração de ESTs a partir de tecido humano foi proposto por Adams e colaboradores (1991). Cerca de 600 ESTs foram geradas para avaliar a eficiência da metodologia na descoberta de novos genes, na identificação de regiões codificadoras e no mapeamento genético. Das sequências geradas, 337 representavam novos genes e 48 possuíam alta similaridade a genes já identificados em outros organismos (supostos genes ortólogos).

Cientistas do TIGR estão entre os grupos inovadores na produção de ESTs em escala maciça (Adams *et al.*, 1995). Este grupo sequenciou cerca de 174 mil ESTs, oriundas de 300 bibliotecas de cDNAs que representavam diversos órgãos, estágios de desenvolvimento, estados patológicos e tipos celulares (linhagens primárias ou imortalizadas). Devido a grande redundância entre as ESTs produzidas, foi possível promover a sobreposição entre estas sequências, originando sequências consenso ou THCs (*Tentative Human Consensus*). Os autores concluem que, apesar do número de genes humanos estar na faixa de 100 a 200 mil genes, as ESTs isoladas neste trabalho representam fragmentos de, ao menos, 50% dos genes expressos.

Paralelamente, Hillier e colaboradores (1995) obtiveram 280.223 ESTs, isoladas de 22 bibliotecas de cDNAs (15 bibliotecas normalizadas e 7 não normalizadas, oriundas de diferentes tecidos e estágios de desenvolvimento), das quais se isolaram cerca de 44 mil sequências, correspondentes a THCs ou *singletons* (sequências únicas, sem possível sobreposição com outras ESTs). A complexidade das bibliotecas foi relativamente baixa (baixa proporção de *singletons*), necessitando de uma maior normalização e consequente aumento na representatividade de transcritos raros.

Projetos que visam a geração de ESTs para a descoberta de novos genes tem sido conduzidos nos mais diferentes organismos. O banco de dados dbEST contém hoje um número superior a 3 milhões de sequências armazenadas, geradas pelo Projeto Genoma Humano (PGH) e por projetos enfocando *Caenorhabditis elegans* (Waterston *et al.*, 1992; McCombie *et al.*, 1992), *Shistosoma mansoni* (Franco *et al.*, 1995; Franco *et al.*, 1997), *Arabidopsis thaliana* (Newman *et al.*, 1994; Cooke *et al.*, 1996), *Drosophila melanogaster* (Berkeley *Drosophila* Genome Project), entre outros 175 organismos diferentes. Os cinco organismos que possuem maior representatividade no dbEST com relação ao número de ESTs depositadas seriam (i) *Homo sapiens*, com 1.6 milhões de sequências (53% do total); (ii) *Mus musculus*, com 700 mil sequências (23%); (iii) *Rattus norvegicus*, com 125 mil sequências (4%); (iv) *C. elegans*, com 101 mil sequências (3%) e (v) *D. melanogaster*, com 86 mil sequências (3%) (dados referentes a agosto de 1999) (Benson *et al.*, 2000). Cerca de 60-70% das sequências depositadas no *GenBank* e 40% dos nucleotídeos individuais são representados por ESTs (Schuler, 1997; Benson *et al.*, 2000).

Dentre os primeiros exemplos de novos genes humanos descobertos pela análise de ESTs depositadas no banco de dados dbEST, podemos citar o caso da telomerase e da proteína Toll. A subunidade catalítica da telomerase humana foi clonada após a verificação de que um gene homólogo, presente em *Schizosaccharomyces pombe*, possuía alta similaridade a uma determinada sequência depositada no banco de dados dbEST (Nakamura *et al.*, 1997). Já o homólogo humano da proteína Toll de *Drosophila melanogaster*, envolvida com a resposta imune anti-fúngica no inseto adulto, foi também clonado e caracterizado após a identificação de uma EST homóloga a porção gênica que codifica um domínio-sinal presente na proteína Toll de *Drosophila* (Medzhitov *et al.*, 1997). Neste caso, se demonstrou que o sistema de defesa Toll/NF- κ B é conservado evolutivamente e que o homólogo humano da proteína Toll pode induzir a expressão de determinadas citocinas.

Técnicas de hibridização em bibliotecas de cDNA e amplificação por RACE-PCR (*Rapid Amplification of Complementary DNA Ends*) para obtenção de cDNAs completos (Frohman *et al.*, 1988; Frohman, 1993) podem ser utilizadas para favorecer a clonagem de novos genes, representados como “etiquetas expressas” no banco dbEST.

Como citado anteriormente, as ESTs também podem ser utilizadas no mapeamento genético de diferentes genomas. Este mapeamento é proporcionado pela hibridização ou

amplificação por PCR, a partir de ESTs, na identificação de YACs, BACs ou outros tipos de vetores ligados aos insertos em estudo. A amplificação por PCR de trechos dos insertos clonados geram “sítios etiquetados por sequências” (STSs) que fazem parte dos *loci* analisados. A sobreposição de diferentes STSs aumenta a complexidade da estrutura genômica proporcionando, por exemplo, a identificação de sítios de controle da expressão gênica (Bentley & Dunham, 1995; Houlgatte *et al.*, 1995; Schuler *et al.*, 1996; Marra *et al.*, 1998).

As ESTs também possibilitam a análise dos padrões de *splicing* em cDNAs já caracterizados, visto que uma razoável quantidade de sequências intrônicas se encontram depositadas no dbEST (Wolfsberg & Landsman, 1997). No entanto, a taxa de RNA nuclear heterogêneo (hnRNA) que é co-purificado com o mRNA total é muito pequena (Lamerdin *et al.*, 1995) e compromete a possível análise de íntrons.

1.7. A técnica de RAP-PCR na geração de “etiquetas de sequências expressas”

Apesar dos resultados promissores, a técnica convencional para a geração de ESTs possui limitações como (i) a alta taxa de sequências pouco informativas, referentes a porções não-traduzíveis dos genes (extremidades 5' e 3' dos transcritos); (ii) o sequenciamento redundante de genes expressos com maior frequência; (iii) a complexidade dos protocolos de normalização e subtração das bibliotecas de cDNA e (iv) a grande quantidade de mRNA necessário para a geração de bibliotecas (5-15 ug de mRNA puro).

A técnica de RAP-PCR (*RNA arbitrarily primed PCR*) é um método variante de iniciação arbitrária da PCR (para revisão veja Perucho *et al.*, 1995). Esta técnica visa a amplificação de sequências parcialmente complementares a *primers* arbitrários, que originarão “impressões digitais” das sub-populações originais de RNA (Welsh *et al.*, 1992).

Nesta técnica, a síntese da primeira fita de cDNA se procede por transcrição reversa e, posteriormente, a síntese da segunda fita é feita pela DNA polimerase, em condições de baixa estringência de amplificação. As moléculas geradas servirão de molde para a amplificação por PCR em alta estringência, favorecendo a geração de “impressões digitais” de RNA (RNA *fingerprinting*) (Welsh & McClelland, 1991; Welsh *et al.* 1992).

Figura 02: Desenho esquemático mostrando os diversos passos envolvidos na produção de perfis de RAP-PCR com *primers* arbitrários.

Mathieu-Daudé e colaboradores (1996) verificaram que, à medida que se aumenta o número de ciclos de uma determinada RAP-PCR, ocorre uma parcial normalização dos genes expressos devido a re-hibridização dos cDNAs mais abundantes, que serão amplificados em menor quantidade, permitindo assim que cDNAs mais raros (pouco concentrados na reação, ou seja, pouco susceptíveis à re-hibridização) sejam amplificados. Devido a estes fatores, os perfis de RAP-PCR tendem a ser parcialmente normalizados em relação ao nível de expressão gênica, permitindo que cDNAs mais raros sejam amplificados (Mathieu-Daudé *et al.*, 1996).

Técnicas de RNA *fingerprinting* como, por exemplo, o *Differential Display*, utilizam *primers* de transcrição reversa compostos por grande quantidade de timina (oligo-dT), possibilitando a hibridização às porções 3' do RNA (no caso, mRNAs e RNAs heterogêneos poliadenilados). Após transcrição reversa e denaturação, são utilizados *primers* arbitrários para amplificação das sequências de cDNA. Esta técnica, entretanto, gera grande quantidade de sequências não-traduzíveis (referentes a porção 3' do RNA) e impede a análise de RNAs não-processados e/ou RNAs estruturais que não contenham a cauda de poliadenilação (Welsh *et al.*, 1992). Neste caso, a técnica de RAP-PCR é mais vantajosa pois possibilita, com maior

frequência, a análise de ORFs (*open reading frames*) presentes no RNA (McClelland *et al.*, 1995).

Até o presente momento, a técnica de RAP-PCR tem sido largamente utilizada na diferenciação de padrões de expressão gênica. A simplicidade e a alta sensibilidade metodológica, onde pequenas quantidades de RNA-molde são utilizados, possibilitam a busca de marcadores com potencial diagnóstico e valor clínico (McClelland *et al.*, 1995). Como exemplo, Tórtola e colaboradores (1998) realizaram estudos por RAP-PCR em tumores coloretais e, após correlação com o tecido normal, obtiveram perfis de bandejamento bastante complexos em mais de 100 casos de carcinoma col retal, dando indicações de que muitas destas bandas correspondem a possíveis marcadores tumorais e/ou fatores envolvidos na diferenciação celular.

A geração de ESTs por RAP-PCR foi recentemente descrita em modelos como *Trypanosoma brucei* (Mathieu-Daudé *et al.*, 1998) e *Schistosoma mansoni* (Dias Neto *et al.*, 1996; Dias Neto, 1997; Dias Neto *et al.*, 1997). No modelo de *T. brucei*, foram obtidas 32 ESTs diferentes, onde 24 ESTs são expressas diferencialmente no ciclo parasitário ou na resposta a elevações de temperatura, e 9 ESTs possuem similaridade a sequências já depositadas nos bancos de dados. Dentro do modelo de *S. mansoni*, ao contrário da análise diferencial realizada no modelo *T. brucei*, houve um maior interesse na padronização metodológica e análise da eficiência global na geração de ESTs.

Franco e colaboradores (1995) empregaram a metodologia convencional na obtenção de ESTs para o desenvolvimento do Projeto Genoma de *Shistosoma mansoni*. Os autores verificaram um nível de redundância de aproximadamente 60% das sequências úteis, ou seja, cerca de 260 das 429 ESTs sequenciadas são repetitivas. O sequenciamento de vetores sem inserto atingiu 11% do total. Com esta abordagem, foram identificados 275 ESTs (64%) referentes à novos genes neste organismo. Utilizando-se a técnica de RAP-PCR dentro do mesmo modelo experimental (Dias Neto, 1997; Dias Neto *et al.*, 1997), verificou-se uma menor taxa de redundância e maior percentual de ESTs referentes a novos genes em *S. mansoni* quando comparada à metodologia convencional. Dias Neto (tese de doutoramento, 1997) analisou um total de 342 ESTs, onde foram identificados 326 ESTs (92%) referentes à novos genes em *S. mansoni*, com um número de clones redundantes equivalente a 10 ESTs (2,9%) e ausência de vetores sem inserto. Desta forma, torna-se evidente a eficiência da técnica de RAP-PCR frente à metodologia convencional. Os resultados obtidos indicam que a RAP-PCR oferece uma baixa

redundância no sequenciamento, baixa ocorrência de artefatos e alta frequência de sequências condizentes a genes novos, indicando uma alta aplicabilidade nos diversos projetos genoma (Dias Neto *et al.*, 1997).

Graças a simplicidade e a alta sensibilidade metodológica da RAP-PCR, onde pequenas quantidades de RNA-molde (0,5 a 5 ng) são suficientes para produzir um complexo perfil de bandas de cDNA, pode-se dizer que a aplicação desta técnica na geração de ESTs em larga escala oferece uma grande oportunidade de descoberta de novos genes no PGH, podendo auxiliar o próprio CGAP na busca pela caracterização molecular dos diversos passos envolvidos com a evolução carcinogênica.

1.8. O câncer de mama como modelo experimental

A mama humana é um tecido composto por células epiteliais, endoteliais e adiposas. A sua porção epitelial está imersa no estroma, formando uma estrutura altamente ramificada (vide figura 03). A principal função das células epiteliais é produzir o leite materno e distribuir este produto láctico para os mamilos.

Durante a vida da mulher, o epitélio mamário responde a vários estímulos químicos, na forma de hormônios e fatores de crescimento. A proliferação celular ocorre em resposta a alterações hormonais na adolescência, durante o ciclo menstrual e na gravidez. Os períodos de proliferação da glândula mamária são os mais susceptíveis a iniciação da tumorigênese. Cerca de 95% dos tumores produzidos na mama se iniciam nas células epiteliais (para revisão vide McKenzie, 1997).

O câncer de mama é o segundo tipo de câncer com maior mortalidade em mulheres nos países desenvolvidos. Em 1996, a *American Cancer Society* estimou que, nos Estados Unidos, cerca de 150 mil mulheres teriam um diagnóstico positivo para câncer de mama e aproximadamente 40 mil seriam vítimas fatais da doença. Estatísticas mostram que uma em cada oito mulheres desenvolverá câncer de mama em algum momento de sua vida (dados fornecidos pela *University of Pennsylvania Cancer Center*). Menos de 1% dos casos de câncer de mama ocorrem em homens.

Figura 03: Anatomia geral da mama. Glândula mamária é formada por ductos altamente ramificados, imersos em tecido adiposo. Ductos mamários interligam o mamilo aos lóbulos e respectivas terminações lobulares, envolvidas na produção do leite materno.

Dentro da perspectiva brasileira, o câncer de mama é o tipo de câncer que mais acomete a população feminina. Dos 284.205 novos casos de câncer com previsão de serem diagnosticados no ano 2000, o câncer de mama será o principal a atingir a população feminina, sendo responsável por 28.340 novos casos e 8.245 óbitos. Observou-se um aumento considerável da taxa populacional de mortalidade por câncer de mama de 1980 a 1997, de 6,14/100.000 para 9,31/100.000 mulheres (vide figura 04). A taxa bruta de óbitos estimados para o ano 2000 é de 9,78/100.000 habitantes e a de casos novos é de 33,58/100.000 habitantes. (dados fornecidos pelo INCA, <http://www.inca.org.br/>).

Dentre os sintomas e/ou sinais físicos mais comuns do câncer de mama, podemos citar (i) inchaço ou aumento da mama e/ou axilas, (ii) eritema e edema local, (iii) ulceração, (iv) expansão venosa com padrão irregular e (v) dores locais. Graças aos atuais avanços tecnológicos, o câncer de mama pode ser detectado em estágios bastante precoces de sua evolução, oferecendo uma maior probabilidade de cura para o paciente. Apesar disso, o exame físico (apalpamento local) e a mamografia ainda são os exames mais utilizados na detecção e prevenção de tumores mamários.

Figura 04: Gráfico ilustrativo mostrando a taxa bruta de mortalidade por câncer (eixo y) de algumas localizações primárias na população feminina brasileira, no período de 1980 a 1996 (eixo x). Esta taxa representa a incidência de óbitos em uma população de 100.000 mulheres (<http://www.inca.org.br/>).

Mulheres com maior risco de desenvolver câncer de mama são aquelas que possuem forte histórico familiar e/ou pessoal de câncer de mama, menarca precoce ou menopausa tardia e gravidez inicial acima dos 30 anos. Terapia com estrógeno a longo prazo, dieta rica em gorduras e uso abusivo de álcool tem sido citados como possíveis fatores de risco, mas sua correlação com a evolução tumoral não é conhecida (dados fornecidos pela *University of Pennsylvania Cancer Center*).

Estudos genéticos em famílias com histórico de câncer de mama mostram que mutações nos genes BRCA1 (cromossomo 17q21) e BRCA2 (cromossomo 13q12-13) são bastante frequentes nos seus respectivos membros (para revisão vide McKenzie, 1997). Mutações germinativas no gene BRCA1 são responsáveis por 50% dos casos hereditários de câncer de mama. Este gene codifica uma proteína de 190 kDa, envolvida com a inibição seletiva da proliferação de células da mama e ovário.

As alterações moleculares mais conhecidas que acometem o câncer de mama são (i) mutações e/ou deleções de genes supressores de tumor como *p53*, *RB-1*, *BRCA-1*, *nm23* e inibidor de CDK4 e (ii) amplificação e/ou superexpressão de oncogenes como *c-myc*, *c-erbB2*, *EGFR*, ciclinas D1 e E (para maiores detalhes vide Dickson & Lippman, 1995). Apesar da

função exata de muitos destes genes não ser conhecida, estes podem ser utilizados como marcadores, com valor prognóstico, na avaliação clínica de tumores de mama.

A evolução tumoral do câncer de mama é acompanhada por alterações qualitativas e quantitativas da expressão gênica (Porter-Jordan *et al.*, 1994). Dentre estas alterações, podemos citar (i) a expressão exacerbada de determinadas ciclinas e CDKs (quinases dependentes de ciclinas), associada a amplificação gênica de ciclina E (responsável pela transição G1-S) (Keyomarsi & Pardee, 1993) e (ii) a diminuição na expressão de certos membros da família S100 (proteínas de ligação ao cálcio), podendo causar alterações no metabolismo de cálcio e aparecimento de focos de calcificação mamário (Lee *et al.*, 1992).

No entanto, restam muitas dúvidas quanto aos genes essencialmente envolvidos na cascata de eventos que se propaga durante a carcinogênese mamária. A descoberta do gene *BRCA-2* é uma indicação deste fato, pois este gene aparece mutado com maior frequência em casos de câncer de mama masculino (independente de estrógeno), evidenciando as peculiaridades que existem nos diferentes quadros clínico-patológicos da doença (para revisão veja Wooster & Stratton, 1995).

A identificação de mudanças na expressão gênica no tecido mamário normal e neoplásico tem proporcionado o isolamento de novos marcadores tumorais (Watson & Fleming, 1994; Ji *et al.*, 1997), com possível utilidade no diagnóstico e/ou tratamento do câncer. Entretanto, até o presente momento, restam muitas dúvidas sobre os mecanismos celulares e moleculares relacionados com a ocorrência deste tipo de câncer e seus diferentes padrões histológicos. Técnicas que proporcionem a descoberta de novos genes humanos fornecerão, num futuro próximo, muitas informações que irão clarear estas e tantas outras dúvidas que restam nos campos da bioquímica e fisiopatologia humana.

2. OBJETIVOS

A partir de tumores malignos de mama, pretendemos gerar 6.000 ESTs pela técnica de RAP-PCR. Estas sequências serão analisadas contra sequências de nucleotídeos e/ou aminoácidos já depositadas nos diversos bancos de dados de acesso remoto (*GenBank, National Institutes of Health, USA*), visando a identificação dos genes humanos expressos.

De posse destes dados, a presente metodologia será avaliada quanto a (i) sua capacidade de normalizar diferentes populações de genes expressos, (ii) sua eficácia na identificação de genes humanos não totalmente sequenciados (em geral, genes raros e/ou pouco expressos) e (iii) seu potencial em gerar sequências informativas, referentes a porções codificadoras dos genes.

De acordo com a eficiência metodológica na identificação de genes pouco expressos, esperamos identificar genes novos ou ainda não descritos neste modelo tumoral. Para validação da técnica, pretendemos caracterizar genes humanos possivelmente correlacionados a processos de carcinogênese.

3. MATERIAIS E MÉTODOS

3.1. Material biológico

Seis amostras de carcinoma de mama, em diferentes estágios de desenvolvimento, foram obtidas em cirurgia no Centro de Tratamento e Pesquisas do Hospital do Câncer - A. C. Camargo, ligado ao Instituto Ludwig de Pesquisa sobre o Câncer - São Paulo/SP. As amostras foram isoladas após a cirurgia e mantidas a -70°C . As amostras teciduais foram manipuladas após o consentimento dos respectivos pacientes.

Informações sobre o paciente (idade e sexo) e devido estadiamento tumoral (escala TNM) (*American Joint Committee on Cancer*, 1993) são tabulados abaixo:

CASO	#001	#002	#003	#004	#005	#006
Nome (iniciais)	AWC	MAM	JT	NAGS	MHMB	MSCM
Sexo (F/M)	F	F	F	F	F	F
Idade (anos)	38	58	31	44	42	49
Estádio ¹	IIB	IIA	IIIB ²	IIB	IIB	IIA
Classificação histopatológica (escala TNM)	T2N1BM0	T2N0Mx	T4DN2M0 ²	T2N1BM0	T2N1BM0	T1N1BM0

1 - Carcinoma *in situ* é considerado como estágio 0; já os casos com metástases à distância, estágio IV. 2 - referente a avaliação clínica da paciente.

3.2. Purificação do mRNA

Um total de 15 preparações de RNA foram obtidas durante este trabalho. Todos os protocolos de purificação de mRNA seguem o mesmo princípio, ou seja, adsorção em resinas ou *beads* magnéticos (macropartículas) contendo oligo-dT.

As extrações iniciais de mRNA, denominadas RBT001 a RBT008, foram realizadas com o kit *QuickPrep*TM (Pharmacia). Dentre estas, apenas as preparações RBT007 e RBT008 são tratadas com DNase I (Gibco-BRL e Promega, respectivamente). O objetivo deste tratamento é a eliminação de DNA genômico como contaminante da preparação de mRNA. Para o tratamento de mRNA com DNase I, adicionam-se 1 a 5 unidades de DNase I para 50 μl de mRNA em

tampão 5,6 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4. A reação se processa por 15 minutos a 37°C. A reação é aquecida por 2 minutos a 95°C e, em seguida, tratada com 3 volumes de RNazol™ e 1/3 volume de clorofórmio. A solução é agitada vigorosamente por 15 segundos e centrifugada a 12000xg por 15 minutos a 4°C. O mRNA, presente na fase aquosa, é precipitado com isopropanol (1:1) por 12-16 horas a -70°C. A solução é centrifugada a 12000xg por 15 minutos a 4°C e, após descarte do sobrenadante, o *pellet* é lavado com etanol a 70%. A solução é novamente centrifugada e, após descarte da solução de lavagem, o *pellet* é seco a temperatura ambiente. O mRNA é ressuspensionado em 50 µl de água deionizada tratada com DEPC (Sigma).

As preparações RBT009 e RBT010 se referem a soluções-estoque de RNA total, obtidos pelo método de Trizol™. Apenas a preparação RBT010 foi tratada com DNase I (Promega). Para o tratamento de RNA total com DNase I, adicionam-se 3 unidades de DNase I para 300 µl de RNA em tampão 5,6 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4. A reação se processa por 15 minutos a 37°C. Em seguida, a solução é tratada com 500 µl de Trizol™ e agitada vigorosamente por 15 segundos. A solução é centrifugada a 12000xg por 15 minutos a 4°C. O mRNA, presente na fase aquosa, é precipitado com isopropanol (1:1) por 12-16 horas a -70°C. A solução é centrifugada a 12000xg por 10 minutos a 4°C e, após descarte do sobrenadante, o *pellet* é lavado com etanol a 70%. A solução é novamente centrifugada e, após descarte da solução de lavagem, o *pellet* é seco a temperatura ambiente. O RNA total é ressuspensionado em 100 µl de água deionizada tratada com DEPC. Este protocolo foi seguido para o tratamento das preparações RBT009 a RBT012.

Para a obtenção das preparações de mRNA RBT011, RBT012 e RBT013, o RNA total foi inicialmente obtido pelo método de Trizol™, tratado com DNase I (Promega). Posteriormente, o mRNA foi purificado através de diferentes protocolos. A purificação de mRNA para estas preparações foi feita, respectivamente, com os kits *MessageMaker*™ (Gibco-BRL), *PolyA Tract*™ (Promega) e *MACS Oligo(dT) Microbeads*™ (Miltenyi Biotec).

O RNA total da preparação RBT014 foi gentilmente cedido pelo *Ludwig Institute, NY, USA*. O RNA total foi tratado com DNase I (Promega). A purificação do mRNA foi feita com o kit *MACS Oligo(dT) Microbeads*™ (Miltenyi Biotec).

A preparação RBT015 foi obtida de maneira semelhante a RBT013. Entretanto, o tempo de reação do RNA total com DNase I foi superior (aproximadamente 45 minutos).

As preparações RBT013 a RBT015 foram fornecidas pelo grupo do Prof. Dr. Luis F. L. Reis (Laboratório de Inflamação, Instituto Ludwig de Pesquisa sobre o Câncer, São Paulo).

Alíquotas de 5-10 µl de mRNA puro (5-10 ng) são estocadas em freezer a -70°C.

Todas as informações sobre cada um das preparações de RNA obtidas são tabuladas a seguir:

	Tumor utilizado	Microdissecção (S/N)	Kit de extração	Seleção por oligo-dT (S/N)	Comentários
RBT001	Caso#001	S	QuickPrep	S	Nenhum
RBT002	Caso#002	N	QuickPrep	S	Nenhum
RBT003	Caso#002	N	QuickPrep	S	Re-purificação de RBT002
RBT004	Caso#002	N	QuickPrep	S	Duplicação do número de lavagens da resina
RBT005	Caso#003	N	QuickPrep	S	Duplicação do número de lavagens da resina
RBT006	Caso#004	N	QuickPrep	S	Duplicação do número de lavagens da resina
RBT007	Caso#004	N	QuickPrep	S	mRNA tratado com DNase
RBT008	Caso#004	N	QuickPrep	S	mRNA tratado com DNase
RBT009	Caso#005	N	Trizol	N	Nenhum
RBT010	Caso#005	N	Trizol	N	RNA total tratado com DNase
RBT011	Caso#005	N	Trizol MessageMaker	S	RNA total tratado com DNase
RBT012	Caso#005	N	Trizol PolyA Tract	S	RNA total tratado com DNase
RBT013	Caso#006	N	Trizol MACS	S	RNA total tratado com DNase
RBT014	Pool ¹	N	MACS	S	RNA total tratado com DNase
RBT015	Caso#002	N	Trizol MACS	S	RNA total tratado com DNase por um período de 45 minutos

1 - O material biológico fornecido para a preparação do RNA total foi proveniente de 8 tumores de mama em diferentes estágios de desenvolvimento (*Ludwig Institute, NY, USA*).

3.3. Validação do mRNA

Para garantir a pureza do mRNA a ser utilizado na geração de bibliotecas de cDNA, são feitas reações de PCR com *primers* específicos para globina (controle para DNA genômico) e região de D-loop (controle para DNA mitocondrial). Quantidades equimolares de DNA genômico são utilizados nas reações de PCR como controle positivo.

Mais recentemente (produção do RNA total para RBT015) foram padronizadas reações de PCR específicas para gliceraldeído 3-P desidrogenase (*GAPDH*, gene constitutivo) e exon 5 do gene supressor de tumor *p53*. Em futuros estudos, estes controles serão utilizados na avaliação da qualidade (grau de degradação) dos estoques de mRNA produzidos.

3.3.1. PCR específica para gene da β -globina

Um alíquota de 1-10 ng de mRNA (solução-estoque e/ou diluição 1:20) é adicionada ao tubo de reação contendo 4 pmoles de cada *primer* (PCO3: 5'-ACACAACTGTGTTCACTAGC-3'; PCO4: 5'-CAACTTCATCCACGTTCCACC-3'); 200 μ M de dNTPs e 1 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 20 μ l). A reação é submetida a um ciclo inicial de denaturação a 95°C por 5 minutos, seguida por 35 ciclos de denaturação a 95°C por 1 minuto, anelamento 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 μ l do produto da reação é aplicado em gel de poliacrilamida a 8% e, após eletroforese, analisado por coloração com prata. A banda referente à β -globina possui aproximadamente 120 pares de bases.

3.3.2. PCR específica para região D-loop de DNA mitocondrial

Um alíquota de 1-10 ng de mRNA (solução-estoque e/ou diluição 1:20) é adicionada ao tubo de reação contendo 20 pmoles de cada *primer* (Mitfw: 5'-AATAACAATTGAATG TCTGCAC-3'; Mitrv: 5'-TTGAGGAGGTAAGCTACATA-3'); 200 μ M de dNTPs e 0,5 unidades 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 μ l). A reação é submetida a 35 ciclos de denaturação a 95°C por 1 minuto, anelamento 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 μ l do produto da reação é aplicado em gel de poliacrilamida a 8% e, após eletroforese, analisado por coloração com prata. A banda referente à região de D-loop possui aproximadamente 1200 pares de bases.

3.4. Síntese da primeira fita de cDNA

Para a síntese do cDNA, uma alíquota de 5-10 µl de mRNA (5 a 50 ng) é misturada com 25 pmoles de *primer* arbitrário e aquecida a 65°C por 10 minutos. Em seguida, são adicionados 5 nmoles de dNTPs e 100 unidades de Superscript™ II Rnase H⁻ transcriptase reversa (Gibco-BRL) em tampão 50 mM Tris-HCl pH 8,3; 75 mM KCl; 3 mM MgCl₂ e 10 mM DTT (volume final de 20 µl). A mistura é incubada a 37°C por 1 hora e, posteriormente, a enzima é denaturada a 95°C por 5 minutos. O cDNA total obtido é estocado a -70°C.

3.5. Obtenção de perfis de expressão por RAP-PCR

Três tipos de perfil de expressão foram obtidos, utilizando-se diferentes estoques de cDNA como molde para a reação. Estes perfis são classificados em (i) perfis X-X, onde o *primer* utilizado na reação de PCR é o mesmo que foi utilizado na síntese de cDNA, (ii) perfis X-Y, onde o *primer* utilizado na reação de PCR é diferente do *primer* utilizado na síntese de cDNA e (iii) perfis X-Xn, onde o *primer* utilizado na reação de PCR é idêntico àquele utilizado na síntese de cDNA, entretanto, possui um nucleotídeo a mais na sua porção 3' final (*template* da reação de RAP-PCR é o produto de amplificação do respectivo perfil X-X).

