TARCÍSIO DE SOUZA PERES

Avaliação de Transcritos Diferencialmente Expressos em Neoplasias Humanas com ORESTES

CAMPINAS

2006

TARCÍSIO DE SOUZA PERES

Avaliação de transcritos diferencialmente expressos em neoplasias humanas com ORESTES

Dissertação de Mestrado apresentada à Pós-graduação da Faculdade de Ciências Médicas da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciências Médicas, área de concentração em Ciências Biomédicas.

ORIENTADOR: *PROF.DR. FERNANDO LOPES ALBERTO*

CAMPINAS

2006

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA FACULDADE DE CIÊNCIAS MÉDICAS DA UNICAMP
Bibliotecário: Sandra Lúcia Pereira – CRB-8ª / 6044

Peres, Tarcísio de Souza
P415a Avaliação de transcritos diferencialmente expressos em neoplasias humanas com ORESTES / Tarcísio de Souza Peres. Campinas, SP : [s.n.], 2006.

Orientador : Fernando Lopes Alberto
Dissertação (Mestrado) Universidade Estadual de Campinas. Faculdade de Ciências Médicas.

1. Bioinformática. 2. Expressão gênica. 3. Câncer. 4. Inferência estatística. 5. Transcrição genética. I. Alberto, Fernando Lopes. II. Universidade Estadual de Campinas. Faculdade de Ciências Médicas. III. Título.

Título em inglês : Evaluation of differential expression profiles across neoplasic human samples using ORESTES (Opening reading frame)

Keywords: • Bioinformatics
• Gene expression
• Neoplasm
. Statistical inference
. Transcription, Genetic

Área de concentração : Ciências Biomédicas
Titulação: Mestrado em Ciências Médicas

Banca examinadora: Prof Dr Fernando Lopes Alberto
Profa. Dra. Helena Paula Brentani
Prof Dr José Barreto Campello

Data da defesa: 30-08-2006

Banca examinadora da Dissertação de Mestrado

Orientador: Prof. Dr. FERNANDO LOPES ALBERTO

Membros:

1. Prof. Dr. FERNANDO LOPES ALBERTO

2. Profa. Dra. HELENA PAULA BRENTANI

3. Prof. Dr. JOSÉ BARRETO CAMPELLO CARVALHEIRA

Curso de pós-graduação em Ciências Médicas, da Faculdade de Ciências Médicas da Universidade Estadual de Campinas.

Data: 30/08/2006

DEDICATÓRIA

Linda Marília, que doou parte de seu perfume

a todas as flores do mundo.

A GOD (RIDLEY M., 2003), acima de todas as coisas.

À minha amada esposa pela fidelidade aos nossos sonhos, companheirismo, carinho, pavê de abacaxi com côco - incentivador memorável de incansáveis jornadas durante a preparação do manuscrito - e por seu apoio irrefutável a todas as etapas deste trabalho.

À dona Joana, mulher virtuosa, que sempre cuidou de seus filhos com muito amor e dedicação, sempre os direcionando através dos caminhos retos.

Ao seu João Tarciso, presente em minhas lembranças. Talvez um lugar-comum do determinismo genético, mas creio que curiosidade científica e aptidão técnica também podem ser explicadas por "*imprinting* paterno".

À Lelê, Dany, Marcelo e Dudinha, dona Yara, seu Ari, Lu e Felipão, pessoas estimadas com as quais pretendo conviver pelo resto de meus dias. À tia Ligia por ter gasto comigo um pouquinho do seu CPE da Cambridge.

Aos "Bohemios" pela amizade, respeito e por acreditarem sempre.

Ao Prof. Dr. Fernando Lopes Alberto - o Ciência - por sua orientação construtiva e visão interdisciplinar, além do incentivo ao raciocínio crítico e busca por elevados padrões de excelência.

À equipe do Hemocentro da UNICAMP, pelo apoio, infra-estrutura e materiais fundamentais neste trabalho. Em especial ao Prof. Dr. Fernando Ferreira Costa; aos bioinformatas Gustavo Lacerda, Tiago Machado e Marcelo Brandão; à Dulcinéia Martins de Albuquerque, referência onipresente nas seções de agradecimentos de teses de Mestrado e Doutorado das melhores instituições do país; ao Anderson Ferreira da Cunha pela amizade e atenção em etapas críticas, Denise por ter acompanhado com paciência minha primeira extração de RNA, Manoela e Hélvia pelo empréstimo dos *primers*, Heloísa pelo DEPC e afins, Flávia pelas dicas sobre a extração, Tiago e Luciana pelo apoio e pipetas exclusivas para RNA.

À Profa. Dra. Sophie Derchain pelo curso diferenciado de Pedagogia e Didática na FCM e à Prof. Dra. Nelci Fenalti Hoehr pelo excelente curso de Metodologia Científica e também por sua visão além do método. À Jovem Pesquisadora, Profa. Dra. Nicola Amanda Conran Zorzeto, à Profa. Dra. Carmem Silva Passos Lima e ao Prof. Dr. José Barreto Campello Carvalheira pela disponibilidade para a Qualificação deste trabalho e por suas sugestões, críticas e comentários. À Márcia e Regina das secretarias da FCM.

Ao (infelizmente) extinto Laboratório de Bioinformática do Instituto de Computação da UNICAMP pelo fornecimento das seqüências do Projeto Genoma Câncer.

Ao Prof. Dr. Fernando Callera e ao Hospital Pio XII de São José dos Campos pela colaboração e também aos pacientes que participaram deste estudo.

À equipe de Métodos Moleculares do Fleury Medicina Diagnóstica e ao Instituto Fleury, pelo apoio em algumas das etapas de validação experimental. À Eloisa de Sá Moreira, cientista brilhante e profissional competente, por seu apoio em momentos fundamentais.

Ao Prof. Dr. Stephen Altschul por seus comentários pertinentes sobre o LyM.

A William Shockley, por ser um péssimo empreendedor, mas ter conseguido montar a equipe pioneira em circuitos integrados. A Robert Noyce, por ter percebido que Shockley era péssimo empreendedor. A Donald Knuth, professor emérito da universidade de Stanford, pelo clássico "The Art of the Computer Programming ".

Ao CNPq pelo apoio finaceiro parcial concedido a este trabalho.

- Senhor, para que um genoma tão grande?

- Para despistar... o segredo mesmo é o barro!

- Senhor, decifraram o Genoma Humano.

- Malditos *hackers*, vou ter que mudar a *password*!

	PÁG.
RESUMO.....	*xxxvii*
ABSTRACT.....	*xli*
1- INTRODUÇÃO.....	45
1.1- Câncer.....	47
1.2- Bioinformática.....	49
2- OBJETIVOS.....	57
3- MATERIAIS E MÉTODOS.....	61
3.1- Montagem dos fragmentos ORESTES.....	63
3.2- Comparação com seqüências conhecidas.....	65
3.3- Determinação dos níveis de expressão.....	67
3.4- Ferramenta para comparação estatística.....	67
3.5- Estratégias para classificação de genes.....	69
3.6- Análises adicionais – perfil de expressão entre os tecidos.....	71
3.7- Desenho de oligonucleotídeos iniciadores.....	72
3.8- Pacientes e Amostras biológicas.....	72
3.9- Extração do RNA Total.....	73
3.10- Síntese do cDNA.....	74
3.11- RT-PCR em Tempo Real.....	75
3.12- Medida da estabilidade gênica (M) e ranking dos genes controle...	76

4- RESULTADOS..... 79

4.1- Avalição da ferramenta estatística (LyM)..... 81

4.2- Análises computacionais dos dados ORESTES..... 85

4.3- Análises experimentais dos genes selecionados..... 90

5- DISCUSSÃO..... 99

6- CONCLUSÕES..... 105

7- REFERÊNCIAS BIBLIOGRÁFICAS..... 109

8- ANEXOS..... 119

Anexo 1- Termo de Informação e Consentimento..... 121

Anexo 2- Código-fonte do programa LyM..... 123

Anexo 3- Os 100 primeiros genes (de cada tecido) da classificação final..... 127

°C	*grau Celcius*
µg	*micrograma*
µl	*microlitro*
ACTB	*ACTin Beta*
ACTG1	*ACTin Gamma 1*
B2M	*Beta-2-Microglobulin*
BLAST	*Basic Local Alignment Seach Tool*
CCD	*Charge Coupled Device*
cDNA	*complementary DNA*
CGAP	*Cancer Genome Anatomy Project*
CNPq	*Conselho Nacional de Desenvolvimento Científico e Tecnológico*
CPE	*Certificate of Proficiency in English*
CT	*Threshold Cycle*
DEPC	*DiEtilPiroCarbonate*
DGED	*Digital Gene Expression Displayer*
DNA	*Deoxiribonucleic Acid*
dNTP	*deoxyriboNucleotide TriPhosphate*
DO	*Densidade Óptica*
dT	*deoxyribose Thymidine*
DTT	*1,4-DiThioThreitol*
E	*Eficiência*
EC	*Expressão Chave*

EST	*Expressed Sequence Tags*
FAPESP	*Fundação de Amparo à Pesquisa do Estado de São Paulo*
FCM	*Faculdade de Ciências Médicas*
GAPDH	*GlycerAldehyde-3-Phosphate DeHydrogenase*
GB	*Gigabyte*
GL	*Gene na Literatutra*
GO	*Gene Ontology*
GOD	*Genome Organizing Device*
HTML	*Hyper Text Markup Language*
IGHG3	*ImmunoGlobulin Heavy Constant Gamma 3*
IGKC	*ImmunoGlobulin Kappa Constant*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
LBI	*Laboratório de BioInformática*
M	*molar*
MHz	*Mega Hertz*
mL	*mililitro*
mM	*milimolar*
NEC	*Nível de Expressão Categorizado*
ng	*nanograma*
nm	*nanômetro*
nmoles	*nanomoles*
nM	*nanomolar*
ORESTES	*Open Reading Frame Expressed Sequence Tags*
PCR	*Polimerase Chain Reaction*

PDGF	*Platelet-Derived Growth Factor*
PDB	*Protein Data Bank*
PERL	*Practical Extraction and Report Language*
PFN1	*Profilin 1*
PG	*Pontuação do Gene*
PHRAP	*Phragment Assembly Program*
PIF	*Protein Research Fundation*
PIR	*Protein Information Resource*
pmoles	*picomoles*
PP	*Peso da Publicação*
PYGB:	*Phosphorylase Glycogen Brain*
qPCR	*RT-PCR quantitativo*
RAM	*Random Access Memory*
RNA	*Ribonucleic Acid*
RNAm	*RNA mensageiro*
RT	*Reverse Transcription*
SAGE	*Serial Analysis of Gene Expression*
Tm	*melting temperature*
UNICAMP	*Universidade Estadual de Campinas*
UV	*Ultra Violeta*
VIL2	*VILlin 2*

PÁG.

Tabela 1- Separação por tecido das seqüências ORESTES do Projeto Genoma Câncer Humano FAPESP-Ludwig..... 64

Tabela 2- Parametrização utilizada na montagem inicial e na segunda montagem, utilizando os valores padronizados. Estes parâmetros dizem respeito à estringência dos alinhamentos..... 65

Tabela 3- Tempo de execução do BLAST para as seqüências consenso ORESTES..... 66

Tabela 4- Bibilotecas SAGE utilizadas para comparação dos fatores de expressão gência entre as ferramentas LyM e DGED..... 68

Tabela 5- Categorização dos valores de expressão para composição do valor de NEC..... 69

Tabela 6- Publicações selecionadas (e busca no "Entrez Gene") para composição de lista de genes de referência e o peso (PP) atribuído à publicação. este peso foi utilizado na composição do valor "GL" para a classificação de um gene cadidato..... 70

Tabela 7- Dados clínicos de cada amostra..... 73

Tabela 8- Número de seqüências consenso resultantes do processo de montagem..... 85

Tabela 9- Número de genes, separados por tecido, após comparação estatística da expressão diferencial..... 87

Tabela 10- Distribuição da expressão gênica em vias relacionadas ao câncer. O número de genes identificado para cada via é apresentado, bem como a porcentagem de genes idenficados que possuem fator de expressão diferencial (em módulo) maior do que cinco..... 89

Tabela 11- Os 100 primeiros genes (ordenados por pontuação), para o tecido "mama"..... 127

Tabela 12- Os 100 primeiros genes (ordenados por pontuação), para o tecido "cólon"..... 131

Tabela 13- Os 100 primeiros genes (ordenados por pontuação), para o tecido "cabeça e pescoço"..... 135

Tabela 14- Os 100 primeiros genes (ordenados por pontuação), para o tecido "pulmão"..... 139

Tabela 15- Os 100 primeiros genes (ordenados por pontuação), para o tecido "estômago"..... 143

Tabela 16- Os 100 primeiros genes (ordenados por pontuação), para o tecido "útero"..... 147

Tabela 17- Os 100 primeiros genes (ordenados por pontuação), para o tecido "testículo"..... 151

Tabela 18- Os 100 primeiros genes (ordenados por pontuação), para o tecido "sistema nervoso central"..... 155

Tabela 19- Os 100 primeiros genes (ordenados por pontuação), para o tecido "próstata"..... 159

PÁG.