Todos os *primers* utilizados na síntese de cDNA e geração dos diferentes perfis de RAP-PCR são listados a seguir:

<i>Primer</i>	Sequência (5'-3')	<i>Primer</i>	Sequência (5'-3')	<i>Primer</i>	Sequência (5'-3')
PBT001	AATCCGTCGAGCAGAGTT	PBT064	CGTCCGAGCCATGGCCAT	PBT126	CGTTTCTTTTCATTNCCAG
PBT002	CCCTTACCCTTACCCTTACCCTAA	PBT065	AGCAGAGCTCATGGTGGG	PBT127	CAACGCATGTGCCGCTCCG
PBT003	TGTTGCTTGTTGGATAGAC	PBT066	CTCCTAGGTTGGCTCTGACTGT	PBT128	CAACGCATGTGCCGCTCCC
PBT004	AGTGGTGTTGGTGGTGATG	PBT067	GTGGCAAGTGGCTCCTGACC	PBT129	GACGGATAGCGGATCTGCA
PBT005	AGTTCTCCCCCTTGGTGG	PBT068	GTCCGCGCCATGGCCATC	PBT130	GACGGATAGCGGATCTGCC
PBT006	AACAGGAAGACCTTCCTC	PBT069	CGCAAATTCCTTCCACTCG	PBT131	TGATGGCGCTGATGACACC G
PBT007	CATTTCTCGCTCTGTCCAC	PBT070	TCATCTTGGGCTGTGTTATC	PBT132	TGATGGCGCTGATGACACC C
PBT008	ACTCCACCATAGATCACTCCC	PBT071	GGGTGGCAAGTGGCTCCTGAC	PBT133	CAGGCAAAAGCTCCTAC
PBT009	GGTGCACGGTCTACGAGACCT	PBT072	ATTTCTTACTGCCTCTTGC	PBT134	CCAATCAACATCCAACCCA TTAGTAAAGATGCCG
PBT010	CCTGTGAGGAACTACTGTCTT	PBT073	ACCTCGTTAGTGCTCCCTGG	PBT135	ACCGCAGGACCATTTGGC
PBT011	GCACTCGCAAGCACCTATCA G	PBT074	GGTGGTAGTGGGGTTGG	PBT136	CGCAGGACCATCGGTATT
PBT012	GTGACCAGGGTACCTTGGCCG CAG	PBT075	AGGAACGGCGGCAAGAGG	PBT137	AATATATCCGGCTATCC

PBT013	CTAGGTGATCAAGATCT	PBT076	CTATCAACTGAGGCGATTCCG	PBT138	ATGCTTTGCTGAATACCG
PBT014	CGAGGTGACGGTATCG	PBT077	GTCAGCATTTCCAGCAGCG	PBT139	GAATCCAGCAAAACATTAC
PBT015	TCTGGAACCTCTCTGGTCTCT CAGCTGG	PBT078	GCAAGTGTCAATCACATGAGCG	PBT140	CGACTAGTCCCAACCATAC
PBT017	GGAGGATCCTGTAGGAATGGT ATCTCG	PBT079	GTCGTGGTTGTTGAAAGTGCC	PBT141	GGAAGGGGATGTTTGTG
PBT018	TGATGGCGCTGATGACACC	PBT080	CTTCAAACGCTGCAACACC	PBT142	TGTATTGGTGGTGTCTC
PBT019	TGATGGCGCTGATGACTCC	PBT081	CAATCACCTTGGAGCCTTTCG	PBT143	GGTGACAATCGTGACAAC
PBT020	GAGTACTGGCGGAAAGACGG	PBT082	CATTGAACAGCAGTGCAGG	PBT144	TAGTCAGCGCCCTG
PBT021	GGGGAGATGATCCGAGCC	PBT083	GCAACCAGCTTACGAATTGCG	PBT145	GAGTGGGTCAAGTGTCTC
PBT022	ACCCTCAGCAAATATCAG	PBT084	GAGTTGGTGTCTGGAGATCG	PBT146	CACTGCAGGGATAGTCA
PBT023	ATGCTGGTCTCTTTGTGG	PBT085	TCGGCGGCAATGACTTCG	PBT147	CCAGTAGTTTGTGATGGGT
PBT024	GAGTCGACAGACATGTTCTGGT	PBT086	CATTGGATAAGCGGATGAGGG	PBT148	CCACGTGATACCGAAAG
PBT025	AGTGTCTGTCAGCCTCCAGG	PBT087	GTAATCGCAACACCTGTCATCC	PBT149	GGAATACAAGATATTTCCG C
PBT026	GCGGAGCTTGCAGTGAG	PBT088	CTTTGCCGTGTGATTGCG	PBT150	GGTGGTGAATAACGATGGC
PBT027	TTGTGTAGCAATCCCTCAGC	PBT089	GGCAATGCTGGAAGACAGG	PBT151	TCAGCTGGACAAGGATGG G
PBT028	CCACAGGCAAGCCAGTTTA	PBT090	GGATACGGGGATATTGAGTTGG	PBT152	CAGTCTAGTGATTTTAGGC G
PBT029	ACACGGCCCTGTATTACTGTG	PBT091	CCTCCTAGCCCTCTTACC	PBT153	AGGGAGTGTCTGTCTTTTG
PBT030	TGAGGAGACGGTACCAGGG	PBT092	CAACGCATGTGCCGCTCC	PBT154	TCGCTTTTGTGTTCAATG
PBT031	GATGACTCCAGCCACGACCT	PBT093	GACGGATAGCGGATCTGC	PBT155	TACGGCACTGTTAGAGC
PBT032	CAGTGTAACCTGGGGGC	PBT094	GCGGTCCAAACTCATCAATGTA TCT	PBT156	GCGTTTTGATGTGAGAGA
PBT033	CCGTTCCTATATTTCTCTGG	PBT095	AATGGCATTTGTTGGGGTA	PBT157	TGTTTAATCTCATTCTGCA A
PBT034	ATTTATAGGGCCATGACCAG	PBT096	TGGTAGTACTGCTCTGGGGG	PBT158	CCTTCCAACATCTCATTCT
PBT035	CTGGCATTCTGGGAGCTTCA	PBT097	GAGTCGATTTTCAAGTTG	PBT159	TCATCCCAGAATAAACCTT
PBT036	CGCGGATCCGGTTCTGTGTTAG G	PBT098	TCGAAAGAGGAGGGTGGTC	PBT160	CCACGGTGTCCATAAA
PBT037	CACAGCCCTGTGCCTCTTCTC	PBT099	ACCAACATTGGCCTCGAC	PBT161	GGTTGAAGTTAGATTCTCT TC
PBT038	GCAGAGTGTGTGCTACTAGA	PBT100	GCGGCCGCCGAATTAACCTCA CTAAAG	PBT162	TAAACCATTCAAATATCCA C
PBT039	CCCAAGGGCTAGAGGAGC	PBT101	ATGATGCGTGCTGCCAAG	PBT163	GCAITTTTGTGAATGATGTA T
PBT040	ACCAGGGTCCCTTGCCCCA	PBT102	ACGACCACCCTCCTCTTTC	PBT164	ATGCACCCATTCCGA
PBT041	CAGCCTTGAAACATTGATGG	PBT103	CGGCCGCGTAATACGACTCACT ATAGGGC	PBT165	CATTCAAATATCCACACAA TAC
PBT042	CGTGTGGTACCACTCCTG	PBT104	TTCCAGCAACTGATCAATTG	PBT166	TGGTTTGAAACAGCACTC
PBT043	GAGGGATCCAGTAAATGCTGC AGTTCAGAGG	PBT105	CCAGGATCCCCCGAGTCAAGA	PBT167	CGTACTGTGGTGTACCCT
PBT044	CCGTGGCATATCATCCCCAAG CTTAGA	PBT106	TACATCCACTGGTACCTACACC AG	PBT168	CACACCCAATACTGCC
PBT045	GAGGGATCCGTGTCGAAAGCA TGAAGGAC	PBT107	CTAGAATTCCAGGGTGTGTG GAATCAGGA	PBT169	GGAAAATTAGCCGACG
PBT046	CACAGACACCCCACTCTATC	PBT108	GCGTTNGTGTCTGGCGGT	PBT170	GGGTGCAGCCAATATG
PBT047	GGGCCACTGTGGCATCATGAT TGG	PBT109	CCTGTCAAGAAATCCTGCCA	PBT171	CAGGGTGCAGCCAAT
PBT048	GGACCTTCCTTGGTGTGTG	PBT110	GGCTCTAANGAAGTCATACC	PBT172	CGGGGAGAAATATCGTTAT
PBT049	AGATTGAACTCATCAGAGCT ACAGGAAC	PBT111	CTCAGGGCTCAGTTGCTTGT	PBT173	CATTGATTTCAACGTGGT
PBT050	GGAGCAGGTGTGCAGAGGG	PBT112	GTGTTTTGCTTTTTGGGGG	PBT174	GGTTGCGGGGGAITGTTGC TGACTGGAGG
PBT051	CTTGATGATATCTCCACTGGAG GC	PBT113	GCACCAACCTTTTTCTCAC	PBT175	GCGGCCGCGGAATTAACCC TCACTAAAG
PBT052	AAGATGGCAAAGCCTCTTCCC TG	PBT114	CTGTGATAGGATGGATGGTGA	PBT176	CGCATAAACGCAGCACAC GCAAAAGGTGGTTAG
PBT053	CCATCTGTGAGAAGCCACAGG	PBT115	CAGTTACCTCTCTTTTTTCTC	PBT177	CGCCCAAACAGCTTCTCC ACTTCTCGG
PBT054	GTCCAGATGTCCAGCTCAC	PBT116	GCGTTGACTGGAGATGAGGT	PBT178	GCGTGGGCTTCTCTTCG GTATACACTC
PBT055	ACCTGAGGAGACGGTGACC	PBT117	CTCCACAGTAATCTTCCT	PBT179	CGATCAAACCTTTTATGT TGCTGGCGGA
PBT056	GGAAAATTCTCCATCAGGAGG CCTGT	PBT118	GGTTGATTCCTGATTTTTCG	PBT180	CCGCTCCTGATGGCTATTT GGCTTGAGTTTCG

PBT057	TCCTGCCTGCCTGGCTGTACTG TAGA	PBT119	CTTCAANCTATGGGGCACAC	PBT181	AATCTACCTCTTTCCGAAG C
PBT058	AGGCAGGAGCAGGAGGAG	PBT120	GTTGGCATAATACAGACTTTG	PBT182	AGCCATCATGGTGTCAAGC T
PBT059	AAGCATTGTTGTGTAGCCCTAA	PBT121	CTTGGGGGGNTGGAGTGAN	PBT183	CGCTCAAGGTTACGCACTC CAGACTCACG
PBT060	CTGGTCCCAGCAGTCAGG	PBT122	GCACCTCATACACAGAAACC	PBT184	TCGAAAGAGGAGGGTGGT C
PBT061	CTGTTGGTGCTGCCGCT	PBT123	CCTGNTTTTGTCCGTTGC	PBT185	GCTTTCGTTACTGTTAATG CCGCGTCTTTC
PBT062	CTCCCCTATCCGGCAC	PBT124	GTGTAGGGAAGGCTAAAGAAA	PBT186	ATTACATCAAACGTGAAGC AATGATTCCGA
PBT063	CCACCTCATTCCCATCTA	PBT125	GGACTTGGAGCCTGTACTTT	PBT187	TTTTTTTTTTTT

3.5.1. Obtenção de perfis X-X

Para a obtenção do perfil de expressão, 0,1-1 µl do cDNA anteriormente obtido (cerca de 0,5 ng) é utilizado como molde na reação de RAP-PCR. A amplificação é realizada após adição de 2,5 nmoles de dNTPs; 14 pmoles de *primer* arbitrário; 0.5 unidade de *Taq* DNA polimerase (Gibco-BRL ou Promega-*sequencing grade*) em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 20 µl). A reação é submetida a um ciclo inicial de denaturação a 95°C por 1 minuto, anelamento a 37°C por 2 minutos e extensão a 72°C por 2 minutos, seguidos por 35 ciclos de 95°C por 45 segundos, 45°C por 1 minuto e 72°C por 90 segundos, com uma extensão final de 72°C por 5 minutos.

3.5.2. Obtenção de perfis X-Y

Para a obtenção do perfil de expressão, 0,1-1 µl do cDNA anteriormente obtido é utilizado como molde na reação de RAP-PCR. A amplificação é realizada após adição de 2,5 nmoles de dNTPs; 1 pmol de *primer* utilizado na síntese de cDNA; 14 pmoles de *primer* diferenciado (para a geração do perfil); 0.5 unidade de *Taq* DNA polimerase (Gibco-BRL ou Promega-*sequencing grade*) em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 20 µl). A reação é submetida a um ciclo inicial de denaturação a 95°C por 1 minuto, anelamento a 37°C por 2 minutos e extensão a 72°C por 2 minutos, seguidos por 35 ciclos de 95°C por 45 segundos, 45°C por 1 minuto e 72°C por 90 segundos, com uma extensão final de 72 °C por 5 minutos (método I).

Como alternativa (método II, modificado de McClelland & Welsh, 1995), a amplificação é realizada após adição de 2,5 nmoles de dNTPs; 10 pmoles de *primer* utilizado na síntese de

cDNA; 80 pmoles de *primer(s)* arbitrário(s) diferenciado(s) (para a geração do perfil); 1 unidade de *Taq* DNA polimerase em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 20 µl). A reação é submetida a um ciclo inicial de denaturação a 94°C por 30 segundos, anelamento a 37°C por 5 minutos e extensão a 72°C por 5 minutos, seguida por 35 ciclos de denaturação a 94°C por 30 segundos, anelamento a 37°C por 1 minuto e extensão a 72°C por 2 minutos, com uma extensão final de 72 °C por 5 minutos.

3.5.3. Obtenção de perfis X-Xn

Para a obtenção do perfil de expressão, 0,1-1 µl do perfil X-X anteriormente obtido é utilizado como molde na reação de RAP-PCR. A amplificação é realizada após adição de 2,5 nmoles de dNTPs; 14 pmoles de *primer* Xn; 1 unidade de *Taq* DNA polimerase (Gibco-BRL ou Promega-*sequencing grade*) em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 20 µl). A reação é submetida a 36 ciclos de denaturação a 95°C por 45 segundos, anelamento a 57°C por 1 minuto e extensão a 72°C por 90 segundos, com uma extensão final de 72°C por 5 minutos.

Os produtos de PCR são resolvidos por eletroforese em gel de poliacrilamida a 8% e visualizados na coloração por prata.

3.5.4. Obtenção de perfis radioativos para controle de qualidade

Para a obtenção do perfil de RAP-PCR radioativo, 0,1-1 µl do cDNA anteriormente obtido (cerca de 0,5 ng) é utilizado como molde na reação de RAP-PCR. A amplificação é realizada após adição de 2,5 nmoles de dNTPs; 14 pmoles de *primer* arbitrário; 0,1 µl de [α -³²P]dCTP (10 µCi/µl) e 0.5 unidade de *Taq* DNA polimerase (Gibco-BRL ou Promega-*sequencing grade*) em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 20 µl). A reação é submetida a um ciclo inicial de denaturação a 95°C por 1 minuto, anelamento a 37°C por 2 minutos e extensão a 72°C por 2 minutos, seguidos por 35 ciclos de 95°C por 45 segundos, 55°C por 1 minuto e 72°C por 90 segundos, com uma extensão final de 72°C por 5 minutos. Os produtos marcados de PCR são resolvidos por eletroforese em gel de

poliacrilamida a 8% e analisados após exposição do gel devidamente seco a filme de raio-X (3 horas de exposição a -70°C).

3.6. *Ligação dos perfis de cDNA a vetores plasmidiais*

Para a produção de insertos com extremidades abruptas e fosforiladas (porção 5'), um volume de 16 µl do produto de PCR obtido (referente ao perfil de cDNA) é tratado com 6,4 unidades de fragmento Klenow de DNA polimerase e 9,5 unidades de T4 polinucleotídeo quinase (Pharmacia), em tampão apropriado (1X *Blunting/Kinasing buffer*). A solução é incubada a 37°C por 45 minutos e, em seguida, tratada com fenol/clorofórmio (v/v) para a denaturação enzimática. A purificação é feita conforme protocolo do fornecedor (*SureCloneTM Ligation Kit*, Pharmacia).

A ligação dos perfis de cDNA é feita no sítio *SmaI* defosforilado do plasmídeo pUC18. A mistura de DNA (7 µl da reação de *blunting/kinasing* para 50 ng de plasmídeo) é corrigida para 0,5 mM rATP; 5 mM DTT; 10 mM Tris-HCl pH 8,0 e 10 mM MgCl₂. Adicionam-se 20 unidades de T4 DNA ligase (Pharmacia) para um volume final de 20 µl de reação. A reação é incubada a 16°C por um período de 4-16 horas.

A eficiência da ligação dos perfis de cDNA em vetor pUC-18 é avaliada por PCR, utilizando-se *primers* específicos ao sítio de clonagem do vetor (M-13F: 5'-TCCCAGTCACGACGT-3'; M-13R: 5'-AACAGCTATGACCATG-3') (Dias Neto *et al.*, 1996b). A reação é feita utilizando-se 0,5 µl de reação de ligação, 3 pmoles de cada *primer*, 200 µM de dNTPs e 0,5 unidades de *Taq* DNA polimerase (Gibco-BRL) em tampão 1,5 mM MgCl₂; 50 mM KCl e 20 mM Tris-HCl pH 8,4 (volume final de 10 µl). A reação se processa por 35 ciclos de denaturação 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 amostra é analisada em gel de agarose a 1% (suplementado com brometo de etídeo) ou poliacrilamida a 8% corado com prata.

3.7. Preparação de células bacterianas competentes

Para a preparação de bactérias competentes são utilizadas linhagens bacterianas de *E.coli* DH5 α ou JM109. Uma alíquota do estoque bacteriano mantido a -70°C é plaqueada em LB sólido (10 g/l de bacto-triptona, 5 g/l de extrato de levedura, 10 g/l de NaCl e 15 g/l de bacto-ágar) e incubada por 16 horas a 37°C para o devido crescimento celular. Cerca de dez colônias isoladas são inoculadas em 260 ml de meio LB líquido (10 g/l de bacto-triptona, 5 g/l de extrato de levedura e 10 g/l de NaCl) e incubadas a 18-22°C sob agitação vigorosa a 200 rpm até atingir uma densidade óptica de 0,6 a 600 nm. Ao atingir esta densidade, o frasco é imediatamente colocado em gelo por 10 minutos e distribuído em seis alíquotas de 42 ml em tubos de centrifuga de 50 ml. Os tubos são centrifugados a 3500 rpm (centrífuga Sorvall RC5B) por 15 minutos a 4°C. Cada precipitado é ressuscitado em 12 ml de solução TB (*Transformation Buffer*: 10mM Pipes, 55mM CaCl₂, 250mM KCl) gelado e, após a ressuspensão, as alíquotas são distribuídas para dois tubos contendo 36 ml cada um. Os tubos são incubados por 10 minutos no gelo e novamente centrifugados conforme descrito acima. O precipitado é ressuscitado em 18,6 ml de TB contendo 1,4 ml de DMSO (solução TB com DMSO a 7%). A suspensão de células é incubada em gelo por 10 minutos e posteriormente aliquoteada em tubos *ependorf*, sendo imediatamente imersas e estocadas em nitrogênio líquido.

3.8. Produção de mini-bibliotecas de cDNA

São aliquoteados 10 μ l de reação de ligação para proceder a transformação de 100 μ l de *E.coli* DH5 α ou JM109 competentes. A transformação celular é feita por choque térmico (Sambrook *et al.*, 1989). Após a transformação, as bactérias são incubadas em 1 ml de meio LB líquido a 37°C, sob agitação, por 1 hora. O plaqueamento é feito em meio LB sólido suplementado com 100 μ g/ml de carbenicilina, 40 μ g/ml de X-gal e 0,5 mM IPTG. As placas contendo as respectivas mini-bibliotecas são mantidas em estufa a 37°C por 12-16 horas.

Foi produzido um número superior a 200 mini-bibliotecas para a geração de ESTs. Na prática, um total de 170 mini-bibliotecas de cDNA foram manipuladas para a obtenção das 6000 ESTs propostas neste trabalho.

3.9. Seleção das colônias de interesse

Os vetores de clonagem da série pUC proporcionam uma seleção prévia de clones por α -complementação (Sambrook *et al.*, 1989). O sítio múltiplo de clonagem (sítio *polilinker*) está inserido na extremidade 5' do gene da β -galactosidase (*lacZ*), presente no vetor. Esta porção gênica codifica um fragmento amino-terminal da enzima, o qual se associa à estrutura enzimática restante, expressa pela própria bactéria, tornando-a ativa. Quando ocorre o processo de clonagem (ligação de inserto ao sítio *polilinker*), a informação gênica é rompida e resulta na expressão de um fragmento amino-terminal incapaz de proceder a α -complementação, causando a inativação enzimática. A atividade da enzima β -galactosidase pode ser avaliada na presença de seu substrato cromogênico X-gal, originando colônias de tonalidade azulada quando a enzima se encontra ativa (ausência de inserto) e colônias de tonalidade branca quando a enzima é inativada (efetivação da clonagem).

Uma amostragem de colônias brancas (em geral, até 100 colônias) é previamente analisada por PCR, utilizando-se os *primers* que flanqueiam o sítio múltiplo de clonagem do pUC18 (M-13F: 5'-TCCCAGTCACGA CGT-3'; M-13R: 5'-AACAGCTATGACCATG-3'). Para tanto, uma alíquota de cada colônia é coletada, com o auxílio de palito estéril, e transferida para a parte inferior dos tubos de reação. Os insertos são amplificados utilizando-se 3 pmoles de cada *primer*, 200 μ M de dNTPs e 0,5 unidades 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 μ l). A reação se processa por 35 ciclos de denaturação 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 a amplificação, as amostras são analisadas em gel de agarose a 1% ou poliacrilamida a 8% corado com prata. Este procedimento permite reduzir o sequenciamento de plasmídeos sem inserto e favorecer o sequenciamento de insertos de diferentes tamanhos (diminuição da redundância global). As colônias contendo plasmídeos de interesse (insertos não-redundantes com tamanho superior a 100 pb) são crescidas a 37°C por 12-16 horas, sob agitação, em 5 ml de meio LB (suplementado com 50-100 μ g/ml de carbenicilina).

Como alternativa, os produtos de PCR gerados a partir de colônias isoladas podem ser diretamente utilizados como molde para a reação de sequenciamento, sem a necessidade de purificação de plasmídeos.

3.10. Mini-preparação de plasmídeos

A preparação dos plasmídeos segue um método modificado de lise alcalina, originalmente descrito por Ozaki & Cseko (1984). O kit utilizado nas mini-preparações corresponde ao *Wizard Plasmid Minipreps* (Promega, USA). O volume final de cada preparação é de 50 µl (cerca de 20 µg de plasmídeo).

3.11. Geração de ESTs

Os plasmídeos purificados ou produtos de PCR de colônia são sequenciados pelo método de terminação da cadeia por 2',3'-didesoxirribonucleotídeos (ddNTPs) (Sanger *et al.*, 1977). Os *primers* M13F ou M13R são utilizados para o sequenciamento. Os laboratórios de pesquisa dos seguintes professores auxiliaram na geração das 6000 ESTs:

Prof. Dr. Andrew J. G. Simpson (Instituto Ludwig/São Paulo) - Grupo IL

Prof. Dr. Sergio V. Almeida (USP/São Paulo) - Grupo QV

Prof. Dr. Fernando F. Costa (Unicamp) - Grupo CM

Prof. Dr. Marcelo R. S. Briones (EPM/São Paulo) - Grupo PM

Prof. Dr. Marco A. Zago (USP/Ribeirão Preto) - Grupo RC

Profa. Dra. Maria A. Nagai (USP/São Paulo) - Grupo MR

3.11.1. Sequenciamento em ALFexpressTM DNA Sequencer (Amersham Pharmacia Biotech)

As reações de sequenciamento são feitas por ciclagem térmica, utilizando-se o kit *ThermosequenaseTM* (Amersham). O programa consiste em uma denaturação inicial a 95°C por 2 minutos, seguida por 25 ciclos de anelamento a 55-60°C por 36 segundos, extensão a 72°C por 84 segundos e denaturação a 95°C por 36 segundos. A extensão final é prolongada por um

período de 5 minutos. O produto final das reações é denaturado conforme protocolo do kit e submetido à eletroforese em gel de poliacrilamida 6% (7M uréia) para o sequenciamento.

3.11.2. Sequenciamento em ABI Prism 377TM DNA Sequencer (Perkin Elmer Applied Biosystems)

As reações de sequenciamento são feitas por ciclagem térmica, utilizando-se o kit *BigDyeTM Terminator Cycle Sequencing Ready Reaction Kit* (Perkin Elmer). O programa consiste em uma denaturação inicial a 96°C por 5 minutos, seguida por 25 ciclos de denaturação a 96°C por 30 segundos, anelamento a 50°C por 15 segundos e extensão a 60°C por 4 minutos. O produto amplificado é precipitado conforme protocolo do kit. O produto é ressuspensionado em 3 µl de tampão de amostra (1:3 de *blue dextran* e formamida). A amostra é submetida à eletroforese em gel denaturante (Perkin Elmer) por um período de 3 a 7 horas para o sequenciamento.

3.11.3. Sequenciamento em MegaBaceTM DNA Sequencer (Amersham Pharmacia Biotech)

As reações de sequenciamento são feitas por ciclagem térmica, utilizando-se o kit *Thermo SequenaseTM II Dye Terminator Cycle Sequencing Kit* (Amersham Pharmacia). O programa é composto por uma série de 40 ciclos de denaturação a 95°C por 30 segundos, anelamento a 55°C por 15 segundos e extensão a 60°C por 1 minuto. O produto amplificado é processado conforme protocolo do kit. A amostra é injetada a uma voltagem de 3 kV por 50 segundos e submetida à eletroforese capilar por um período de 100 minutos a uma voltagem de 9 kV para o sequenciamento.

3.12. Validação de mini-bibliotecas de cDNA

A validação é realizada a partir do sequenciamento parcial de 10 a 20 clones de cada mini-biblioteca gerada. Para um bom aproveitamento de sequências novas e de baixa redundância, foram estabelecidos os seguintes critérios de controle de qualidade:

- a) quantidade de elementos repetitivos inferior a 10% (submissão ao programa *RepeatMaker*, acesso remoto em <http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>);
- b) número de sequências correspondentes a DNA mitocondrial e/ou RNA ribossomal inferior a 10% (análise no Genbank através do programa BLAST, acesso remoto em <http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast?Iform=0>);
- c) número de sequências correspondentes a genes bacterianos inferior a 5% (representatividade deve ser baixa);
- d) sequências com tamanho superior a 100 pares de bases;
- e) redundância entre sequências da mesma mini-biblioteca inferior a 25%.

Mini-bibliotecas de cDNA que respeitam os critérios acima mencionados são utilizadas na geração de ESTs em maior escala (sequenciamento de 50 a 200 clones por mini-biblioteca).

3.13. Análise de similaridade das ESTs frente a bancos de dados de acesso remoto

As ESTs obtidas são editadas (remoção de sequências do plasmídeo e *primers*) e analisadas frente a um banco de dados específico para elementos repetitivos (*RepeatMasker homepage*). Sequências que contém tais elementos são "mascaradas", podendo ser descartadas caso o percentual de elemento repetitivo na sequência seja muito alto (acima de 30%), o que inviabiliza as submissões subsequentes. Posteriormente, é feita a análise de similaridade das ESTs obtidas frente a sequências nucleotídicas (BLASTn) e de aminoácidos (BLASTX) depositadas no *Genbank* (NCBI, USA). Sequências que possuem um valor de E superior a 10^{-15} , quando analisadas contra sequências de DNA (BLAST(nr) e/ou dbEST), ou um valor de E superior a 10^{-5} , quando analisadas contra proteínas (BLASTX), são consideradas sequências novas. A redundância das ESTs frente a genes e/ou ESTs já conhecidas são analisadas no banco de dados do Unigene (acesso remoto em <http://www.ncbi.nlm.nih.gov/UniGene/index.html>).

As submissões contra bancos de dados para cada uma das ESTs geradas segue a seguinte ordem: (i) similaridade a proteínas humanas; (ii) similaridade a *contigs* humanos (Unigene); (iii) similaridade a ESTs humanas sem similaridade ao Unigene; (iv) similaridade a proteínas não-humanas; (v) similaridade a ESTs não-humanas; (vi) similaridade a DNA genômico (bancos de dados HTGS e GSS). As informações para a composição de cada banco de dados para busca de

similaridade foram importadas dos bancos de dados internacionais (*GenBank*, *EMBL* e *DDBJ*) e devidamente organizadas *in loco* no Laboratório de Bioinformática do Instituto Ludwig de Pesquisa sobre o Câncer / São Paulo.

As referidas submissões foram realizadas no laboratório de Bioinformática do Instituto Ludwig / São Paulo, sob coordenação do Prof. Dr. Sandro J. de Souza.

3.14. Obtenção da sequência completa de novos genes humanos

Algumas sequências expressas, com alta similaridade a genes já descritos em outros organismos, serviram de ferramenta para a obtenção do cDNA completo de genes ortólogos humanos. Para tanto, a abordagem inicialmente utilizada foi baseada na técnica de RACE-PCR. Posteriormente, uma série de *primers* foram desenhados, a partir de sequências de DNA e/ou ESTs humanas já descritas com similaridade significativa a estes genes ortólogos, de modo a proporcionar a amplificação de vários fragmentos gene-específicos que, após devida montagem, resultariam na sequência completa dos referidos genes.

3.14.1. Amplificação por RACE-PCR (*Rapid Amplification of Complementary DNA Ends*)

A técnica de RACE-PCR visa a amplificação gênica completa a partir da construção de *primers* baseados em porções gênicas centrais e previamente conhecidas (Frohman *et al.*, 1988; Frohman, 1993). Para tal finalidade, é utilizado um *pool* de cDNA, isolado de glândula mamária normal, e flanqueado por adaptadores idênticos em ambas extremidades, como molde de amplificação (*Marathon-Ready cDNA*, Clontech). A amplificação é realizada após a adição de 5 ng de cDNA em um tubo *epENDORF* contendo 5 pmoles de *primer* gene-específico para SEMA6B (SEM2F: 5'-CTTGGAGTCATGTTTCACGG-3') ou NOTCH2 (NOT2F: 5'-TCACACTGGTTGTTGATATAATCCC-3'), 5 pmoles de *primer* correspondente à extremidade flanqueadora (AP1, *Marathon-Ready cDNA kit*), 1,0 unidade de *rTth* DNA Polimerase, 800 µM de cada dNTP, 1,5 mM Mg(OAc)₂ e 7,5 µl de tampão *3.3X PCR buffer* (*GeneAmp XL PCR kit*, Perkin-Elmer), num volume final de 25µl. A reação é submetida a 35 ciclos de denaturação a 94°C por 40 segundos, anelamento a 60°C por 1 minuto e extensão a 72°C por 2 minutos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por dois

ciclos a 68°C e 64°C, utilizando as mesmas condições de denaturação e extensão). Esta metodologia proporciona a amplificação de fragmentos de cDNA gene-específicos (ao redor de 800 pb) a serem clonados e utilizados posteriormente como sondas em análises de Northern blot.

3.14.2. Amplificação por primers desenhados a partir da análise racional dos bancos de dados

A sequência nucleotídica completa dos genes murinos *Notch2* e *SemaZ*, com alta similaridade às sequências expressas IL-BT000-001 e IL-BT000-002 (e respectivos produtos de RACE-PCR), é submetida à análise de similaridade pelo programa BLAST (<http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=1>) contra os bancos de dados de DNA (HTGS) e ESTs (dbEST) humanas. A similaridade observada permite o desenho de *primers* gene-específicos a partir de sequências nucleotídicas humanas já disponibilizadas. Os *primers* gene-específicos desenhados e utilizados nas reações de RT-PCR (e sequenciamento) são listados a seguir.

Sema 6B (5'-3')	Notch2 (5'-3')	Notch2 (5'-3')
SEM2-2F: AGGTAGTTAACTCCATCGCAATC	NOT2F: TCACACTGGTGTGTGATATAATCCC	NOT14F: CTCCTTCTCTTGCATGTGCC
SEM2F: CTTGGAGTCATGTTTCACGG	NOT2R: ACCTGTCTGAGCCAGTATTGTG	NOT15F: GCAAGTGCCAGGCAGGATTT
SEM2R: GGGATGCTCTTCACAGCTACT	NOT3R: GGGATTATATCAACAACCAAGTGTGA	NOT15R: CATGGAGACCGACTGCAGAGA
SEM3R: CCGTGAAACATGACTCCAAG	NOT4R: CTTGGCAAGGTAGGGAGGCTAG	NOT16F: TACGTCAACAGTTACACTTGCA
SEM4F: GCCATGCAGACCCCGCGAGC	NOT8F: TGATGCCACCTGAACAACCTG	NOT16R: GCATCTGGGGACACACATCGAC
SEM4R: CCACG TGGCATGCATGGTCAG	NOT8R: CTCAAATCTGAGCTGCCTCC	NOT17F: TGTGAACACTTGTGCCAGCAC
SEM5F: GAGGAGTTTGAGACCTACCGGC	NOT9F: GATGATGAAGGGCCCCAG	NOT-18F: GCATCTGCAATCCCGGTACA
SEM5-2R: GCCGGTAGGTCTCAAACCTCCTC	NOT9R: TCTGTGATGTCTCGATTGGC	NOT-18R: TTCACCTGTGAGTAGCAGCTGG
	NOT10R: ACACCTGACCATGGTGGCAAG	NOT19F: AGGAGGCGACCGAGAAGATG
	NOT11F: CCAGTTTGCCAGATTGATATTG	NOT19R: CCTCTCAGTCCATACCCAC
	NOT11R: CAGCTGTGGACAAGTGAAATG	NOT20F: GAAGATGCCCCGCCCTGC
	NOT12F: AGTATCCTGTGCCATTGTGG	NOT20R: CTCATAGGTATCCCGGCTGA
	NOT12R: TCCTGTAAACCCTGAGGCAC	NOT21F: CCAGCTGCTACTCACAGGTGAA
	NOT13F: AGGCCAGTACTGTGACAGCC	NOT21R: ACCTGCCTTCCCCTCTGG

Cerca de 100 ng de mRNA, isolado de tumores humanos de mama, é convertido a cDNA por transcrição reversa utilizando-se o kit *ThermoscriptTM RT-PCR System* (Gibco-BRL, Life Technologies, Gaithersburg, USA).

Para a amplificação da isoforma menor de semaforina 6B (*Sema 6B_s*, 1.8 Kb), 5 ng of cDNA são adicionados a tubo *ependorf* contendo 5 pmoles de cada *primer* (SEM4F e SEM4R), 1,0 unidade de *rTth* DNA Polimerase, 800 µM de cada dNTP, 1,5 mM Mg(OAc)₂ e 7,5 µl de

tampão *3.3X PCR buffer* (*GeneAmp XL PCR kit*, Perkin-Elmer), num volume final de 25µl. As condições de ciclagem térmica seguem conforme descrito no item 3.14.1.

Para a amplificação da isoforma maior de semaforina 6B (*Sema 6B_L*, 2.3 Kb), são utilizados *primers* específicos para a porção 3' codificadora ausente na isoforma menor mas presente no clone genômico CTB-50L17 referente ao cromossomo 19 (Genbank AC011498) (SEM5-2R e SEM5F), em conjunto com os primers SEM4F e SEM4R, respectivamente. Condições de reação seguem conforme descrito acima. A reação é submetida a 35 ciclos de denaturação a 94°C por 40 segundos, anelamento a 66°C por 1 minuto e extensão a 72°C por 3 minutos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por dois ciclos a 70°C e 68°C, utilizando as mesmas condições de denaturação e extensão). Esta amplificação resulta, respectivamente, na geração de fragmentos sobreponíveis de 1.5 e 0.9 Kb, referentes às porções 5' e 3' do cDNA.