Figura 1- Distribuição das seqüências ORESTES em relação à posição do gene. (DIAS ET AL, 2000)..... 49

Figura 2- Interação entre Biologia e Tecnologia (GIBAS E JAMBECK, 2001)..... 51

Figura 3- Comparação das seqüências dos genes v-sis e PDGF..... 52

Figura 4- Crescimento anual do Genbank (azul). Atualmente (julho DE 2006), o número nucleotídeos depositados é de 130 bilhões..... 53

Figura 5- Os cinco trabalhos mais citados na literatura científica. levantamento feito pelo Thompson ISI Web of Science (1983-2002)..... 54

Figura 6- Estrutura de um arquivo em formato phd. Além de informações sobre a reação de sequenciamento e cabeçalhos específicos, é informada a seqüência (nucleotídeos) e o *phred score* individual (primeiro número à direita de cada nucleotídeo)..... 63

Figura 7- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre estômago normal e carcinoma de estômago..... 81

Figura 8- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre mama normal e carcinoma de mama..... 82

Figura 9- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre cólon normal e adenocarcinoma de cólon..... 82

Figura 10- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre pulmão normal e adenocarcinoma de pulmão..... 83

Figura 11- Resultados da comparação entre LyM e DGED para a avaliação da diferença de expressão gênica entre mama normal e carcinoma de mama, por meio dos fatores arbitrários 4 (a), 8(b), 16(c), 32(d) e 64(e)..... 84

Figura 12- Resultado da análise por BLAST no Tecido "mama". A lista possui uma descrição das proteínas presentes e seus respectivos níveis de expressão para tecido normal e tumoral. O quadro em detalhe mostra a distribuição das 543 seqüências ORESTES para a proteína "TO2955" em tecido normal (o termo "*Contig*" refere-se ao identificador da seqüência consenso e o termo "*Reads*" é o número de fragmentos orestes presentes em cada seqüência consenso)..... 86

Figura 13- Arquivo em formato HTML com os genes classificados. À direita, o menu "*Links"* para acesso à informações adicionais de cada gene..... 87

Figura 14- Detalhamento de uma lista (pulmão) para o gene *HRAS*. Ele é o décimo elemento da classificação. A posição é indicada pelo número "10" e a cor verde indica que ele está mais expresso em tecido tumoral. ao passar o mouse sobre a posição, aparece um texto (caixa amarelo claro) indicando: o valor de pontuação (86,275), o número de seqüências ORESTES para este gene em condição normal (0), o número de seqüências ORESTES para este gene em condição tumoral (5) e o fator de expressão diferencial calculado pelo programa LyM (4,25). todas as demais informações formam obtidas por meio da ferramenta Gene Entrez..... 88

Figura 15- Vias relacionadas ao fenômeno neoplásico (HAHN E WEINBERG, 2002). Em destaque, trechos das vias para as quais foram identificados genes presentes na lista de genes produzida a partir de dados ORESTES..... 90

Figura 16- Estabilidade dos genes para as amostras estudadas em tecido normal e tumoral (estômago). A estabilidade média (M) é mostrada no eixo vertical. Os genes mais à direita (eixo horizontal) são os mais estáveis. Os genes de controle tradicionalmente utilizados (*GAPDH, ACTB E B2M*) não apresentaram as menores médias de estabilidade..... 91

Figura 17- Curvas de amplificação do gene *GAPDH* para amostras normais (a) e tumorais (b*)*. A variação entre as amostras pode ser observada através da variação dos CTs (intersecção entre a linha verde horizontal e a respectiva curva de amplificação)..... 92

Figura 18- Curvas de amplificação do gene *IGHG3* para amostras normais (a) e tumorais (b). a variação entre as amostras, observada através da variação do CT, É mais discreta do que a variação de genes habitualmente utilizados como normalizadores como *GAPDH* (Figura 17)..... 93

Figura 19- Curvas de amplificação do gene *B2M* para amostras normais (a) e tumorais (b). O Gene *B2M* teve a maior estabilidade entre os genes de referência "padrão" que foram testados *(B2M, GAPDH e ACTB)*..... 94

Figura 20- Curvas de amplificação do gene *ACTB* para amostras normais (a) e tumorais (b)..... 95

Figura 21- Curvas de dissociação do gene *JUN* para amostras normais (a,b) e tumorais (b). O pico das curvas correspondem à temperatura (representada no eixo x) de dissociação do fragmento amplificado (*amplicon)*. A morfologia e a coincidência das curvas em torno da temperatura esperada para o *amplicon* (neste caso cerca de 78°c) indica ausência de amplificação de produtos inespecíficos..... 96

Figura 22- Variação par-a-par para o conjunto de genes avaliados. A menor variação encontra-se em $V_{4/5}$, indicando que o número ótimo de genes de controle para as amostras estudadas é quatro..... 97

Figura 23- Expressão relativa (normal vs tumor) em estômago para a análise ORESTES e para a análise por RT-PCR em Tempo Real. Os genes utilizados como controle foram os quatro genes mais estáveis para as amostras utilizadas: *PFN1, IGHG3, ACTG1 e JUN*..... 97

RESUMO

PERES, T. S. **Avaliação de transcritos diferencialmente expressos em neoplasias humanas com ORESTES.** Campinas, 2006. Dissertação (Mestrado) – Faculdade de Ciências Médicas, Universidade Estadual de Campinas.

Durante todo o século XX, a pesquisa do câncer se desenvolveu de maneira sistemática, porém os últimos 25 anos foram notadamente caracterizados por rápidos avanços que geraram uma rica e complexa base de conhecimentos, evidenciando a doença dentro de um conjunto dinâmico de alterações no genoma. Desta forma, o entendimento completo dos fenômenos moleculares envolvidos na fisiopatologia das neoplasias depende do conhecimento dos diversos processos celulares e bioquímicos característicos da célula tumoral e que, porventura, a diferenciem da célula normal (GOLUB e SLONIM, 1999).

Nesse trabalho buscamos o melhor entendimento das vias moleculares no processo neoplásico por meio da análise dos dados do Projeto Genoma Humano do Câncer (CAMARGO, 2001) com vistas à identificação de genes diferencialmente expressos nas neoplasias dos seguintes tecidos: mama, cólon, cabeça e pescoço, pulmão, sistema nervoso central, próstata, estômago, testículo e útero. A metodologia de geração dos transcritos utilizada pelo Projeto Genoma Humano do Câncer é conhecida como ORESTES (DIAS et al, 2000).

Inicialmente, os dados de seqüenciamento (fragmentos ORESTES) foram agrupados por meio de uma técnica conhecida em Bioinformática como "montagem", utilizando o pacote de programas de computador PHRED/PHRAP (EWING e GREEN P., 1998). A comparação de cada agrupamento com seqüências conhecidas (depositadas em bases públicas) foi realizada por meio do algoritmo BLAST (ALTSCHUL et al, 1990). Um subconjunto de genes foi selecionado com base em critérios específicos e submetido à avaliação de seus níveis de expressão em diferentes tecidos com base em abordagem de inferência Bayesiana (CHEN et al, 1998), em contraposição às abordagens mais clássicas, como testes de hipótese nula (AUDIC e CLAVERIE, 1997). A inferência Bayesiana foi viabilizada pelo desenvolvimento de uma ferramenta computacional escrita em linguagem PERL (PERES et al, 2005).

Com o apoio da literatura, foi criada uma lista de genes relacionados ao fenômeno neoplásico. Esta lista foi confrontada com as informações de expressão gênica, constituindo-se em um dos parâmetros de um sistema de classificação (definido para a seleção dos genes de interesse). Desta forma, parte da base de conhecimento sobre câncer foi utilizada em conjunto com os dados de expressão gênica inferidos a partir dos fragmentos ORESTES. Para contextualização biológica da informação gerada, os genes foram classificados segundo nomenclatura GO (ASHBURNER et al, 2000) e KEGG (OGATA et al, 1999). Parte dos genes apontados como diferencialmente expressos em pelo menos um tecido tumoral, em relação ao seu equivalente normal, integram vias relacionadas ao fenômeno neoplásico (HAHN e WEINBERG, 2002). Dos genes associados a estas vias, 52% deles possuíam fator de expressão diferencial (em módulo) superior a cinco.

Finalmente, dez entre os genes classificados foram escolhidos para confirmação experimental dos achados. Os resultados de qPCR em amostras de tecido gástrico normal e neoplásico foram compatíveis com com os dados de expressão gênica inferidos a partir dos fragmentos ORESTES.

ABSTRACT

PERES, T. S. **Evaluation of differential expression profiles across neoplasic human samples using ORESTES.** Campinas, 2006. Dissertação (Mestrado) – Faculdade de Ciências Médicas, Universidade Estadual de Campinas.

The XXth century showed the development in cancer research in a systematic way, most notably in the last 25 years that were characterized by rapid advances that generated a rich and complex body of knowledge, highlighting the disease within a dynamic group of changes in the genome. The complete understanding of the molecular phenomena involved in the physiopathology of neoplasia is based upon the knowledge of the varied cellular and biochemical processes which are characteristic of the tumor and which make it different from the normal cell (GOLUB e SLONIM, 1999)

In this work, we investigated the molecular pathways in the neoplasic process through data analyses of the cDNA sequences generated on the Human Cancer Genome Project (CAMARGO, 2001). The following neoplasias were included: breast, colon, head and neck, lungs, central nervous system, prostate gland, stomach, testicle and womb. The methodology of generation of transcripts used by the Genome Project of Human Cancer is known as ORESTES (DIAS et al, 2000).

Initially, the sequence of data (ORESTES fragments) were grouped and assembled according to similarity scores. For this purpose, we used the package of computer programs PHRED/PHRAP (EWING e GREEN P., 1998). The resulting consensus sequences, each representing a cluster, were compared to known sequences (deposited in public databanks) through the BLAST algorithm (ALTSCHUL et al, 1990). A subgroup of genes was selected based on specific criteria and their levels of expression in different tissues were evaluated by a bayesian inference approach (CHEN et al, 1998), as compared to more classical approaches such as null hypothesis tests (AUDIC e CLAVERIE, 1997). The bayesian inference tool was represented as a PERL script developed for this work.

A list of genes, putatively related to the neoplasic phenotype, was created with the support of the literature. This list was compared to the gene expression information, becoming one of the parameters of a ranking system (defined for the selection of genes of interest). Therefore, part of the knowledge related to cancer was used together with the data of gene

expression inferred from ORESTES fragments. For a more accurate understanding of the molecular pathways involved in the generated information, the genes were classified according to the Gene Ontology (ASHBURNER et al, 2000) and KEGG (OGATA et al, 1999) nomenclatures. Additional global analyses by pathways related to the neoplasic phenomenon (HAHN e WEINBERG, 2002) demonstrated differential expression of the selected genes. About 52% of the genes in this pathways were differentially expressed in tumor tissue with at least a 5-fold.

Finally, ten genes were selected for experimental validation (*in vitro*) of the findings with real-time quantitative PCR, confirming in silico results.

1- INTRODUÇÃO

1.1- Câncer

De acordo com a Organização Mundial de Saúde, câncer é o crescimento e a propagação descontrolada de células que pode afetar praticamente qualquer tecido do corpo. Anualmente, mais de 11 milhões de pessoas são diagnosticadas com câncer e estimativas demonstram uma expansão para 16 milhões de novos casos/ano até 2020. Além disso, câncer é a causa de sete milhões de óbitos anuais, representando 12,5% de todas as mortes mundiais (WORLD HEALTH ORGANIZATION AND INTERNATIONAL UNION AGAINST CANCER, 2005).

A associação entre câncer e genética foi feita pela primeira vez no século XIX, por meio da observação de aberrações mitóticas em 13 amostras de carcinoma diferentes. Foi postulado que estas alterações (anáfases multipolares) geravam assimetrias na distribuição de material genético nas células neoplásicas (VON HANSEMANN, 1890). Trabalhos posteriores e diversos pesquisadores demonstraram que fatores cancerígenos (tais como radiação ionizante) atuavam como mutagênicos. Uma consistente anormalidade cromossômica foi encontrada em 1960 (NOWELL e HUNGERFORD, 1960) em pacientes com Leucemia Mielóide Crônica – o cromossomo *Philadelphia* – o que deu sustenção à associação entre câncer e doença genética. Estudos contemporâneos e posteriores seguiram a mesma abordagem e promoveram descobertas importantes, tais como aquelas que levaram à descrição dos conceitos de gene supressor de tumor (HARRIS, 1971) e oncogenes (LEVINSON et al, 1978). Adicionalmente, estudos na década de 70 relacionados à origem clonal de tumores permitiram o desenvolvimento da visão de que o câncer é um processo em múltiplas etapas: um processo dinâmico avançando entre estágios quantitativamente diferentes. O conceito de evolução Darwiniana também foi incorporado às descobertas (NOWELL, 1976), solidificando o paradigma de que câncer pode ser entendido como resultado de alterações no genoma selecionadas ao logo do tempo.

Simplificadamente, expressão gênica (LEWIN, 1980) é o processo de conversão da informação genética codificada no DNA em um produto final de um gene (isto é, uma proteína ou qualquer um dos vários tipos de RNA). O perfil de expressão gênica é o nível e a duração da expressão de um ou mais genes selecionados em uma célula ou tecido particular. Evidências iniciais de que o perfil de expressão gênica poderia

distinguir tipos distintos de neoplasias surgiram com estudos de leucemias linfóides e mielóides agudas. A estratégia adotada envolveu a identificação de genes "preditivos" de classe e, por meio do perfil de expressão diferencial destes genes, foi possível distinguir entre os dois tipos de leucemia aguda. Adicionalmente, o método conseguiu prever o tipo de resposta à quimioterapia (GOLUB e SLONIM, 1999). Este trabalho revela a abordagem analítica para classificação tumoral baseada em expressão gênica que foi utilizada em estudos posteriores em linfomas (ALIZADEH e EISEN, 2000), câncer de mama (VAN'T VEER et al, 2002), e outras neoplasias.

Em 1999, foi lançado no Brasil o Projeto Genoma Humano do Câncer (BONALUME, 1999), uma parceria entre a FAPESP e o Instituto Ludwig. O objetivo deste projeto foi o seqüenciamento de genes expressos em tumores considerados importantes no contexto de saúde pública do Estado de São Paulo. O projeto gerou cerca de um milhão e duzentas mil seqüências oriundas de genes expressos em tumores e em tecido normal. Para identificação e seqüenciamento dos genes expressos foi utilizada metodologia denominada ORESTES, cuja diferença fundamental observada em relação aos seqüenciamentos convencionais de ESTs é a determinação de regiões centrais dos transcritos (DIAS et al, 2000). Isto é explicado pelo fato de que a principal característica do método é a geração de ESTs por meio de reações de PCR de baixa estringência, além do uso de iniciadores não degenerados e selecionados arbitrariamente. A reação com estas características permitiu a confecção de seqüências que privilegiavam a parte central do transcrito em detrimento de seqüenciamentos tradicionais de ESTs, que privilegiam as extremidades 5´ ou 3´(Figura 1). Estas diferenças sugerem que as duas técnicas podem fornecer informações complementares para mapeamento genômico, fato que foi explorado para caracterizar regiões não mapeadas no genoma (SOGAYAR et al, 2004).

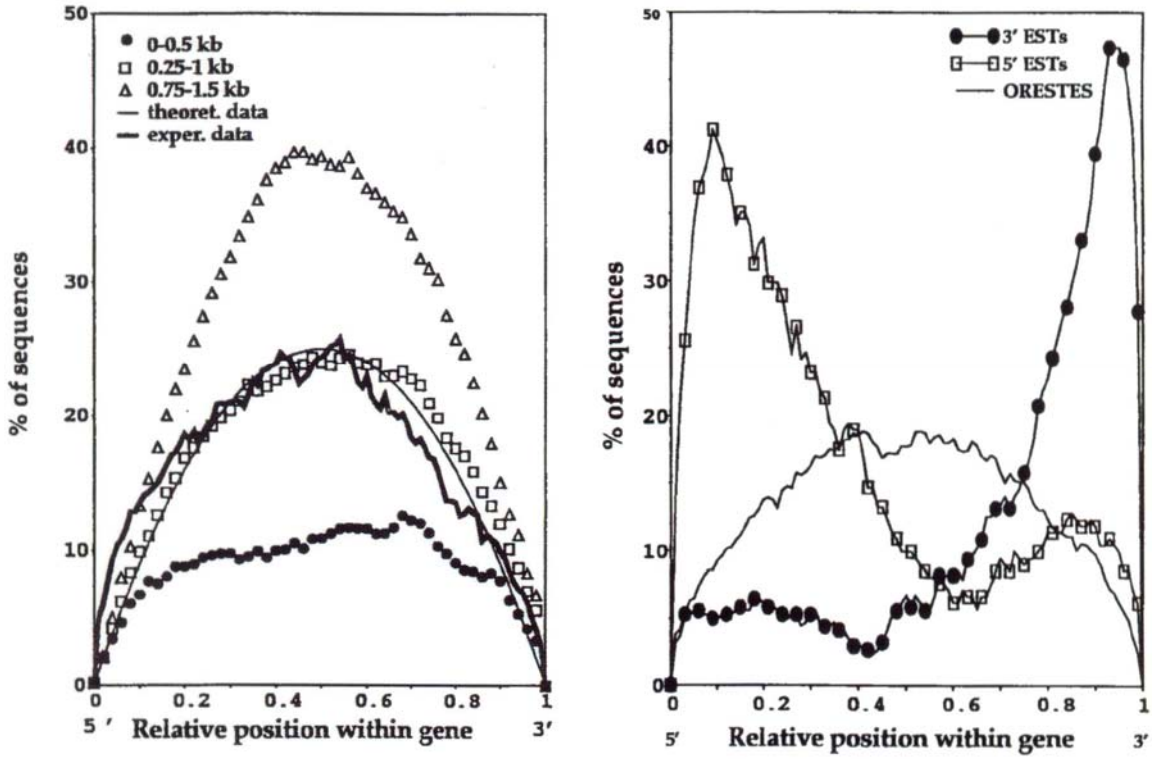


Figura 1- Distribuição das seqüências ORESTES em relação à posição do gene. (DIAS et al, 2000)

Após o término do projeto, as seqüências foram depositadas em bases públicas (CAMARGO, 2001).

1.2- Bioinformática

Bioinformática é um campo da ciência no qual Biologia, Química, Matemática Aplicada, Ciências da Computação e Estatística se fundem em uma única disciplina com o objetivo de entender e resolver problemas biológicos. O caráter multidisciplinar da Bioinformática está presente desde o seu surgimento: Margaret Oakley Dayhoff (1925-1983), considerada a fundadora da área, foi pioneira em estudos integrados de Química, Matemática, Biologia e Ciência da Computação.

Os termos Bioinformática e Biologia Computacional são utilizados indistintamente, embora existam esforços de vários grupos e diversos autores para uma classificação mais precisa, dado que tipicamente o segundo termo tem foco no desenvolvimento de métodos computacionais e o primeiro, em teste de hipóteses e descobertas (HUERTA et al, 2000). Neste trabalho adotaremos o termo Bioinformática.

Em julho de 2006, o número de páginas na Internet que fazem referência ao termo "*bioinformatics*" é de 101.000.000, enquanto o número de páginas que fazem referência à "*molecular biology*" é de 99.000.000. Curiosamente, estas quantidades são muito similares e a soma total das páginas é aproximadamente o número de páginas que fazem referência ao termo "*genetics*" (212.000.000). Para ilustrar a ordem de grandeza presente nos números apresentados, é possível comparar com número de páginas que fazem referência à temas populares da Internet como "*The Simpsons*" (26.900.000) e "*The Da Vinci Code*" (41.000.000).

Uma tarefa comum em projetos de Bioinformática é o uso de ferramentas matemáticas para extrair informação relevante de dados não homogêneos ou "ruidosos" - em Teoria da Informação a palavra "ruído" denota um sinal ou informação não pertinente no contexto ou sem significado (SHANNON C.E. e WEAVER W., 1949) – que podem ser produzidos por técnicas biológicas de larga escala tais como projetos de seqüenciamento genômico ou análises de microarranjos de cDNA. Recentemente, as tecnologias de geração de dados biológicos têm se tornado mais produtivas e robustas (MONTELIONE e ANDERSON, 1999; SCHENA et al, 1995; WADA et al, 1983), gerando uma crescente quantidade de informação a ser compreendida e correlacionada aos fenômenos biológicos de interesse (WESTON e HOOD, 2004). Este paradigma tem tornado a evolução científica mais eficiente, evidenciando a necessidade de desenvolvimento de ferramentas cada vez mais complexas para análise e apontando a Bioinformática como disciplina fundamental no apoio às descobertas (Figura 2). Este paradigma também tem se refletido nas grades curriculares (HONTS, 2003) das Universidades de todo o mundo. As linhas de pesquisa mais conhecidas em Bioinformática são: Análise de Seqüências, Evolução Biológica, Análise de Expressão Gênica, Análise Regulatória, Predição de Estruturas, Genômica Comparativa e Modelagem de Sistemas Biológicos.

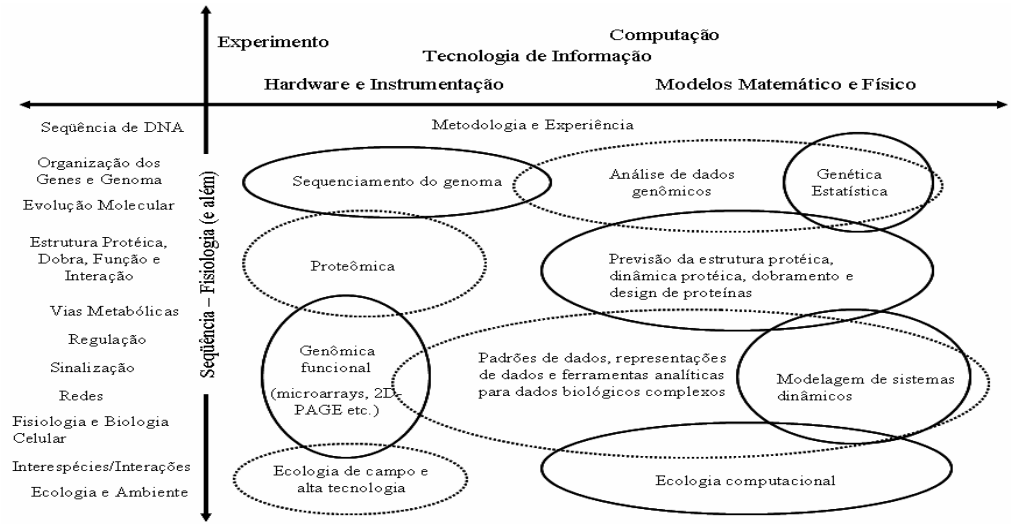


Figura 2- Interação entre Biologia e Tecnologia (GIBAS e JAMBECK, 2001)

Na linha de pesquisa sobre Análise de Seqüências, destacam-se a interpretação de cromatogramas (produzidos por seqüenciadores automatizados) e a montagem e comparação de seqüências, que representam as tarefas mais comumente realizadas.

Como o seqüenciamento automatizado é susceptível a erros, a qualidade de uma base refere-se, entre outros parâmetros relevantes, à morfologia do pico detectado no seqüenciador. Um dos algoritmos mais utilizados para análise destes dados é conhecido como Phred, que utiliza séries de Fourier para identificar o nucleotídeo (A,C,T,G) por meio do pico detectado e atribuir a ele um valor de qualidade que é conhecido como *phred score* (EWING e GREEN P., 1998). Matematicamente, o *phred score* é a probabilidade de erro da interpretação de um pico:

$$\mathbf{Q = -10 \log10(Pe)}$$

Onde Q e Pe são, respectivamente, o valor de qualidade e a probabilidade de erro para um nucleotídeo atribuído a um pico em particular.

O processo de montagem de seqüências refere-se ao pareamento e fusão de fragmentos seqüenciados para a reconstrução das seqüências originais. Existem várias ferramentas computacionais desenvolvidas para esta finalidade tais como PHRAP, TIGR

Assembler (POP e KOSACK, 2004) e CAP3 (HUANG e MADAN, 1999), que utilizam um algoritmo guloso conhecido como “a menor superseqüência comum”. Algoritmos gulosos são meta heurísticas que escolhem o ótimo local em cada estágio com a intenção de atingir o ótimo global (KNUTH, 1969). O programa PHRAP é um dos mais utilizados e busca similaridades entre os fragmentos - tentando agrupá-los em um mosaico (etapa de alinhamento) - respeitando os valores de qualidade (*phred score*), construindo assim a “seqüência consenso”. Ao final das interações do algoritmo é gerada uma relação de seqüências consenso e das seqüências que não foram alinhadas.

A primeira comparação relevante entre seqüências biológicas foi feita por meio da associação de dois genes aparentemente não relacionados: v-sis (*Simian Sarcoma Virus Oncogene*) e um fator de crescimento denominado PDGF. A similaridade descoberta (DOOLITTLE, 1983; WATERFIELD, 1983) entre as duas seqüências (Figura 3) contruibuiu para demonstrar que oncogenes encontrados em retrovírus codificam componentes da maquinaria regulatória de crescimento da célula.

```
v-sis: 6 QGDPIPEELYKMLSGHSIRSFDDLQRLLQGDSGKEDGAELDLNMTRSHSGGELESLARGK 65
                QGDPIPEELY+MLS HSIRSFDDLQRLL GD G+EDGAELDLNMTRSHSGGELESLARG+
PDGF : 10 QGDPIPEELYEMLSDHSIRSFDDLQRLLHGDPGEEDGAELDLNMTRSHSGGELESLARGR 69

v-sis: 66 RSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ 125
                RSLGSL++AEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ
PDGF : 70 RSLGSLTIAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ 129

v-sis: 126 CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCEIVAAARAVTRSPGTSQEQR 185
                CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCE VAAAR VTRSPG SQEQR
PDGF : 130 CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCETVAAARPVTRSPGGSQEQR 189

v-sis: 186 AKTTQSRVTIRTVRVRRPPKGKHRKCKHTHDKTALKETLGA 226
                AKT Q+RVTIRTVRVRRPPKGKHRK KHTHDKTALKETLGA
PDGF : 190 AKTPQTRVTIRTVRVRRPPKGKHRKFKHTHDKTALKETLGA 230
```

Figura 3- Comparação das seqüências dos genes v-sis e PDGF.

O problema de homologia entre seqüências de caracteres foi explorado em meados do século XX e o conceito de "distância de edição" - o qual pode ser entendido como um valor de similaridade entre duas seqüências - foi definido (LEVENSHTEIN, 1966). O conceito foi inicialmente empregado na identificação de padrões no reconhecimento de voz (VINTSYUK T.K., 1968) e, posteriormente, foi criado o primeiro algoritmo para comparação de seqüências biológicas (NEEDLEMAN e WUNSCH, 1970). Em virtude do crescimento acelerado (Figura 4) das bases de dados com seqüências biológicas – em 2005 o GenBank (BENSON et al, 2005) comemorou a marca de 100 bilhões de pares de bases depositadas - foram desenvolvidos algoritmos cada vez mais velozes e, em 1990, foi publicado um algoritmo de alinhamento local que se tornou o mais utilizado para comparação de seqüências biológicas chamado BLAST (ALTSCHUL et al, 1990).

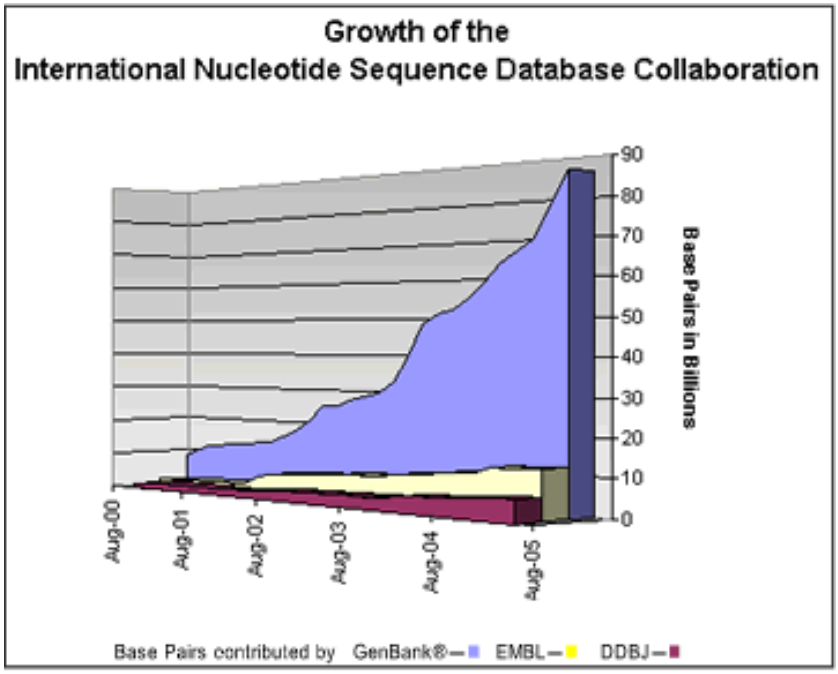


Figura 4- Crescimento anual do Genbank (azul). Atualmente (julho de 2006), o número nucleotídeos depositados é de 130 bilhões.