A amplificação de fragmentos de cDNA que representam ambas isoformas de *Sema 6B* é obtida utilizando-se os *primers* SEM2R ou SEM3R, em combinação com o *primer* SEM4R. Condições de reação seguem conforme descrito acima. A reação é submetida a 37 ciclos de denaturação a 94°C por 30 segundos, anelamento a 60°C por 40 segundos e extensão a 72°C por 2 minutos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por dois ciclos a 68°C e 64°C, utilizando as mesmas condições de denaturação e extensão). Os produtos de PCR específicos para *Sema 6B_S* e *Sema 6B_L*, obtidos com primers SEM2R e SEM4R, correspondem a bandas de 1.0 e 1.5 Kb, respectivamente.

Para a amplificação específica de fragmentos sobreponíveis de cDNA, correspondentes ao gene ortólogo humano *Notch2*, foram utilizados *primers* gene-específicos conforme listado na tabela anterior. As condições de reação seguem conforme descrito acima. A reação é submetida a 35 ciclos de denaturação a 94°C por 40 segundos, anelamento a 66°C por 1 minuto e extensão a 72°C por 3 minutos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por três ciclos a 70°C e 68°C, utilizando as mesmas condições de denaturação e extensão).

3.14.3. Sequenciamento e montagem de cDNAs

Os produtos de PCR gerados, correspondentes aos genes humanos em caracterização, são analisados após eletroforese em gel de agarose (suplementado com brometo de etídeo) e purificados com o kit *Geneclean II* (BIO 101, Inc.). Produtos com um tamanho inferior a 1.0 Kb são clonados conforme descrito no item 3.6, utilizando-se um volume de 5-10 µl do produto de PCR obtido. Produtos com tamanho superior são geralmente sequenciados sem prévia clonagem. O procedimento de sequenciamento segue conforme descrito no item 3.11.2, utilizando-se 10-40 ng de produto de PCR ou 200-500 ng de plasmídeo como molde.

A montagem de sequências referentes a cada gene específico foram realizadas manualmente, através da sobreposição de sequências de cDNA vizinhas, ou através do programa PhredPhrap (Gordon *et al.*, 1998).

3.15. Caracterização preliminar de novos genes humanos

3.15.1. Ensaios quantitativos de RT-PCR

A amplificação de ambas isoformas do marcador tumoral NABC1 é realizada por ensaios de RT-PCR. Cerca de 100 ng de mRNA foi utilizado na transcrição reversa utilizando-se o kit *ThermoscriptTM RT-PCR System* (Gibco-BRL, Life Technologies, Gaithersburg, USA). Para um volume final de 20 µl de reação de PCR, foram utilizados 5 ng de cDNA, 4 pmoles de cada *primer* (1F: 5'-AGAAGGACTGGAGACTGCAAAG-3'; 1R: 5'-TAAGGTCAGCTGAAGTGG TGG-3'), 0.5 unidades de enzima *Taq* DNA polimerase (Gibco-BRL), 125 µM de dNTPs, 1,5 mM MgCl₂ e 2 µl de tampão de PCR 10X (Gibco-BRL). A reação é submetida a 35 ciclos de denaturação a 94°C por 30 segundos, anelamento a 62°C por 35 segundos e extensão a 72°C por 50 segundos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por 2 ciclos a 69°C e 65°C, utilizando as mesmas condições de denaturação e extensão). Os produtos de PCR com tamanho aproximado de 170 bp e 305 bp se referem às isoformas NABC1 e NABC1_5B, respectivamente. Reações-controle são realizadas em paralelo, utilizando mRNA e DNA genômico como molde.

A re-amplificação dos produtos de PCR com um primeiro *primer* específico para a isoforma NABC1_5B e um segundo *primer* comum para ambas isoformas é feita conforme descrito acima, utilizando-se os *primers* 1F and 2R (5'-CACCCCTTATGCCTAAAGAATTTTC-3'). A reação é submetida a 40 ciclos de denaturação a 94°C por 30 segundos, anelamento a 58°C por 35 segundos e extensão a 72°C por 50 segundos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por 2 ciclos a 66°C e 62°C, utilizando as mesmas condições de denaturação e extensão). A isoforma NABC1_5B é representada por um banda específica de 270 bp, a ser sequenciada para confirmar a existência desta isoforma.

Ensaio de RT-PCR quantitativos são realizados utilizando-se 1 µg de RNA total isolado de tecido tumoral e normal de pacientes independentes. A fase exponencial da reação de amplificação foi estabelecida para NABC1 (28 ciclos), NABC1_5B (40 ciclos) e GAPDH (23 ciclos). A reação de RT-PCR para NABC1 segue conforme descrito para a amplificação de ambas isoformas (vide acima). Reações para NABC1_5B e GAPDH seguem conforme descrito à re-amplificação de NABC1_5B (vide acima), usando o número estabelecido de ciclos para a devida quantificação de cada sub-população de mRNA.

Ensaio de RT-PCR quantitativos específicos para NOTCH 2 seguem conforme descrito para a isoforma NABC1, utilizando-se os *primers* NOT2F e NOT2R (item 3.14.2). A fase exponencial da reação de amplificação foi estabelecida em 32 ciclos.

Todas as reações são analisadas, após eletroforese, em gel de poliacrilamida a 8% corado com prata.

3.15.2. Mapeamento genético

O mapeamento físico dos loci gênicos de interesse é realizado por PCR utilizando o painel de média resolução Stanford G3, contendo 83 fragmentos ordenados de DNA isolados de linhagens celulares híbridas de ser humano e hamster (*Research Genetics*, USA). A amplificação é realizada após a adição de 0,5 µl de cada DNA-molde do painel a um respectivo tubo *eppendorf* contendo 4 pmoles de cada *primer* gene-específico (SEM2R e SEM2-2F ou NOT2R e NOT2F), 0,5 unidades de *Taq* DNA polimerase (Gibco-BRL), 125 µM de cada dNTP, 1,5 mM MgCl₂ e 2 µl de tampão de PCR 10X (Gibco-BRL). A reação é submetida a 35 ciclos de denaturação a 94°C por 30 segundos, anelamento a 63°C por 35 segundos e extensão a 72°C por

50 segundos, precedida por uma programação *touchdown* (um ciclo de anelamento a 72°C, seguido por 2 ciclos a 69°C e 66°C, utilizando as mesmas condições de denaturação e extensão). As reações de PCR são analisadas após eletroforese em gel de agarose a 1% (suplementado com brometo de etídeo). A presença ou ausência de bandejamento gene-específico é convertida em uma sequência de 83 dígitos, numerados como 1 (presença) ou 0 (ausência de banda) que, após submissão por acesso remoto a *SHGC RHSEVER* (*Stanford Human Genome Center Rhserver*, <http://www-shgc.stanford.edu/>), indica a localização do gene analisado.

3.15.3. Ensaios de Northern blot

Membranas de nylon contendo alíquotas de 2 µg de poly(A)⁺ RNA, isolado de diferentes tecidos humanos (*Human 12-Lane Multiple Tissue Northern Blot*, Clontech, Palo Alto, CA), são hibridizadas a 65°C com sondas radioativamente marcadas com ³²P (kit Rediprime™ II, Amersham Pharmacia Biotech), específicas para *Sema 6B*, *Notch2* ou *GAPDH*, em tampão contendo 250 mM Na₂HPO₄ (pH 7.2), 1 mM EDTA, 7% (p/v) SDS e 1% (p/v) BSA por um período de 12 a 16 horas (protocolo descrito por Church & Gilbert, 1984).

Alíquotas de 8 a 15 µg de RNA total isolado de tecido mamário humano (amostras tumorais e normais) e linhagens celulares de câncer de mama T-47D (ATCC HTB-133), MDA-MB-231 (ATCC HTB-26) e MCF7 (ATCC HTB-22) são submetidas a eletroforese em gel denaturante de agarose a 1% (2,2 M formaldeído) e transferidas para membrana de nylon Hybond-N (Amersham, Arlington Heights, IL). Hibridização foi realizada a 42°C com sondas específicas, marcadas com ³²P (*Sema-6B* ou *GAPDH*) em tampão contendo 50% de formamida, 2X SSPE (0,75 M NaCl; 0,05 M NaH₂PO₄ e 4 M EDTA pH 7,4), 5X reagente de Denhart (0,05 g/ml de Ficoll 400; 0,05 g/ml de polivinilpirrolidona e 0,05 g/ml de BSA), 5X dextran-sulfato, 1% SDS and 150 mg/ml de DNA de esperma de salmão, por um período de 16-20 horas (protocolo alterado do descrito por Sambrook *et al.*, 1989). O bandejamento obtido com a sonda gene-específica (*Sema 6B*) é quantificado por densitometria (Ultrascan XL-Pharmacia) e o valor mensurado para a sua intensidade é normalizado em função da expressão de *GAPDH* (relação SEMA/*GAPDH*). Este ensaio foi realizado no laboratório da Profa. Dra. Maria Mitzi Brentani (Departamento de Radiologia, FM/USP).

Alíquotas de 10 µg de RNA total isolados de linhagens celulares de glioblastoma humano

T98G (ATCC CRL-1690) e A172 (ATCC CRL-1620) não-tratadas ou tratadas com ácido retinóico a 10^{-5} M (*all-trans-retinoic acid*, ATRA), por diferentes períodos de tempo, são manipuladas e analisadas por Northern blot conforme descrito por Sambrook *et al.* (1989). Uma sonda específica para fosfoproteína ribossomal ácida PO (36B4) é utilizada como controle interno da análise. Este ensaio foi realizado no laboratório da Profa. Dra. Mari Cleide Sogayar (Departamento de Bioquímica, IQ/USP).

4. RESULTADOS

4.1. Critérios de qualidade para preparação de mRNA

Para avaliar o grau de contaminação com DNA genômico, diferentes estoques de mRNA foram submetidos a uma PCR-controle para a amplificação específica do gene da β -globina.

Preferencialmente, uma alíquota de cada estoque foi diluída em uma proporção de 1:20, o que representa a quantidade de mRNA utilizado na amplificação por RAP-PCR nas condições experimentais.

Devido à grande sensibilidade da técnica, esta foi repetida 3 a 5 vezes para cada preparação de mRNA (fig. 05). Como prova desta sensibilidade, pode-se verificar no poço 6 da figura 5B que a própria reação de controle negativo (supostamente sem *template* para a reação) fornece bandejamento característico para porção gênica da β -globina (ao redor de 120 bp). Preparações que originaram bandejamento em mais de 1 experimento independente foram descartadas (RBT003 a RBT005).

A maior parte das preparações de mRNA utilizadas na síntese de cDNA e geração de ESTs (exceto RBT001, RBT002 e RBT015) foi submetida a PCR-controle para β -globina. Em 3 reações independentes, as preparações RBT006 e RBT013 (poços 5 e 10, fig. 05C, respectivamente) originaram, em apenas um caso, fraco bandejamento na faixa de 120 pares de bases. A partir destes resultados, pode-se afirmar que os estoques RBT006, RBT011, RBT013 e RBT014 não apresentam contaminação significativa com DNA genômico.

Figura 05: Imagem esquadrinhada de géis de poliacrilamida a 8% corados com prata, correspondentes a reações de PCR específicas para gene da β -globina. (A) perfil de PCR para diferentes amostras de mRNA (solução-estoque). P- marcador de peso molecular (ϕ X-174 HaeIII); 1- controle positivo (DNA genômico); 2 a 5- produto de PCR obtido com 1.0 μ l de preparação de mRNA (RBT003 a RBT006, respectivamente). (B) perfil de PCR para diferentes amostras de mRNA (diluição 1:20). P- marcador de P.M.; 1- controle positivo; 2 a 5- produto de PCR obtido com 1.0 μ l das preparação de mRNA RBT003 a RBT006, respectivamente; 6- controle negativo (sem DNA). (C) perfil de PCR para amostras de mRNA (diluição 1:20, em triplicata) utilizadas na síntese de cDNA. P- marcador de P.M.; 1- controle positivo; 2, 3 e 4- produto de PCR obtido com 1.0 μ l de RBT006; 5, 6 e 7- produto de PCR obtido com 1.0 μ l de RBT011; 8, 9 e 10- produto de PCR obtido com 1.0 μ l de RBT013; 11, 12 e 13- produto de PCR obtido com 1.0 μ l de RBT014; 14- controle negativo. Bandeamento esperado (ao redor de 120 bp) é indicado pelas setas. Condições de reação descritas no item 3.3.1.

Para avaliar o grau de contaminação com DNA mitocondrial, procedeu-se à amplificação específica para a região do D-loop mitocondrial (região conservada, não-expressa) a partir das preparações de mRNA que forneceram resultado negativo na amplificação de β -globina (fig. 06).

Figura 06: Imagem esquadrinhada de gel de poliacrilamida a 8% corado com prata, correspondente a reações de PCR específicas para D-loop mitocondrial. P- marcador de peso molecular (*100 bp ladder*); 1- controle negativo (sem DNA); 2- controle positivo (DNA genômico); 3, 5, 7, 9 e 11- produto de PCR obtido com 1.0 μ l de preparação de mRNA (RBT002, RBT006, RBT011, RBT013 e RBT014, respectivamente); 4, 6, 8, 10 e 12- produto de PCR obtido com diluição 1:20 de preparação de mRNA (RBT002, RBT006, RBT011, RBT013 e RBT014, respectivamente). Bandejamento esperado (ao redor de 1200 bp) é indicado pelas seta. Condições de reação descritas no item 3.3.2.

A preparação RBT002, apesar de não ter sido analisada na amplificação específica para globina, foi utilizada na geração de mini-bibliotecas de cDNA. Conforme mostra a figura 06, as preparações RBT002 e RBT006 apresentam contaminação com DNA mitocondrial. No caso de RBT002, esta contaminação é evidente tanto na amostra concentrada (poço 3, fig. 06) quanto na diluída (poço 4, fig. 06), mostrando que o volume de mRNA utilizado na síntese de cDNA e posterior amplificação por RAP-PCR "carrega" uma quantidade mensurável de DNA mitocondrial contaminante, a ser inevitavelmente utilizado como *template* na amplificação em baixa estringência. Já no caso de RBT006, esta contaminação é mais evidente na amostra concentrada, ocorrendo um bandejamento muito fraco na amostra diluída (fig. 06, poços 5 e 6, respectivamente). Ou seja, a quantidade de DNA mitocondrial contaminante a ser utilizada como *template* nas reações de RAP-PCR é questionável, devido ao fraco bandejamento obtido na

amostra diluída. As preparações RBT011, RBT013 e RBT014 não mostraram amplificação para região de D-loop, evidenciando a ausência de contaminação com DNA mitocondrial.

Apesar de ter sido utilizada na geração de mini-bibliotecas de cDNA, a preparação RBT008 não foi analisada quanto a contaminação com DNA genômico ou mitocondrial devido à sua baixa concentração e limitado volume da preparação.

4.2. Produção de perfis de cDNA por RAP-PCR

Sete preparações de mRNA (RBT002, RBT006, RBT008, RBT011, RBT013, RBT014 e RBT015, respectivamente) foram utilizadas na síntese de cDNA e posterior amplificação por RAP-PCR. A maior parte dos perfis gerados correspondem ao padrão X-X (mesmo *primer* utilizado tanto na reação de transcrição reversa quanto na RAP-PCR). Estes perfis mostram, em geral, um bandejamento complexo acima de 200 bp, independentemente do tipo de *primer* utilizado (fig. 07). Quando comparado a um perfil-controle produzido a partir de DNA genômico (controle positivo), os perfis de RAP-PCR estão deslocados para um menor peso molecular (fig. 07 A e B).

Os controles negativos iniciais da RAP-PCR foram feitos a partir do produto da reação de transcrição reversa na ausência da enzima (transcriptase reversa). Ou seja, foi usado mRNA em quantidade equimolar ao cDNA usado como *template* nas amostras. Devido à reprodutibilidade das reações-controle, aliada a maior confiabilidade nos estoques recentes de mRNA, optamos por utilizar reações sem *template* como controle negativo da RAP-PCR (vide figuras 07 C e D).

Figura 07: Imagem esquadrihada de géis de poliacrilamida a 8% corados com prata, correspondentes a perfis X-X de reações de RAP-PCR para geração de mini-bibliotecas de cDNA. (A) e (B) Perfis iniciais obtidos com *primers* arbitrários a partir de cDNA sintetizado com RBT001. P- marcador de peso molecular (100 bp ladder); 1-branco (sem DNA e/ou *primer*); 2, 5, 8, 11, 14 e 17- produtos de PCR obtidos a partir de 1.0 µl (10 ng) de DNA genômico com *primers* TC2, MIO4, IE, PAVII, CEPL1 e GT1R, respectivamente; 3, 6, 9, 12, 15 e 18- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de mRNA (controle negativo) com *primers* na referida ordem; 4, 7, 10, 13, 16 e 19- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de cDNA com *primers* na referida ordem. *Primers* utilizados: TC2 (5'-AGGAGTGACGGTTGATCAGT-3'), MIO4 (5'-GAGTCGATTTCAGGTTG-3'), IE (5'-CCCGCTCCTCCTGAGCACCC-3'), PAVII (5'-CGCTATTGATCAGCATGGCGG GTTAATCAT-3'), CEPL1 (5'-GCGGTCCAAACTCATCAATGTATCT-3') e GT1R (5'-TGGTAGTACTGCTCTGGGGG-3'). (C) e (D) Perfis obtidos com *primers* aleatórios a partir de cDNA sintetizado com RBT013. P- marcador de peso molecular (100 bp ladder); 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 e 23- controle negativo (sem *template*) para *primers* PBT133 a PBT144, respectivamente; 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 e 24- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de cDNA com *primers* na referida ordem. Os perfis obtidos (exceto com PBT136) geraram as mini-bibliotecas BT194 a BT204. Condições de reação descritas no item 3.5.1.

Verificamos, na prática, que uma importante fonte de contaminação de DNA exógeno se refere a solução de *primers* utilizados na produção dos perfis (veja maiores detalhes na figura 10). Perfis complexos de bandejamento na reação de controle negativo resultam na inutilização do respectivo *primer* para a geração de mini-bibliotecas de cDNA. Perfis de cDNA pouco complexos e/ou similares ao controle negativo (vide poços 3 e 4, fig. 07 A) são descartados.

Na figura 07 C, podemos verificar que o perfil X-X obtido com o *primer* PBT136 (poço 8) possui uma banda, de aproximadamente 1000 bp, que se destaca no perfil. Este tipo de perfil tem grande chance de originar uma mini-biblioteca redundante de cDNA, onde a maioria dos seus clones teriam a mesma sequência clonada. Por este motivo, perfis que apresentam uma preferência de amplificação para determinadas bandas são descartados no momento da clonagem.

Os perfis X-Y gerados para a possível produção de mini-bibliotecas de cDNA são mostrados na figura 08. Estes perfis, obtidos com a utilização de *primers* diferentes para a síntese de cDNA e amplificação por RAP-PCR, possuem uma tendência em gerar bandejamento mais intenso na faixa de peso molecular inferior a 200 bp (vide fig. 08 A).

No sentido de produzir perfis X-Y com um maior percentual de bandas acima da faixa de 200 bp, foram alteradas a concentração de *primer* na reação e a temperatura de anelamento para a ciclagem térmica (baixa estringência em todos os ciclos de PCR). Os perfis obtidos se mostraram mais concentrados e levemente enriquecidos com bandas acima de 200 bp (fig. 08 B). Combinações de 2 *primers* simultâneos para a produção do perfil de RAP-PCR foram igualmente obtidos (poços 6-7 e 12-13, fig. 08 B) e, aparentemente, não há diferenças no padrão de bandejamento característico quando comparados aos demais perfis X-Y.

Os perfis X-Y obtidos para os diferentes controles negativos (inferiores ao peso molecular de 100 bp) se referem a interações e/ou amplificações inespecíficas ocorridas entre os *primers* utilizados na reação.

Figura 08: Imagem esquadrinhada de géis de poliacrilamida a 8% corados com prata, correspondentes a perfis X-Y de reações de RAP-PCR para geração de mini-bibliotecas de cDNA. (A) Perfis iniciais obtidos com *primers* aleatórios a partir de cDNA sintetizado com *primer* CWS1L (preparação RBT002). P- marcador de peso molecular (100 bp ladder); 1, 4, 7, 10 e 12- produtos de PCR obtidos a partir de 1.0 µl (10 ng) de DNA genômico com *primers* PSA, A2, GT2, CWAL e CW1, respectivamente; 2, 5, 8- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de mRNA (controle negativo) com *primers* PSA, A2 e GT2, respectivamente; 3, 6, 9, 11 e 13- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de cDNA com *primers* PSA, A2, GT2, CWAL e CW1, respectivamente. *Primers* utilizados: PSA (5'-GATGACTCCAGCCACGACCT-3'), A2 (5'-AGATTCGAACTCATCAGAGCTACAGGAAC-3'), GT2 (5'-AGATTCGAAGTCCAGATGCTCCAGCTCAC-3'), CWAL (5'-CTGATCTAGGTCTGCAGAAAAC-3'), CW1 (5'-CGTAACCGAAATCGGTTGAAC-3') e CWS1L (5'-AATGGCATTGTTGGGGTA-3'). Condições de reação descritas no item 2.5.2 (método I). (B) Perfis obtidos com *primers* aleatórios a partir de cDNA sintetizado com *primer* PBT102 (preparação RBT014). P- marcador de peso molecular (100 bp ladder); 1- branco (presença de *primer* PBT102); 2, 4, 6, 8, 10 e 12- controle negativo (sem *template*) para *primers* PBT132, PBT131, PBT132/PBT131 (1:1), PBT128, PBT127 e PBT129/PBT127 (1:1), respectivamente; 3, 5, 7, 9, 11 e 13- produtos de PCR obtidos a partir de 1.0 µl (0.5 ng) de cDNA com *primers* na referida ordem. Poços 6-7 e 12-13 representam amplificações de RAP-PCR feitas com 2 *primers* aleatórios simultaneamente, em quantidades equimolares. Condições de reação descritas no item 3.5.2 (método II).

Para um maior aproveitamento dos perfis X-Y padronizados até o momento, se torna necessária a seleção por tamanho das bandas acima de 200 bp (mais informativas na análise de similaridade em bancos de dados). Este passo envolve a eletroforese do perfil em gel de agarose com a posterior purificação do material selecionado. Devido ao baixo rendimento na obtenção do perfil selecionado, aliado ao tempo dispensado para a devida purificação, este tipo de perfil foi pouco utilizado na geração de mini-bibliotecas de cDNA.

Os perfis X-X_n, gerados a partir da re-amplificação de produtos de RAP-PCR em alta estringência, correspondem a outra metodologia abordada na produção de mini-bibliotecas de cDNA. Conforme mostra a figura 09, os perfis obtidos são parcialmente complexos, pois pode-se

observar, em geral, uma maior nitidez na definição de bandas quando comparado aos perfis X-X. A maioria dos perfis possui um bandejamento superior a 300 bp.

Para a produção de perfis X-Xn, é realizada a re-amplificação de perfis X-X com respectivos *primers* cuja sequência é idêntica ao *primer* utilizado no perfil X-X inicialmente obtido, mas acrescido de 1 nucleotídeo na sua extremidade 3' (por exemplo, PBT127 é idêntico a PBT092, acrescido de uma guanina na sua porção 3'). Foram gerados perfis X-Xn com o mesmo *primer* utilizado na amplificação inicial (geração de perfil X-X) apenas como controle das reações a serem obtidas (figura 09).

Figura 09: Imagem esquadrinhada de géis de poliacrilamida a 8% corados com prata, correspondentes a perfis X-Xn de reações de RAP-PCR para geração de mini-bibliotecas de cDNA. (A) e (B) Perfis obtidos a partir da re-amplificação de produtos de RAP-PCR, em alta estrigência. P- marcador de peso molecular (*100 bp ladder*); 1, 4, 7, 10, 13, 16, 19 e 22- controle negativo (*sem template*) para *primers* PBT129/PBT130 (1:1), PBT092, PBT128, PBT127, PBT127/PBT128 (1:1), PBT093, PBT129, PBT130, respectivamente; 2, 17, 20 e 23- produtos de re-amplificação obtidos a partir de 1.0 µl de produto de RAP-PCR (*template* obtido com *primer* PBT093 a partir da preparação RBT014) com *primers* PBT129/PBT130 (1:1), PBT093, PBT129 e PBT130, respectivamente; 3, 18, 21 e 24- produtos de re-amplificação obtidos a partir de 1.0 µl de produto de RAP-PCR (*template* obtido com *primer* PBT093 a partir da preparação RBT011) com *primers* PBT129/PBT130 (1:1), PBT093, PBT129 e PBT130, respectivamente; 5, 8, 11 e 14- produtos de re-amplificação obtidos a partir de 1.0 µl de produto de RAP-PCR (*template* obtido com *primer* PBT092 a partir da preparação RBT014) com *primers* PBT092, PBT128, PBT127 e PBT127/PBT128 (1:1), respectivamente; 6, 9, 12 e 15- produtos de re-amplificação obtidos a partir de 1.0 µl de produto de RAP-PCR (*template* obtido com *primer* PBT092 a partir da preparação RBT011) com *primers* PBT092, PBT128, PBT127 e PBT127/PBT128 (1:1), respectivamente. Combinações de 2 *primers* aleatórios simultâneos, em quantidades equimolares, para a produção de perfis X-Xn também foram realizados (PBT127/PBT128 e PBT129/PBT130). Condições de reação descritas no item 3.5.3.

Neste caso, seria esperado que os perfis X-Xn gerados com PBT092 e PBT093 (poços 5-6 e 17-18, figura 09) tivessem uma maior complexidade quando comparados aos demais, pois não há a seleção de *templates* amplificáveis por nucleotídeo adicional no *primer* (no caso de PBT127 a PBT130), o que ocasiona uma menor estringência na amplificação e consequente aumento na geração de sequências de cDNA. Na prática, entretanto, diferenças na complexidade dos perfis X-Xn obtidos foi dificilmente observada pela análise dos géis de poliacrilamida.

A combinação de primers (2 ou mais) idênticos ao primer utilizado na RAP-PCR original, acrescidos de diferentes nucleotídeos na suas respectivas porções 3', pode ser uma alternativa para o aumento na complexidade dos perfis X-Xn (vide poços 2-3 e 14-15, fig. 09) e possível geração de mini-bibliotecas de cDNA com baixa redundância entre seus clones.

4.3. Critérios de qualidade para produção de perfis por RAP-PCR

Para a geração dos diferentes perfis de cDNA, foram otimizados os cuidados na manipulação de soluções de *primers*, para assim minimizar a contaminação com DNA exógeno no momento da RAP-PCR. Isto porque, após o estabelecimento dos critérios de qualidade para a preparação de mRNAs, não seria aceitável o sequenciamento de mini-bibliotecas com alto grau de elementos repetitivos e/ou similaridade a regiões não-expressas do genoma (conteúdo não-transcrito).

Neste sentido, foi realizada uma reação-teste de RAP-PCR radioativa (marcação com [α - 32 P] dCTP) para geração de perfis X-X com 3 *primers* diferentes: (i) PBT087 (*primer* controle), responsável pela geração de mini-bibliotecas adequadas para a produção de ESTs; (ii) PBT048, responsável pela geração de mini-bibliotecas com alto grau de elementos repetitivos (SINEs e LINEs) e (iii) PBT084, responsável pela geração de mini-bibliotecas com alto nível de sequências de DNA mitocondrial e RNA ribossomal. Estes três *primers* foram utilizados na produção de perfis radioativos a partir de cDNA sintetizado com as preparações RBT002, RBT006, RBT011, RBT013 e RBT014. Este tipo de experimento foi feito para verificar se as sequências produzidas eram provenientes de contaminação (preparação de mRNA e/ou solução de *primer*) ou amplificação característica ao *primer* utilizado. Como mostra a figura 10, observou-se que o *primer* PBT084 origina perfis relativamente complexos nos diferentes controles negativos (reação de transcrição reversa na ausência de transcriptase reversa). Este

resultado indica contaminação nesta solução de *primer*, pois não foram observados perfis semelhantes nos controles negativos referentes aos *primers* PBT087 e PBT048.

Figura 10: Imagem esquadrinhada de autoradiografia referente a gel de poliacrilamida a 8% contendo perfis X-X de reações de RAP-PCR radioativa (marcação com $[\alpha\text{-}^{32}\text{P}]\text{dCTP}$). Três diferentes *primers* (PBT87, PBT48 e PBT84) foram utilizados para amplificar 1 μl de cDNA (0,5 a 1 ng) sintetizado com o respectivo *primer* e diferentes preparações de mRNA (RBT002, RBT006, RBT011, RBT013 e RBT014). **B** se refere ao sinal de fundo (*background*, sem *template*); + (positivo) se refere a amostra e - (negativo) se refere ao controle negativo da respectiva amostra (reação de transcrição reversa na ausência de enzima transcriptase). Condições de reação descritas no item 3.5.4.

Os resultados obtidos com o *primer* PBT048 (alto grau de elementos repetitivos nas mini-bibliotecas geradas) parecem ser específicos ao *primer* utilizado, ou seja, a sequência do *primer* possibilita, nas condições de reação impostas, a amplificação preferencial de sequências com alta similaridade a elementos repetitivos do DNA (presentes, em geral, nas regiões não-traduzíveis do mRNA).

Desta maneira, controles negativos contendo apenas os *primers* utilizados na respectivas reações-amostra foram realizados rotineiramente, em paralelo com a amostra ou até previamente à síntese de cDNA, evitando assim o desperdício de mRNA. Como mostrado no item 4.2, a

análise em gel de poliacrilamida a 8% corado com prata (experimento não-radioativo) foi utilizada a contento, reproduzindo, com sensibilidade, resultados semelhantes aos obtidos no teste de RAP-PCR radioativa.

4.4. Geração de mini-bibliotecas de cDNA

Os perfis de cDNA foram clonados no vetor pUC-18. A eficiência de ligação foi analisada por PCR a partir de *primers* específicos ao sítio de clonagem no vetor (fig. 11 A). Perfis de PCR que mostram forte bandejamento na faixa de 140 a 160 bp indicam a ligação de vetor pUC a *primers* da RAP-PCR (banda de pUC-controle corresponde a 120 bp). Apesar da baixa incidência, resultados deste tipo descartam o respectivo perfil de cDNA clonado, pois há grande possibilidade de, após a transfecção, obtermos grande número de clones contendo plasmídeos ligados somente a sequências de *primers*.

Aparecimento de banda na faixa de 120 bp indica a presença de vetor pUC re-ligado na reação de ligação. Na prática, a intensidade deste bandejamento é diretamente proporcional ao aparecimento de clones azuis (expressão de β -galactosidase) no plaqueamento celular. Após a devida transfecção, os produtos de ligação descritos (figura 11 A) forneceram um número de clones adequado para a geração de ESTs (50 a 200 colônias por placa).

Conforme mostra a figura 11 B, clones isolados de cada mini-biblioteca são amplificados por PCR e analisados em gel. Reações pouco amplificadas ou com aparecimento de bandas inespecíficas são descartadas. Caso o padrão de bandas, para uma série de clones de mini-bibliotecas em validação, seja específico e concentrado, este produto de PCR pode ser utilizado no sequenciamento direto dos insertos para a geração de ESTs.

Figura 11: Imagem esquadrihada de géis de poliacrilamida a 8% corados com prata, correspondentes a reações de PCR com *primers* específicos de vetor pUC-18. (A) Análise por PCR da eficiência de ligação dos produtos de RAP-PCR em plasmídeo pUC-18. P- marcador de peso molecular (*100 bp ladder*); 1- controle positivo (vetor pUC-18); 2 a 10- produto de ligação de perfis de RAP-PCR em pUC-18 (correspondente a mini-bibliotecas BT187 a BT193, BT143 e BT151, respectivamente); 11- controle negativo (sem *template*). Condições de reação descritas no item 2.6. (B) Análise por PCR para seleção e sequenciamento de clones isolados (mini-biblioteca BT002). P- marcador de peso molecular (*1 Kb ladder*); 1- controle negativo (sem *template*); 2- controle positivo (vetor pUC-18); 3 a 12- clones 1 a 10 da referida mini-biblioteca. Cerca de 3 µl de cada produto de PCR foi utilizada para a eletroforese em gel. Bandeamento característico de vetor pUC-18 (ao redor de 120 bp) é indicado pelas setas. Condições de reação descritas nos itens 3.6 e 3.9, respectivamente.

4.5. Critérios de qualidade para a utilização de mini-bibliotecas de cDNA na geração de ESTs (validação de mini-bibliotecas)

Conforme exemplificado anteriormente, determinados *primers* amplificam sequências de maneira preferencial e, em alguns casos, estas sequências podem não corresponder à porções traduzíveis do gene. Aliado a este fato, determinadas preparações de mRNA (por exemplo, RBT015) podem possuir um enriquecimento na população de mRNAs mais abundantes (*housekeeping genes*) devido a uma provável degradação, o que dificulta a normalização parcial proporcionada pela RAP-PCR.