O BLAST compara uma seqüência desconhecida com uma base de dados de seqüências conhecidas. O algoritmo é relativamente veloz porque é baseado em métodos heurísticos (não garantindo portanto o melhor resultado, mas resultados que atendam a um determinado critério). O BLAST atribui um valor estatístico à similaridade entre as seqüências denominado *e-value* (valor esperado), que descreve o número de resultados que se poderia esperar ao acaso numa procura em um banco de dados de determinado tamanho. Esse valor decresce exponencialmente na proporção do aumento do escore resultante da similaridade entre duas seqüências. Um "valor esperado" de *zero* equivale a dizer que duas seqüências são idênticas. Um "valor esperado" de *um* para um resultado significa afirmar que em banco de dados com o tamanho em questão há a chance de encontro de uma seqüência com escore semelhante ao acaso. A ferramenta BLAST tornou-se tão utilizada que o artigo original da ferramenta tornou-se um dos três trabalhos mais citados na literatura científica (Figura 5).

Most-Cited Papers, 1983-2002

Rank	Paper	Citations
1	Chomczynski, N. Sacchi, **"Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction,"** *Analyt. Biochem.*, 162(1): 156-9, 1987.	49,562
2	A.P. Feinberg, B. Vogelstein, **"A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity,"** *Analyt. Biochem.*, 132(1): 6-13, 1983.	20,609
3	S.F. Altschul, *et al.*, **"Basic Local Alignment Search Tool,"** *J. Molec. Biol.*, 215(3): 403-10, 1990.	15,306
4	G. Grynkiewicz, M. Poenie, R.Y. Tsien, **"A new generation of CA-2+ indicators with greatly improved fluorescence properties,"** *J. Biol. Chem.*, 260(6): 3440-50, 1985.	14,357
5	J. Devereux, P. Haeberli, O. Smithies, **"A comprehensive set of sequence-analysis programs for the VAX,"** *Nucleic Acids Res.*, 12(1): 387-95, 1984.	13,056

SOURCE: Thomson ISI Web of Science

Figura 5- Os cinco trabalhos mais citados na literatura científica. Levantamento feito pelo Thompson ISI Web of Science (1983-2002).

Após a comparação com BLAST, cada seqüência pode ter similaridade com zero ou mais seqüências (as similaridades foram rotuladas "*blast hits"*). Esta redundância adicional existe em função da característica heurística do algoritmo, porém novas abordagens demonstram-se promissoras e com maior confiabilidade (ENGELHARDT et al, 2005).

Como os processos biológicos são multifatoriais, os estudos com um grande número de amostras ou com um elevado número de cenários têm, na maioria das vezes, custo impraticável. Assim sendo, métodos estatísticos são ferramentas fundamentais para maximizar a extração de informação de um conjunto de dados a respeito de um determinado fenômeno em estudo. Um destes métodos é a inferência estatística sobre um processo aleatório por meio da observação de seu contexto por um período finito de tempo. A inferência bayesiana é um tipo de inferência estatística na qual as observações do fenômeno estudado são usadas para atualizar a probabilidade de que uma dada hipótese possa ser verdadeira (BOLSTAD, 2004).

O entendimento de qualquer fenômeno ou mecanismo permite também sua reprodutibilidade ou previsibilidade. Assim sendo, ensaios laboratoriais (análises *in vitro* e *in vivo*) são fundamentais para a confirmação de modelos computacionais (análises *in silico*) e aperfeiçoamento dos mesmos, além de um entendimento mais abrangente do próprio fenômeno. Um dos métodos considerado padrão-ouro para análise de expressão gênica é o PCR quantitativo em tempo real. Fundamentalmente, esta técnica permite a mensuração do número de cópias de uma determinada molécula de ácido nucléico em uma amostra por meio de fluorescência. Assim sendo, ela permite realizar ensaios comparativos de expressão gênica (normal *vs* tumor), através da utilização de genes de referência. Para validar a expressão estável presumida para um dado gene que se pretende utilizar como referência, é necessário conhecimento prévio de medida confiável para normalizar este gene, com a finalidade de remover qualquer variação não específica. Foi proposta (VANDESOMPELE et al, 2002) uma medida que determina a estabilidade de expressão de genes com base em níveis de expressão não normalizados. Esta medida baseia-se no princípio de que a razão de expressão de dois genes de referência ideais é idêntica em todas as amostras, independentemente da condição experimental ou tipo celular. Assim, a

variação real das razões de expressão de dois genes de referência, na prática, reflete o fato de que um (ou ambos) os genes não é (são) constantemente expressos, com variação crescente na razão correspondendo a uma estabilidade de expressão decrescente. Neste contexto, a medida de estabilidade (*M*) de um gene controle foi definida como a variação par-a-par média de um gene particular com outros genes controles. Genes com valor de *M* mais baixo têm expressão mais estável. A análise de estabilidade também permite a verificação do número ideal de genes de referência para um dado conjunto de amostras, através das variações par-a-par entre os genes.

2- OBJETIVOS

No presente trabalho, tivemos por objetivos:

- Desenvolvimento de ferramenta computacional de análise estatística, para avaliação do nível de expressão gênica presente nas diferentes bibliotecas de ORESTES.

- Identificação de genes diferencialmente expressos em tumores de mama, cólon, cabeça e pescoço, pulmão, sistema nervoso central, próstata, estômago, testículo e útero utilizando a análise de dados de ORESTES produzidos pelo Projeto Genoma Humano do Câncer FAPESP-Ludwig (CAMARGO, 2001).

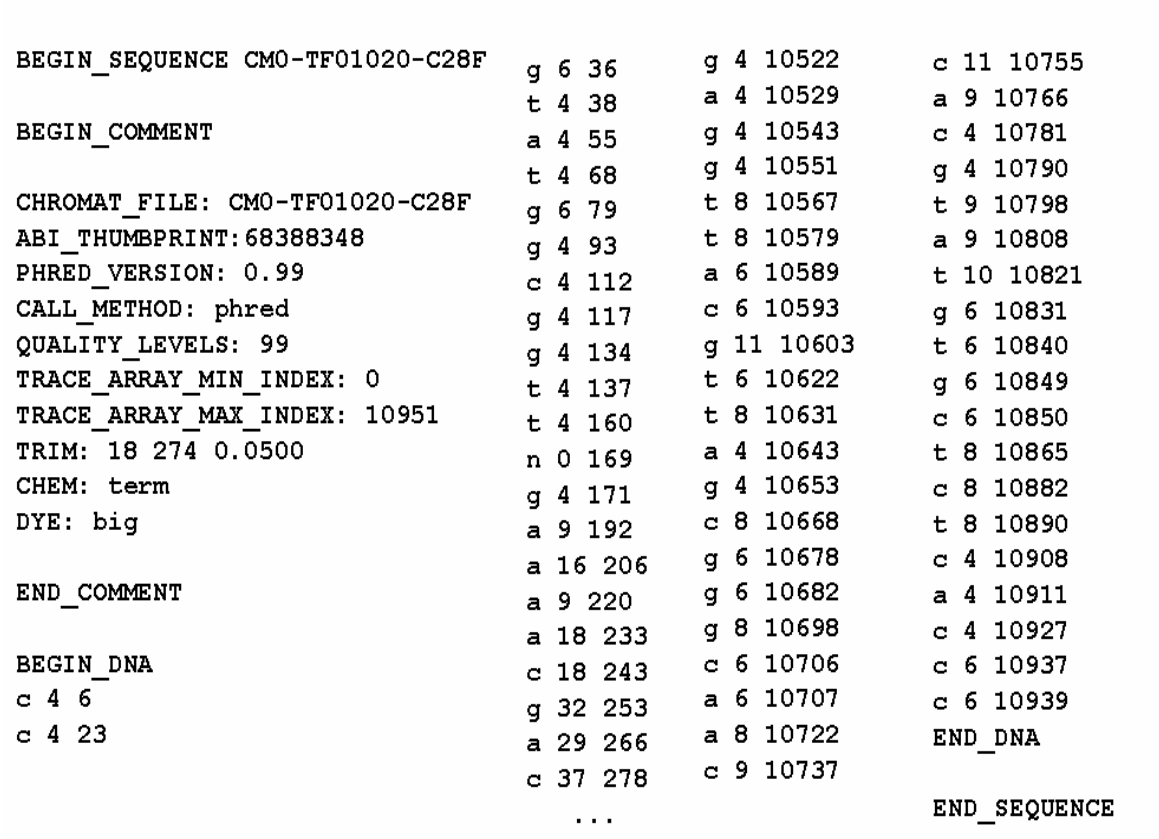
- Análise de expressão dos genes envolvidos em processos bioquímicos relacionados ao câncer.

- Confirmação pela técnica de PCR em tempo real das diferenças de expressão de alguns dos genes identificados.

MATERIAIS E MÉTODOS

3.1- Montagem dos fragmentos ORESTES

No início deste trabalho, foram fornecidos pelo Laboratório de Bioinformática da UNICAMP 1.188.862 fragmentos ORESTES originados do Projeto Genoma Câncer Ludwig-FAPESP, por meio de arquivos de qualidade denominados phds (Figura 6). Foram selecionados aqueles que estavam representados tanto em tecido normal como em tecido neoplásico (84%), os quais foram agrupados por tecido (Tabela 1). Neste estudo, estas seqüências constituem-se no objeto primário para avaliação da relação existente entre os transcritos humanos e o fenômeno neoplásico.



```
BEGIN_SEQUENCE CM0-TF01020-C28F

BEGIN_COMMENT

CHROMAT_FILE: CM0-TF01020-C28F
ABI_THUMBPRINT:68388348
PHRED_VERSION: 0.99
CALL_METHOD: phred
QUALITY_LEVELS: 99
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 10951
TRIM: 18 274 0.0500
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
c 4 6
c 4 23
```

```
g 6 36
t 4 38
a 4 55
t 4 68
g 6 79
g 4 93
c 4 112
g 4 117
g 4 134
t 4 137
t 4 160
n 0 169
g 4 171
a 9 192
a 16 206
a 9 220
a 18 233
c 18 243
g 32 253
a 29 266
c 37 278
  ...
```

```
g 4 10522
a 4 10529
g 4 10543
g 4 10551
t 8 10567
t 8 10579
a 6 10589
c 6 10593
g 11 10603
t 6 10622
t 8 10631
a 4 10643
g 4 10653
c 8 10668
g 6 10678
g 6 10682
g 8 10698
c 6 10706
a 6 10707
a 8 10722
c 9 10737
```

```
c 11 10755
a 9 10766
c 4 10781
g 4 10790
t 9 10798
a 9 10808
t 10 10821
g 6 10831
t 6 10840
g 6 10849
c 6 10850
t 8 10865
c 8 10882
t 8 10890
c 4 10908
a 4 10911
c 4 10927
c 6 10937
c 6 10939
END_DNA

END_SEQUENCE
```

Figura 6- Estrutura de um arquivo em formato phd. Além de informações sobre a reação de sequenciamento e cabeçalhos específicos, é informada a seqüência (nucleotídeos) e o *phred score* individual (primeiro número à direita de cada nucleotídeo).

Tabela 1- Separação por tecido das seqüências ORESTES do Projeto Genoma Câncer Humano FAPESP-Ludwig

Tecido	*Numero de seqüências*
mama normal	68969
mama tumoral	91596
cólon normal	72468
cólon tumoral	97426
cabeça e pescoço normal	18776
cabeça e pescoço tumoral	190357
pulmão normal	21961
pulmão tumoral	45387
sistema nervoso central normal	86274
sistema nervoso central tumoral	57140
próstata normal	79741
próstata tumoral	23262
estômago normal	13879
estômago tumoral	40736
testículo normal	28700
testículo tumoral	3465
útero normal	22691
útero tumoral	38456

Para a montagem destas seqüências, foi utilizado o pacote PHRAP, disponível gratuitamente para uso acadêmico. A primeira tentativa de montagem apresentou dois problemas:

1) Os parâmetros de qualidade exigidos (Tabela 2) para a montagem geraram poucas seqüências consenso;

2) A versão padrão PHRAP é limitada a 64000 seqüências. Alguns tecidos possuíam mais seqüências do que o número limite (Tabela 1);

Estes problemas iniciais foram abordados basicamente de duas maneiras. A montagem foi realizada com parâmetros padronizados (Tabela 2) e o *software* foi recompilado, permitindo sua execução com mais de 64000 seqüências. Como resultado da montagem, as seqüências se agruparam em consensos (Tabela 8) que foram submetidos à comparação com bases de dados de proteínas.

Tabela 2- Parametrização utilizada na montagem inicial e na segunda montagem, utilizando os valores padronizados. Estes parâmetros dizem respeito à estringência dos alinhamentos.

Parâmetro	***Primeira Montagem***	***Segunda Montagem***
penalty	9	2
Minmatch	20	14
gap_init	9	2
gap_ext	9	1
repeat_stringency	0,97	0,95

3.2- Comparação com seqüências conhecidas

A ferramenta BLAST foi utilizada para a comparação das seqüências consenso geradas com um banco de seqüências de proteínas. Para este tipo de comparação o BLAST tem uma variante específica denominada "Blastx", com os parâmetros especiais:

-e 0.00001: define 10^{-5} como valor máximo para o "valor esperado";

-b 5 : apresenta no máximo 5 alinhamentos no relatório final;

-v 10: apresenta no máximo 10 descrições por comparação no relatório final;

O banco de proteínas era composto por 2.314.886 seqüências (seqüências codificantes não redundantes depositadas no GenBank, PDB, SwissProt, PIR, PRF). O recurso de hardware utilizado para a comparação foi uma *SUN BLADE*® 1000, com dois processadores de 750MHz (*UltraSparc*® III – 64 bits) e três GB de memória RAM (Sun Microsystems Inc., Santa Clara, CA). Também foi utilizado um computador *Pentium*® II biprocessado (400MHz) com um GB de memória RAM (Intel, Santa Clara, CA). A duração das comparações com o BLAST utilizando este equipamento pode ser observada a seguir (Tabela 3) e foi proporcional ao número de seqüências existentes em cada tecido.