De posse destas informações, procedeu-se rotineiramente ao sequenciamento de 10 a 20 clones de cada mini-biblioteca para verificar a incidência de elementos repetitivos, sequências de DNA mitocondrial e/ou RNA ribossomal e taxa de redundância dos clones sequenciados. Caso

os critérios de qualidade pré-estabelecidos sejam respeitados, procede-se ao sequenciamento de 50 a 200 clones da respectiva mini-biblioteca.

Para a análise de elementos repetitivos, foi utilizado o programa *RepeatMasker* (disponibilizado por acesso remoto, <http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>) para o rastreamento de tais sequências nas ESTs geradas para a validação de cada mini-biblioteca. Conforme mostra a figura 12, após a submissão a este programa por acesso remoto, é originada uma listagem com os possíveis elementos repetitivos presentes e seu percentual na mini-biblioteca. No caso da mini-biblioteca BT009 (total de 11 sequências), foi obtido um percentual total de 1,72% em elementos repetitivos. Como este valor está dentro dos critérios de validação, baseado na análise de elementos repetitivos, a mini-biblioteca BT009 está aprovada para o sequenciamento de um maior número de clones. A análise de ESTs geradas durante a padronização da técnica de RAP-PCR (total de 111 sequências, referentes a uma mini-biblioteca teste) mostra que o percentual de elementos repetitivos é alto (34,34%) e o próprio conteúdo de G-C das sequências é relativamente baixo (na faixa de 40%), indicando uma provável contaminação com DNA genômico (critérios de qualidade para a preparação de mRNA não haviam sido estabelecidos).

file name:	BT009	file name:	C2-C2/test
sequences:	11	sequences:	111
total length:	5741 bp	total length:	40830 bp
GC level:	52.95 %	GC level:	40.63 %
bases masked	99 bp (1.72 %)	bases masked	14599 bp (35.76 %)

	number of elements*	length occupied	percentage of sequence
SINES:	0	0 bp	0.00 %
ALUs	0	0 bp	0.00 %
MIRs	0	0 bp	0.00 %
LINEs:	0	0 bp	0.00 %
LINE1	0	0 bp	0.00 %
LINE2	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
MaLRs	0	0 bp	0.00 %
Retrov.	0	0 bp	0.00 %
MER4_group	0	0 bp	0.00 %
DNA elements:	1	99 bp	1.72 %
MER1_type	0	0 bp	0.00 %
MER2_type	1	99 bp	1.72 %
Mariners	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		99 bp	1.72 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	0	0 bp	0.00 %
Low complexity:	0	0 bp	0.00 %

	number of elements*	length occupied	percentage of sequence
SINES:	21	2860 bp	7.00 %
ALUs	13	1969 bp	4.82 %
MIRs	8	891 bp	2.18 %
LINEs:	26	6008 bp	14.71 %
LINE1	15	4236 bp	10.37 %
LINE2	11	1772 bp	4.34 %
LTR elements:	14	3101 bp	7.59 %
MaLRs	7	1149 bp	2.81 %
Retrov.	5	1249 bp	3.06 %
MER4_group	2	703 bp	1.72 %
DNA elements:	12	2050 bp	5.02 %
MER1_type	8	1084 bp	2.65 %
MER2_type	4	966 bp	2.37 %
Mariners	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		14019 bp	34.34 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	6	396 bp	0.97 %
Low complexity:	6	168 bp	0.41 %

Figura 12: Laudo de análise de elementos repetitivos para duas mini-bibliotecas independentes produzido pelo programa *RepeatMasker* (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>). Baixo percentual de elementos repetitivos favorece validação da mini-biblioteca BT009 (lado esquerdo). Mini-biblioteca C2-C2/test (lado direito) não está de acordo com critério de qualidade estabelecido para validação (percentual de 34,34% de *total interspersed repeats*).

Conforme mostra a tabela a seguir, foram selecionadas 5 mini-bibliotecas não validadas e validadas para exemplificar os critérios de qualidade estabelecidos na seleção de mini-bibliotecas de cDNA. Baseado nos resultados de análise de similaridade em bancos de dados (submissão em BLAST e Unigene), as mini-bibliotecas BT013 e BT016 não foram aprovadas devido ao alto percentual de sequências contendo elementos repetitivos (30,5% e 25,6%, respectivamente), característicos de DNA genômico e regiões não-traduzidas de genes expressos. As mini-bibliotecas BT055 e BT076 não foram aprovadas devido ao alto número de sequências mitocondriais obtidas (27% e 56%, respectivamente), o que dificultaria a geração de novas

sequências humanas. Já a BT186 não foi aprovada devido à taxa de redundância dos clones sequenciados (equivalente a 63,6%), ou seja, há uma baixa complexidade de sequências na referida mini-biblioteca.

Os resultados mostrados na figura 13 para algumas mini-bibliotecas validadas (BT006, BT032, BT057, BT070 e BT077) se baseiam no número total de ESTs obtidas e analisadas quanto ao grau de similaridade a DNA mitocondrial, RNA ribossomal, elementos repetitivos e redundância global da mini-biblioteca. Verificou-se que, no geral, não houve grandes variações do percentual de elementos repetitivos, sequências mitocondriais e ribossomais quando analisamos 10-20 clones, possibilitando uma extrapolação para a geração de 100 clones. Um relativo aumento na redundância das mini-bibliotecas é admissível pois, com o aumento de clones a serem sequenciados, há uma maior probabilidade das sequências de cDNA serem sequenciadas mais de uma vez. Mesmo assim, valores de redundância acima de 25% indicam que o sequenciamento da referida mini-biblioteca deve ser cessado (por exemplo, BT070), pois existe a probabilidade de sequenciar clones repetidos com uma frequência crescente.

	MINI-BIBLIOTECAS NÃO-VALIDADAS					MINI-BIBLIOTECAS VALIDADAS				
Biblioteca	BT013	BT016	BT055	BT076	BT186	BT006	BT032	BT057	BT070	BT077
Mitocondria	0 (0.0%)	0 (0.0%)	3 (27%)	5 (56%)	1 (10%)	6 (5.5%)	4 (3.2%)	9 (8.6%)	6 (7.9%)	10 (6%)
rRNA	0 (0.0%)	1 (11%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (6.4%)	1 (0.8%)	3 (2.9%)	0 (0.0%)	0 (0.0%)
Contigs	0 (0.0%)	0 (0.0%)	2 (18%)	2 (22%)	5 (50%)	21 (19%)	3 (2.4%)	20 (19%)	16 (21%)	51 (31%)
ESTs não-humanas	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.9%)	0 (0.0%)	6 (5.7%)	0 (0.0%)	3 (1.9%)
DNA	2 (22%)	1 (11%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (5.5%)	3 (2.4%)	7 (6.7%)	1 (1.3%)	5 (3.1%)
No matches	2 (22%)	3 (33%)	4 (36%)	1 (11%)	2 (20%)	42 (38%)	67 (54%)	16 (15%)	17 (22%)	17 (11%)
Repetições	3 (33%)	2 (22%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	8 (7.3%)	2 (1.6%)	3 (2.9%)	9 (12%)	4 (2.5%)
ESTs sem similaridade ao Unigene	0 (0.0%)	0 (0.0%)	1 (9.0%)	1 (11%)	0 (0.0%)	7 (6.4%)	3 (2.4%)	5 (4.8%)	5 (6.6%)	8 (4.9%)
Proteínas humanas	2 (22%)	2 (22%)	1 (9.0%)	0 (0.0%)	2 (20%)	12 (11%)	16 (13%)	32 (30%)	22 (29%)	64 (40%)
Proteínas não-humanas	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	25(20%) 2	4 (3.8%)	0 (0.0%)	0 (0.0%)
Total de clones gerados	9	9	11	9	10	110	124	105	76	162
Redundância	11.11%	12.50%	0.0%	0.0%	63.64%	23.71%	10.92%	10.75%	27.14%	19.74%
Elementos repetitivos ¹	30.52%	25.63%	12.31%	6.31%	4.13%	12.21%	2.45%	5.26%	13.70%	4.42%

Figura 13: Tabela contendo critérios de validação de mini-bibliotecas de cDNA (similaridade a rRNA, DNA mitocondrial, elementos repetitivos e redundância global). 1- dados fornecidos pela submissão no banco de dados RepeatMasker, referentes a total interspersed repeats (SINEs, LINEs, LTRs e MERs); 2- número significativo de ESTs com similaridade a sequências bacterianas.

Para mini-bibliotecas validadas que não mostraram aumento expressivo na redundância de seus clones, foi estabelecido o sequenciamento de um número crescente de clones (variável entre 20 a 200 clones). O acompanhamento contínuo das sequências geradas foi adaptado para a manutenção dos critérios de qualidade estabelecidos na validação das mini-bibliotecas. Como exemplo, a mini-biblioteca BT070 respeitou este critérios até atingir o número aproximado de 76 clones sequenciados, enquanto que a BT077 ainda respeita os critérios estabelecidos mesmo atingindo um número de 162 clones sequenciados (fig. 13).

Um total de 85 mini-bibliotecas foram bloqueadas neste procedimento de validação. Este valor equivale a 50% do número global de mini-bibliotecas de cDNA geradas (total de 170 mini-bibliotecas).

4.6. Geração de 6000 "etiquetas de sequências expressas" em câncer de mama

Um total de 170 mini-bibliotecas de cDNA foi suficiente para a geração de 6000 ESTs no modelo de câncer de mama. O número de ESTs geradas por mini-biblioteca é mostrado a seguir:

Mini-biblioteca	Total de sequências	Mini-biblioteca	Total de sequências	Mini-biblioteca	Total de sequências	Mini-biblioteca	Total de sequências
BT001	066	BT046	124	BT091	175	BT142	025
BT002	119	BT047	011	BT092	040	BT143	002
BT003	141	BT048	010	BT093	008	BT144	087
BT004	008	BT049	009	BT094	100	BT145	010
BT005	011	BT050	013	BT095	013	BT146	014
BT006	110	BT051	010	BT096	024	BT147	010
BT007	008	BT052	008	BT097	038	BT148	026
BT008	006	BT053	012	BT098	011	BT149	012
BT009	163	BT054	022	BT099	005	BT150	017
BT010	002	BT055	011	BT100	029	BT152	023
BT011	194	BT056	028	BT101	012	BT153	025
BT012	018	BT057	105	BT102	012	BT154	008
BT013	009	BT058	012	BT103	060	BT155	015
BT015	034	BT059	010	BT104	012	BT156	014
BT016	009	BT060	004	BT105	096	BT157	014
BT017	014	BT061	011	BT106	063	BT158	010
BT018	018	BT062	021	BT107	038	BT159	011
BT019	188	BT063	006	BT108	102	BT160	013
BT020	014	BT065	150	BT109	011	BT161	042

BT021	014	BT066	051	BT111	005	BT162	014
BT022	159	BT067	061	BT112	010	BT163	010
BT023	016	BT068	106	BT113	028	BT164	010
BT024	015	BT069	010	BT114	002	BT165	007
BT025	016	BT070	076	BT117	001	BT166	009
BT026	009	BT071	012	BT118	011	BT168	010
BT027	033	BT072	009	BT119	045	BT169	008
BT028	009	BT073	013	BT120	003	BT170	008
BT029	335	BT074	012	BT121	004	BT171	010
BT030	008	BT075	019	BT122	053	BT181	003
BT031	087	BT076	009	BT123	035	BT182	007
BT032	124	BT077	162	BT124	084	BT183	007
BT033	023	BT078	010	BT125	032	BT184	006
BT034	014	BT079	016	BT126	043	BT185	007
BT035	035	BT080	011	BT127	016	BT186	009
BT036	004	BT081	009	BT128	038	BT187	006
BT037	073	BT082	010	BT129	005	BT188	007
BT038	100	BT083	010	BT130	006	BT189	009
BT039	118	BT084	011	BT132	042	BT190	002
BT040	095	BT085	010	BT133	077		
BT041	116	BT086	075	BT137	018	Total	6029
BT042	012	BT087	012	BT138	035		
BT043	100	BT088	009	BT139	013		
BT044	010	BT089	014	BT140	031		
BT045	011	BT090	006	BT141	028		

Dentro do número de 6000 ESTs geradas, temos a seguinte distribuição global:

Sequências anotadas	3350 (55,6%)
Sequências <i>no matches</i>	1509 (25,0%)
Elementos repetitivos	549 (9,1%)
Sequências mitocondriais	428 (7,1%)
Sequências ribossomais	193 (3,2%)
Total	6029

Onde: (i) sequências anotadas representam todas as ESTs obtidas que possuem similaridade significativa a sequências de DNA e/ou aminoácidos já depositadas nos bancos de dados estudados; (ii) sequências *no matches* correspondem a ESTs sem similaridade significativa aos diferentes bancos de dados; (iii) elementos repetitivos representam ESTs que possuem identidade a sequências disponibilizadas pelo programa *RepeatMasker*, o qual abrange SINES

(*small interspersed repeats*), LINEs (*long interspersed repeats*) e outros elementos descritos até o momento; (iv) sequências mitocondriais correspondem a ESTs que possuem identidade a sequências de DNA mitocondrial e (v) sequências ribossomais representam ESTs que possuem identidade a sequências de RNA ribossomal.

Nota-se que 25% das ESTs geradas não possuem similaridade alguma a outros genes e/ou ESTs já conhecidas. Ou seja, cerca de 25% das sequências se referem a novos genes humanos ou melhor, condizem a genes já identificados e não totalmente sequenciados.

No número total de sequências anotadas (3350 ESTs) estão inclusas (i) ESTs que formam *clusters* com sequências do Unigene (*contigs*, equivalente a 1065 ESTs), (ii) ESTs sem similaridade aos *clusters* do Unigene (398 sequências), (iii) ESTs similares a sequências codificadoras de proteínas humanas (1490 sequências), (iv) ESTs similares a sequências codificadoras de proteínas não-humanas (91 sequências) e (v) ESTs similares a sequências não-expressas de DNA (306 sequências). Ou seja, 44,5% das ESTs anotadas (24,7% do total de ESTs geradas) possuem similaridade a mRNAs que codificam proteínas humanas já descritas.

4.7. Análise computacional de normalização de bibliotecas de cDNA

Para avaliar se a técnica utilizada favorece a identificação de genes pouco expressos (40-45% do mRNA total das células), promovendo assim uma redução na abundância dos genes mais expressos, foi analisado um número representativo de ESTs não-redundantes (793 ESTs), que possuem similaridade a genes humanos totalmente sequenciados, quanto ao tamanho dos *clusters* presentes no *Unigene database*. Esta análise se baseia na similaridade das ESTs a *clusters* presentes neste banco de dados. Quando este *cluster* for composto por um grande número de ESTs conhecidas, isto será um indicativo de que o gene representado é altamente expresso, pois foi capaz de gerar várias sequências expressas. Ou seja, processos que visam a normalização de populações de cDNA teriam a capacidade de gerar sequências que possuem similaridade a *clusters* de pequeno tamanho no *Unigene*. Para maior confiabilidade, bibliotecas de cDNA não-normalizada (NCI CGAP Br1.1) e normalizada (NCI CGAP Br2) foram analisadas em paralelo. *Unigene clusters* compostos por um número inferior a 50 sequências indicam que a respectiva biblioteca de cDNA contém sequências correspondentes a genes de baixo nível de expressão. Como mostrado na figura 14, o maior percentual de sequências equivalentes a *clusters* de

pequeno tamanho (cerca de 25%) correspondem às ESTs geradas no presente projeto (denominadas também de ORESTES, *Open Reading Frame Expressed Sequence Tags*, curva amarela). Bibliotecas não-normalizada (curva azul) e normalizada (curva vermelha) possuem percentuais de 10 e 18% das sequências que correspondem a *clusters* pequenos, respectivamente. Deste modo, pode-se afirmar que a metodologia aplicada está favorecendo a normalização das bibliotecas de cDNA, de modo a enriquecer o conteúdo de transcritos raros e não totalmente conhecidos.

Figura 14: Comparação da abundância de transcritos (tamanho do *Unigene cluster*) em três diferentes bibliotecas de cDNA (não-normalizada, normalizada e ORESTES), provenientes de tumor de mama. O número de sequências correspondentes a cada tamanho de *cluster* é indicado no eixo y.

4.8. Análise posicional das ESTs geradas ao longo dos respectivos genes humanos

Para verificar qual a posição mais frequente dos genes transcritos à qual as ESTs se referem, foi novamente analisado um número representativo de ESTs não-redundantes (793 ESTs), que possuem similaridade a genes humanos totalmente sequenciados. A distribuição posicional à qual as referidas ESTs tem similaridade, ao longo de um gene hipotético, está discriminada na figura 15 (curva verde). Para determinação da curva teórica (curva azul), foi

feita um cálculo baseado na probabilidade de um *primer* arbitrário anelar em cada lado de um gene hipotético. Assim, se o tamanho do gene é considerado 1 e a distância da extremidade 3' do cDNA à região de anelamento é considerada como S, a probabilidade de um determinado *primer* interagir com o *template* seria S e 1-S. Consequentemente, a probabilidade de amplificação com um mesmo *primer* arbitrário seria S(1-S). Neste caso, a chance de um determinado *amplicon* corresponder a porção média do gene seria $0.5 \times 0.5 = 0.25$ (25%). De acordo com a comparação feita entre a curva experimental e a curva teórica, pode-se concluir que a maior parte das ESTs geradas neste projeto corresponde a porção mais central dos transcritos humanos, sendo consideradas sequências altamente informativas e relevantes para o processo de descoberta gênica pois se referem às porções codificadoras dos genes expressos. Em virtude deste fato, tais ESTs foram denominadas ORESTES.

Figura 15: Análise posicional de sequências expressas não-redundantes que possuem similaridade a genes já conhecidos (experimental), ao longo de um gene hipotético. Pode-se perceber que o percentual de sequências equivalentes à porção central do gene é muito próximo ao valor esperado (curva teórica).

4.9. Avaliação global de redundância das sequências expressas

Dentre as 6029 ORESTES geradas neste projeto, um total de 3700 sequências expressas correspondem a *singletons* (sequências únicas, sem possível sobreposição com outras ESTs). As 2329 ORESTES restantes podem ser separadas em (i) 178 sequências duplicatas (sequências expressas idênticas presentes em clones diferentes) e (ii) 2151 sequências agrupadas em 557 *contigs* (conjunto de sequências sobrepostas umas às outras). Os *contigs* gerados são distribuídos, de acordo com o número de ORESTES constituintes, da seguinte maneira:

Número de <i>contigs</i> gerados	Número de ESTs por <i>contig</i>	Número de <i>contigs</i> compostos por ESTs de mesma biblioteca	Número de <i>contigs</i> gerados	Número de ESTs por <i>contig</i>	Número de <i>contigs</i> compostos por ESTs de mesma biblioteca
334	02	278 (80,8%)	01	19	00 (0,0%)
89	03	70 (78,9%)	02	25	00 (0,0%)
39	04	26 (66,7%)	01	27	00 (0,0%)
28	05	18 (64,3%)	01	29	00 (0,0%)
17	06	09 (52,9%)	02	31	00 (0,0%)
08	07	06 (75,0%)	01	33	00 (0,0%)
02	08	02 (100%)	01	39	00 (0,0%)
03	09	02 (66,7%)	01	44	00 (0,0%)
06	10	03 (50,0%)	01	47	00 (0,0%)
01	11	00 (0,0%)	01	54	00 (0,0%)
04	14	00 (0,0%)	01	59	00 (0,0%)
02	15	00 (0,0%)	01	79	00 (0,0%)

Dentro da composição global dos *contigs* gerados, foi observado que 414 *contigs* (74,4% do total) são formados por sequências expressas pertencentes a uma mesma mini-biblioteca de cDNA, evidenciando uma possível redundância dos clones sequenciados. Conforme tabela acima, observa-se que, de maneira geral, quanto maior o número de ORESTES que compõem um determinado *contig*, menor a probabilidade destas sequências serem referentes a uma única mini-biblioteca. Ou seja, há uma maior probabilidade de montagem crescente (em extensão) de *contigs* na presença de números maiores de ORESTES constituintes (acima de 11 sequências expressas) e uma menor chance destas sequências serem redundantes entre si.

A estimativa global de redundância das sequências expressas, geradas neste projeto, é de 29,4%. Este valor é derivado do percentual complementar ao número total de *singletons* (3700 ESTs) e *contigs* (557 ESTs) dividido pelo montante global de ESTs geradas (1-4257/6029).

Todos os resultados descritos neste ítem foram provenientes da interpretação dos dados gerados por análise computacional no programa PhredPhrap (Gordon *et al.*, 1998).

4.10. Avaliação do percentual de singletons e potencial codificador das sequências “no matches”

Dentre as sequências *no matches* obtidas (total de 1509 ORESTES), 1184 ORESTES correspondem a *singletons*, o que equivale a 78,5% das sequências *no matches* (19,6% do número total de ORESTES geradas). Sem considerar o número de *singletons* obtidos dentre as sequências anotadas (total de 3350 ORESTES), o percentual gerado se iguala àquele obtido por Adams e colaboradores (1995) na análise de 292.878 ESTs convencionais (58.384 *singletons*, equivalente a 19,9% do total) e supera o percentual obtido por Hillier e colaboradores (1996) (44.335 *singletons* em 280.223 ESTs, provenientes de 15 bibliotecas normalizadas e 7 bibliotecas não-normalizadas de cDNA), o qual equivale a 15,8%.

Das 325 ORESTES *no matches* restantes, 17 ORESTES se referem a sequências duplicatas e 308 ORESTES foram distribuídas em 115 *contigs*. Estes *contigs* específicos podem ser distribuídos, quanto ao número de ESTs constituintes, da seguinte maneira:

Número de <i>contigs</i> gerados	Número de ESTs por <i>contig</i>
80	02
16	03
07	04
06	05
03	06
02	07
01	10

Entretanto, as sequências *no matches* que constituem cada um dos respectivos *contigs* citados pertencem a uma mesma biblioteca de cDNA, indicando que estas sequências tem grande chance de serem redundantes. Este fato é confirmado com a análise de sequências “com alta probabilidade de serem duplicatas” (*near duplicate reads*) no programa PhredPhrap (Gordon *et al.*, 1998), o qual evidenciou a existência de 86 sequências *no matches* que estão agrupadas em *contigs*. Considerando-se então que cada *contig* citado possa corresponder a uma única EST não-

redundante (a qual seria um *singleton*), teríamos um total real de 1299 *singletons* (1184+115) e uma redundância de 13,9% (325-115/1509) entre as sequências *no matches*.

A predição do potencial codificador (presença de *open reading frame*) das sequências *no matches* obtidas foi realizado com o programa ESTScan (Iseli *et al.*, 1999). Este programa tem a capacidade de avaliar e reconstruir regiões codificadoras a partir de sequências de qualidade questionável, utilizando para tanto algoritmos baseados na composição dos hexanucleotídeos presentes (análise de *codon usage* e regiões de *splicing*), conteúdo de G-C e extensão das sequências. Do total de 1509 sequências, 475 ORESTES apresentaram tal potencial, o que corresponde a 31,5% das sequências *no matches* (7,9% do total de sequências geradas).

4.11. Distribuição parcial de sequências expressas similares a genes humanos conhecidos em distintas classes gênicas

Um total de 254 ORESTES, referentes a proteínas e mRNAs humanos já conhecidos, foram tabuladas de acordo com a classe gênica pertencente. Este número de sequências expressas representa cerca de 25% das sequências não-redundantes com alta similaridade a proteínas e mRNAs humanos já conhecidos.

1. Enzimas envolvidas no metabolismo celular

Sequência	Similaridade	Sequência	Similaridade
CM-BT006-152	6-PHOSPHOFRUCTOKINASE, LIVER TYPE (PHOSPHOFRUCTOKINASE 1).	PM-BT107-012	SELENOPHOSPHATE SYNTHETASE 2.
CM-BT006-150	CYTOCHROME C OXIDASE POLYPEPTIDE II.	RC-BT097-018	ACID SPHINGOMYELINASE-LIKE PHOSPHODIESTERASE.
CM-BT006-141	ADENYLATE KINASE 2A	RC-BT122-084	VACUOLAR ATP SYNTHASE 16 KD PROTEOLIPID SUBUNIT.
CM-BT006-137	NADH-CYTOCHROME B5 REDUCTASE (FRAGMENT).	IL-BT170-004	ATP SULFURYLASE/APS KINASE 2.
CM-BT006-140	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE	RC-BT144-088	ALPHA-AMYLASE 2B PRECURSOR (ALPHA-AMYLASE CARCINOID).
CM-BT094-122	ACONITATE HYDRATASE, MITOCHONDRIAL PRECURSOR (ACONITASE).	MR-BT038-094	ISOCITRATE DEHYDROGENASE [NADP], MITOCHONDRIAL PRECURSOR (OXALOSUCCINATE DECARBOXYLASE).
CM-BT066-073	DOLICHOL-PHOSPHATE-MANNOSE SYNTHASE.	MR-BT041-049	UBIQUINOL-CYTOCHROME C REDUCTASE COMPLEX CORE PROTEIN 2 PRECURSOR.
IL-BT044-004	GLYCOGEN SYNTHASE KINASE 3ALPHA	PM-BT057-322	FARNESYL PYROPHOSPHATE SYNTHETASE
QV-BT032-084	HISTIDINE AMMONIA-LYASE (HISTIDASE).	QV-BT011-146	MALTASE-GLUCOAMYLASE-LIKE (ALPHA-GLUCOSIDASE).
QV-BT032-080	NICOTINAMIDE NUCLEOTIDE TRANSHYDROGENASE PRECURSOR.	QV-BT065-009	ORNITHINE DECARBOXYLASE ANTIZYME
QV-BT077-008	SUPEROXIDE DISMUTASE [CU-ZN].	RC-BT133-054	HNK-1 SULFOTRANSFERASE.
QV-BT032-002	XANTHINE DEHYDROGENASE.	RC-BT108-018	ATP-CITRATE LYASE.
RC-BT031-147	BETA-GALACTOSIDE ALPHA-2, 3-SIALYLTRANSFERASE (CMP-N-ACETYLNEURAMINATE-BETA-GALACTOSAMIDE-ALPHA-2, 3-SIALYL- TRANSFERASE).	RC-BT119-020	GLUTAMATE DECARBOXYLASE
RC-BT029-288	PROTEIN ARGININE N-METHYLTRANSFERASE 2.	RC-BT132-024	GLUCOSE-6-PHOSPHATE ISOMERASE (PHOSPHOGLUCOSE ISOMERASE).

RC-BT029-168	PROTEIN DISULFIDE ISOMERASE-RELATED PROTEIN PRECURSOR (PDIR).	RC-BT105-056	PYRIDOXINE KINASE (PYRIDOXAL KINASE).
CM-BT092-002	68 KDA TYPE I PHOSPHATIDYLINOSITOL-4-PHOSPHATE 5-KINASE ALPHA (1-PHOSPHATIDYLINOSITOL-4-PHOSPHATE KINASE).	RC-BT105-067	ACYL-COA DESATURASE (STEAROYL-COA DESATURASE) (FATTY ACID DESATURASE).
CM-BT094-153	PUTATIVE PEROXISOMAL ANTIOXIDANT ENZYME (LIVER TISSUE 2D-PAGE SPOT 71B).	RC-BT105-070	INOSITOL 1,4,5-TRISPHOSPHATE 3-KINASE ISOENZYME (FRAGMENT).
CM-BT094-149	SERINE/THREONINE-PROTEIN KINASE SGK (SERUM/GLUCOCORTICOID-REGULATED KINASE).	RC-BT105-076	TYROSINE-PROTEIN KINASE JAK1 (JANUS KINASE 1) (JAK-1).
QV-BT011-023_1	PLASMA GLUTATHIONE PEROXIDASE PRECURSOR (GSHPX-P).	RC-BT133-032	FATTY ACID SYNTHASE.
QV-BT077-144	URACIL DNA GLYCOSYLASE	PM-BT103-023	CGMP-INHIBITED 3',5'-CYCLIC PHOSPHODIESTERASE B (CGIPDE1).
IL-BT111-008	CERAMIDE GLUCOSYLTRANSFERASE (GLUCOSYLCERAMIDE SYNTHASE) (FRAGMENT).	IL-BT027-020	ALDOLASE A
IL-BT091-002	CARBONIC ANHYDRASE (CARBONATE DEHYDRATASE) PRECURSOR.	IL-BT034-001	CARBOXYPEPTIDASE H PRECURSOR (CPH) (CARBOXYPEPTIDASE E).
IL-BT067-039	MRNA FOR LACTATE DEHYDROGENASE-A.	IL-BT038-005	RIBONUCLEASE 6 PRECURSOR.
IL-BT070-016	NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4.	IL-BT105-001	HEXOKINASE I (NICOTINAMIDE-NUCLEOTIDE ADENYLYLTRANSFERASE).
IL-BT038-003	CATHEPSIN Z PRECURSOR.	QV-BT077-075	PROTOPORPHYRINOGEN OXIDASE (PPO).
IL-BT108-007	GLUTATHIONE TRANSFERASE A4-4 (GLUTATHIONE S-ARALKYLTRANSFERASE).	RC-BT029-170_1	CARTILAGE GLYCOPROTEIN-39 PRECURSOR (GP-39) (39 KD SYNNOVIAL PROTEIN) (CHITINASE-3 LIKE 1).
RC-BT031-073	GLYCOGEN PHOSPHORYLASE, BRAIN FORM.	RC-BT133-044	CATHEPSIN B PRECURSOR (APP SECRETASE).
CM-BT086-053	ATPASE, CA++ TRANSPORTING, CARDIAC MUSCLE, SLOW TWITCH 2		

2. Proteínas envolvidas na síntese e reparo de DNA e divisão celular

Sequência	Similaridade	Sequência	Similaridade
RC-BT029-254	60S RIBOSOMAL PROTEIN L17 (L23).	QV-BT011-170	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (60 KD CHAPERONIN) (HSP-60).
RC-BT015-047	CYCLIN T2B.	QV-BT009-056	INTEGRIN ALPHA-E PRECURSOR (MUCOSAL LYMPHOCYTE-1 ANTIGEN) (CD103 ANTIGEN).
RC-BT031-113	ELONGATION FACTOR 1-ALPHA 1 (EF-1-ALPHA-1) (ELONGATION FACTOR TU) (EF- TU).	RC-BT133-052	CLEAVAGE SIGNAL-1 PROTEIN (CS-1).
RC-BT091-132	90 KD HEAT SHOCK PROTEIN	RC-BT133-050	C-ERB-2 PROTEIN
RC-BT091-068	CYCLIN-D BINDING MYB-LIKE PROTEIN.	RC-BT133-085	PREGNANCY-ASSOCIATED PLASMA PROTEIN A
RC-BT091-017	SMALL NUCLEAR RIBONUCLEOPROTEIN SM D1 (SM-D1) (SM-D AUTOANTIGEN).	QV-BT022-174	C-MYC ONCOGENE (EXON 1 SUBUNIT).
RC-BT091-001	ZINC FINGER PROTEIN 74.	QV-BT022-135	MITOGEN ACTIVATED PROTEIN KINASE ACTIVATED PROTEIN KINASE-3.
CM-BT092-004	CELL DIVISION CYCLE 42 (GTP-BINDING PROTEIN, 25KD)	QV-BT019-030	P120 CATENIN ISOFORM 2B
CM-BT040-016	A-RAF PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE (ONCOGENE PKS2).	QV-BT019-103_1	P97.
CM-BT006-079	HMG BOX FACTOR SOX-13	QV-BT032-025	NIDOGEN (ENACTIN)
CM-BT094-177	U1 SMALL NUCLEAR RIBONUCLEOPROTEIN C.	QV-BT021-006	GTPASE-ACTIVATING PROTEIN
CM-BT094-130	RIBOSOMAL PROTEIN L6	QV-BT088-003	HSP27 ERE-TATA-BINDING PROTEIN (SCAFFOLD ATTACHMENT FACTOR).
CM-BT094-125	TBP-ASSOCIATED FACTOR TAFII150.	QV-BT088-002	INTERFERON-INDUCIBLE PROTEIN 1-8U.
CM-BT094-111	NUCLEOPHOSMIN-ANAPLASTIC LYMPHOMA KINASE FUSION PROTEIN	RC-BT125-004	RAS-LIKE PROTEIN
CM-BT006-080	DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (RPB1).	RC-BT108-021	NK-TUMOR RECOGNITION PROTEIN (NATURAL-KILLER CELLS CYCLOPHILIN- RELATED PROTEIN).
CM-BT043-089	APOPTOSIS REGULATOR BCL-X.	RC-BT108-011	C-MYC BINDING PROTEIN MM-1.
CM-BT043-056	BETA-2-MICROGLOBULIN	RC-BT126-045	HUMAN APO-3 mRNA.
CM-BT086-062	BETA4-INTEGRIN.	RC-BT119-047	HUMAN CALCYCLIN GENE.
CM-BT086-044	TFIIB RELATED FACTOR HBRF.	RC-BT119-048	mRNA FOR MDC3
CM-BT086-045	INTEGRIN BETA-6 SUBUNIT PRECURSOR	RC-BT119-042	SWI/SNF COMPLEX 155 KDA SUBUNIT (BAF155) mRNA.
CM-BT086-040	TRANSCRIPTION FACTOR IL-4 STAT mRNA	RC-BT127-007	RAS-RELATED PROTEIN RAB-13.
CM-BT094-125	TATA BINDING PROTEIN ASSOCIATED FACTOR.	RC-BT029-025	MYOCYTE-SPECIFIC ENHANCER FACTOR 2A (SERUM RESPONSE FACTOR-LIKE PROTEIN 1).
CM-BT006-165	60S ACIDIC RIBOSOMAL PROTEIN P0 (L10E).	RC-BT132-021	CYSTATIN B (LIVER THIOL PROTEINASE INHIBITOR).
CM-BT066-090	PROBABLE RNA-DEPENDENT HELICASE P68 (DEAD-BOX PROTEIN P68).	RC-BT031-076	ELONGIN A.
CM-BT066-088	ARP2/3 COMPLEX 21 KD SUBUNIT (P21-ARC).	IL-BT150-014	40S RIBOSOMAL PROTEIN S4, X ISOFORM

			(SCR10).
CM-BT012-012	TUMOR SUSCEPTIBILITY PROTEIN.	RC-BT068-004	TRANSLATION REPRESSOR NAT1.
IL-BT027-022	MUCIN 1 PRECURSOR (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN).	RC-BT105-008	60S RIBOSOMAL PROTEIN L12.
IL-BT148-003	CIS-GOLGI SNARE p28	RC-BT105-052	LARGE PROLINE-RICH PROTEIN BAT2 (HLA-B-ASSOCIATED TRANSCRIPT 2).
IL-BT059-009	POSSIBLE GLOBAL TRANSCRIPTION ACTIVATOR SNF2L4 (BRAHMA PROTEIN HOMOLOG 1).	RC-BT105-079	CDC25C ASSOCIATED PROTEIN KINASE C-TAK1.
IL-BT128-010	PROTHYMOSIN ALPHA	PM-BT103-019	ELONGATION FACTOR 1-ALPHA
IL-BT070-009	SIGNALING LYMPHOCYTIC ACTIVATION MOLECULE (SLAM).	PM-BT103-033	SPINDLE POLE BODY PROTEIN SPC97 HOMOLOG.
IL-BT078-002	INITIATION FACTOR 4A-LIKE NUK-34 (HA0659).	RC-BT108-077	MITOSIS-SPECIFIC CHROMOSOME SEGREGATION PROTEIN SMC1 HOMOLOG.
IL-BT081-007	NUCLEAR ANTIGEN H731.	RC-BT097-034	TRANSFORMATION-SENSITIVE PROTEIN IEF SSP 3521.
IL-BT081-006	BB1=MALIGNANT CELL EXPRESSION-ENHANCED GENE/TUMOR PROGRESSION-ENHANCED GENE.	RC-BT113-002	TRANSCRIPTION FACTOR LZIP-ALPHA.
IL-BT082-002	RIBOSOMAL PROTEIN S13	IL-BT159-006	THREONYL-TRNA SYNTHETASE, CYTOPLASMIC (THREONINE--TRNA LIGASE) (THRRS).
IL-BT067-043	SPLICING FACTOR.	PM-BT107-009	ZINC FINGER PROTEIN 140 (ZNF140).
IL-BT067-023	SERINE/THREONINE KINASE RICK.	PM-BT107-011	POLYADENYLATE-BINDING PROTEIN 1 (POLY(A) BINDING PROTEIN 1).
IL-BT095-016	NF-KAPPA-B TRANSCRIPTION FACTOR SUBUNIT	QV-BT140-020	TRANSLATIONAL ACTIVATOR GCN1 (FRAGMENT).
IL-BT062-010	HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN A3 (HNRNP A3).	QV-BT141-018	DNA-BINDING PROTEIN B
IL-BT062-005	POLYADENYLATE-BINDING PROTEIN 1 (POLY(A) BINDING PROTEIN 1).	RC-BT153-025	METAXIN 2 (MTX2).
IL-BT094-003	MAC-2 BINDING PROTEIN PRECURSOR.	RC-BT133-091	PEROXISOMAL ASSEMBLY PROTEIN PEX3 (PEROXIN-3).
IL-BT070-050	GLUTAMINYL-TRNA SYNTHETASE (GLUTAMINE--TRNA LIGASE).	MR-BT039-202	INTEGRIN SUBUNIT ALPHA 10 PRECURSOR.
IL-BT070-062	SPLICEOSOME ASSOCIATED PROTEIN 114 (SAP 114).	MR-BT039-191	ALPHA-CP1.
IL-BT002-015	MDM2 PROTEIN (P53-ASSOCIATED PROTEIN).	MR-BT039-176	FIBULIN-1, ISOFORM C PRECURSOR.
IL-BT009-003	CREB-BINDING PROTEIN.	MR-BT041-110	CAVEOLIN-1.
IL-BT011-009	SP100-HMG NUCLEAR AUTOANTIGEN.	MR-BT041-101	CHROMODOMAIN-HELICASE-DNA-BINDING PROTEIN 2 (CHD-2).
IL-BT011-005	PUTATIVE TRANSCRIPTION FACTOR CA150.	MR-BT041-099	SYNAPTOGYRIN 2.
IL-BT011-002	INTEGRIN, ALPHA 4 (ANTIGEN CD49D, ALPHA 4 SUBUNIT OF VLA-4 RECEPTOR)	MR-BT041-097	SPLICING FACTOR PRP8.
IL-BT007-001	GTP-BINDING PROTEIN.	PM-BT056-301	MITOTIC CENTROMERE-ASSOCIATED KINESIN.
IL-BT099-005	M-PHASE PHOSPHOPROTEIN 4 (FRAGMENT).	PM-BT056-297	CELL SURFACE GLYCOPROTEIN CD44
IL-BT100-008	CCAAT-BINDING TRANSCRIPTION FACTOR I SUBUNIT A (ENHANCER FACTOR I SUBUNIT A).	PM-BT057-310	NON-FUNCTIONAL FOLATE BINDING PROTEIN.
IL-BT061-004	TRANSCRIPTION INITIATION FACTOR TFIID 55 KD SUBUNIT (TAFII55).	QV-BT065-136	RHO-ASSOCIATED, COILED-COIL CONTAINING PROTEIN KINASE P160ROCK.
MR-BT038-082	C-MYB PROTEIN.	QV-BT065-135	BG-X PROTEIN.
MR-BT039-232	HUMAN INTERCRINE-ALPHA (HIRH) mRNA.	IL-BT108-011	UV EXCISION REPAIR PROTEIN PROTEIN RAD23 HOMOLOG B (P58).
QV-BT124-029	P53		

3. Proteínas envolvidas na adesão celular e transdução de sinal

Sequência	Similaridade	Sequência	Similaridade
RC-BT091-162	CHONDROADHERIN.	QV-BT009-116	LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 1 PRECURSOR (APOLIPOPROTEIN E RECEPTOR).
RC-BT091-152	LAMININ-BINDING PROTEIN	RC-BT029-025	MYOCYTE-SPECIFIC ENHANCER FACTOR 2A (SERUM RESPONSE FACTOR-LIKE PROTEIN 1).
RC-BT091-100	LIPOCORTIN II	RC-BT068-018	CADHERIN-8 PRECURSOR.
RC-BT091-074	FIBROBLAST GROWTH FACTOR RECEPTOR 4.	RC-BT105-038	INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 3 PRECURSOR
RC-BT029-266	FIBROMODULIN PRECURSOR (FM) (COLLAGEN-BINDING 59 KD PROTEIN).	RC-BT105-044	sp P50995 ANXA_HUMAN [HOMO SAPIENS] (ANX11) ANNEXIN XI (CALCYCLIN-ASSOCIATED ANNEXIN 50) (CAP-50) (56 KD AUTOANTIGEN).
RC-BT029-164	GABA-A RECEPTOR PI SUBUNIT.	RC-BT105-078	CADHERIN-RELATED TUMOR SUPPRESSOR HOMOLOG PRECURSOR (FAT PROTEIN HOMOLOG).
CM-BT092-028	SEVEN TRANSMEMBRANE DOMAIN PROTEIN.	RC-BT105-080	NA+/NUCLEOSIDE COTRANSPORTER.

CM-BT092-019	FOLATE RECEPTOR	RC-BT133-020	CONNECTIVE TISSUE GROWTH FACTOR (FRAGMENT).
CM-BT040-062	COMPLEMENT COMPONENT C4A.	RC-BT133-023	METALLOPAINSTIMULIN
CM-BT040-059	COMPLEMENT C4B PRECURSOR.	QV-BT065-063	DUAL SPECIFICITY MITOGEN-ACTIVATED PROTEIN KINASE KINASE 2 (MAP KINASE KINASE 2) (ERK ACTIVATOR KINASE 2).
CM-BT094-151	IMMUNOGLOBULIN KAPPA LIGHT CHAIN	QV-BT065-064	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, ALPHA CHAIN PRECURSOR.
CM-BT094-148	HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, B-27 ALPHA CHAIN PRECURSOR	PM-BT103-007	CELL SURFACE ANTIGEN
CM-BT054-091	RABKINESIN6.	PM-BT103-014	REGULATOR OF G-PROTEIN SIGNALING 12 (RGS12).
CM-BT043-018	INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 5 (IGFBP5) mRNA	MR-BT137-024	NEUROTROPHIC TYROSINE KINASE, RECEPTOR, TYPE 1
IL-BT053-008	ESTROGEN REGULATED LIV-1 PROTEIN.	RC-BT113-005	NEUROBLAST DIFFERENTIATION ASSOCIATED PROTEIN AHNAK (DESMOYOKIN) (FRAGMENTS).
IL-BT161-008	DYNEIN LIGHT CHAIN 1 (PROTEIN INHIBITOR OF NEURONAL NITRIC OXIDE SYNTHASE).	IL-BT155-016	POLYUBIQUITIN.
IL-BT113-010	MATRIX GLA-PROTEIN PRECURSOR (MGP).	IL-BT164-007	NEUROPEPTIDE Y RECEPTOR TYPE 1 (NPY1-R).
IL-BT070-002	NUCLEOPORIN (FRAGMENT).	IL-BT165-007	G PROTEIN-COUPLED RECEPTOR RDC1 HOMOLOG.
IL-BT094-009	SODIUM CHANNEL BETA-1 SUBUNIT PRECURSOR.	IL-BT190-001	CD27L RECEPTOR PRECURSOR
IL-BT094-002	HEPATOMA-DERIVED GROWTH FACTOR (HDGF).	RC-BT133-084	MATRILIN-4.
IL-BT070-032	INTESTINAL MEMBRANE A4 PROTEIN (DIFFERENTIATION-DEPENDENT PROTEIN A4).	RC-BT133-082	IG ALPHA-1 CHAIN C REGION.
IL-BT032-009	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 1-ALPHA/BETA (TRANSCRIPTION FACTOR ISGF-3 COMPONENTS P91/P84).	QV-BT077-159	INTERFERON ALPHA/BETA SIGNALING PATHWAY-RELATED PROTEIN TYROSINE KINASE.
IL-BT033-006	LAMININ BETA-1 CHAIN PRECURSOR (LAMININ B1 CHAIN).	QV-BT022-170_1	ONCOSTATIN-M SPECIFIC RECEPTOR BETA SUBUNIT.
IL-BT108-006	SECRETORY CARRIER-ASSOCIATED MEMBRANE PROTEIN 2.	QV-BT022-116_1	EPIDERMAL GROWTH FACTOR RECEPTOR
MR-BT039-093	PROCOLLAGEN-LYSINE, 2-OXOGLUTARATE 5-DIOXYGENASE PRECURSOR (LYSYL HYDROXYLASE).	QV-BT032-142	PHOSPHOTYROSYL PHOSPHATASE ACTIVATOR.
MR-BT041-085	INTERLEUKIN 10 RECEPTOR	QV-BT032-134	TRANSDUCIN BETA-1 SUBUNIT
PM-BT046-252	CALCYCLIN (PROLACTIN RECEPTOR ASSOCIATED PROTEIN).	QV-BT079-003_1	FIBRONECTIN 1
PM-BT046-240	GEPHYRIN (PUTATIVE GLYCINE RECEPTOR-TUBULIN LINKER PROTEIN).	QV-BT009-075	ALZHEIMER'S DISEASE AMYLOID BETA PROTEIN 704 (FRAGMENT).
PM-BT057-310	NON-FUNCTIONAL FOLATE BINDING PROTEIN.	QV-BT124-124	mRNA FOR RETICULOCALBIN.
PM-BT057-309	PRO ALPHA 1(I) COLLAGEN.	QV-BT065-102	SODIUM/POTASSIUM-TRANSPORTING ATPASE ALPHA-1 CHAIN (SODIUM PUMP) (NA+/K+ ATPASE).

4. Proteínas de citoesqueleto, transporte e motilidade celular

Sequência	Similaridade	Sequência	Similaridade
RC-BT091-155	TUBULIN ALPHA-3 CHAIN.	MR-BT039-176	FIBULIN-1, ISOFORM C PRECURSOR.
RC-BT091-115	CYTOKERATIN 18	MR-BT041-110	CAVEOLIN-1.
RC-BT091-094	CALMODULIN mRNA.	QV-BT065-121_1	PROTEIN KINASE C, MU TYPE (NPKC-MU).
RC-BT029-208	SMOOTHIELIN.	QV-BT124-125	TUBULIN ALPHA-4 CHAIN
CM-BT092-007	GELSOLIN PRECURSOR (ACTIN-DEPOLYMERIZING FACTOR).	QV-BT009-075	ALZHEIMER'S DISEASE AMYLOID BETA PROTEIN 704 (FRAGMENT).
IL-BT070-008	ACTIN, CYTOPLASMIC 1 (BETA ACTIN).	RC-BT133-066	KERATIN, TYPE II CYTOSKELETAL 8.
IL-BT058-010	THROMBOSPONDIN3 (FRAGMENT).	RC-BT144-005	FIBRILLIN
IL-BT058-006	PROFILIN I.	QV-BT077-034	PRECURSOR LACTOFERRIN.
IL-BT070-046	OSTEOPONTIN PRECURSOR (BONE SIALOPROTEIN 1) (UROPONTIN).	RC-BT105-055	TETRACYCLINE TRANSPORTER-LIKE PROTEIN mRNA.
IL-BT037-009	MYOSIN REGULATORY LIGHT CHAIN.	QV-BT018-001	BONE PROTEOGLYCAN II PRECURSOR (DECORIN).
IL-BT011-001	SLOW MYBP-C.	QV-BT019-148	HEPARAN SULFATE PROTEOGLYCAN
IL-BT107-012	DYNACTIN, 150 KD ISOFORM (FRAGMENT).	RC-BT105-042	MYOSIN HEAVY CHAIN, NONMUSCLE TYPE A (CELLULAR MYOSIN HEAVY CHAIN, TYPE A).
IL-BT097-003	FERRITIN LIGHT CHAIN.	PM-BT103-025	BETA-FILAMIN.
RC-BT097-004	ACTIN, ALPHA CARDIAC.	RC-BT108-093	EZRIN (CYTOVILLIN).
RC-BT128-030	TROPOMYOSIN ALPHA CHAIN (SKELETAL MUSCLE)	QV-BT100-006	TITIN (FRAGMENT).

4.12. Identificação de novos genes humanos (ortólogos e parálogos) por análise de similaridade

Dentre as sequências anotadas que possuem similaridade a *clusters* do Unigene e/ou sequências codificadoras de proteínas humanas e não-humanas, temos ORESTES similares a genes humanos total ou parcialmente conhecidos (genes parálogos) e similares a genes conhecidos em outros organismos (genes ortólogos) que correspondem a prováveis novos genes humanos.

Os conceitos de ortologia e paralogia provêm de análises de homologia entre sequências gênicas. Por definição, homologia se refere a inferências baseadas em ancestralidade comum. Se a ancestralidade comum entre duas sequências pode ser traçada até um evento inicial de especiação, diz-se que estas estão relacionadas por ortologia (ortólogos). Se a ancestralidade comum de duas sequências pode ser traçada até um evento de duplicação gênica, então a relação entre elas refere-se a paralogia (parálogos) (Fitch, 1970).

Uma amostragem significativa de prováveis novos genes humanos identificados por análise de similaridade a bancos de dados de acesso remoto (*GenBank*) é mostrada na tabela a seguir.

EST	Tamanho (bp)	Similaridade (organismo)	Score* (bits)	E value*	Provável função
IL-BT000-001	253	notch 2 (rato)	178	7e-45	Proteína envolvida com diferenciação e destino celular
IL-BT000-002	280	semaphorin Z (rato)	115	7e-24	guia de motilidade axonal
PM-BT057-308	329	NABC1 (humano)	112	1e-25	desconhecida (superexpresso em tumores de mama)
IL-BT001-003	142	coronin-like (rato)	81.5	6e-16	proteína de ligação à actina
QV-BT003-122	418	F31E3.2 (<i>C.elegans</i>)	45.7	2e-05	quinase
QV-BT009-013	455	SPAC17G6.16C (<i>S.pombe</i>)	55.0	3e-07	fator de clivagem e poli-adenilação
QV-BT009-041	248	cadherin-like (camundongo)	109	3e-24	receptor de superfície celular
QV-BT009-057_2	232	CREB-binding protein (camundongo)	46.9	2e-05	co-ativador transcriptional
IL-BT026-007	437	Miranda (<i>D.melanogaster</i>)	47.6	4e-05	envolvido em divisões mitóticas assimétricas
MR-BT035-075	315	ZK675.2 (<i>C.elegans</i>)	86.2	6e-17	reparo de DNA
IL-BT054-002	306	T23B12.6 (<i>C.elegans</i>)	101	2e-21	proteína de canal de potássio
QV-BT009-027	100	AA521676 (camundongo)	125	1e-27	tirosina-fosfatase

IL-BT081-009	233	dynein-like (humano)	60.9	1e-09	proteína motora de microtúbulos
IL-BT054-005	276	calpain-like (humano)	145	7e-35	tiol-protease ativada por cálcio
IL-BT078-010	311	R29368_2 (humano)	59.7	6e-09	receptor hormonal
QV-BT009-027_1	219	Robo 1 (humano)	64.0	1e-10	receptor de difusão axonal
MR-BT038-041	74	KIAA0220 (human)	49.6	3e-06	desconhecida
QV-BT009-082	592	K09C4.8 (<i>C.elegans</i>)	131	5e-30	sulfatase
IL-BT111-007	616	alpha II spectrin (humano)	56.6	3e-15	proteína de membrana
IL-BT108-004	235	cortactin-binding protein (humano)	67.9	1e-11	proteína citoesquelética
RC-BT108-033	441	nuclear ribonucleoprotein (humano)	162	1e-40	proteína de ligação ao mRNA
RC-BT108-028	471	3-7 protein (humano)	153	6e-38	proteína com domínio <i>zinc-finger</i>
IL-BT097-010	349	KIAA0378 (humano)	108	2e-23	desconhecida
CM-BT094-112	180	D9651.6 (<i>S.cerevisiae</i>)	52.3	3e-07	helicase
RC-BT091-096	676	eyelid (<i>D.melanogaster</i>)	108	4e-23	proteína de ligação ao DNA
QV-BT089-006	243	polybromo 1 (<i>G.gallus</i>)	50.0	3e-06	metiltransferase
CM-BT086-007	226	C03A3.2 (<i>C.elegans</i>)	62.1	5e-10	desconhecida
CM-BT086-034	397	erbb2 receptor (humano)	62.8	2e-15	receptor
IL-BT078-005	186	KIAA0771 (humano)	74.1	1e-13	proteína de ligação a p53
QV-BT065-092	352	KIAA0798 (humano)	124	3e-29	proteína com domínio <i>zinc-finger</i>
PM-BT057-386	202	PGI protein (humano)	51.9	4e-07	desconhecida (marcador de células transformadas)
PM-BT056-298	417	M01b12.5 (<i>C.elegans</i>)	49.6	1e-05	desconhecida
MR-BT041-048	293	F16F9.4 (<i>C.elegans</i>)	56.6	4e-08	enzima lipolítica
MR-BT041-042	505	U76754 (camundongo)	67.1	7e-11	desconhecida
MR-BT041-057	493	CPG2-like protein (rato)	53.9	7e-07	gene regulado pela luz

* Os valores de *Score* e *E value* foram obtidos a partir da análise de similaridade contra bancos de dados específicos para proteínas (submissão em BLASTX, <http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=1>).

4.13. Caracterização preliminar de novos genes humanos identificados por análise de similaridade

4.13.1. *NABC1_5B* - uma nova isoforma do gene *NABC1*

Análise mais detalhada da sequência expressa PM-BT057-308 (item 4.12) mostra uma alta similaridade (*Score* = 112 bits, *E value* = 1e-25) ao gene humano *NABC1*. O alinhamento desta sequência, devidamente traduzida, com a estrutura primária da proteína *NABC1* mostra uma porção interna de 135 bp com provável potencial codificador mas sem similaridade a referida proteína, indicando a existência de um exon adicional (denominado exon 5B) no gene *NABC1* (fig. 16A).

Através de ensaios de RT-PCR, a partir de mRNA isolado de tumores mamários, foi possível proceder à amplificação de ambas isoformas. Análise densitométrica entre as bandas correspondentes à isoforma já descrita (*NABC1*) e à isoforma recém-identificada (denominada *NABC1_5B*) mostra uma aparente diferença nos seus níveis de expressão, onde a intensidade de bandejamento para *NABC1* é cerca de 10 vezes superior àquela observada em *NABC1_5B* (fig. 17A). Reações de PCR utilizando DNA genômico e mRNA como molde foram realizadas em paralelo como controle interno da reação. A re-amplificação específica da isoforma *NABC1_5B* foi obtida com a utilização de um *primer* comum a ambas isoformas e um *primer* específico à isoforma em estudo (*nested PCR*, fig. 17B). O produto de PCR específico para *NABC1_5B* (tamanho estimado de 270 pb) foi clonado e sequenciado para confirmar a existência desta nova isoforma e seu grau de similaridade com a sequência expressa de origem. Os resultados obtidos confirmam a existência de uma nova isoforma do marcador tumoral *NABC1*.

A partir do alinhamento entre a sequência parcial da isoforma *NABC1_5B* e os clones genômicos de BAC contendo grandes porções do cromossomo 20 (GenBank AC004501 e AC00522), foi possível obter a organização genômica de *NABC1* (fig. 16B). O gene *NABC1* é constituído por 12 exons e o locus gênico abrange uma região de 114 Kb no cromossomo 20 humano (fig. 16B).

Ensaio quantitativo de RT-PCR utilizando quatro amostras tumorais e normais de mama confirmam uma superexpressão da isoforma conhecida *NABC1* (fig. 17C) conforme

descrito anteriormente (Collins *et al.*, 1998). A nova isoforma NABC1 foi fracamente identificada em apenas uma das amostras tumorais analisadas.

Ensaio preliminares de RT-PCR com diferentes tipos tumorais (cólon, estômago, cabeça/pescoço e mama) mostraram uma tendência não-esperada de baixa expressão de ambas isoformas em tumores de cólon (resultados não mostrados). Ensaio quantitativos de RT-PCR com quatro amostras tumorais e normais de cólon mostraram uma baixa expressão de ambas isoformas (fig. 17D), confirmando os resultados previamente obtidos.

Estes resultados indicam uma ação tecido-específica das diferentes isoformas de *NABC1* durante a progressão tumoral em tecidos mamários e colorretais.

A

```

1  MGNQMSVQRVEDQENEPAEITYQDNASALNGVFVVVSTHTVQHLEEVDLGISVKTQNVATSSPETTEISAVADANGKNLGEAKPEAPAAKSRFFLMLSRPVPGRTGDAQADSSLSGVK 120
121 LDVSSNKAPANKDPSESWTLPVAAGPGQDTKTPGHAPAQDKVLSAARDPTLLPPETGGAGGEAPSPKPKDSFFDKFLDKGQEKVPFGDSQQEAKRAEHQDKVDEVPGLSGGQDDVPAG 240
241 KDIVDGKEKEGQELGTADCSVEGDPPEGLETAKDQDSQAAAIAENNNSTMSFTTKTIVSPNKAETKKDPEDT-----GAEKSP 315
DPEGLETAKDQDSQAAAIAENNNSTMSFTTKTIVSPNKAETKKDPEDTAPAAESVCPGAGCKTSEIOAGCTKKKHLDSPRLGLAFKKTFRHEGAKSP
316 TTSADLKSDKANFTTSQETQGAGNKGCPNPSGHTQSVTTPEPAKEGTKEKSGPTSLPLGKLFWKKSVEDSVPTGAENNVVCPESPVEIIKSKVESALQTVDLNEGDAPEPTEAKLKRE 435
TTSADLKSDKANFT
436 ESKPRTSILMAFLRQMSVKGDGGITHSEEEINGKDSQCQTSDEKTTTPPEPEPTGAPQKKGEGSSKDKKSAEMNKQKSNKQEAKEPAQCTEQATVDTNSLQNGDKLKQKPEKQQSLGG 555
556 FFKGLGPKRMLDAQVQTDVPSIGPVGKPK 584

```

B

Figura 16: A) Alinhamento de aminoácidos entre estrutura primária de NABC1 e EST traduzida, gerada a partir de tumor de mama (Genbank AJ904532). O intervalo observado entre resíduos 309 e 310 de NABC1 é consequente da presença de um novo segmento traduzido proveniente da EST gerada, que corresponde a um novo exon, não identificado anteriormente, denominado exon 5B. B) Estrutura esquemática da organização genômica do locus NABC1. Limites entre introns e exons foram definidos comparando-se o alinhamento do cDNA de NABC1 e a referida EST com dois clones genômicos do cromossomo 20, derivados do projeto genoma humano (Genbank AC004501 e AC005220). Assumiu-se que todos os exons remanescentes estão presentes em ambos isoformas de NABC1. O locus gênico completo abrange aproximadamente 114 Kb.

A **B** **C** **mama****D** **coloreto**

Figura 17: Imagem esquadrinhada de géis de poliacrilamida a 8%, corados com prata, para análise por RT-PCR das isoformas do gene NABC1. **(A)** Amplificação de ambas isoformas por *primers* específicos para NABC1, utilizando-se mRNA (canaleta 02), DNA genômico (canaleta 03) ou cDNA (canaleta 04), isolados de tecido tumoral de mama, como molde. mRNA (canaleta 02) foi utilizado como controle interno, correspondendo a mesma reação de PCR, mostrada na canaleta 04, sem a reação prévia de transcrição reversa. Os produtos de PCR com tamanho aproximado de 170 bp e 305 bp se referem às isoformas NABC1 e NABC1_5B, respectivamente. **(B)** Re-amplificação dos produtos de PCR gerados em (A), com um primeiro *primer* específico para a isoforma NABC1_5B e um segundo *primer* comum para ambas isoformas, utilizando-se 0,1 ul das respectivas reações com mRNA (canaleta 02), DNA genômico (canaleta 03) e cDNA (canaleta 04) como molde. A isoforma NABC1_5B é representada por um banda específica de 270 bp. **(C)** RT-PCR quantitativa utilizando-se 1µg de RNA total isolado de tecido mamário tumoral (T) e normal (N) de 4 pacientes independentes. Amostras 1, 2, 3 e 4 correspondem a carcinoma ductal invasivo (estadio I), carcinoma mucinoso (estadio I), carcinoma ductal invasivo (estadio IIB) e carcinoma ductal invasivo (estadio I), respectivamente. A fase exponencial da reação de amplificação foi estabelecida para NABC1 (28 ciclos), NABC1_5B (40 ciclos) e GAPDH (23 ciclos). **(D)** RT-PCR quantitativa utilizando-se 1µg de RNA total isolado de tecido coloretal tumoral (T) e normal (N) de 4 pacientes independentes. Amostras 1, 2, 3 e 4 correspondem a adenocarcinoma reto-sigmóide (estadio IV), adenocarcinoma reto-sigmóide (estadio II), adenocarcinoma reto-sigmóide (estadio I) e adenocarcinoma mucinoso de cólon ascendente (estadio IV), respectivamente. Condições de reação seguem conforme descrito no item 3.15.1. Amostras teciduais foram obtidas após cirurgia no Hospital A. C. Camargo (São Paulo, SP) e imediatamente congeladas em nitrogênio líquido. Análise histopatológica de rotina foi utilizada para devido diagnóstico. Amostras de carcinoma mamário (estádios I e IIB) e adenocarcinoma coloretal (estádios I, II e IV) foram submetidas a microdissecção. Material tumoral foi mantido livre de estroma circundante e tecido normal. Respektivas amostras normais foram retiradas de regiões de, ao menos, 5 cm de distância da margem tumoral. O RNA total dos respectivos tecidos foi isolado por método de isotiocianato de guanidina (*Trizol reagent*, Gibco-BRL) e digerido com DNase livre de RNase (Gibco-BRL). A purificação de mRNA foi feita conforme protocolo do kit utilizado (*MACS Oligo-dT Microbeads*, Miltenyi Biotec, CA, USA).

4.13.2. Gene humano semaforina 6B (*Sema 6B*)

Análise mais detalhada da sequência expressa IL-BT000-002 (item 4.12) mostra uma alta similaridade (*Score* = 115 bits, *E value* = 7e-24) ao gene de semaforina VIb (*Mus musculus*) e semaforina Z (*Rattus norvegicus*), indicando que esta EST corresponde a uma nova semaforina humana.

Sequências de cDNA correspondentes a este novo gene foram inicialmente geradas por RACE-PCR. De modo a promover uma maior agilidade na conclusão da sequência gênica completa, foram construídos, em paralelo, *primers* baseados em ESTs, já disponibilizadas (GenBank AI825934 e T15790), com alta similaridade às extremidades 3' e 5' do cDNA de semaforina Z (*primers* SEM4R and SEM4F, fig. 18). Reações de PCR baseadas nestes *primers*, utilizando-se um *pool* de cDNA referente a tumores mamários, permitiram a obtenção de um produto de PCR de 1.8 Kb, correspondente a um novo membro humano da classe VI de semaforinas denominado semaforina 6B.

Esta isoforma de semaforina 6B (GenBank AF216389, isoforma menor, *Sema 6B_S*) possui uma ORF de 1.551 pb (identidade de 84% à sequência nucleotídica de semaforina Z) compreendida por 13 exons, e uma porção 3'-UTR de 250 pb sem similaridade alguma a outras semaforinas. Esta isoforma codifica uma proteína de 517 aminoácidos (fig. 20).

O mapeamento do locus gênico de *Sema 6B* foi realizado por PCR, com os *primers* SEM2R and SEM2-2F (fig. 18), utilizando-se o painel de média resolução Stanford G3 (*Research Genetics*, USA). Esta amplificação resulta na geração de um fragmento gene-específico de 0.8 Kb contendo uma sequência intrônica do gene. O locus gênico de *Sema 6B* está ligado a STS SHGC-1476 (lod_score = 5.6), localizada no cromossomo 19 humano.

Análise de similaridade no programa BLAST entre semaforina 6B e semaforina Z no banco de dados HTGS indicou a possível existência de uma nova isoforma, gerada por *splicing* alternativo do gene *Sema 6B*, contendo exons adicionais na porção 3' do gene. Para avaliar esta possibilidade, foram realizados ensaios de RT-PCR com *primers* específicos para esta porção 3' com potencial codificador, ausente na sequência de cDNA de *Sema 6B_S* mas presente no clone genômico CTB-50L17, referente ao cromossomo 19 humano (Genbank AC011498) (*primers* SEM5F and SEM5-2R, fig. 18). Os resultados obtidos confirmaram a existência de uma isoforma maior de semaforina 6B (*Sema 6B_L*, GenBank AF293363) com uma ORF de 2.061 pb, contendo

4 exons adicionais e um sítio aceptor de splicing no meio do último exon (exon 17) na região 3' da sequência do mRNA (fig. 19). Ao contrário da isoforma *Sema 6B_S*, esta nova isoforma codifica uma proteína de 687 aminoácidos contendo prováveis domínios transmembrânico e citoplasmático, conforme verificado na análise do banco de dados Pfam 5.2 (<http://pfam.wustl.edu/>) (fig. 20).