Tabela 3- Tempo de execução do BLAST para as seqüências consenso ORESTES.

Tecido	***Tempo Total BLAST (horas)***
mama normal	142
mama tumoral	162
cólon normal	127
cólon tumoral	185
cabeça e pescoço normal	32
cabeça e pescoço tumoral	311
pulmão normal	45
pulmão tumoral	81
sistema nervoso central normal	171
sistema nervoso central tumoral	122
próstata normal	52
próstata tumoral	50
estômago normal	26
estômago tumoral	76
testículo normal	60
testículo tumoral	8
útero normal	51
útero tumoral	82

Após o BLAST, os seguintes critérios foram adotados para reduzir o número de falsos positivos, incluindo apenas seqüências válidas em nosso estudo:

- possuir similaridade com pelo menos uma seqüência (assim sendo, as seqüências chamadas de "*no hits*" foram excluídas da amostra).

- possuir *e-value* (valor esperado) menor do que 10^{-5}

- não possuir em sua descrição os termos: "ribosomal", "unnamed", "similar to", "PREDICTED", "unknown", "FLJ", "KIAA", "LOC", "DKFPZP".

3.3- Determinação dos níveis de expressão

Para cada proteína encontrada, foi contado o número de seqüências que originaram o consenso e que possuíam similaridade com a dada proteína, conforme definição anterior. Desta forma, foi criada uma lista genes (na prática, identificadores de proteínas nas bases de dados citadas) com seus respectivos níveis de expressão definidos por meio do resultado das contagens para cada tecido.

3.4- Ferramenta para comparação estatística

Para determinação da expressão diferencial dos genes entre condições normal e tumoral de cada tecido desenvolvemos um programa em PERL denominado LyM (PERES et al, 2005), que utiliza uma abordagem bayesiana para avaliação da significância estatística das diferenças de expressão. Esta abordagem foi inicialmente utilizada para a análise de expressão gênica com dados de SAGE no processo de ativação de basófilos (CHEN et al, 1998), evitando assim o uso de simulações aleatórias - que exigem um esforço computacional considerado impraticável em alguns casos (LAL et al, 1999) - requeridas por métodos mais clássicos, tais como simulações de Monte Carlo (PORTER et al, 2001). LyM foi implementado (Anexo 2) com um algoritmo de busca binária (KNUTH, 1969) para encontrar o melhor fator que representa a diferença de

expressão entre a condição normal e tumoral de um gene, levando em consideração o número total de genes presentes em cada tecido. Algoritmos de busca binária são tipicamente O(*logn*) reduzindo assim o tempo utilizado nas buscas (a notação "O" foi introduzida pela teoria dos números (BACHMANN, 1894) e O(*logn*) significa que a relação entre o tamanho "n" dos dados de entrada e o tempo de computação para estes dados é de ordem logarítmica). Adicionalmente, o programa possui estratégias de armazenamento temporário para reutilização de valores calculados. Para validação, LyM foi comparada com uma ferramenta desenvolvida pelo CGAP chamada DGED, que utiliza a mesma abordagem estatística (LASH et al, 2000). A diferença substancial entre os dois programas é que DGED analisa os genes por meio de um fator diferencial de expressão arbitrariamente escolhido, emitindo um valor de confiança para cada gene, considerando tal fator; enquanto a LyM emite o melhor fator diferencial possível por meio de busca binária, baseado num valor de confiança arbitrário. Para a comparação entre as duas ferramentas, foi utilizado um conjunto de dados de expressão gênica compreendendo quatro bibliotecas (mama, cólon, estômago e pulmão) nas condições normal e tumoral, totalizando 495.577 seqüências (Tabela 4). Para o programa DGED foram escolhidos arbitrariamente os fatores: 2, 4, 8, 16, 32 e 64. Para LyM o valor de confiança selecionado foi de 95% (ou p = 0,05). A comparação final entre os dois programas foi feita através da razão entre o fator calculado pelo LyM e o fator arbitrário para o DGED.

Tabela 4- Bibilotecas SAGE utilizadas para comparação dos fatores de expressão gência entre as ferramentas LyM e DGED

Biblioteca	***Número de seqüências (SAGE tags)***
Breast_carcinoma_epithelium_AP_DCIS7	89184
Breast_normal_stroma_AP_1	79152
Colon_adenocarcinoma_CL_Caco2	60682
Colon_normal_B_NC1	49610
Lung_adenocarcinoma_MD_L10	86887
Lung_normal_CL_L15	42226
Stomach_carcinoma_B_G189	63075
Stomach_normal_epithelium_B_body1	24761

3.5- Estratégias para classificação de genes

Após o cálculo da diferença de expressão entre as condições normais e tumorais nos tecidos, foi definido um sistema de pontuação para permitir que genes relacionados ao câncer na literatura pudessem ser considerados na avaliação. Ou seja, além do padrão de expressão gênica, o fator "gene previamente relacionado com câncer" teve seu papel nesta análise. A pontuação de cada gene (**PG**) foi definida da seguinte maneira:

$$\mathbf{PG = 3 * NEC + 2 * GL + EC}$$

Onde:

- **NEC** é o Nível de Expressão Categorizado. O valor de NEC é o valor da diferença de expressão de um gene divido por 10 e a este resultado é adicionado um valor complementar, conforme a categoria da diferença de expressão (Tabela 5). Deste modo, um gene com diferença de expressão de 35 tem NEC = 35/10 + 2 + 1 + 1 + 2 +4 + 5 = 18,5.

Os valores de categorização foram definidos após simulações com os dados de mama. O objetivo da simulação era atingir uma ponderação equilibrada, levando em conideração os demais fatores da equação apresentada na página anterior.

Tabela 5- Categorização dos valores de expressão para composição do valor de NEC.

Fator de expressão diferencial	***Acréscimo NEC***
>1	2
>2	1
>5	1
>7	2
>10	4
>30	5
>50	5

- **EC** é uma constante (EC = 20) atribuída a genes que possuam uma "Expressão Chave" em sua descrição: "cancer", "tumor", "onco", "apoptosis", "death", "metastasis", "oma", "proliferation". As expressões chave foram escolhidas arbitrariamente e testadas individualmente para confirmar a sua relevância de acordo com o objetivo proposto.
- **GL** é valor relativo ao aparecimento do Gene na Literatura.

Foi criada uma lista de genes presentes em publicações relacionadas a câncer (Tabela 6), as quais foram selecionadas considerando os seguintes critérios: publicações recentes em revistas indexadas e com alto fator de impacto, revisões de literatura de câncer, atlas ou "enciclopédias" de referência, e publicações que norteram produtos de indústrias fornecedoras de insumos à pesquisa em Ciência da Vida.

Para cada publicação foi atribuído um peso "PP". O valor de GL é a soma de todos os valores de "PP" de um determinado gene candidato. Adicionalmente, foi incluída uma lista de genes por meio de busca por palavra-chave utilizando a ferramenta Entrez Gene (WHEELER et al, 2005) com a expressão "tumor *or cancer and human*". Para ilustrar, um gene presente nas duas primeiras referências (Tabela 6) tem GL = 14 + 5 = 19.

Tabela 6- Publicações selecionadas (e busca no "Entrez Gene") para composição de lista de genes de referência e o peso (PP) atribuído à publicação. Este peso foi utilizado na composição do valor "GL" para a classificação de um gene cadidato.

Lista de referência	***PP***
(FUTREAL et al, 2004)	14
(DORKELD et al, 1999)	5
(WHITE, 2004)	4
(NILSSON e CLEVELAND, 2003)	4
(HANAHAN e WEINBERG, 2000)	4
(REINHOLD et al, 2003)	2
Busca por expressão no "Entrez Gene"	1

Após a definição dos critérios apresentados, foi desenvolvido um segundo programa em PERL que efetuou o cálculo da pontuação de classificação para cada gene candidato, separado por tecido. Este programa gerou a lista final com os genes classificados e ordenados de modo ascendente, separados por tecido, em formato HTML.

Além disso, o programa fez a associação dos genes com as nomenclaturas GO (ASHBURNER et al, 2000) e KEGG (OGATA et al, 1999) para detalhamento funcional, atribuindo também *hyperlinks* específicos de cada gene - obtidos por meio do serviço *LinkOut* (WHEELER et al, 2005) - para uma avaliação mais aprofundada dos achados. Finalmente, todos os genes classificados foram contextualizados por meio de vias importantes no processo neoplásico (HAHN e WEINBERG, 2002).

Os genes listados receberam ainda um filtro adicional para seleção dos candidatos a confirmação experimental:

- Haver já disponibilidade de ensaio padronizado de qPCR no nosso laboratório
- Ter sido detectado por ORESTES em todos os tecidos investigados.

3.6- Análises adicionais – perfil de expressão entre os tecidos

Com o objetivo de realizar uma avaliação global dos genes comuns a todos os tecidos, eles foram submetidos a abordagens estatísticas adicionais (agrupamento hierárquico, agrupamento por k-medianas, análise de componentes principais, mapas auto- organizáveis e redes de relevância) para identificação de grupos com perfil de expressão distinto ou grupos de genes correlacionados. Para os agrupamentos, foi utilizado o método "*average linkage*" e a distância padrão utilizada foi a distância euclidiana. As análises foram realizadas com o auxílio dos *softwares* TM4 (SAEED et al, 2003) e R (IHAKA e GENTLEMAN, 1996).

3.7- Desenho de oligonucleotídeos iniciadores

Este trabalho utilizou inciadores já disponíveis em nosso laboratório para os seguintes genes: *ACTB*, *ACTG1*, *B2M*, *GAPDH*, *IGKC*, *IGHG3*, *JUN*, *PFN1*, *PYGB* e *VIL2*. Para desenho destes iniciadores foram utilizados os *softwares* Primer Express® (Applied Biosystems, Foster City, CA) e Gene Runner® (Hastings Software Inc., Hudson, NY). Foram selecionados os iniciadores que melhor se adequavam aos critérios:

- Amplicon com tamanho mínimo de 70 e máximo de 90 pares de bases;
- Par de iniciadores em éxons distintos;
- Proximidade da cauda poli A;
- Ausência de tendência a formação de estruturas secundárias (*hairpin loops*, *inner loops* e *dimers*) nas condições da reação;
- Especificidade (por meio de comparação BLAST);

3.8- Pacientes e amostras biológicas

Para os experimentos de validação dos achados com qPCR, foram utilizadas amostras de tecido humano, obtidas dos ambulatórios de cirurgia do Hospital PIO XII, em São José dos Campos. As amostras foram coletadas no momento da biópsia diagnóstica. O diagnóstico foi confirmado com base nos achados clínicos e anatomopatológicos (KLEIHUES e SOBIN, 2000). Os pacientes concordaram com o estudo e assinaram Termo de Informação e Consentimento (Anexo 1), aprovado pela Comissão de Ética da Faculdade de Ciências Médicas da UNICAMP. Todas as etapas descritas foram realizadas sob supervisão do Prof. Dr. Fernando Callera, diretor clínico do Hospital PIO XII.

Tabela 7- Dados clínicos de cada amostra.

Amostra	*Sexo*	*Idade*	*Diagnóstico*	*Cirurgia realizada*
7	M	66	Tumor de Estômago	Gastrectomia
14	F	81	Tecido Normal	-
51	F	25	Mucosa Normal	Gastrectomia Subtotal
61	M	76	Tumor de Estômago	Gastrectomia total
62	F	56	Cancêr Gástrico	Gastrectomia subtotal+Linfadenectomia
66	F	25	Adenocarcinoma Gástrico	Gastrectomia Subtotal

Após a cirurgia, as amostras formam encaminhadas em embalagem térmica com gelo seco ao Hemocentro da UNICAMP e foram mantidas em nitrogênio líquido até o momento da extração do RNA (temperatura de aproximadamente -196°C).

3.9- Extração do RNA Total

As amostras de tecido ainda congeladas foram maceradas manualmente. O RNA total das amostras de pacientes e indivíduos normais foi isolado utilizando Trizol® *Reagent* (Invitrogen Inc., Carlsbad, Califórnia), uma solução monofásica de fenol e isocianato de guanidina. Durante a homogeneização, o reagente Trizol® mantém a integridade do RNA enquanto rompe as células e dissolve os componentes celulares. Para extração de RNA total cada amostra foi homogeneizada em Trizol® na proporção de 1,0 mL / 1,0 x 10^7 células (CHOMCZYNSKI e SACCHI, 1987). A solução de Trizol® foi incubada por cinco minutos à temperatura ambiente para completa dissolução dos complexos nucleoprotéicos. Para cada 1,0 mL de Trizol® utilizado na amostra foi realizado o seguinte procedimento: acrescentaram-se às amostras 200 µl de clorofórmio e estas foram em seguida agitadas vigorosamente. Nova incubação foi realizada por mais cinco minutos à temperatura ambiente e, então, as amostras foram submetidas à centrifugação por 15

minutos a 13.200 rpm, em temperatura de 4°C. A fase orgânica, na qual se encontra o RNA, foi recuperada, transferida para um novo tubo e procedeu-se à precipitação do RNA por adição de 500 µl de isopropanol gelado, incubação à temperatura ambiente por 10 minutos e centrifugação a 13.200 rpm por 15 minutos em temperatura de 4°C. Após a precipitação, o *pellet* de RNA foi lavado com etanol 75% em água destilada tratada com dietilpirocarbonato (DEPC) para remoção do excesso de sal e, novamente centrifugado, a 11.800 rpm por cinco minutos, em temperatura de 4°C. As amostras de RNA total foram ressuspendidas em água com DEPC e incubadas por 10 minutos a 55°C. Adicionaram-se 200 µL de clorofórmio, agitando-se vigorosamente esta suspensão e incubando-a por cinco minutos à temperatura ambiente. Em seguida, a suspensão foi centrifugada a 13.500 rpm durante 15 minutos a 4°C. O sobrenadante contendo o RNA total foi coletado e foram adicionados 500 µL de isopropanol. Esta suspensão foi incubada por 10 minutos à temperatura ambiente e centrifugada por 13.500 rpm, durante 10 minutos a 4°C. Em seguida, o *pellet* de RNA total foi coletado e a este foi adicionado 1,0 mL de etanol 75% (em água destilada tratada com DEPC). Após agitação em vórtex, esta suspensão foi centrifugada a 11.800 rpm por cinco minutos a 4°C. Em seguida, o *pellet* foi coletado, ressuspendido em água destilada tratada com DEPC e estocado a -80^{o}C. A quantificação do RNA obtido em solução aquosa foi realizada por meio de leitura da densidade óptica (DO) de uma alíquota da amostra no espectrofotômero de luz UV *NanoDrop*® (NanoDrop Technologies, Inc., Wilmington, DE). Nestas condições, uma unidade de DO equivale a 40 µg/mL de RNA. A relação entre as leituras realizadas a 260 e 280 nm foi utilizada como parâmetro na estimativa do grau de contaminação do RNA por proteínas, e resultou, invariavelmente, entre 1,7 e 2,0.