A expressão de ambas isoformas foi avaliada por RT-PCR, utilizando-se os *primers* específicos SEM2R e SEM4R (fig. 21). Para garantir a especificidade dos produtos gerados, uma alíquota da referida reação foi re-amplificada com o *primer* interno SEM3R e SEM4R (*nested PCR*) (fig. 21, canaleta 04). As duas bandas geradas revelam a existência concomitante de ambas isoformas, apesar da isoforma *Sema 6B_L* ser nitidamente mais abundante que *Sema 6B_S*.

A expressão gênica tecido-específica de *Sema 6B* foi avaliada por análise de Northern blot em membranas contendo alíquotas de poli(A⁺) RNA de diferentes tecido humanos adultos (fig. 22A). Um forte bandejamento, com tamanho estimado de 4.5 Kb, foi visualizado em mRNA de cérebro. O referido bandejamento foi fracamente visualizado em mRNA derivado de coração, baço, placenta e pulmão. Este bandejamento está possivelmente relacionado com a isoforma maior *Sema 6B_L*, visto que ensaios anteriores de RT-PCR indicam que a isoforma *Sema 6B_S* é de menor abundância (fig. 21).

Para avaliar o possível envolvimento da expressão de *Sema 6B* em processos metastáticos, como já observado para outras semaforinas (Christensen *et al.*, 1998; Martin-Satué & Blanco, 1999; Eckhardt *et al.*, 1999), foi verificado se a expressão de *Sema 6B* é regulada por agentes anti-tumorais em linhagens celulares de glioblastoma humano (T98G e A172) com capacidade invasiva *in vitro*. Neste contexto, as linhagens celulares em estudo foram tratadas com hormônios glicocorticóides ou ácido retinóico (ATRA) por períodos crescentes de tempo e o RNA total correspondente foi isolado para análise de Northern blot. Como resultado, a expressão de *Sema 6B* não foi regulada pelo tratamento com glicocorticóides (dados não mostrados). Entretanto, o tratamento com ácido retinóico promoveu a inibição gradativa da expressão de *Sema 6B* em função do tempo de tratamento (fig. 22B). A expressão de *Sema 6B* foi inibida de 2,5 e 7 vezes na linhagem celular T98G tratada com ATRA por 24 e 72 horas, respectivamente. De modo semelhante, a expressão de *Sema 6B* foi inibida de 2,5 e 10 vezes na linhagem celular A172 tratada com ATRA por 48 e 72 horas, respectivamente. Nos ensaios realizados com a linhagem T98G, é possível visualizar uma banda adicional, com um tamanho

estimado de 3.8 Kb, que aparentemente representa *Sema 6B_S* e também é inibida pelo tratamento com ATRA (fig. 22B). Estes resultados indicam um possível papel oncogênico de *Sema 6B* na progressão de tumores cerebrais.

Para verificar se o aumento da expressão de *Sema 6B* ocorre em outros processos carcinogênicos como o carcinoma de mama, conforme já observado com *M-SemaH* (Martin-Satué & Blanco, 1999), foi examinada a quantidade de transcrito produzido em tumores mamários humanos e tecidos normais correspondentes por análise de Northern blot (fig. 23). Uma banda específica de 4.5 Kb foi visualizada nos quatro casos analisados, sendo que três dentre as quatro amostras tumorais mostraram uma leve queda de expressão de *Sema 6B*. Paralelamente, foi analisada a expressão de *Sema 6B* em linhagens celulares humanas de câncer de mama. A linhagem metastática MDA-MB-231, que apresenta níveis não-detectáveis de expressão de receptor de estrógeno e altos níveis de expressão de ativador de plasminogênio (relacionado com potencial metastático em câncer de mama) (Ranson *et al.*, 1998), mostrou baixos níveis de expressão de *Sema 6B* quando comparada às linhagens celulares não-metastáticas T-47D e MCF7 (fig. 23). Dentre as linhagens não-metastáticas, MCF7 mostrou um nível de expressão de *Sema 6B* menor que a linhagem T-47D, a qual apresenta baixos níveis de produção de ativador de plasminogênio (Ranson *et al.*, 1998).

Estudos adicionais sobre o papel biológico das diferentes isoformas gênicas e sua importância na progressão tumoral tecido-específica devem ser realizados para um melhor entendimento funcional do gene.

Sema 6B

Figura 18: Desenho esquemático da sequência de cDNA de *Sema 6B_L* (isoforma maior de semaforina 6B), mostrando os diferentes *primers* utilizados na obtenção da sequência completa de ambas isoformas (*Sema 6B_S* e *Sema 6B_L*). As posições de anelamento de cada *primer* são mostradas na figura. Condições de reação para a obtenção do sequenciamento completo de cada isoforma são descritas no item 3.14.2.

Figura 19: Desenho esquemático da organização genômica do locus SEMA 6B. Os limites entre introns e exons foram definidos conforme análise de alinhamento entre cada um das isoformas obtidas (Sema 6B_L e Sema 6B_S) e o clone genômico CTB-50L17, referente ao cromossomo 19 (GenBank AC011498). O cDNA de *Sema 6B_L* contém 4 exons adicionais na região 3' e utiliza um sítio aceitador de *splicing* no meio do exon 17 (descrito como exon 17b). Um sítio doador de *splicing* é utilizado pela isoforma the *Sema 6B_S* no meio do exon 12 (descrito como exon 12a). O locus gênico completo abrange aproximadamente 42 Kb.

Sema6BL	MQT PRASPPRPAL-LLLLLLLG-GAHLFPPEPPPLSVAPRDYLNHYPVFVSGSGPGRITP	58
Sema6BS	*****_*****_*****	58
SemaZ	*W***P*****-F*****LRVT***D*****S*****	59
SemaVib	*W***VP*****SFF*****-VTY*****S*****A	59
Sema6BL	AEGADDINIQRVLRVNRTLFIGDRDNLRYVELEPPTSTELRYQRKLTWRSNPSDINVCRM	118
Sema6BS	*****	118
SemaZ	***E*****Q*****S*****D****	119
SemaVib	***E*****Q*****S*****D****	119
Sema6BL	KGKQEGECRNFVKVLLLRDESTLFVCGSNAPNFPVCANYSIDTLQPVGDNISGMARCPYDP	178
Sema6BS	*****	178
SemaZ	*****I*****M***LL*****	179
SemaVib	*****I*****M***LL***S*****	179
Sema6BL	KHANVALFSDGMLFTATVTDFLAIDAVIYRSLGDRPTLRTVKHDSKNFKEPYFVHAVEWG	238
Sema6BS	*****	238
SemaZ	*****	239
SemaVib	*****	239
Sema6BL	SHVYFFFREIAMEFNYLEKVVVSRVARVCKNDVGSGSPRVLEKQWTSFLKARLNCVPGDS	298
Sema6BS	*****	298
SemaZ	*****	299
SemaVib	*****	299
Sema6BL	HFYFNVLOAVTGVVSLGGRPVVLAVFSTPSNSIPGSAVCAFDLTQVAAVFEGRFREQKSP	358
Sema6BS	*****	358
SemaZ	*****I*****MN*****	359
SemaVib	*****I*****MN*****	359
Sema6BL	ESIWTPVPEDQVPRPRPGCCAAPGMQYNASSALPDDILNFVKTHPLMDEAVPSLGHPWI	418
Sema6BS	*****	418
SemaZ	*****N***E*****S***	419
SemaVib	*****E*****S***	419
Sema6BL	LRTIMRHQLTRVAVDVAGPWCNQTVVFLGSEAGTVLKFLVRPNASTSGTSGLSVFLEEF	478
Sema6BS	*****-----	469
SemaZ	V***I*****I*****V*****K***V***T*P*I****	479
SemaVib	V*****I*****I*****K***V***T*P*I****	479
Sema6BL	ETYPDRCE R P G - G G E T G Q R L L D S G L L S L E L D A A S G G L L A A F P R C V R V P V A R C Q Q Y S G C	537
Sema6BS	-----*R-----*--*VGHA*	479
SemaZ	*****G*SSSA*W*-----*****L***	534
SemaVib	*****G*SSA*W*-----*****L***	534
Sema6BL	MKNICIGSQDPYCGWAPDGS C I F L S P G T R A A F E Q D V S G A S T S G L G D C T G L L R A S L S E D R A G	597
Sema6BS	R-V*VHERR---S*W*Q---R-*-----R-----	496
SemaZ	*****R**S*T*****D***	594
SemaVib	*****R**S*T*****D***	594
Sema6BL	LVS V N P L V T S S V A A F V V R A V V S G F S V G W F V G L R R E R L A R R K D K E A I L A H G A G E A V L S V S	657
Sema6BS	-----*LS-----	499
SemaZ	***L*****G*****GS*****	654
SemaVib	***I*****G*****GS*****	654
Sema6BL	RLGERRAQGPSRRWGFOKARGSPRCRLGV-----	687
Sema6BS	-----P*****	517
SemaZ	*****GT*T*G*GGAGGGPG*P*EAL*APLMQNGWTKAALLHGGPHDLDSGLLPTPEQT	714
SemaVib	*****GT**G*GGAGGGPG*P*EAL*APLMQNGWTKAALLHGGPHDLDTGLLPTPEQT	714
Sema6BL	-----	687
Sema6BS	-----	517
SemaZ	PLPQKRLPTTHPHAHALGPRAWDHSALLSASASTSLLLLAHTRAPEQPVPPTESGPESR	774
SemaVib	PLPQKRLPTPHPHAHALGSRWDHSALLSASASTSLLLLAPARASEQPQVPAEPGPESR	774
Sema6BL	-----	687
Sema6BS	-----	517
SemaZ	LCAPRSCRASHPGDFPLTPHASPDRRRVVSAPTGPLDSSSVGDDLPGPWSPPATSSSLRRP	834
SemaVib	LCAPRSCRASHPGDFPLTPHASPDRRRVVSAPTGPLDP-SVGDLPGPWSPPATSSSLRRP	833
Sema6BL	-----	687
Sema6BS	-----	517
SemaZ	GPHGPPTAALRRTHTFNSGEARPGGHRPRRHAPADSTHLLPCGTGERTAPPVP	887
SemaVib	GPHGPPTAALRRTHTFNSGEARPGGHRPRRHPPADSTHLLPCGTGERTAPPVP	886

Figura 20: Sequência de aminoácidos das isoformas de semaforina 6B (Sema6B_L and Sema6B_S), alinhadas com genes ortólogos da classe VI de semaforinas, semaphorin Z (SemaZ) and semaphorin VIb (SemaVIb). Resíduos de aminoácidos conservados são indicados por asteriscos. O provável domínio sema está mostrado em negrito. Porções proteicas definidas como sequência-sinal e domínio transmembrânico estão sublinhadas. O presente alinhamento múltiplo foi realizado pelo programa CLUSTAL W (1.7) (<http://www.clustalw.genome.ad.jp/>). A análise de domínios proteicos foi feita através de submissão ao banco de dados Pfam 5.2 (<http://pfam.wustl.edu/>).

Figura 21: Imagem esquadrihada de gel de agarose a 1%, suplementado com brometo de etideo, contendo alíquotas de reações de RT-PCR específicas para o gene semaforina 6B (isoformas *Sema 6B_S* e *Sema 6B_L*). A amplificação de ambas isoformas, realizada com os *primers* SEM2R e SEM4R, é indicada na canaleta 2 (sem *template*) e canaleta 3 (cDNA de tumor de mama como *template*). O produto de PCR mostrado na canaleta 3 é re-amplificado com os *primers* SEM3R and SEM4R (*nested PCR*) e evidenciado na canaleta 4. Os fragmentos de PCR específicos para as isoformas *Sema 6B_S* e *Sema 6B_L* são indicados na figura. O marcador de peso molecular (*100 bp ladder*) é mostrado na canaleta 1. Condições de reação são descritas no item 3.14.2.

Figura 22: Análise de *Northern blot* específica para semaforina 6B em diferentes tecidos humanos e linhagens celulares de glioblastoma humano. **(A)** Hibridização em membrana contendo poli(A)⁺ RNA isolado de diferentes tecidos humanos normais (Human 12-Lane Multiple Tissue Northern Blot, Clontech, Palo Alto, CA). GAPDH foi utilizado como controle interno da análise. **(B)** Hibridização com RNA total (10 µg) isolado de linhagens celulares de glioblastoma humano (T98G e A172) não-tratadas (0) ou tratadas com 10⁻⁵M de ácido retinóico (*all-trans-retinoic acid*, ATRA) por diferentes períodos de tempo (24, 48 and 72 h). 36B4 (acidic ribosomal phosphoprotein PO) foi utilizada como controle interno para a quantificação de RNA. A expressão relativa (*relative expression*, eixo y) de Sema 6B foi mensurada como um sinal densitométrico da relação Sema 6B/36B4 obtido com *Phosphor image storm* (*Molecular Dynamics*). Os níveis de expressão de células não-tratadas com ácido retinóico foi padronizado como valor 1. Linhagens T98G (ATCC CRL-1690) e A172 (ATCC CRL-1620) foram mantidas em meio DMEM (*Dulbecco's modified Eagle's medium*) suplementado com 2% (v/v) de soro fetal bovino. RNA total foi isolado utilizando-se o método de isotiocianato de guanidina (Trizol reagent, Gibco-BRL). Condições experimentais são descritas no item 3.15.3.

Figura 23: Análise de *Northern blot* específica para semaforina 6B em tecidos mamários humanos e linhagens celulares de câncer de mama. Hibridização de sondas específicas (Sema 6B e GAPDH) a RNA total isolado de tecido tumoral (T) e normal (N) (n=4) e linhagens tumorais de mama (T-47D, MDA-MB-231, MCF7). As bandas específicas foram quantificadas por densitometria e respectivos valores foram normalizados em função da expressão de GAPDH (relação de sinal densitométrico Sema 6B/GAPDH). Os valores densitométricos normalizados para cada amostra são os seguintes: 1N = 0.66; 1T = 0.3; 2N = 0.93; 2T = 0.2; 3N = 1.3; 3T = 1.3; 4N = 0.5; 4T = 0.22; T-47D (5) = 1.9; MB-231 (6) = 0.58; MCF7 (7) = 0.8. Condições de reação seguem conforme descrito no item 3.15.3. Amostras tumorais isoladas dos pacientes 1, 2, 3 and 4 correspondem a carcinoma de mama ductal invasivo em estádios IIB, IIIA, IIB and IIB, respectivamente. Análise histopatológica de rotina foi utilizada para devido diagnóstico. Material tumoral foi devidamente microdissecado, sendo mantido livre de estroma circundante e tecido normal. Respektivas amostras normais foram retiradas de regiões de, ao menos, 5 cm de distância da margem tumoral. O RNA total dos respectivos tecidos foi isolado por método de isotiocianato de guanidina (*Trizol reagent*, Gibco-BRL) e digerido com DNase livre de RNase (Gibco-BRL). A purificação de mRNA foi feita conforme protocolo do kit utilizado (*MACS Oligo-dT Microbeads*, Miltenyi Biotec, CA, USA). Condições experimentais são descritas no item 3.15.3.

4.13.3. Gene humano *Notch 2*

Análise mais detalhada da sequência expressa IL-BT000-001 (item 4.12) mostra uma alta similaridade (*Score* = 178; *E value* = 7e-45) ao gene murino *Notch 2* (*Rattus norvegicus*). Sequências de cDNA correspondentes a este novo gene foram inicialmente geradas por RACE-PCR. De modo a promover uma maior agilidade na conclusão da sequência gênica completa, foram construídos, em paralelo, *primers* baseados em ESTs já disponibilizadas e sequências genômicas de alta qualidade (GenBank AC025241 e AL359752) após devida análise de similaridade à sequência transcrita de *Notch 2* murino (GenBank M93661) (fig. 24). Reações de PCR baseadas nestes *primers*, utilizando-se um *pool* de cDNA isolado de tumores mamários, geraram diferentes produtos de PCR gene-específicos que, após montagem pelo programa computacional PhredPhrap, permitiram a obtenção de grande parte do cDNA humano de *Notch 2* (ao redor de 6.4 Kb) (fig. 25). O produto final desta montagem foi alinhado a uma sequência de cDNA já disponibilizada nos bancos de dados (GenBank AL133036), referente à porção 3' do referido transcrito (extensão aproximada de 5.0 Kb) (fig. 24), resultando na sequência completa do cDNA de *Notch 2* (GenBank AF315356). O cDNA completo de *Notch 2* humano (11.189 bp) contém uma ORF de 7416 pb que codifica uma proteína de 2471 aminoácidos (fig.26).

Utilizando o painel de média resolução Stanford G3 (*Research Genetics, Inc.*), foi realizado o mapeamento do locus gênico de *Notch 2*, o qual se localiza no cromossomo 1 em uma região próxima ao marcador D1S2696 (posição 337 do mapa).

A expressão gênica tecido-específica de *Notch 2* foi avaliada por análise de Northern blot em membranas contendo alíquotas de poli(A⁺) RNA de diferentes tecido humanos adultos (fig. 27A). Uma banda gene-específica com tamanho estimado de 11 Kb foi visualizada em mRNAs derivados de placenta, pulmão e leucócitos de sangue periférico.

Estudos preliminares para avaliar o papel de *Notch 2* em processos tumorigênicos, conforme já descrito por Larsson e colaboradores (1994), foram realizados utilizando-se tumores de mama como modelo experimental. Ensaio quantitativos de RT-PCR indicaram que tumores mamários femininos de estadiamento mais avançado (acima de estadio II) possuem uma expressão levemente aumentada de *Notch 2* quando comparados a tumores mais precoces (estadio I) (fig. 27B). Tal resultado não foi observado em linhagens celulares, visto que a

expressão de *Notch 2* sofre uma aparente diminuição em um *pool* de linhagens tumorais, quando comparado a linhagens normais (fig. 27B, canaletas 1 e 2).

Estudos mais aprofundados sobre a função deste novo gene no desenvolvimento humano, assim como em órgãos humanos em que ocorre sua maior expressão (placenta, pulmão e leucócitos) precisam ser realizados. Novos ensaios quantitativos de RT-PCR, com uma amostragem mais representativa de tumores de mama em diferentes estadiamentos, devem ser feitos para se verificar a reprodutibilidade e a veracidade dos resultados preliminares aqui mostrados.

Notch 2

Figura 24: Desenho esquemático da sequência de cDNA de *Notch 2* humano, mostrando os diferentes primers (e respectivo fragmento de cDNA com similaridade à porção 3' do transcrito, GenBank AL133036) utilizados na obtenção de sua sequência completa, estimada em 11 Kb. As posições de anelamento de cada primer são mostradas na figura. Condições de reação para a obtenção desta sequência completa são descritas no item 3.14.2.

Figura 25: Imagem esquadrihada de gel de agarose a 1%, suplementado com brometo de etídeo, contendo alíquotas de reações de RT-PCR específicas para o gene *Notch 2* humano. 1 - marcador de peso molecular (*1 Kb ladder*); 2 - alíquota de 2,5 μ l de reação realizada com *primers* NOT19F e NOT21R; 3 - alíquota de 2,5 μ l de reação realizada com *primers* NOT19F e NOT10R; 4 - alíquota de 2,5 μ l de reação realizada com *primers* NOT19F e NOT15R; 5 - alíquota de 2,5 μ l de reação realizada com *primers* NOT13F e NOT16R; 6 - alíquota de 2,5 μ l de reação realizada com *primers* NOT16F e NOT2F; 7 - alíquota de 2,5 μ l de reação realizada com *primers* NOT15F e NOT2F; 8 - alíquota de 2,5 μ l de reação realizada com *primers* NOT2R e NOT4R. Condições de reação são descritas no item 3.14.2.

Sequência completa do cDNA de *Notch 2* humano

gacacgcgaagaagatgccgcacctgcgcctcgctctgctgtggcgctgctggcgctctggctgtgctgcgcgacccccgcgatgcatgtcagtgctcgagatggctatgaa
M P A A L R P A L L W A L L A L A L L A L L C A T T P A H A L Q C R D G Y A E
ccctgtgtaaatgaagaatgtgtgttacctaccacaatggcagcaggaactcaaatgtccagaagcctcttgggggaatattglaacatcgagacccctgcyagaag
P C V N E G M C V T Y H N G T G Y C K C P E G F L G E Y C Q H R D P C E K
aacccgtgcgcaaatgggtggactgtgtgtggccagccatgtggggaaagccagctgcccagtgtgcctcagggtttacaggagaggaactgccagttacgacatcctcat
N R C Q N G G T C V A A Q A M L G K A T C R C A S G F T G E D C Q Y S T S H
ccatgtcttgtgtctcgaccctgctgaaatggcgccacatgccatatgctcagccgggatacctatgagtgcacctgtcaagtcgggtttacaggtaaggagtgccaatgg
P C F V S R P C L N G G T C H M L S R D T Y E C T C Q V G F T G G K E C Q W
accgatgctgctgtctctacctgtgcaaatggaagtacctgtaccactgtggccaacagttctctgtcaaatgctctcagagcttcaaggggcaaaatgtgagact
T D A C L S H P C A N G S T C T T V A N Q F S C K K C L T G F T G Q K C E T
gatgtcaatgagtgtagactccaggacactgcagcatgtgtggccacctgctcaacctgcttgcctaccagtgccatgctcagggtctcacaggccagtgactgt
D V N E C D I P G H C Q H G G T C L N L P G S Y Q C Q C L Q G F T G Q Y C
gacagcctgtatgtgcctgtgcacctgcctctgtgtcaatggaggacacctgtgcgcagactggtagtctcaacttttagtgtaaacgtgcctccagggtttgaagggaagc
D S L Y V P C A P S P C V N G G T C R Q T G D F T F E C N C L P G F A E G S
acctgtgagaggaatattgatgactgcctaacacacagtggtcgaaatggagggtttgtgtggatgggtgcaacctacaactgcgcgtgtgccccaactggacagga
T C T E R C A N I D D G C P N H R C Q N G G V C V D G V N T Y N C R C R C P P Q W T G
cagttctgcacagaggatgtgatgactgctgcagcccaatgctgtcaaaatgggggcacctgtgccaaccgcaatggaggtatggctgtgtgtcaacggc
Q F C T E D V D E C L L Q P N A C Q N G G T C A C N R N G G Y G C V C V N G
tggagtgagatgactgcagtgagaacattgatgattgtgcttgcctctactccaggctccacctgcagccgtgcagcctgctcctctctgtgcatgtgccagag
W S G D D C S E N I D D C A F A S C T P G S T C I D R V A S F S C M C P E
gggaaggcagggtctcctgtgtcatctggatgatgcatgcatcagcaatccttgcacaagggggacctgtgtgacaccaacccccaaatgggcaatatattgcaactgc
G K A A G L L C H L D D A C I S N P C H K G A L C D T N P L N G Q Y I C T C
cccaaggctacaaagggtgactgcacagaagatgtgatgaatggccatagccaactcctgtgagcatgcaggaataattgtgaacaggatggcgctctc
P Q G Y K G A D C T E D V D E C A M A N S N P C E H A G K C V N T D G A F
cactgtgagtgctgaagggttatgcaggacactgtgtgagatggacatcaatgagtgccatgcagacccctgcagaatgatgtcattctgtgataagattggaggc
H C E C L K G Y A G R P C E M D I N E C H S D P C Q N D A T C T C L D K I G G
ttcacatgtctgtgcatgccagggttcaaagggtgtgcatgtgtaattagaataaatgaatgtcagagcaacccctgtgtgaacaatgggcagtggtgtgataaagtcaat
F T C L C M P G F K G V H C E L E I N E C Q S N P C V N N G Q C V D K V N
cgtttccagtgctgtcctcctgtgttccactggccaggtttgcagattgatattgatgactgtccagttccgtgtctgaatggggcaagtgtatgcatacccg
R F Q C L C P P G F T G P V C Q I D A T D D C S S T P C L N G A K C I D H P
aatggctatgaatgccagttgtgccacaggtttcaactgtgtgtgtgagggagaactgcacaactgtgaccccgatccttgccacactggtagctgtcaggtatggat
N G Y E A C Q C A T G T G T G V L C E E N I D N C D P D P C C H G Q C T Q D G I
gattcctacacctgcatctgcaatcccgggtacatggcgccatctgcagtgaccagattgatgaatgttacagcagcccttgctgaacgatggctgcgtgcatgacctg
D S Y T C I C N P G Y M G A I C S D Q I D E C Y S S P C L N D G R C I D L
gtcaatggtaccagtgcaactgcagccaggcagctcaggggttaattgtgaatttaatttgatgactgtgcagtaacctgtatcatagtaactgtgatggagc
V N G Y Q C N C Q P G T S G V N C E I N F D D G C A S N P C I H G I C M D G
attaatgcctacagttgtgtctgctcaccaggaattcacagggcagagatgaacattgatgactgtgcctccaatccctgtgcgcaagggtgcaacatgtatcaac
I N R Y S C V C S P G F T G Q R C N I D I D E C A S N P C R K G A T C I N
ggtgtgaatgggttccgctgtatatgcccgaggggaccccatccacagctgctactcacaggtgaacgaatgctgagcaatccctgcatccatggaaactgtactgga
G V N G F R G T C I P E G P H P S C Y S Y Q V N E C L L S N P C I H G N C T G
ggctcagtgatataagtgctctgtgactgcaggtgggttggtacactgtaagtggaacaaatgaatgctccttgcgaatccatgcagaaatggagaaactgtgaac
G L S G Y K C L C D A G W V G I N C E V D K N E C L S N P C Q N G N G G T C D
aatcctggtgaatggatacaggtgtacttgcaagaagggtttaaaggctataactgccaggtgaatatgtagtgactgtgcctcaatccatgccaacaggacacctgc
N L V N G Y R C T C K K G F K G Y N C Q V N I D E C A S N P C L N Q G G T C
ttgtgacataaagggtcactacgtgtgctgccatacacaggcaagaattgtcagacagatttggtcctcctgttcccaaacctgttgagaatgctgctgtt
F D D I S G Y T C H C V L P Y T T G K N C Q T V L A P C A C S P N P C E N A V A
tgcaaaagatgccaaaattttgagagttataactgtctgtgtgctcctggtgcgaagtgacaggtgtaccattgacagatgagctgtatctccaagccctgcatgaac
C K E S P N F E S Y T C T C L C A P G W Q G Q R C T I D I D E C I S K P C C M N
catggctctgccataacacccagggcagctacatgtgtgaatgtccaccaggctcagttgtagtgcagtgaggagacattgatgactgcttgccttgccttgcctg
H G L C H N T Q G S Y M C E C P P G F S G M D C E E D I D D C L A N P C Q
aatggaggttctgtatgtagtgagtgtaactttctcctgcctcctcgggtttcaactggggataaagtcgcagacagacatgaatgagtgctgtgagtaagccctgt
N G S C M D G V N T F S C L C L P G F T G D K C Q T D M N E C L S E P C
aagaatggagggacctgctctgactacgtcaacagttacacttgcaagtgccaggcaggattttagtgaggtccattgtgagaacaacatcaatgagtgactgagagctcc
K N G G T C S D Y V N S Y T C K C Q A G F D G V H C E N N I N E C T E S S
tgtttcaatgggtgcacatgtgttgatggatataactccttctctgtgtcctgtgggtttcaactggatcctctgcctccatgagatcaatgaatgcagctctcat
C F N G G T C V D G I N S F S C L C P V G F T G S F C L H E I N E C S S H
ccatgctgaatgagggaagctgtgtgtatgctcctgggtacctaccgtgcagctgcctcctgggtcactgctggaaacactgcagacactggatgactgtgactgtgact
P C L N E G T C V D G L G T Y R C S C P L G Y T G K N C Q T L V N L C S R
tctccatgtaaaaacaaagggtacttgcgttcagaaaaagcagagtgccagtgccatgtgccatctggatgggtgggtgcttattgtgactgtgccaatgtctctgtgac
S P C K N K G T C T V Q K K A E S Q C L C P S G W A G A A Y C D V P N V S C D
atagcagctccaggaggtgtgctgttgaaactgtgccagcactcaggtgtgctcatcaatgtgcgaacacgactattgtcagtgccccctgggctactggtg
I A A S R R G G V L V E H L C Q H S G V C I N A G N T H Y C Q C P L G Y T G
agctactgtgaggagcaactgtagtgctgcgtccaaacctgccagcaggggcaacatgcagtgacttattgtggatacagatgcaggtgtgtccagggtatcag
S Y C E E Q L D E C A S N P C Q H G A T C S D F I G G Y R C E G T V P G Y Q
ggtgtcaactgtgagtagaagtgagtgagtgccagaatcagccctgccagaatggaggacactgtattgacctgtgaaccattcaagtgctctgtgccaccaggcact
G V N C E Y E V D E C Q N Q P C Q N G G T C I D L V N H F G T C A V A S N M F D G
cggggcctactctgtgaagagaacattgatgactgtgcgggggtccccactgcttaattggtggtcagtgcatgtaggtagggaggtacagttgtgcagcctgtgcgtgctgctgct
R G L L C E E N I D D C A R G P H C L N G G Q C M D R I G G Y S C R C L P
ggcttctgtggggagcgttgtgaggaggacatcaacagtgctctccaaacctgcagctgtaggggcagctggagctgtatacagctcaccaatgactacactgtgtgt
G F A G E R C E G D I N E C L S N P C S S E G S L D C I Q L T N D Y L C V
tgccgtagtgccttactggccggcactgtgaaacctctgctgagatgtgtgtcccagatgcctgctgaatggagggaactgtgtgtgtggccagtaacatgctgatggt
C R S A F T G R H C E T F V D V C P Q M P C L N G G T C A V A S N M F D G
ttacttgcgttgcctcccccgggtatttccggggcaaggtgccagagcagctgtgacaagtgaaatgtagggaaggggcaggtgtgtgcacacccctctggaccggc
F I C C R C P P P G F S G A R C Q S S C G Q V K C R K G E Q C V H T A S G P R
tgcttctgccccagctccccgggactgcagtgagctgtgccagttagccctgccagcagggggcagctgccacctcagcgcagcctccttattactcctgcagtg
C F C P S P R D C E S G C A S S P C Q H G G S C H P Q R Q P P Y Y S C Q Q

102

gatgaggaggacaacactgcctttcaaactttttactgatctcttagattttaagaactcttgaattgtgtggtatctaataaagggaaggtaagatggataatcacttt
ctcatttgggttctgaattggagactcagtttttatgagacacatcttttatgccatgtatagatcctccctgctattttgggtttattttattgttataaatgctttc
tttctttgactcctctctgctgcctttgggtaggtttttgtttgtttattgtctcctctgtttgttttaagcatcattttcttatgtgaggtggggaagggaa
aggtatgagggaaagagagtctgagaattaaaatattttagtataagcaattggctgtgatgctcaaatccattgcatcctcttattgaatttgccaatttgtaattttg
cataataaagaaccaagggtgtaattgtttgttgagaggtggttttagggattttggccctaaccaatacattgaatgtatgatgactatttgggaggacacatttatgtac
ccagaggccccactaataagtggtactatggttacttcttgtgtacatttctcttaaaagtgaattatctgtttgtatgagaaaccagtaaccaataaaaatgacc
gcataattcctgactaaacgtagtaaggaaaaatgcacactttgttttacttttccgtttcattctaaaggtagttaagatgaaatttataatgaaagcattttatcaciaa
ataaaaaagggttgccaagctcagtggtgtgtattttttattttccaatactgcatccatggcctggcagtggttaacctcatgatgtcataatttgcagagagcaaatt
ttcttttctttctgaatcccaaaagcctagcaccaaaacttcttttttcttctttaaattagatcataaataaatgatcctggggaaaaagcatctgtcaaataggaaac
atcacaaaactgagcactcttctgtgcactagccatagctggtgacaaaacagatggttgctcagggacaagggtgccttccaatggaaatgcgaagtagttgctatagcaag
aattgggaactgggatataagtcataatattaattatgctgttatgtaaagattggtttgtaacattccttaagtgaatttgtgtagaacttaataacaggattataa
aataatattttgtgtataaatttgttataagttcacattcatatattttataaagtcagttagatatttgaacatgaaaaaaaaa

Figura 26: Alinhamento da sequência nucleotídica completa do cDNA de *Notch 2* humano (11.189 pb) com respectiva estrutura primária da proteína traduzida. Códon de iniciação e terminação são sublinhados.

A

Cérebro
Coração
Músculo esquelético
Cólon
Timo
Baço
Rins
Fígado
Intestino liso
Placenta
Pulmão
Leucócitos

11.0 Kb ▶

NOTCH2

1.3 Kb ▶

GAPDH

Detailed description: This figure shows a Northern blot analysis of NOTCH2 and GAPDH mRNA expression. The top panel, labeled 'NOTCH2', shows a band at 11.0 Kb. The bottom panel, labeled 'GAPDH', shows a band at 1.3 Kb. The tissues analyzed are Cérebro, Coração, Músculo esquelético, Cólon, Timo, Baço, Rins, Fígado, Intestino liso, Placenta, Pulmão, and Leucócitos. NOTCH2 expression is detected in most tissues, with strong bands in Intestino liso, Placenta, and Leucócitos. GAPDH expression is used as a loading control and is detected in all tissues, with strong bands in Cérebro, Coração, and Músculo esquelético.

B

		* * 0 0				* * * *								* *						
N	T	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
NOTCH2
GAPDH

Figura 27: Caracterização do gene *Notch 2* humano. **(A)** Análise de *Northern blot* específica para *Notch 2* em diferentes tecidos humanos. Hibridização em membrana contendo poli(A)⁺ RNA isolado de diferentes tecidos humanos normais (Human 12-Lane Multiple Tissue Northern Blot, Clontech, Palo Alto, CA). GAPDH foi utilizado como controle interno da análise. Condições experimentais são descritas no item 3.15.3. **(B)** RT-PCR quantitativa utilizando-se 1µg de RNA total isolado de tecido mamário tumoral (T) e normal (N) de 9 pacientes independentes (amostras 3 a 20) e de *pools* de linhagens celulares (amostras 1 e 2). Canaletas indicadas por asteriscos (*) representam amostras de carcinoma ductal invasivo de estadiamentos IIA e IIB e respectivos tecidos normais. Canaletas indicadas por círculos (o) representam amostras de carcinoma ductal invasivo de estadiamento IIB e respectivo tecido normal, isolados de paciente do sexo masculino. Demais canaletas contendo amostras tumorais (canaletas 7, 9, 11 e 17) correspondem a carcinoma ductal invasivo com estadiamento inicial (estadio I). A fase exponencial da reação de amplificação foi estabelecida para NOTCH2 (32 ciclos) e GAPDH (23 ciclos). Condições de reação seguem conforme descrito no item 3.15.1. Amostras teciduais foram obtidas após cirurgia no Hospital A. C. Camargo (São Paulo, SP) e imediatamente congeladas em nitrogênio líquido. Análise histopatológica de rotina foi utilizada para devido diagnóstico. Amostras tumorais foram submetidas a microdissecção, sendo mantidas livres de estroma circundante e tecido normal. Respectivas amostras normais foram retiradas de regiões de, ao menos, 5 cm de distância da margem tumoral. O RNA total dos respectivos tecidos foi isolado por método de isotiocianato de guanidina (*Trizol reagent*, Gibco-BRL) e digerido com DNase livre de RNase (Gibco-BRL).

5. DISCUSSÃO

5.1. Estabelecimento do número total de ESTs geradas e modelo experimental de estudo

Estudos preliminares no modelo de câncer de mama (item 5.2) mostraram que em uma amostragem de 403 ESTs produzidas, um total de 194 (48%) não apresentavam similaridade às sequências depositadas nos bancos de dados (*E value* superior 10^{-5}). Considerando-se que tais sequências expressas sejam correspondentes a prováveis novos genes humanos, a extrapolação para um número de 6000 ESTs indicaria que, aproximadamente, 3000 sequências estariam dentro deste contexto. Este número, apesar de conter uma suposta redundância, auxiliaria na validação da técnica de RAP-PCR como uma importante ferramenta na identificação de prováveis novos genes humanos.

Aproximadamente 10% dos 80.000 a 100.000 genes humanos estimados foram completamente sequenciados e caracterizados até o presente momento. Dos 90% restantes, muitos genes humanos já foram identificados através de técnicas convencionais de geração de ESTs e consequente *Gene clustering* (*Unigene database*), porém não totalmente sequenciados e caracterizados. Até o ano de 1999, haviam sido identificados aproximadamente 60.000 *clusters* correspondentes a gene humanos, oriundos de projetos ligados ou não ao CGAP (*Cancer Genome Anatomy Project*). A partir destas informações, podemos afirmar que muitas das ESTs sem similaridade aos bancos de dados (*no matches*), geradas no presente trabalho, podem corresponder a novos genes humanos mas, na sua grande maioria, condizem a genes já identificados e não totalmente sequenciados. Estas sequências geradas não encontram similaridade aos bancos de dados internacionais pois correspondem, na sua maioria, a porções gênicas mais centrais (*coding regions*), enquanto que a maior parte das ESTs convencionais se referem às extremidades 5' e 3' dos genes.

Sabe-se que, durante o processo de carcinogênese, ocorrem várias alterações na expressão gênica da célula. Estas alterações podem ser ocasionadas, por exemplo, pelo aumento ou diminuição da transcrição de diversos genes envolvidos com o processo. Considerando-se que (i) os genes pouco expressos representam 40 a 45% do mRNA total da maioria das células humanas (Bishop *et al.*, 1974; Davidson & Britten, 1979); (ii) a maior parte dos genes desconhecidos se referem a genes pouco expressos; (iii) os genes pouco expressos em células

normais podem ter seus níveis de mRNA alterados durante a transformação maligna e (iv) projetos como o CGAP (*Cancer Genome Anatomy Project, NCBI, NIH*), responsável pela geração de milhares de ESTs em modelos tumorais, tem favorecido um grande avanço na descoberta de novos genes, podemos concluir que a escolha de um tumor maligno como modelo experimental, no caso o carcinoma de mama, para a geração de ESTs é uma opção adequada para o presente estudo.

5.2. Padronização metodológica e análise comparativa da qualidade de preparações de mRNA

Através de uma técnica variante de *RNA fingerprinting* (RAP-PCR), rotineiramente utilizada na análise de expressão diferencial em diferentes organismos, produzimos 6029 ESTs no modelo de câncer de mama. Isto equivale, aproximadamente, ao sequenciamento de 1,9 milhões de nucleotídeos, distribuídos em sequências com tamanho médio de 313 bases.

As ESTs foram geradas a partir do sequenciamento parcial de clones aleatórios presentes em 170 mini-bibliotecas de cDNA. Estas mini-bibliotecas foram produzidas a partir da síntese de cDNA com *primers* arbitrários. Foram utilizadas 15 preparações de mRNA (validadas e não-validadas) para a síntese de cDNA. As purificações de mRNA foram procedidas a partir de 6 tumores de mama (diferentes estadiamentos histopatológicos) e uma preparação de RNA total, proveniente de 8 tumores em diferentes estágios de desenvolvimento.

A padronização da síntese de cDNA e produção dos perfis iniciais de RAP-PCR para a geração de mini-bibliotecas foi realizada com a preparação de mRNA RBT001. Esta preparação não foi avaliada segundo os critérios de qualidade estabelecidos para a geração de ESTs proposta neste trabalho. Estes critérios surgiram à medida que foram verificadas limitações no aproveitamento de mini-bibliotecas informativas para a identificação de sequências codificadoras expressas.

Durante a padronização da técnica de RAP-PCR para a geração de ESTs, obtivemos 2 mini-bibliotecas de cDNA a partir de perfis X-X. Destas mini-bibliotecas, foram geradas 403 ESTs que, após análise em BLAST, forneceram os seguintes resultados:

ESTs com alta similaridade ao <i>database</i> ($E < 10^{-25}$)	083 (21%)
ESTs com similaridade parcial ao <i>database</i> ($10^{-25} < E < 10^{-5}$)	031 (08%)
ESTs sem similaridade ao <i>database</i> ($E > 10^{-5}$)	194 (48%)
ESTs correspondentes a sequências Alu (<i>small interspersed repeats</i>)	052 (13%)
Sequências de DNA mitocondrial	017 (04%)
Sequências de RNA ribossomal	026 (06%)

Entretanto, uma análise mais detalhada destas ESTs mostrou que o seu conteúdo de G-C era relativamente baixo (em torno de 40%) e muitas das sequências geradas correspondiam a outros elementos repetitivos além de sequências Alu (vide análise parcial no laudo C2-C2/test, figura 12). Em geral, tais elementos aparecem com maior frequência no DNA genômico e em porções não-traduzíveis de mRNA, sendo pouco informativos para a identificação de *coding regions*. De posse destes dados, concluímos que o alto percentual de ESTs sem similaridade ao *database* estava sendo falseado pela presença de DNA genômico contaminante.

Novas preparações de mRNA foram providenciadas, seguindo diferentes protocolos de purificação, na presença ou ausência de DNase I. O primeiro critério de qualidade, estabelecido a partir da preparação RBT003, foi a análise do grau de contaminação com DNA genômico nos diferentes estoques de mRNA por reação de PCR específica para o gene da β -globina. Este tipo de reação foi utilizada devido à sua grande sensibilidade, sendo capaz de gerar bandejamento a partir da amplificação de um único gene contaminante de globina. As preparações RBT003, RBT004, RBT005, RBT009, RBT010 e RBT012 foram descartadas através desta análise.

Estabelecido este primeiro critério, foi observado que determinadas mini-bibliotecas, independentemente do *primer* utilizado na RAP-PCR, originavam um alto percentual de sequências de DNA mitocondrial. Desta maneira, procedemos a avaliação de uma possível contaminação das preparações de mRNA com DNA mitocondrial através de reação de PCR utilizando *primers* específicos para a região não-expressa de D-loop mitocondrial. Como resultado, verificamos que as preparações RBT002 e RBT006, previamente utilizadas na produção de mini-bibliotecas de cDNA, estavam parcialmente contaminadas com DNA mitocondrial. As demais preparações de mRNA, utilizadas na geração de ESTs, estavam de acordo com este novo critério de validação estabelecido.

A porcentagem total de sequências contendo elementos repetitivos, nas diferentes mini-bibliotecas geradas, foi analisada de acordo com a preparação de mRNA originária. Esta análise foi válida para (i) avaliar a qualidade do mRNA utilizado e (ii) inferir a posição das ESTs ao longo do respectivo mRNA-molde (alto percentual de repetições é mais característico em porções 5' e/ou 3' não-traduzíveis). Tal análise segue na tabela a seguir:

				Análise de elementos repetitivos (RepeatMasker)				
	Número de bases sequenciadas (bp)	Conteúdo total de G-C (%)	Número de ESTs geradas	SINEs (%)	LINEs (%)	LTR (%)	DNA (%)	Total de elementos repetitivos (%)
RBT002	146.997	45,72	479	5,35	9,21	2,94	1,36	19,34
RBT006	493.182	48,50	1576	4,33	5,24	1,78	0,65	12,18
RBT008	148.276	46,71	612	3,94	4,75	5,79	0,55	15,20
RBT011	397.008	47,75	1134	3,62	4,38	1,48	0,18	09,79
RBT013	198.217	47,43	848	2,93	3,13	1,31	0,37	07,84
RBT014	316.511	51,03	878	2,06	1,94	0,99	0,45	05,45
RBT015	186.991	45,62	502	5,04	5,49	0,97	0,98	12,47
Total	1.887.182	-----	6.029					

Como pode se observar, obtivemos uma diminuição constante na presença de elementos repetitivos nas mini-bibliotecas geradas (com exceção da preparação RBT015). Valores iniciais de 19,3% de sequências correspondentes a elementos repetitivos (RBT002) diminuíram à medida que otimizamos os critérios de validação na preparação de mRNA e produção de mini-bibliotecas, atingindo percentuais inferiores a 10% do total de sequências geradas. Tal observação também é possível na análise do conteúdo de G-C das ESTs geradas, onde há uma tendência de aumento neste valor nas sequências produzidas a partir de mRNAs validados (por exemplo, RBT014). Ou seja, a diminuição no percentual de elementos repetitivos e elevação no conteúdo de G-C das sequências geradas está diretamente relacionado ao aumento gradativo na qualidade das preparações de mRNA e produção crescente de ESTs correspondentes à região codificadora dos genes expressos.

A preparação RBT015 mostrou uma alta qualidade no que se refere à ausência de contaminação com DNA genômico ou mitocondrial. Entretanto, devido ao extensivo tratamento com DNase I (45 minutos de reação), a preparação de RNA total se tornou susceptível a degradação enzimática. Como o método de purificação de mRNA é baseado na seleção das

moléculas pela cauda poli-A, houve um enriquecimento de fragmentos de mRNA contendo porções 3' não-traduzíveis. Isto justifica o baixo conteúdo de G-C das sequências e o relativo aumento no percentual de elementos repetitivos, mais característicos em regiões não-codificadoras do mRNA. Análise detalhada das ESTs geradas com a preparação RBT015 mostra que tais sequências tem grande similaridade à porções 3' não-codificadoras de ESTs e/ou mRNAs já depositados nos bancos de dados de acesso remoto.

5.3. Eficiência dos diferentes tipos de perfis de cDNA na geração de sequências expressas informativas e não-redundantes

Na produção de mini-bibliotecas de cDNA, utilizamos com maior frequência perfis X-X de RAP-PCR. De fato, foram feitas 151 mini-bibliotecas com perfis X-X (89%), 11 com perfis X-Y (6,4%) e 8 com perfis X-Xn (4,6%). Isto porque, na análise geral, os perfis X-X se mostram mais complexos que os perfis X-Xn e seu bandejamento característico tem um tamanho superior àquele obtido nos perfis X-Y.

Quando se compara o bandejamento de perfis X-Xn, obtidos na figura 09, com aqueles obtidos na figura 07 (perfil X-X), é possível observar uma melhor distinção das bandas geradas no perfil X-Xn, indicando uma menor complexidade do produto de reação. Na geração de ESTs, as mini-bibliotecas produzidas com perfis X-Xn forneceram altas taxas de redundância, conforme mostrado a seguir:

Mini-bibliotecas	Total de ESTs geradas	Redundância (%)
BT132	042	52,63
BT133	077	56,15
BT137	018	61,04
BT138	035	80,00
BT139	013	53,85
BT140	031	53,85
BT141	028	28,57
BT142	025	43,75

Uma possível explicação para a redundância das mini-bibliotecas obtidas com perfis X-Xn vem da própria metodologia utilizada na produção destes perfis (*nested* RAP-PCR, descrito em McClelland & Welsh, 1995). A re-amplificação, em alta estringência, do perfil X-X com um

primer semelhante àquele utilizado na síntese do cDNA, contendo apenas um nucleotídeo a mais na sua porção 3' final, promove a amplificação preferencial das sequências que possuam uma alta similaridade ao *primer* contendo esta base adicional. A presença desta base adicional ocorre em 1/16 das sequências geradas no perfil inicial, nas respectivas porções 5' da dupla fita de DNA (chance de 25% desta base estar presente em uma das porções finais do produto de PCR). Isto acarretará a parcial normalização das sequências selecionadas, elevando a concentração das sequências pouco amplificadas (*background*) na RAP-PCR inicial. No entanto, ocorrerá uma redução da ordem de 16 vezes na complexidade do perfil original. Isto quer dizer que, se o perfil X-X inicial possui 400 sequências diferentes de cDNA, a re-amplificação deste perfil irá produzir apenas 25 sequências, causando uma alta redundância na mini-biblioteca produzida com o perfil re-amplificado.

Como alternativa ao *nested* RAP-PCR, foram produzidos perfis X-X_n com 2 *primers* simultaneamente, cada um contendo um nucleotídeo diferente na porção 3'. Seguindo o mesmo princípio, é esperada a amplificação seletiva de 1/4 das sequências de cDNA do perfil inicial. Portanto, mini-bibliotecas produzidas com estes perfis deverão ter uma taxa de redundância inferior aos perfis X-X_n convencionais. De fato, este resultado é observado nas mini-bibliotecas BT141 e BT142, as quais possuem uma menor taxa de redundância (29% e 44%, respectivamente) em relação às demais. Entretanto, estas taxas estão acima dos critérios de qualidade estabelecidos, inviabilizando a utilização dos perfis X-X_n na geração de mini-bibliotecas validadas.

Os perfis X-Y possuem, em geral, um bandejamento com tamanho inferior ao obtido nos perfis X-X (vide figuras 7 e 8). Isto porque os *primers* utilizados na amplificação em baixa estringência tem uma maior probabilidade de se anelar em porções mais internas àquelas delimitadas pelos *primers* envolvidos na síntese do cDNA, produzindo perfis inferiores a 200 bp (este valor é usado como *cut-off* pois, após a edição das ESTs, as sequências geradas tem um tamanho superior a 150 bp). Como, em geral, os perfis X-Y são mais complexos abaixo de 200 bp, as bandas de maior tamanho tem uma baixa concentração no perfil. Esta observação, aliada a purificação e precipitação deste material (*size-selection* em gel de agarose), faz com que a quantidade final das referidas sequências seja baixa, dificultando o processo de clonagem e transformação celular. Mesmo procedendo a alterações na produção dos perfis X-Y (maior concentração de *primer* e maior tempo de anelamento na PCR), tivemos dificuldades na

clonagem deste material. Devido a estes fatores, os perfis X-Y foram pouco utilizados na produção de mini-bibliotecas de cDNA.

5.4. Análise de similaridade das sequências expressas geradas

As submissões estabelecidas para a análise de similaridade das sequências expressas geradas foram precedidas pelo “mascaramento” de sequências contendo total ou parcialmente elementos repetitivos (9,1% das sequências geradas) e remoção de sequências com similaridade significativa a DNA mitocondrial (7,1%) e RNA ribossômico (3,2%), pois não contribuem para o processo de descoberta gênica e montagem de genes humanos parcialmente conhecidos. As submissões iniciais foram realizadas frente a bancos de dados referentes a proteínas humanas (24,7% das sequências geradas) e *contigs* humanos (17,7% das sequências geradas), ou seja, genes humanos total ou parcialmente conhecidos. Os genes humanos parcialmente conhecidos são disponibilizados pelo database *Unigene* na forma de *clusters* (grupamentos de ESTs já depositadas referentes a um mesmo gene), devidamente alinhados para a formação de *contigs* humanos.

As sequências expressas sem similaridade a proteínas e/ou *contigs* humanos foram então analisadas frente a bancos de dados de ESTs humanas ausentes do *Unigene* (6,6% das sequências geradas). A eventual similaridade a este banco de dados poderia corresponder a genes humanos pouco expressos, pois não existem *clusters* pré-estabelecidos para a montagem de *contigs* humanos. Sequências sem similaridade ao referido banco de dados foram posteriormente analisadas frente a bancos de proteínas e ESTs não-humanas (1,5% das sequências geradas). Os novos genes ortólogos humanos, descritos no item 4.12, foram identificados durante este procedimento. Por outro lado, algumas sequências expressas com similaridade a genes bacterianos foram igualmente geradas, sendo provenientes, em geral, de eventuais contaminações nas soluções de *primers* utilizadas.

Sequências sem similaridade aos bancos de dados acima citados foram comparadas a sequências de DNA genômico. Sequências similares a DNA genômico (5% das sequências geradas) podem estar relacionadas a genes raros, cujas sequências expressas ainda não foram identificadas, e eventuais contaminações com DNA genômico nas preparações de mRNA.

Um percentual comparável ao número de ESTs com similaridade a proteínas humanas foi obtido para sequências sem similaridade aos bancos de dados (25% do número total gerado) dentro do presente trabalho. Em virtude do grande número de sequências de cDNA e DNA humanas já conhecidas e a gradativa dificuldade de se obter sequências humanas expressas sem similaridade a ESTs ou clusters já depositados, foi realizada uma avaliação do potencial codificador das referidas sequências *no matches* frente a programas computacionais apropriados (programa ESTScan, Iseli *et al.*, 1999). Do total de sequências (1509 ESTs), 31,5% apresentou a capacidade de apresentar porções nucleotídicas referentes a ORFs (*open reading frames*). Apesar deste percentual ser aparentemente baixo, análise semelhante foi realizada para uma amostragem significativa de sequências expressas não-redundantes similares a genes humanos já conhecidos (793 ESTs), das quais 68% exibem regiões codificadoras identificáveis pelo programa ESTScan. Desta maneira, estima-se que metade do total de sequências *no matches* geradas neste trabalho correspondem a genes humanos expressos. Como um número considerável de sequências expressas (2055 ESTs, 34% do total produzido) foram geradas a partir de preparações de mRNA sem prévio tratamento com DNase (RBT002 e RBT006), pode-se inferir que muitas sequências expressas não possuem similaridade aos bancos de dados pois se referem a sequências genômicas ainda não descritas e/ou disponibilizadas no período da análise.

5.5. Normalização parcial das mini-bibliotecas de cDNA geradas

Em virtude da heterogeneidade nos níveis de expressão gênica em uma célula, tem surgido técnicas apropriadas que permitem equilibrar parcialmente a representatividade dos respectivos transcritos. A idéia original básica é utilizar a hibridização em fase líquida para reduzir a representatividade dos transcritos mais abundantes e consequente aumento na concentração dos transcritos mais raros. A população dos diversos transcritos é equilibrada e usada para a construção de bibliotecas normalizadas com a hibridização e subtração dos clones mais abundantes.

Bonaldo e colaboradores (1996) descreveram processos de normalização e subtração que permitiram acelerar os processos de descoberta de genes dos projetos genoma. As bibliotecas são construídas com cerca de 1 µg 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 dos genes mais abundantes, e o aumento da frequência dos clones mais raros. Para isto os plasmídeos são convertidos em fita simples e hibridizados 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 hibridização é feita uma cromatografia em coluna de hidroxapatita, onde os ácidos nucleicos que permaneceram como fita simples são separados dos ácidos nucleicos 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 uma biblioteca normalizada. Infelizmente, a construção de bibliotecas de cDNA normalizadas requer a disponibilidade de, pelo menos, um micrograma de mRNA, um fator limitante em situações onde a quantidade de material biológico é pequena. Este fator, associado à complexidade metodológica, faz com que o uso desta técnica seja restrito.

A eficiência de processos de normalização pode ser exemplificada nos resultados de Lamerdin e colaboradores (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, apenas dois clones continham RNA ribossomal, dois continham DNA mitocondrial, um representava contaminação com *E. coli* e apenas um correspondia a 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 frequência de ESTs (79,5%) que se mostraram distintas de genes humanos já seqüenciados, demonstrando o universo de genes analisados com a redução dos genes mais abundantes.

A utilização da técnica de RAP-PCR para a geração de seqüências expressas é uma alternativa favorável para o presente estudo, visto que (i) existem limitações na dimensão do modelo experimental em estudo (carcinoma de mama) e, conseqüentemente, na quantidade de mRNA a ser purificado e (ii) permite uma parcial normalização em relação ao nível de expressão gênica, favorecendo a amplificação de cDNAs mais raros (Mathieu-Daudé *et al.*, 1996). No entanto, para avaliarmos a efetiva capacidade de normalização desta técnica, foram analisadas as

bibliotecas de cDNA geradas por RAP-PCR frente bibliotecas normalizadas e não-normalizadas de câncer de mama já disponibilizadas nos bancos de dados (NCI CGAP Br2 e NCI CGAP Br1.1, respectivamente). Esta comparação foi baseada na análise de similaridade a ESTs já depositadas que compõem *clusters* do Unigene. Quanto maior este *cluster*, maior o número de sequências depositadas e, como consequência, mais abundantemente expresso é este gene. Conforme resultado obtido, o maior percentual de sequências expressas obtidas se referem a *clusters* de pequeno tamanho (cerca de 25% das sequências), ou seja, constituídos por um baixo número de ESTs e provavelmente relacionados a genes pouco expressos. Este valor é superior àquele obtido com bibliotecas normalizadas (18%) e não-normalizadas (10%), evidenciando a capacidade desta metodologia em obter informações a respeito de genes pouco expressos e/ou raros. Uma justificativa adicional para esta característica metodológica se refere a complexidade gênica das sub-populações de mRNA, onde a amplificação em baixa estringência é dependente da sequência transcrita e não de sua abundância. Como existe uma alta representatividade de transcritos raros (15 mil a 20 mil genes, equivalentes a 40 a 45% do mRNA total) em células somáticas, a amplificação se torna menos redundante.

Além destas características metodológicas, perfis de cDNA gerados por RAP-PCR compostos por bandeamentos pouco complexos (número nitidamente contável de bandas) evidenciavam uma maior redundância entre os *amplicons* gerados. Sendo assim, a maioria dos perfis de cDNA com estas características (tipicamente observado em perfis X-Xn) não foram utilizados para a geração de ESTs.

5.6. Análise posicional e redundância global das sequências expressas geradas

A amplificação, por RT-PCR, de qualquer porção correspondente a um dado transcrito requer o anelamento do(s) *primer(s)* em ambos os lados da porção de interesse. A chance de cada um destes eventos ocorrer é proporcional ao tamanho das sequências em ambos os lados da referida porção. Ou seja, se o tamanho de um transcrito é estimado como 1 e a distância da extremidade 3' do cDNA à porção a ser amplificada é referida como S, então a probabilidade de interação do(s) *primer(s)* a ambos os lados do cDNA-molde será S e (1-S). Consequentemente, a probabilidade de amplificação desta porção de cDNA hipotética será S.(1-S). Utilizando-se *primers* arbitrários, a chance de se promover a amplificação de porções centrais dos genes é de

0,5 X 0,5 (25%), enquanto que a amplificação de porções a 1/10 de distância de qualquer extremidade dos referidos transcritos é de 0,1 X 0,9 (9%). A partir destes conceitos, foi obtida uma curva teórica simétrica ao redor da porção transcrita central. As condições de baixa estringência características da técnica de RAP-PCR permitiram a obtenção de uma curva experimental (gerada a partir de ESTs não-redundantes com similaridade a genes humanos totalmente sequenciados) muito semelhante à curva teórica, comprovando que existe uma maior probabilidade de geração de sequências expressas correspondentes a porções mais centrais dos genes. Em comparação às metodologias convencionais de geração de ESTs, a partir da extremidades 5' e/ou 3' não-traduzíveis, as sequências expressas geradas pela presente metodologia são muito mais informativas no processo de descoberta de novos genes expressos pois abrangem regiões gênicas codificadoras. Em virtude deste fato, as ESTs geradas neste trabalho foram denominadas ORESTES (*Open Reading Frame ESTs*).

A estimativa global de redundância das sequências expressas, geradas neste trabalho, é de 29,4% (item 4.9). Na inovadora análise de 83 milhões de nucleotídeos expressos, Adams e colaboradores (1995) obtiveram 292.878 ESTs convencionais, distribuídas em 58.384 *singletons* e 29.599 THCs. Considerando-se que sequências expressas com significativa similaridade a um mesmo gene são redundantes entre si, podemos proceder ao mesmo cálculo de redundância para as ESTs geradas por Adams e colaboradores (1995). Este valor de redundância equivale a 70%, derivado do percentual complementar ao número total de *singletons* e THCs dividido pelo montante global de ESTs geradas no referido trabalho ($1 - 87.983 / 292.878$). Comparando este resultado àquele obtido na análise de redundância das ORESTES, pode-se observar uma queda superior a duas vezes do valor de redundância entre as sequências expressas, evidenciando novamente a eficiência desta metodologia na normalização parcial de sub-populações de mRNA.

5.7. Utilização prática das sequências expressas geradas na caracterização de novos genes humanos

Em consequência da grande competição no ramo da Genômica, relacionado ao PGH, pela identificação e caracterização de novos genes humanos, além da necessidade de avaliarmos a utilização prática das ESTs geradas neste projeto (ORESTES), procedemos à caracterização de (i) uma nova isoforma do gene *NABC1* (denominada *NABC1_5B*), gerado a partir de um processamento de *splicing* alternativo (Correa *et al.*, 2000) e (ii) dois novos genes humanos, denominados Semaforina 6B e Notch 2, supostamente relacionados com o processo de embriogênese e desenvolvimento celular (Correa *et al.*, 2001).

Estes genes foram inicialmente identificados por análise de similaridade das ESTs geradas a genes já descritos em diferentes organismos (vide item 4.12).

5.7.1. *NABC1_5B* - uma nova isoforma do marcador tumoral *NABC1*

Vários trabalhos tem mostrado que uma determinada região do cromossomo 20 (20q13) sofre amplificação gênica em uma grande variedade de tumores e ainda, está fortemente associada com a agressividade tumoral (Kallioniemi *et al.*, 1994; Iwabuchi *et al.*, 1995; Schlegel *et al.*, 1995; Collins *et al.*, 1998). Aumento no número de cópias gênicas da região 20q13 tem sido evidenciado em 40% das linhagens celulares de tumor de mama e 18% dos tumores primários de mama (Kallioniemi *et al.*, 1994). Recentemente, Collins e colaboradores (1998) identificaram dois novos genes nesta região, denominados *ZNF217* e *NABC1*, os quais são amplificados e superexpressos em várias linhagens celulares e tumores mamários.

O gene *NABC1* (*Novel Amplified sequence in Breast Carcinoma 1*) origina um transcrito de cerca de 3.0 Kb. Sequenciamento completo do cDNA revela uma ORF de 1752 bp, a qual codifica uma proteína de 584 aminoácidos sem similaridade significativa a qualquer outra proteína conhecida.

Análise mais detalhada da ORESTES PM-BT057-308 (item 4.12) mostrou uma alta similaridade (*Score* = 112 bits, *E value* = 1e-25) ao gene *NABC1*. Entretanto, esta sequência expressa contém uma porção interna de 135 bp com potencial codificador mas sem similaridade ao transcrito, evidenciando a existência de um novo exon no gene (fig. 16A). A partir desta

informação, foi confirmada a existência desta nova isoforma por RT-PCR e, a partir do alinhamento de clones de BAC contendo grandes porções do cromossomo 20, foi possível obter a organização genômica de *NABC1* (fig. 16B).

Ensaio quantitativo de RT-PCR utilizando amostras tumorais e normais de mama confirmam uma superexpressão da isoforma conhecida *NABC1* (fig. 17C). A nova isoforma *NABC1* foi fracamente identificada em apenas uma das amostras tumorais analisadas. Por outro lado, ensaios quantitativos realizados com amostras tumorais e normais de cólon mostraram uma baixa expressão de ambas isoformas (fig. 17D).

Apesar da amostragem utilizada para os ensaios quantitativos de RT-PCR ser pequena ($n=4$), os resultados obtidos indicam que, por um lado, a superexpressão de *NABC1* em tumores de mama revela uma possível ação oncogênica mas, por outro lado, o mesmo não ocorre em tumores de cólon, onde uma baixa expressão poderia indicar que este gene teria uma ação supressora de tumor. No entanto, estudos mais aprofundados, como o número de cópias gênicas em amostras tumorais e normais de cólon e mama (*comparative genomic hybridization*), precisam ser realizados para uma possível conclusão a respeito da função gênica, visto que a região em que este gene está localizado é relativamente instável e possui um complexo padrão de rearranjo.

5.7.2. *Semaforina 6B (Sema 6B) - um novo gene humano da família de semaforinas*

As semaforinas correspondem a uma família gênica muito conservada entre as espécies, também conhecidas como moléculas de orientação axonal. Elas são expressas em diferentes regiões do sistema nervoso durante o desenvolvimento e possuem um importante papel no estabelecimento de projeções axonais. A primeira semaforina identificada, Semaforina III (também conhecida como Colapsina-1 ou SemaD) causa colapso de cones de crescimento, retração axonal e repulsão em axônios sensoriais ou simpáticos em cultura (para revisão vide Yu & Kolodkin, 1999).

As semaforinas são divididas em 7 classes, compostas por semaforinas secretadas e transmembrânicas.

Eckhardt e colaboradores (1997) identificaram um cDNA de semaforina murina denominada *SemaVIIb*. Essa molécula contém um domínio semaforina (*Sema domain*)

extracelular e não possui o domínio imunoglobulina ou repetições de trombospondina que estão normalmente presentes em outras semaforinas transmembrânicas de vertebrados. Durante o desenvolvimento, *SemaVIIb* é expressa em sub-regiões do sistema nervoso e é particularmente encontrada na musculatura esquelética. No animal adulto, o mRNA de *SemaVIIb* é encontrado ubiquamente. Ensaio de ligação *in vitro* mostraram que SemaVIIb liga especificamente o domínio SH3 do proto-oncogene c-src, sugerindo uma possível função na sinalização intracelular através de uma cascata de transdução relacionada a c-src.

Recentemente, Christensen e colaboradores (1998) identificaram um novo membro da família de semaforinas em linhagens celulares metastáticas de adenocarcinoma mamário murino, através da técnica de *Differential Display* (DD). Essa molécula, denominada M-SemaH, é menos abundante em tecidos normais do que em tumores. Esse trabalho foi o primeiro a relacionar a expressão de um gene da classe de semaforinas com o processo de progressão tumoral. Os resultados obtidos sugerem que M-SemaH poderia agir, de maneira semelhante, nos processos de metástase e desenvolvimento axonal, ou seja, atuando como um “guia” de motilidade.

Aplicações de DD na comparação de linhagens celulares de adenocarcinoma pulmonar humano mostraram que o mRNA de semaforina E se encontrava superexpresso em linhagens metastáticas quando comparadas à linhagens normais (Martín-Satué & Blanco, 1999). Esse resultado reforça a idéia de que as semaforinas tenham algum envolvimento com processos metastáticos.

Análise mais detalhada da ORESTES IL-BT000-002 (item 4.12) mostra uma alta similaridade (*Score* = 115 bits, *E value* = 7e-24) ao gene de semaforina VIIb (*Mus musculus*) e semaforina Z (*Rattus norvegicus*). A partir desta informação, aliada a técnicas de RACE-PCR e similaridade a outras sequências do dbEST, obtivemos a sequência completa de duas isoformas geradas por *splicing* alternativo deste novo gene ortólogo humano denominado Semaforina 6B (*Sema 6B*), correspondente a classe VI de semaforinas humanas.