3.10- Síntese do cDNA

As reações de transcrição reversa foram realizadas a partir de 1 µg de RNA, ao qual foram adicionados, inicialmente, 25 pmoles de oligo dT, 50 ng de iniciadores aleatórios (*random hexamers*) e 10 nmoles de cada dNTP em volume total de 14 µL. A adição de *random hexamers* à reação garante uma maior eficácia na síntese de

cDNA em regiões distantes da cauda poli-A das moléculas de RNAm. Neste estudo, esta medida foi adotada com a finalidade de minimizar o impacto de uma possível degradação das amostras de RNA. Os tubos contendo 14 µL de reação foram incubados a 65°C por 5 minutos e em gelo por um minuto de forma a estender a molécula de RNAm, desestabilizando sua estrtura secundária. Para início do processo de transcrição reversa, foram adicionados 4,0µL de tampão, 1,0 ul de DTT 0,1 M e 1,0µL de *SuperScript* III® 200 U/µL (Invitrogen Inc., Carlsbad, Califórnia), totalizando 20 µL de reação. Os tubos foram foram incubados a 25°C por 5 minutos, para anelamento dos iniciadores aleatórios, e a 50°C por 60 minutos, para síntese do cDNA. Finalmente, a reação foi incubada a 70°C durante 15 minutos para inativação da enzima de transcrição reversa.

3.11- RT-PCR em Tempo Real

Esta técnica foi realizada em equipamento *GeneAmp*® 7500 *Sequence Detector System* com o agente fluorescente *SybrGreen*® I (Applied Biosystems Inc., Foster City, CA). O método é baseado na detecção do produto de amplificação por meio da medida do aumento na fluorescência, conduzida pelo sistema óptico e detectada por equipamento digital que, em nosso caso, tratava-se de câmara CCD (amplitude de detecção de 530 a 580 nm). A fluorescência aumenta proporcionalmente à geração de produtos durante a reação, conseqüência da ligação do agente *SybrGreen*® I à fita dupla de DNA, o que causa aumento de emissão de luz em até 100 vezes para uma mesma concentração de *SybrGreen*® I livre em solução. O ciclo de PCR em que a fluorescência de determinada amostra atinge o chamado limiar de fluorescência, ou *treshold*, foi, em nosso caso, denominado CT. O *threshold* corresponde ao nível de fluorescência detectado inequivocamente acima do ruído de fundo. (GINZINGER, 2002). Dessa forma, o CT é inversamente proporcional ao número de cópias do segmento-alvo de DNA presentes no início da reação.

Por meio da análise das curvas de amplificação, tendo como parâmetros a intensidade de fluorescência (eixo y) *versus* o número de ciclos (eixo x), a linha de base - que representa o ruído de fundo - é definida de forma a abranger os ciclos de PCR nos quais o sinal fluorescente é detectado, mas está abaixo do limite de quantificação do

instrumento. Os parâmetros "linha de base" e *threshold* podem ser ajustados para cada amplicon, mas devem ser idênticos para todas as amostras na mesma corrida, tornando-as comparáveis. O *treshold*, por exemplo, deve ser posicionado na fase de acúmulo exponencial do produto amplificado, em que a eficiência da reação é máxima, permitindo quantificação mais precisa.

Em cada reação foram empregados 12,5 µL de SYBR® Green PCR Master Mix (Applied Biosystems Inc., Foster City, CA), 1,0 µL de cDNA e 150 a 300 nM de iniciadores em volume final de 25 µL. As condições de qPCR foram 50ºC por 2 minutos, 95ºC por 10 minutos, seguidos por 40 ciclos de amplificação de 95ºC por 15 segundos e um passo combinado de extensão e anelamento de 60^0C durante 1 minuto. Neste trabalho foram utilizados iniciadores disponíveis em nosso laboratório para os quais a reação de qPCR em tempo real já havia sido padronizada e as condições ideais para valores próximos de 100% (média de 99,97%) de eficiência da reação – incluindo a concentração dos iniciadores - foram determinadas. Na fase de padronização das reações, foi feita análise do produto de amplificação em gel de agarose. Reações controle, sem adição de cDNA, foram incluídas para cada conjunto de iniciadores. Ao final da reação de amplificação foi realizado ciclo de variação crescente de temperatura (de 60ºC a 95ºC) e monitoramento da fluorescência de forma a avaliar a curva de dissociação (Tm) do produto de PCR (*melting curve analysis*). Este resultado, expresso em graus Celsius, é característico de cada produto de PCR, dependente de seu tamanho e composição química, podendo revelar a presença de mais de um produto em determinada amostra, o que poderia corresponder à amplificação inespecífica ou à excessiva presença de interferentes na reação de amplificação, como dímeros de iniciadores que eventualmente concatenam-se naquelas condições, produzindo quimeras.

3.12- Medida da estabilidade gênica (M) e ranking dos genes controle

Para validação dos resultados, os valores de CT de cada amostra para cada gene controle foram transformados em quantidades (Q), valores de entrada do aplicativo geNorm (VANDESOMPELE et al, 2002), através da equação:

$$Q = E^{(CTmin - CTamostra)}$$

Onde “E” é a eficiência de amplificação (2 = 100%) e CTmin, o menor valor de CT entre as amostras, correspondente à amostra com maior expressão. A análise inicial das amostras demostrou que os genes tradicionalmente usados como referenciais (GAPDH, ACTB e B2M) possuíam grande variação, e assim sendo, optamos por testar todos os genes incluídos nos ensaios para verificar suas estabilidades. O programa geNorm também permite a determinação do número ótimo de genes (os mais estáveis, que minimizam a variação inter-ensaios) a serem utilizados para a quantificação relativa. Neste estudo, as variações par-a-par indicaram a necessidade de quatro genes como referenciais, resultando em uma normalização mais precisa e confiável, comparada ao uso de apenas um. Em seguida, foi realizado o cálculo da expressão normalizada de cada gene através da divisão entre o valor Q do gene de interesse e o fator de normalização calculado pelo geNorm.

Para as comparações entre tecido normal e tumoral foram utilizadas a média geométrica entre as amostras de tecido normal e a media geométrica entre as amostras de tecido tumoral. A razão de expressão tumor/normal foi comparada qualitativamente à razão de expressão diferencial evidenciada pela avaliação dos dados ORESTES.

4- RESULTADOS

4.1- Avalição da ferramenta estatística (LyM)

Para a comparação entre os fatores de expressão de cada programa (LyM e DGED), foram gerados gráficos, separados por tecido (condições normais e tumor). A comparação se deu pela razão entre os fatores diferenciais de expressão gênica para cada programa.

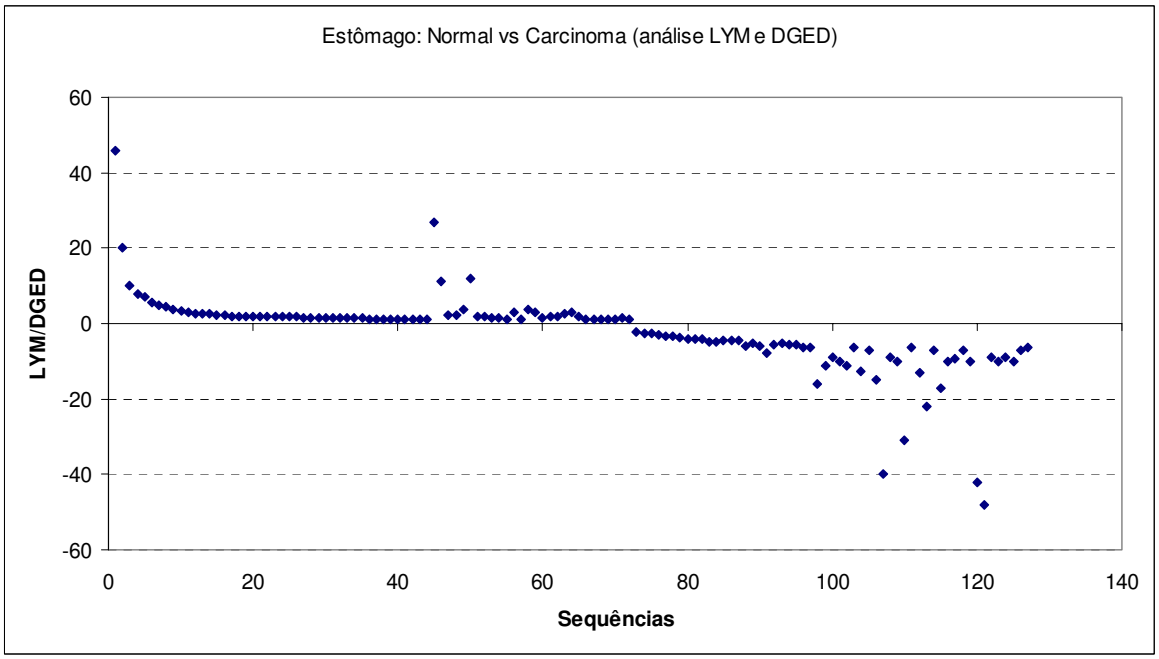


Figura 7- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre estômago normal e carcinoma de estômago.

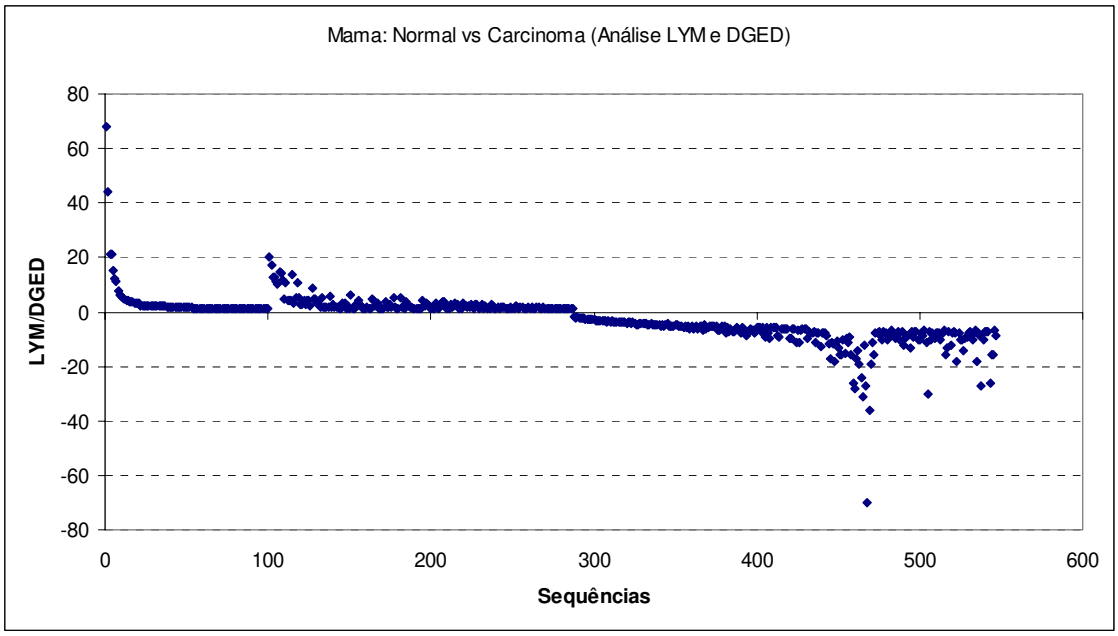


Figura 8- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre mama normal e carcinoma de mama.

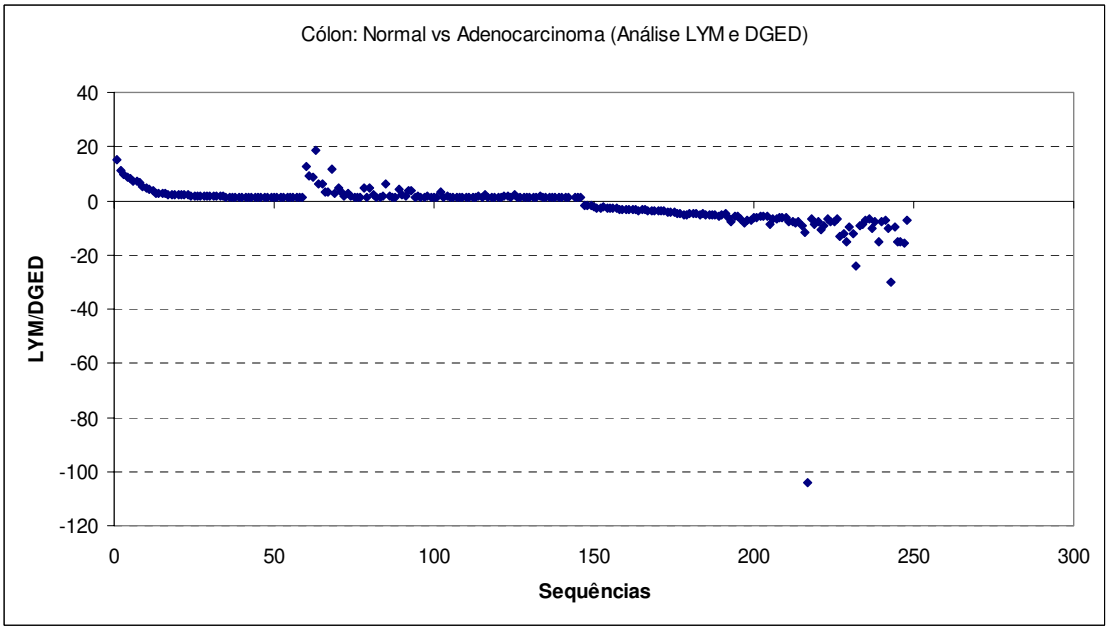


Figura 9- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre cólon normal e adenocarcinoma de cólon.

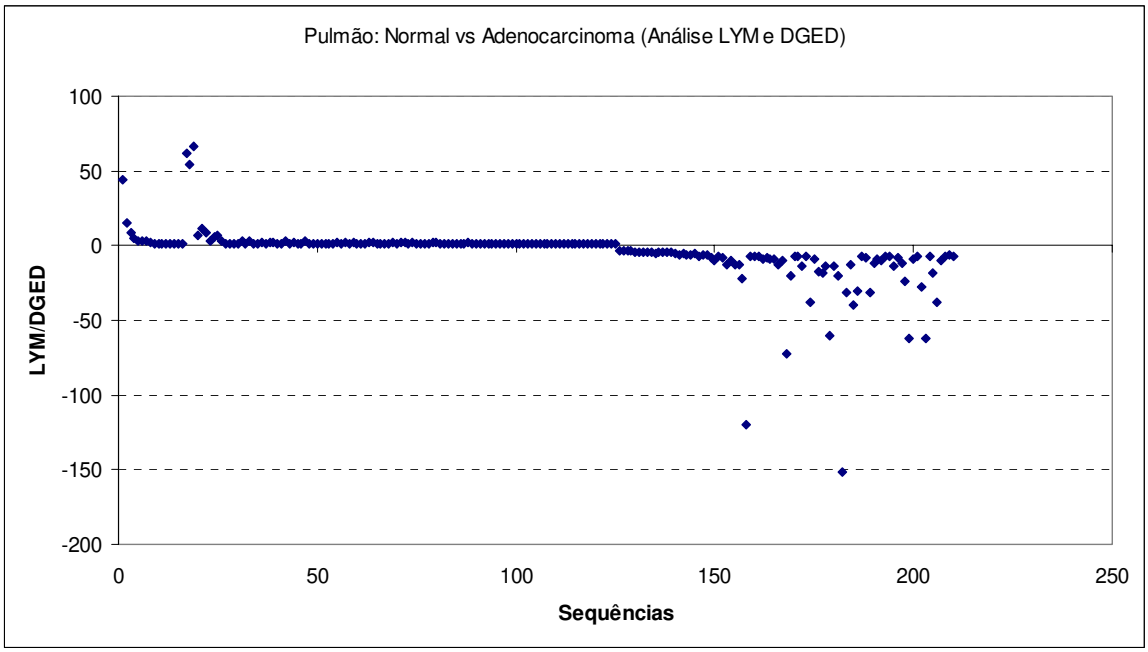
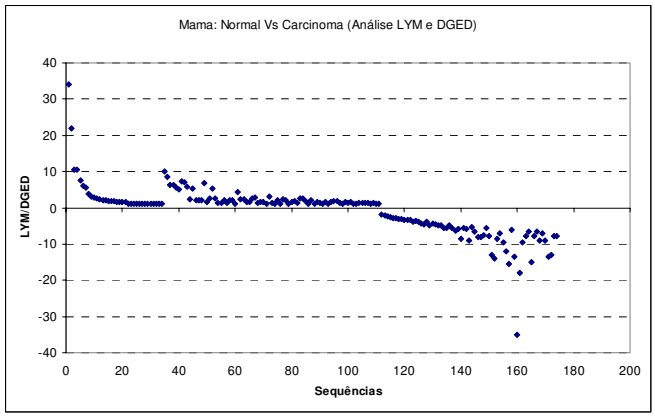
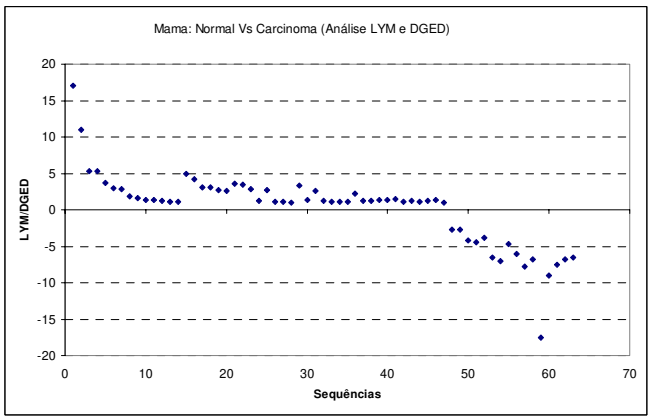


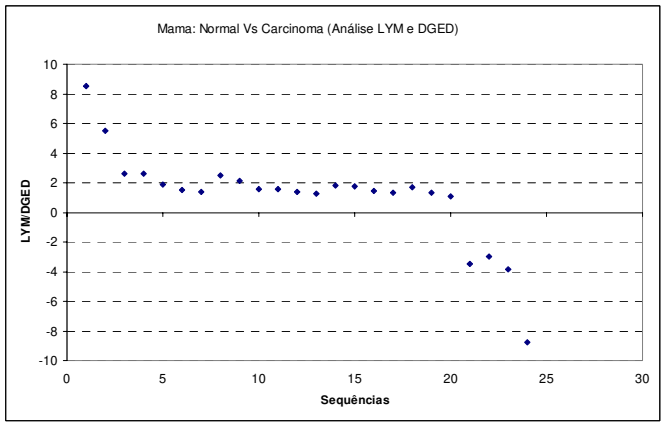
Figura 10- Resultados da comparação entre LyM e DGED (fator diferencial igual a dois) para a avaliação da diferença de expressão gênica entre pulmão normal e adenocarcinoma de pulmão.



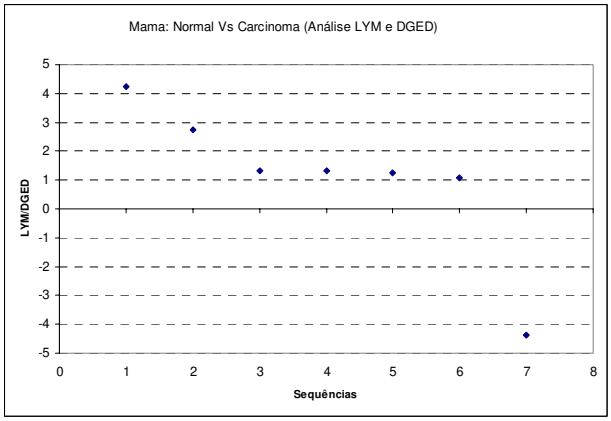
(a)



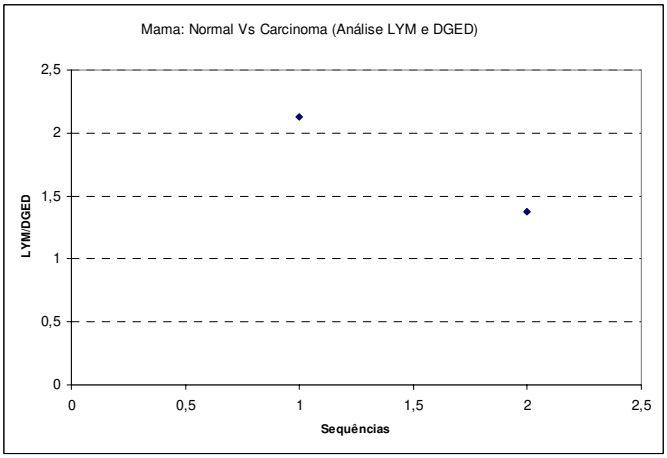
(b)



(c)



(d)



(e)

Figura 11- Resultados da comparação entre LyM e DGED para a avaliação da diferença de expressão gênica entre mama normal e carcinoma de mama, por meio dos fatores arbitrários 4 (a), 8(b), 16(c), 32(d) e 64(e).

4.2- Análises computacionais dos dados ORESTES

Após a etapa de montagem, os fragmentos ORESTES se agruparam conforme a tabela a seguir (Tabela 8):

Tabela 8- Número de seqüências consenso resultantes do processo de montagem.

Tecido	*Número de seqüências consenso*
mama normal	8528
mama tumoral	9746
cólon normal	7641
cólon tumoral	11121
cabeça e pescoço normal	1906
cabeça e pescoço tumoral	18661
pulmão normal	2670
pulmão tumoral	4858
sistema nervoso central normal	10262
sistema nervoso central tumoral	7334
próstata normal	3098
próstata tumoral	2975
estômago normal	1548
estômago tumoral	4569
testículo normal	3598
testículo tumoral	482
útero normal	3062
útero tumoral	4899

Estas seqüências consenso (Tabela 8) foram comparadas com os bancos de proteínas citados anteriormente por meio do programa BLAST. Após a análise por BLAST, o número de seqüências por proteína foi contabilizado, conforme arquivo HTML a seguir (Figura 12):

BLAST HIT	*NORMAL*	*TUMOR*
gb|AAA36138.1 Hermes antigen gp90 homing receptor precursor	4	2
emb|CAI41750.1 complement component 4A [Homo sapiens]	7	5
gb|AAL54607.1 cytochrome c oxidase subunit I [Homo sapiens]	190	12
gb|AAK17836.1 ATP synthase 6 [Homo sapiens] >gi|32348679|gb|AAP...	563	198
emb|CAB82418.1 hypothetical protein [Homo sapiens] >gi|11276938...	28	1
gb|AAW30001.1 structure protein NSP5b3a [Homo sapiens]	2	2
gb|AAH51800.1 CLTC protein [Homo sapiens]	2	4
gb|AAH23006.1 HSPCA protein [Homo sapiens]	61	2
ref|NP_112576.1 SH3 domain binding glutamic acid-rich protein l...	9	7
gb|AAP29640.1 envelope glycoprotein [Homo sapiens] >gi|30026466...	4	3
ref|NP_000601.3 CD44 antigen isoform 1 precursor [Homo sapiens]	4	1
emb|CAI41726.1 B-factor, properdin [Homo sapiens] >gi|55961819|...	2	2
gb|AAO25525.1 large anti-HSV-glycoprotein D single chain antibo...	5	5
gb|AAK17586.1 cytochrome c oxidase subunit I [Homo sapiens]	5	10
pir||T02955 probable cytochrome P450 monooxygenase - maize (frag...	543	488
ref|NP_996670.1 MORF-related gene 15 isoform 2 [Homo sapiens] >...	3	6
sp|Q37649 Cytochrome c oxidase polypeptide II >gi|343...	1	2
ref|NP_077182.1 NADH dehydrogenase (ubiquinone) 1, subcomplex u...		38
gb|AAP72967.1 glutathione S-transferase pi [Homo sapiens] >gi|5...		2
gb|AAH62583.1 MLL5 protein [Homo sapiens]		3
pir||T02995 unspecific monooxygenase (EC 1.14.14.1) - common tob...	5	493
gb|AAF29584.1 PRO0974 [Homo sapiens]		17
emb|CAB70757.2 hypothetical protein [Homo sapiens]		2

Contig	Reads	e-value
51	1	2e-10
52	1	1e-16
68	1	1e-06
76	1	1e-12
4878	3	6e-10
7723	10	3e-12
7872	12	2e-06
7980	14	1e-15
8096	17	7e-12
8129	18	7e-16
8321	37	3e-15
8444	80	8e-13
8469	108	2e-12
8507	240	2e-15

Figura 12- Resultado da análise por BLAST no tecido "mama". A lista possui uma descrição das proteínas presentes e seus respectivos níveis de expressão para tecido normal e tumoral. O quadro em detalhe mostra a distribuição das 543 seqüências ORESTES para a proteína "TO2955" em tecido normal (o termo "*Contig*" refere-se ao identificador da seqüência consenso e o termo "*Reads*" é o número de fragmentos ORESTES presentes em cada seqüência consenso).

Após o uso de filtros específicos, conforme apresentado na seção de métodos, as proteínas foram relacionadas aos seus respectivos genes. Para cada tecido, os genes resultantes (Tabela 9) foram submetidos ao programa LyM para avaliação estatística das diferenças de expressão entre as condições normal e tumoral. Após esta avaliação, foi gerada a classificação dos genes (conforme detalhado na seção 3.5) e o resultado fornecido em formato HTML (Figura 13). Os 100 primeiros genes presentes nesta classificação final, separados por tecido, estão presentes no Anexo 3.

Tabela 9- Número de genes, separados por tecido, após comparação estatística da expressão diferencial.

Tecido	*Número de Genes*
Mama	3835
Cólon	4266
Cabeça e pescoço	4637
Pulmão	2049
Sistema nervoso	3473
Prostata	1449
Estômago	1629
Testículo	1091
Útero	1984

Figura 13- Arquivo em formato HTML com os genes classificados. À direita, o menu "*Links*" para acesso à informações adicionais de cada gene.

Figura 14- Detalhamento de uma lista (pulmão) para o gene *HRAS*. Ele é o décimo elemento da classificação. A posição é indicada pelo número "10" e a cor verde indica que ele está mais expresso em tecido tumoral. Ao passar o *mouse* sobre a posição, aparece um texto (caixa amarelo claro) indicando: o valor de pontuação (86,275), o número de seqüências ORESTES para este gene em condição normal (0), o número de seqüências ORESTES para este gene em condição tumoral (5) e o fator de expressão diferencial calculado pelo programa LyM (4,25). Todas as demais informações formam obtidas por meio da ferramenta Gene Entrez.

Excluindo-se a redundância de genes entre os tecidos (Tabela 9), o número total de genes classificados é de 9261. Destes, 1015 (11%) participam de vias relacionadas ao fenômeno neoplásico (HAHN e WEINBERG, 2002). Notadamente, cerca de 52% deles apresentam fator de expressão diferencial (em módulo) maior do que cinco em tecido tumoral (Tabela 10). As vias que possuem genes diferencialmente expressos identificados pela análise de dados ORESTES foram destacadas (Figura 15).