Estas isoformas, denominadas *Sema 6B_S* (isoforma menor, GenBank AF216389) e *Sema 6B_L* (isoforma maior, GenBank AF293363) correspondem, respectivamente, a prováveis semaforinas secretada e transmembrânica (figuras 19 e 20). Processamento pós-transcricional por *splicing* alternativo em semaforinas de classe VI foi somente observado no gene *Sema Y* (semaforina Y murina) (Kikuchi *et al.*, 1999), resultando em duas isoformas com uma diferença de 96 pb em seus domínios extracelulares, especificamente em uma região de ligação ao receptor

neuropilina-1. A expressão das isoformas de *Sema Y* é aparentemente controlada de uma maneira tecido e estágio-dependente.

Ensaio de *Northern blot* mostram que o mRNA de Semaforina 6B (aparentemente, a isoforma transmembrânica *Sema 6B_L*) é altamente expresso em cérebro humano (fig. 22A). A alta expressão local de *Sema 6B* evidencia a importância deste gene no funcionamento e/ou organização cerebral, o que é característico de sua família gênica.

Para avaliar o possível papel de *Sema 6B* no processo de progressão tumoral, conforme descrito anteriormente com outras semaforinas (Christensen et al., 1998; Martín-Satué & Blanco, 1999), foram quantificados os níveis de transcrito em linhagens celulares de glioblastoma humano, com capacidade invasiva *in vitro*, frente a ação de agentes anti-tumorais glicocorticóides e retinóides. O glioblastoma multiforme é um tumor altamente maligno e refratário a diferentes tipos de terapias, principalmente devido a sua natureza invasiva no sistema nervoso. Hormônios glicocorticóides são largamente utilizados como agentes anti-inflamatórios e anti-tumorais, constituindo a única quimioterapia disponível para o tratamento de gliomas e glioblastomas. Retinóides inibem intensamente a proliferação e migração de glioblastoma multiforme humano em culturas celulares primárias, possibilitando a sua aplicação em ensaios clínicos visando o tratamento de gliomas humanos malignos (Bouterfa *et al.*, 2000).

O tratamento com ácido retinóico, por períodos de tempo crescentes, em linhagens celulares de glioblastoma humano promove uma inibição progressiva na expressão de *Sema 6B* em ambas linhagens (fig. 22B). Este resultado pode indicar um possível papel oncogênico de *Sema 6B* na progressão de tumores cerebrais humanos.

No intuito de verificar se a expressão aumentada de *Sema 6B* ocorre igualmente em outros processos tumorigênicos, conforme observado para *M-SemaH* (Martín-Satué & Blanco, 1999), foram realizados ensaios de *Northern blot* com amostras de RNA total isoladas de tecidos tumorais e normais de mama, correspondente ao modelo experimental da qual foi gerada a EST inicial (câncer de mama). Curiosamente, três das quatro amostras tumorais analisadas, assim como linhagens celulares metastáticas de câncer de mama, apresentaram uma menor expressão de *Sema 6B* quando respectivamente comparadas a tecidos normais de mama e linhagens não-metastáticas. Apesar da amostragem analisada ter sido limitante, é possível concluir que o papel de *Sema 6B* na progressão tumoral é tecido-dependente, tendo uma importância aparentemente diferenciada na evolução de tumores cerebrais e mamários. O efetivo papel deste gene e suas

diferentes isoformas no desenvolvimento do sistema nervoso e na progressão tumoral tecido-específica necessita ser melhor avaliado para corroborar os resultados aqui obtidos.

5.7.3. *Notch 2 - um novo gene humano com provável envolvimento na diferenciação e destino celular*

A cascata de transdução envolvida com genes da família Notch regula um enorme complexo de interações celulares durante o desenvolvimento animal, como o controle do crescimento, diferenciação e formação de padrões teciduais (Kimble *et al.*, 1998). Em *D. melanogaster* e vertebrados, esta via controla a decisão de destino celular entre células epidermais e neurônios, assim como uma série de outras decisões tecido-específicas.

Por outro lado, disfunções nesta via de sinalização tem sido identificadas em vários distúrbios humanos, incluindo leucemia, carcinoma cervical e de cólon, câncer de mama, mal de Alzheimer, derrame e demência (Ellisen *et al.*, 1991; Robbins *et al.*, 1992; Joutel *et al.*, 1996; Gridley, 1997).

Modelos murinos para estudos de tumorigênese mamária como resultado da ação do vírus MMTV (*mouse mammary tumor virus*) tem mostrado que genes da família *FGF* (*fibroblast growth factor*), *wnt* e *NOTCH* são utilizados como sítios de integração proviral, ocorrendo mutagênese insercional nos respectivos genes (Callahan, 1996; Tekmal & Keshava, 1997). Camundongos transgênicos, contendo porções LTR (*long terminal repeats*) de MMTV flanqueadas pela sequência gênica de *int-3/notch*, tem demonstrado que um aumento de expressão do referido locus mutado pode causar alterações no desenvolvimento mamário normal e hiperproliferação do epitélio glandular (Jhappan *et al.*, 1992).

Larsson e colaboradores (1994) identificaram, por hibridização *in situ* e mapeamento genético, os *loci* dos genes humanos *NOTCH1*, 2 e 3, localizados nas posições 9q34, 1p13-p11 e 19p13.2-p13.1, respectivamente. Estas posições se referem a regiões de translocação associada a diferentes tipos de neoplasias. Como exemplo, a região cromossômica 1p13-p11 sofre mutações na síndrome mielodisplásica e policitemia, além de eventuais translocações em quadros de leucemia mieloblástica aguda (Cuticchia & Pearson, 1993).

Análise mais detalhada da ORESTES IL-BT000-001 (item 4.12) mostra uma alta similaridade (Score = 178; E value = 7e-45) ao gene murino *Notch 2* (*Rattus norvegicus*)

(Weinmaster *et al.*, 1992). A partir desta informação, aliada a técnicas de RACE-PCR e similaridade a outras sequências do dbEST, obtivemos a sequência completa do respectivo cDNA (tamanho estimado de 11.189 bp), o qual se refere ao novo gene humano *Notch 2* (GenBank AF315356). Em virtude do sequenciamento incompleto do cromossomo 1 (grande quantidade de *gaps* nos clones genômicos disponibilizados nos bancos de dados) pelo PGH, não foi possível descrever a organização do locus gênico de *Notch 2*.

Este gene apresenta uma maior expressão em placenta, pulmão e leucócitos (fig. 27A). A alta expressão no tecido placentário poderia ser o resultado da importância desta família gênica durante os processos de decisão do destino celular na embriogênese. Em paralelo, a expressão de *Notch 2* em leucócitos periféricos do sangue evidencia a sua importância neste tipo celular, uma vez que eventos mutacionais em genes da família *Notch* estão correlacionados à evolução do quadro patológico de leucemia (Ellisen *et al.*, 1991; Cuticchia & Pearson, 1993; Larsson *et al.*, 1994).

Para avaliar se a expressão do gene humano *Notch 2* era aumentada em tumores mamários, conforme observado com o gene murino *int-3* na transformação do epitélio mamário (Robbins *et al.*, 1992; Jhappan *et al.*, 1992), foram realizados ensaios quantitativos de RT-PCR a partir de RNA total isolado de tumores mamários humanos em estádios I e II e respectivos tecidos normais. Todas as amostras de RNA isoladas de tumores com estadiamento mais avançado, com exceção da amostra de carcinoma mamário de origem masculina (fig. 27B, canaleta 5), evidenciaram uma expressão levemente aumentada do transcrito *Notch 2* quando comparadas às amostras provenientes de tumores de estadiamento mais inicial. Este resultado não foi confirmado na análise de linhagens celulares tumorais e normais, apesar do fato que as referidas linhagens não foram estudadas individualmente, mas agrupadas em *pools*. É provável que o padrão de expressão gênica de uma ou mais linhagens normais e/ou tumorais, presentes nos respectivos *pools*, apresente certas peculiaridades que acabem prejudicando a comparação entre linhagens tumorais e normais. Apesar deste resultado ser preliminar, necessitando de uma maior amostragem de tumores em diferentes estadiamentos (inclusive estadio III e IV), indica que a expressão aumentada de *Notch 2* acompanha a progressão tumoral do tecido mamário.

Utilizando o cDNA completo do gene *Notch 2* humano para a análise posicional de sequências expressas com alta similaridade a este gene, disponibilizadas no dbEST, podemos verificar novamente que as ORESTES se alinham às porções mais centrais do gene com uma

frequência maior do que as ESTs geradas por outros projetos independentes (CGAP e MGC) (fig. 28). Este dado reforça a idéia de que as sequências expressas, geradas no presente trabalho, são altamente informativas pois correspondem, na sua maioria, a regiões codificadoras de genes humanos expressos.

Figura 28: Alinhamento posicional de sequências expressas disponibilizadas no banco de dados dbEST (GenBank) ao longo do cDNA completo referente ao gene *Notch 2* humano. Pode-se observar o alinhamento exclusivo de ORESTES na porção gênica mais central.

5.8. Conclusões finais

Utilizando o carcinoma de mama como modelo de estudo, foi realizada a geração, sequenciamento e anotação de 6.029 ESTs, equivalente a 1,9 milhões de nucleotídeos. Foi estabelecida uma rotina experimental que ofereceu um maior rigor na obtenção de sequências informativas para a proposta de descoberta gênica. As diversas preparações de mRNA utilizadas foram gradualmente melhoradas quanto a ausência de contaminantes (DNA genômico e mitocondrial, RNA ribossômico), as bibliotecas de cDNA geradas foram minuciosamente validadas de modo a manter uma baixa ocorrência de elementos repetitivos (SINEs e LINEs) e o método de geração de perfis de cDNA foi avaliado em diferentes situações, onde tamanhos aceitáveis de ESTs (acima de 200 bp) e menor taxa de redundância se tornaram mais convenientes para a informatividade gênica. O total global de mini-bibliotecas produzidas no presente trabalho compuseram cerca de 90% das mini-bibliotecas de cDNA utilizadas na conclusão e consequente validação do projeto-piloto do Genoma Humano do Câncer, proposto no final de 1998.

Em virtude dos resultados promissores na obtenção de mini-bibliotecas de cDNA parcialmente normalizadas e a geração de sequências expressas referentes a porções codificadoras de genes humanos expressos, foi proposto, no início do ano de 1999, o Projeto Genoma Humano do Câncer (HCGP), com o objetivo de gerar, pela presente metodologia, um número estimado de 1 milhão de ESTs, de modo a abranger o transcriptoma global das células humanas.

Os dados gerados pelo HCGP, finalizado em 22 de dezembro de 2000, tem auxiliado no processo de descoberta gênica como, por exemplo, a anotação de genes não previamente preditos no cromossomo 22 (De Souza *et al.*, 2000), além da criação de projetos paralelos com o intuito de obter sequências completas de cDNA, devida análise funcional e correlação com diferentes quadros patológicos por análises de *microarrays*.

Conforme mostrado no presente trabalho, cada sequência expressa pode ser suficientemente informativa para o processo de caracterização gênica. Uma grande quantidade de ESTs descritas neste trabalho merecem atenção mais continuada, de modo a auxiliar no entendimento bioquímico e fisiológico do organismo a partir dos diversos genes que compõem o genoma humano.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., Venter, J.C. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R.F., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., Venter, J.C. (1992). Sequence identification of 2375 human brain genes. *Nature* **355**, 632-634.

Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C., Venter, J.C. (1993). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4**, 373-380.

Adams, M.D., Kerlavage, A.R., Fleischman, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., Sutton, G., Blake, J.A., Brandon, R.C., Chiu, M.W., Clayton, R.A., Cline, R.T., Cotton, M.D., Hughes, J.E., Fine, L.D., FitzGerald, L.M., FitzHugh, W.M., Fritchman, J.L., Geoghagen, N.S.M., Glodek, A., Gnehm, C.L., Hanna, M.C., Hedblom, E., Hinkle Jr., P.S., Kelley, J.M., Klimek, K.M., Kelley, J.C., Liu, L.I., Marmaros, S.M., Merrick, J.M., Palanques, R.F.M., McDonald, L.A., Nguyen, D.T., Pellegrino, S.M., Phillips, C.A., Ryder, S.E., Scott, J.L., Saudek, D.M., Shirley, R., Small, K.V., Spriggs, T.A., Utterback, T.R., Weidman, J.F., Li, Y., Barthlow, R., Bednarik, D.P., Cao, L., Cepeda, M.A., Coleman, T.A., Collins, E.J., Dimke, D., Feng, P., Ferrie, A., Fischer, C., Hastings, G.A., He, W.W., Hu, J.S., Huddleston, K.A., Greene, J.M., Gruber, J., Hudson, P., Kim, A., Kozak, D.L., Kunsch, C., Ji, H., Li, H., Yu, G.L., Ruben, S.M., Dillon, P.J., Fannon, M.R., Rosen, C.A., Haseltine, W.A., Fields, C., Fraser, C.M., Venter, J.C. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377** (suppl.), 3-174.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor Miklos, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacle, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter, J.C. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-95.

American Joint Committee on Cancer (1993). *Handbook for staging of cancer*. J. B. Lippincott Company, Philadelphia, USA.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L. (2000). GenBank. *Nucleic Acids Res.* **28**, 15-18.

Bentley, D.R. & Dunham, I. (1995). Mapping human chromosomes. *Curr. Opinion Genet. Dev.* **5**, 328-334.

Bishop, J.O., Morton, J.G., Rosbasch, M., Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* **250**, 199-204.

Bodmer, W. & McKie, R. (1994). *The Book of Man: the Human Genome Project and the quest to discover our genetic heritage*. Oxford University Press, New York, USA.

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

Bouterfa, H., Picht, T., Kess, D., Herbold, C., Noll, E., Black, P.M., Roosen, K., Tonn, J.C. (2000). Retinoids inhibit human glioma cell proliferation and migration in primary cell cultures but not in established cell lines. *Neurosurgery* **46**, 419-430.

Britten, R.J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. USA* **93**, 9374-9377.

Brown, T.A. (1999). *Genética: um enfoque molecular*. Guanabara Koogan, Rio de Janeiro, Brasil.

Callahan, R. (1996). MMTV-induced mutations in mouse mammary tumors: their potential relevance to human breast cancer. *Breast Cancer Res Treat.* **39**, 33-44.

Christensen, C.R.L., Klingelhöfer, J., Tarabykina, S., Hulgaard, E.F., Kramerov, D., Lukanidin, E. (1998). Transcription of a novel mouse semaphorin gene, M-semaH, correlates with the metastatic ability of mouse tumor cell lines. *Cancer Res* **58**, 1238-44.

Church, G.M., Gilbert, W. (1984). Genomic sequencing. *Proc Natl Acad Sci USA*. **81**, 1991-1995.

Collins, C., Rommens, J.M., Kowbel, D., Godfrey, T., Tanner, M., Hwang, S.I., Polikoff, D., Nonet, G., Cochran, J., Myambo, K., Jay, K.E., Froula, J., Cloutier, T., Kuo, W.L., Yaswen, P., Dairkee, S., Giovanola, J., Hutchinson, G.B., Isola, J., Kallioniemi, O.P., Palazzolo, M., Martin, C., Ericsson, C., Pinkel, D., Gray, J.W. (1998). Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. *Proc Natl Acad Sci USA* **95**, 8703-8.

Collins, F. & Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* **262**, 43-46.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., *et al.* (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**, 682-689.

Cooke, R., Raynal, M., Laudie, M., Grellet, F., Delseny, M., Morris, P.C., Guerrier, D., Giraudat, J., Quigley, F., Clabault, G., Li, Y.F., Mache, R., Krivitzky, M., Gy, IJ., Kreis, M., Lechamy, A., Parmentier, Y., Marbach, J., Fleck, J., Clement, B., Philipps, G., Herve, C., Bardet, C., Tremousaygue, D., Hofte, H. *et al.* (1996). Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non-redundant ESTs. *Plant J*. **9**, 101-124.

Cooper, G.M. (1997). *The Cell: a molecular approach*. ASM Press. Washington, D.C., USA.

Correa, R.G., de Carvalho, A.F., Pinheiro, N.A., Simpson, A.J., de Souza, S.J. (2000). NABC1 (BCAS1): alternative splicing and downregulation in colorectal tumors. *Genomics* **65**, 299-302.

Correa, R.G., Sasahara, R.M., Bengtson, M.H., Katayama, M.L.H., Salim, A.C.M., Mitzi Brentani, M., Sogayar, M.C., de Souza, S.J., Simpson, A.J.G. (2001). Human semaphorin 6B [(HSA)SEMA6B], a novel human class 6 semaphorin gene – alternative splicing and all-trans-retinoic acid- dependent downregulation in glioblastoma cell lines. *Genomics* (in press).

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., Saurin, W., Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genet.* **25**, 235-238.

Cuticchia, A.J. & Pearson, P.L. (1993). *Human Gene Mapping*. Johns Hopkins University Press, Baltimore, MD, USA.

Davidson, E.H. & Britten, R.J. (1979). Regulation of gene expression: possible role of repetitive sequences. *Science* **204**, 1052-1059.

De Souza, S.J., Camargo, A.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A., Andrade, L.E., Carrer, H., El-Dorry, H.F., Espreafico, E.M., Habr-Gama, A., Giannella-Neto, D., Goldman, G.H., Gruber, A., Hackel, C., Kimura, E.T., Maciel, R.M., Marie, S.K., Martins, E.A., Nobrega, M.P., Paco-Larson, M.L., Pardini, M.I., Pereira, G.G., Pesquero, J.B., Rodrigues, V. (2000). Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc Natl Acad Sci USA* **97**, 12690-12693.

Dias Neto, E., Harrop, R., Oliveira, R.C., Pena, S.D.J., Wilson, R.A., Simpson, A.J.G. (1996a). The Schistosome genome project: RNA arbitrarily primed PCR allows the accelerated generation of expressed sequence tags. *Mem. Inst. Oswaldo Cruz* **91**, 655-657.

Dias Neto, E., Rahal, P., Moura, R.P., Caballero, O., Villa, L.L., Simpson, A.J.G. (1996b). DNA ligation quality control by competitive PCR. *Trends Genet.* **12** (9), 341-342.

Dias Neto, E., Harrop, R., Oliveira, R.C., Wilson, R.A., Pena, S.D.J., Simpson, A.J.G. (1997). Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the efficient generation of ESTs from nanogram quantities of mRNA. *Gene* **186**, 135-142.

Dias Neto, E. (1997). Desenvolvimento de metodologias de PCR em baixas condições de estringência para a análise genotípica: “*Shistosoma mansoni*” como modelo. *Tese de Doutorado*, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Brasil.

Dias Neto, E., Garcia Correa, R., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva, W. Jr., Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., Carvalho, A.F., Matsukuma, A., Baia, G.S., Simpson, D.H., Brunstein, A., de Oliveira, P.S., Bucher, P., Jongeneel, C.V., O'Hare, M.J., Soares, F., Brentani, R.R., Reis, L.F., de Souza, S.J., Simpson, A.J. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci USA*. **97**, 3491-6.

Dickson, R.B. & Lippman, M.E. (1995). Molecular basis of breast cancer. In: *The molecular basis of cancer*, pp.358-384, W.B. Saunders Company, Philadelphia, USA.

Dunham, I., Shimizu, N., Roe, B.A., Chisoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., O'Brien, K.P. *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402**, 489-95.

Eckhardt, F., Behar, O., Calautti, E., Yonezawa, K., Nishimoto, I., Fishman, M.C. (1997). A novel transmembrane semaphorin can bind c-src. *Mol Cell Neurosci* **9**, 409-19.

Ellisen, L.W., Bird, J., West, D.C., Soreng, A.L., Reynolds, T.C., Smith, S.D., Sklar, J. (1991). TAN-1, the human homolog of the *Drosophila* notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell*. **66**, 649-661.

Ewing, B. & Green, P. (2000). Analysis of expressed sequence tags indicates 35.000 human genes. *Nature Genet.* **25**, 232-234.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool.* **19**, 99-113.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Franco, G.R., Adams, M.D., Soares, M.B., Simpson, A.J.G., Venter, J.C., Pena, S.D.J. (1995). Sequencing and identification of expressed *Schistosoma mansoni* genes by random selection of cDNA clones from a directional library. *Gene* **152**, 141-147.

Franco, G.R., Rabelo, E.M.L., Azevedo, V., Pena, H.B., Ortega, J.M., Santos, T.M., Meira, W.S.F., Rodrigues, N.A., Dias, C.M.M., Harrop, R., Wilson, A., Saber, M., Abdel-Hamid, H., Faria, M.S.C., Margutti, M.E.B., Parra, J.C., Pena, S.D.J. (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res.* **4**, 231-240.

Frohman, M.A. (1993). Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: thermal RACE. *Methods Enzymol.* **218**, 340-356.

Frohman, M.A., Dush, M.K., Martin, G.R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85**, 8998-9002.

Gardiner, K. (1996). Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet.* **12**, 519-524.

Green, P. (1997). Against a whole-genome shotgun. *Genome Res.* **7**, 410-417.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996). Life with 6000 genes. *Science* **274**, 563-567.

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

Gridley, T. (1997). Notch signaling in vertebrate development and disease. *Mol Cell Neurosci* **9**, 103-108.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M., Schudy, A. (2000). The DNA sequence of human chromosome 21. *Nature* **405**, 311-9.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfsing, T., Schellenberg, K., Soares, M.B., Tan, F., Meg, J.T., Trevaskis, E., Underwood, K., Wohldman, P., Waterston, R., Wilson, R., Marra, M. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6**, 807-828.

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

Iseli, C., Jongeneel, C.V., Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Ismb*. 138-148.

Iwabuchi, H., Sakamoto, M., Sakunaga, H., Ma, Y.Y., Carcangiu, M.L., Pinkel, D., Yang-Feng, T.L., Gray, J.W. (1995). Genetic analysis of benign, low-grade, and high-grade ovarian tumors. *Cancer Res* **55**, 6172-80.

Jhappan, C., Gallahan, D., Stahle, C., Chu, E., Smith, G.H., Merlino, G., Callahan, R. (1992). Expression of an activated Notch-related int-3 transgene interferes with cell differentiation and induces neoplastic transformation in mammary and salivary glands. *Genes Dev.* **6**, 345-355.

Ji, H., Liu, Y.E., Jia, T., Wang, M., Liu, J., Xiao, G., Joseph, B.K., Rosen, C., Shi, Y.E. (1997). Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. *Cancer Res.* **57**, 759-764.

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., Alamowitch, S., Domenga, V., Cecillion, M., Marechal, E., Maciazek, J., Vayssiere, C., Cruaud, C., Cabanis, E.A., Ruchoux, M.M., Weissenbach, J., Bach, J.F., Bousser, M.G., Tournier-Lasserre, E. (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* **383**, 707-710.

Kallioniemi, A., Kallioniemi, O.P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H.S., Pinkel, D., Gray, J.W., Waldman, F.M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci USA* **91**, 2156-60.

Keyomarsi, K. & Pardee, A.B. (1993). Redundant cyclin overexpression and gene amplification in breast cancer cells. *Proc. Natl. Acad. Sci. USA* **90**, 1112-1116.

Kikuchi, K., Chedotal, A., Hanafusa, H., Ujimasa, Y., de Castro, F., Goodman, C.S., Kimura, T. (1999). Cloning and characterization of a novel class VI semaphorin, semaphorin Y. *Mol Cell Neurosci* **13**, 9-23.

Kimble, J., Henderson, S., Crittenden, S. (1998). Notch/LIN-12 signaling: transduction by regulated protein slicing. *Trends Biochem Sci.* **23**, 353-357.

Kirkness, E.F. & Venter, J.C. (1995). Receptor cloning: high-throughput sequencing of cDNA tags for identification of novel genes. *Methods in Neuroscience* **25**, 126-141.

Lambais, M.R., Goldman, M.H., Camargo, L.E., Goldman, G.H. (2000). A genomic approach to the understanding of *Xylella fastidiosa* pathogenicity. *Curr Opin Microbiol.* **3**, 459-462.

Lamerdin, J.E., Athwal, R.S., Kansara, M.S., Sandhu, A.K., Patanjali, S.R., Weissman, S.M., Carrano, A.V. (1995). Chromosomal localization and expressed sequence tag generation of clones from a normalized human adult thymus cDNA library. *Genome Res.* **5**, 359-367.

Larsson, C., Lardelli, M., White, I., Lendahl, U. (1994). The human NOTCH1, 2, and 3 genes are located at chromosome positions 9q34, 1p13-p11, and 19p13.2-p13.1 in regions of neoplasia-associated translocation. *Genomics* **24**, 253-258.

Lee, S.W., Tomasetto, C., Swisshelm, K., Keyomarsi, K., Sager, R. (1992). Down-regulation of a member of the S100 gene family in mammary carcinoma cells and reexpression by azadeoxycytidine treatment. *Proc. Natl. Acad. Sci. USA* **89**, 2504-2508.

Lewin, B. (1997). *Genes VI*. Oxford University Press, New York, USA.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J. (2000). Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239-240.

Marra, M.A., Hillier, L., Waterston, R.H. (1998). Expressed sequence tags - establishing bridges between genomes. *Trends Genet.* **14**, 4-7.

Martín-Satué, M. & Blanco, J. (1999). Identification of semaphorin E gene expression in metastatic human lung adenocarcinoma cells by mRNA differential display. *J Surg Oncol* **72**, 18-23.

Mazzarella, R. & Schlessinger, D. (1997). Duplication and distribution of repetitive elements and non-unique regions in the human genome. *Gene* **205**, 29-38.

McKenzie, K.E. (1997). Breast cancer: molecular epidemiology. In: *Encyclopedia of cancer*, vol.I, pp.173-183, Academic Press, San Diego, California, USA.

Mathieu-Daudé, F., Welsh, J., Vogt, T., McClelland, M. (1996). DNA rehybridization during PCR: the 'C₀t effect' and its consequences. *Nucleic Acids Res.* **24**, 2080-2086.

Mathieu-Daudé, F., Welsh, J., Davis, C., McClelland, M. (1998). Differentially expressed genes in the Trypanosoma brucei life cycle identified by RNA fingerprinting. *Mol. Biochem. Parasitol.* **92**, 15-28.

McClelland, M., Mathieu-Daudé, F., Welsh, J. (1995). RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends Genet.* **11**, 242-246.

McClelland, M. & Welsh, J. (1995). RNA fingerprinting using arbitrarily primed PCR. In: *PCR Primer - a laboratory manual*, pp. 213-234, Cold Spring Harbor Laboratory Press, New York, USA.

Medzhitov, R., Hurlburt, P.P., Janeway Jr., C.A. (1997). A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature* **388**, 394-397.

Nakamura, T.M., Morin, G.B., Chapman, K.B., Weinrich, S.L., Andrews, W.H., Lingner, J., Harley, C.B., Cech, T.R. (1997). Telomerase catalytic subunit homologs from fission yeast and human. *Science* **277**, 955-959.

Newman, T., de Bruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M. *et al.* (1994). Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol.* **106**, 1241-1255.

Ozaki, L.S. & Czeko, Y.M.T. (1984). Genomic DNA cloning related techniques. In: *Genes antigens of parasites. A laboratory manual*, pp.165-185, Morel, C.M., UNDP-World Bank-Who and Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

Perucho, M., Welsh, J., Peinado, M.A., Ionov, Y., McClelland, M. (1995). Fingerprinting of DNA and RNA arbitrarily primed polymerase chain reaction: applications in cancer research. *Methods Enzymol.* **254**, 275-291.

Porter-Jordan, K. & Lippman, M.E. (1994). Overview of the biological markers of breast cancer. *Hematol. Oncol. Clin. North Am.* **8**, 73-100.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28**, 141-145.

Ranson, M., Andronicos, N.M., O'Mullane, M.J., Baker, M.S. (1998). Increased plasminogen binding is associated with metastatic breast cancer cells: differential expression of plasminogen binding proteins. *British J. Cancer* **77**, 1586-1597.

Robbins, J., Blondel, B.J., Gallahan, D., Callahan, R. (1992). Mouse mammary tumor gene int-3: a member of the notch gene family transforms mammary epithelial cells. *J Virol.* **66**, 2594-2599.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989). *Molecular Cloning, a laboratory manual*. Cold Spring Harbor Laboratory Press, New York, USA.

Sanger, F., Nicklen, S., Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.

Schlegel, J., Stumm, G., Scherthan, H., Bocker, T., Zirngibl, H., Ruschoff, J., Hofstadter, F. (1995). Comparative genomic in situ hybridization of colon carcinomas with replication error. *Cancer Res* **55**, 6002-5.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tomé, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannikulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J.R., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J.B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matisse, T.C., McKusick, K.B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H.C., Page, D.C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D.K., Soderlund, C., Sun, W.L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M.D., Auffray, C., Walter, N.A.R., Brandon, R., Dehejia, A., Goodfellow, P.N., Houlgatte, R., Hudson Jr., J.R., Ide, S.E., Iorio, K.R., Lee, W.Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M.H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J.C., Sikela, J.M., Beckmann, J.S., Weissenbach, J., Myers, R.M., Cox, D.R., James, M.R., Bentley, D., Deloukas, P., Lander, E.S., Hudson, T.J. (1996). A gene map of the human genome. *Science* **274**, 540-546.

Simpson, A.J.G. & Perez, J.F. (1998). ONSA, the São Paulo Virtual Genomics Institute. *Nat. Biotechnol.* **16**, 795-796.

Simpson, A.J., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R., Alves, L.M., Araya, J.E., Baia, G.S., Baptista, C.S., Barros, M.H., Bonaccorsi, E.D., Bordin, S., Bove, J.M., Briones, M.R., Bueno, M.R., Camargo, A.A., Camargo, L.E., Carraro, D.M., Carrer, H., Colauto, N.B., Colombo, C., Costa, F.F., Costa, M.C., Costa-Neto, C.M., Coutinho, L.L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A.P., Ferreira, A.J., Ferreira, V.C., Ferro, J.A., Fraga, J.S., Franca, S.C., Franco, M.C., Frohme, M., Furlan, L.R., Garnier, M.,

Goldman, G.H., Goldman, M.H., Gomes, S.L., Gruber, A., Ho, P.L., Hoheisel, J.D., Junqueira, M.L., Kemper, E.L., Kitajima, J.P., Krieger, J.E., Kuramae, E.E., Laigret, F., Lambais, M.R., Leite, L.C., Lemos, E.G., Lemos, M.V., Lopes, S.A., Lopes, C.R., Machado, J.A., Machado, M.A., Madeira, A.M., Madeira, H.M., Marino, C.L. The genome sequence of the plant pathogen *Xylella fastidiosa*. (2000). *Nature* **406**, 151-7.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., Collins, F.S. (1999). The Mammalian Gene Collection. *Science* **286**, 455-457.

Tekmal, R.R. & Keshava, N. (1997). Role of MMTV integration locus cellular genes in breast cancer. *Front Biosci.* **2**, 519-526.

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

Tórtola, S., Capellà, G., Marcuello, E., Günther, K., Aiza, G., Masramon, L., Reymond, M.A., Peinado, M.A. (1998). Analysis of differential gene expression in human colorectal tumor tissues by arbitrarily primed-PCR: a technical assessment. *Lab. Invest.* **78**, 309-317.

Touchman, J.W., Bouffard, G.G., Weintraub, L.A., Idol, J.R., Wang, L., Robbins, C.M., Nussbaum, J.C., Lovett, M., Green, E.D. (1997). 2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries. *Genome Res.* **7**, 281-292.

Venter, J.C., Smith, H.O., Hood, L. (1996). A new strategy for genome sequencing. *Nature* **381**, 364-366.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R. Green, P. Shownkeen, R., Halloran, N. *et al.* (1992). A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1**, 114-123.

Watson, M.A. & Fleming, T.P. (1994). Isolation of differentially expressed sequence tags from human breast cancer. *Cancer Res.* **54**, 4598-4602.

Weber, J.L. & Myers, E.W. (1997). Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401-409.

Weinmaster, G., Roberts, V.J., Lemke, G. (1992). Notch2: a second mammalian Notch gene. *Development* **116**, 931-941.

Welsh, J. & McClelland, M. (1991). Genomic fingerprints produce by PCR with consensus tRNA gene primers. *Nucleic Acids Res.* **19**, 861-866.

Welsh, J., Chada, K., Dalal, S., Cheng, R., Ralph, D., McClelland, M. (1992). Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res.* **20**, 4965-4970.

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

Wooster, R. & Stratton, M.R. (1995). Breast cancer susceptibility: a complex disease unravels. *Trends Genet.* **11**, 3-5.

Yu, H.H., Kolodkin, A.L. (1999). Semaphorin signaling: a little less per-plexin. *Neuron* **22**, 11-4.