Tabela 10- Distribuição da expressão gênica em vias relacionadas ao câncer. É apresentado o número de genes identificado nos elementos (genes, fatores ou complexos) presentes no circuito molecular (Figura 15), bem como a porcentagem de genes idenficados que possuem fator de expressão diferencial (em módulo) maior do que cinco.

Elementos	***Número de genes***		***(F/T)%***
	Total (T)	***|Fator| > 5 (F)***	
TGFB	18	13	72,2%
CYCE-CDk2	35	23	65,7%
PP2A	31	18	58,1%
RAS	38	22	57,9%
MYC	65	37	56,9%
E2Fs	78	43	55,1%
CYCD-CD4	40	22	55,0%
MEK	68	37	54,4%
MAPK	125	67	53,6%
AKT	147	77	52,4%
TGFBR	37	19	51,4%
Frizzled	6	3	50,0%
RSK	4	2	50,0%
GRB2-SOS	106	52	49,1%
WNT	92	43	46,7%
PI3K	90	41	45,6%
Dishevelled	16	5	31,3%
SMADs	7	2	28,6%

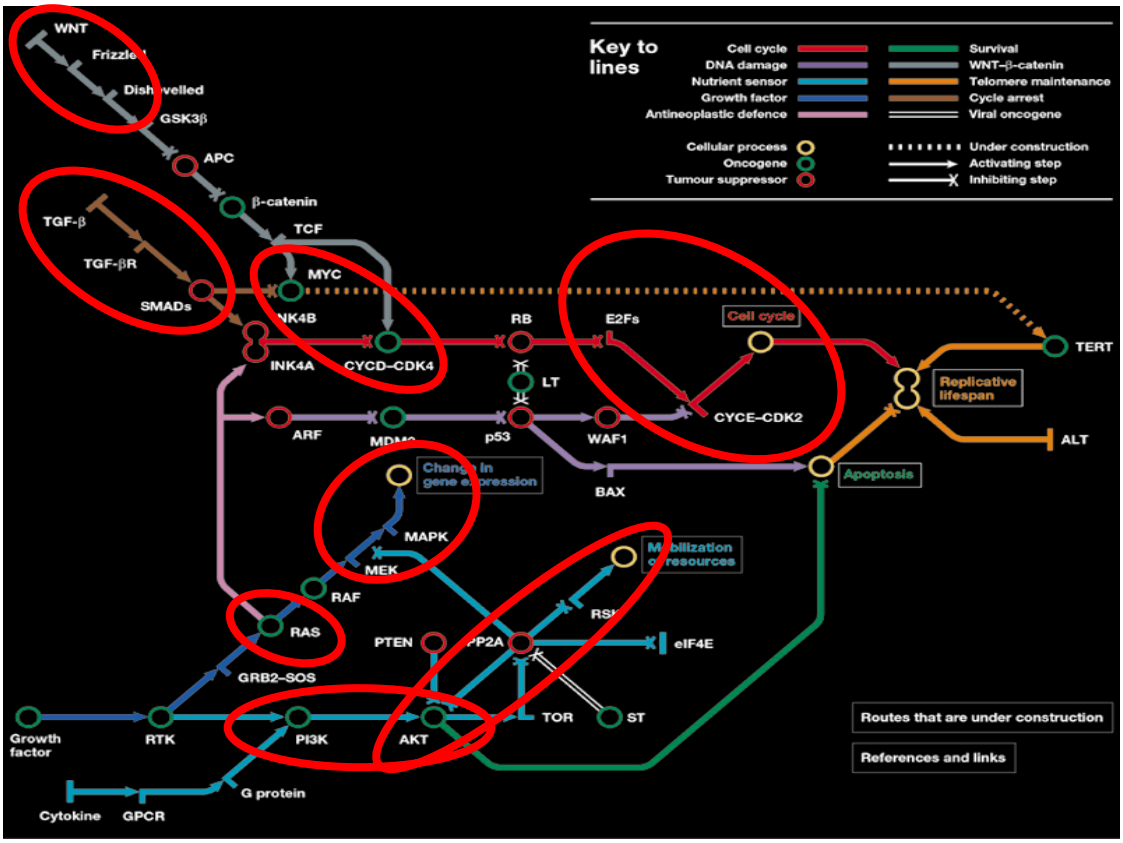


Figura 15- Vias relacionadas ao fenômeno neoplásico (HAHN e WEINBERG, 2002). Em destaque, trechos das vias deste circuito molecular para as quais foram identificados genes presentes na lista de genes produzida a partir de dados ORESTES.

4.3- Análises experimentais dos genes selecionados

A estabilidade de todos os genes foi testada entre as amostras, utilizando o programa geNorm. A seguir (Figura 16), é apresentado um gráfico com a estabilidade de cada gene:

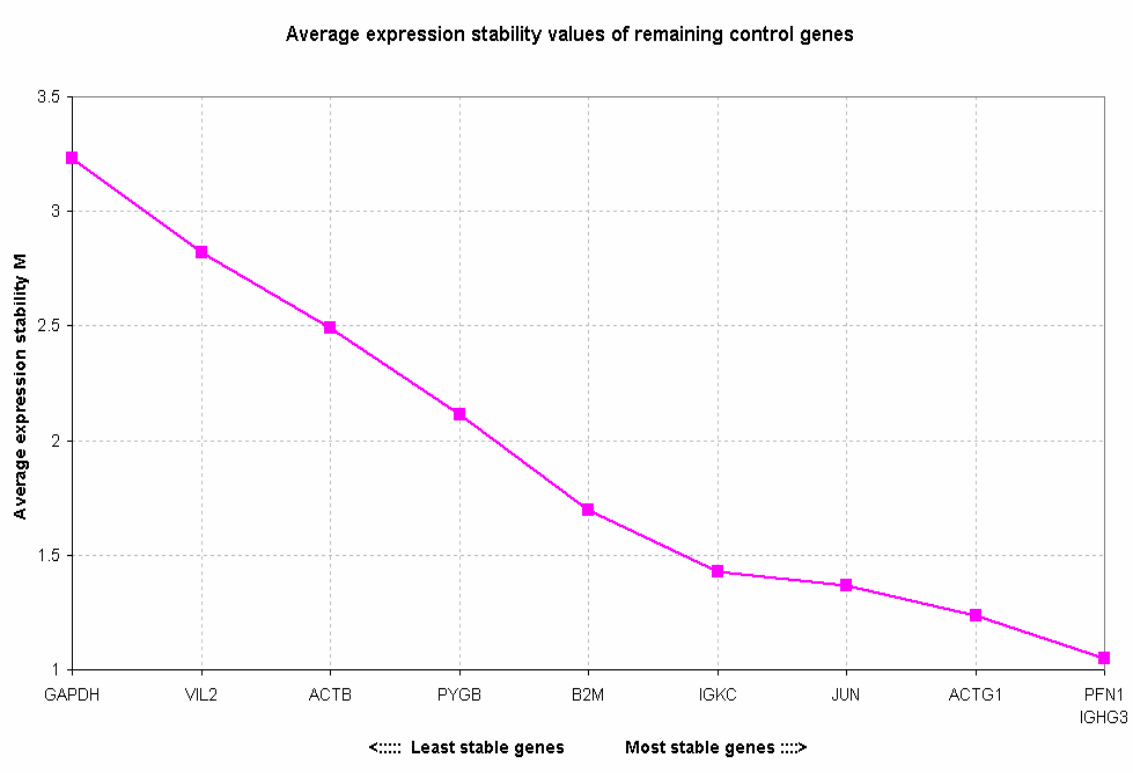
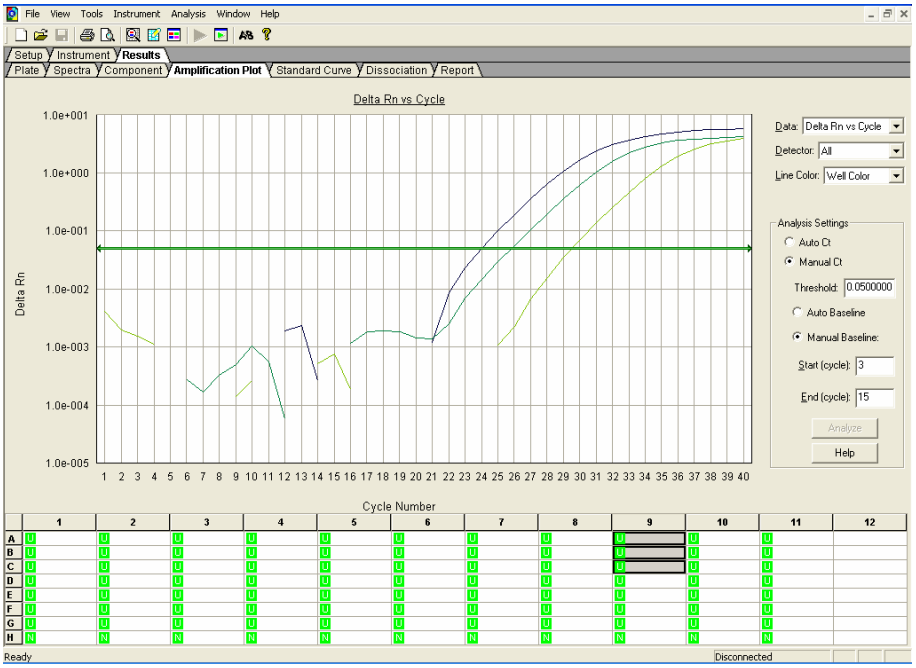
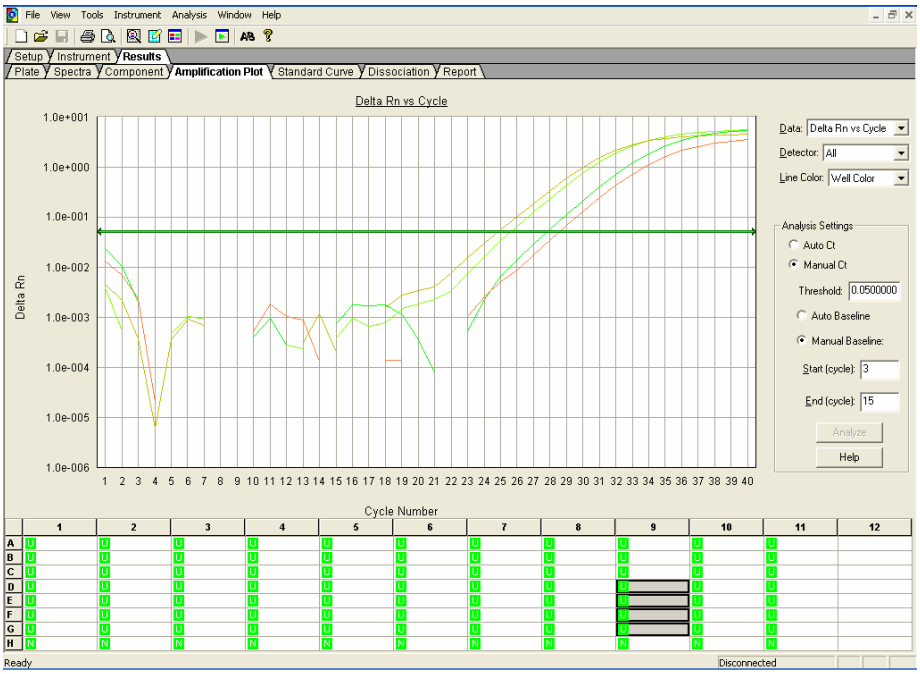


Figura 16- Estabilidade dos genes para as amostras estudadas em tecido normal e tumoral (estômago). A estabilidade média (M) é mostrada no eixo vertical. Os genes mais à direita (eixo horizontal) são os mais estáveis. Os genes de controle tradicionalmente utilizados (*GAPDH, ACTB e B2M*) não apresentaram as menores médias de estabilidade.

A instabilidade dos genes tradicionalmente utilizados como controle também foi constatada pela simples observação dos CTs das curvas de amplificação das amostras. Os CTs destes genes menos estáveis, como *GAPDH*, possuíam maior amplitude de variação entre amostras para um mesmo tecido do que os genes mais estáveis, como *IGHG3*:

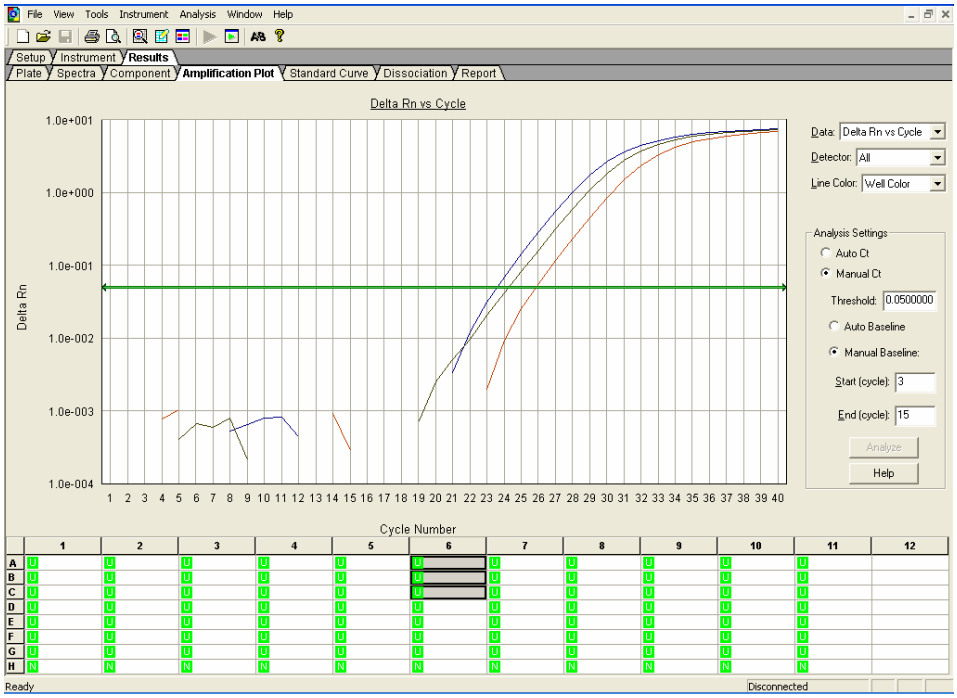


(a)

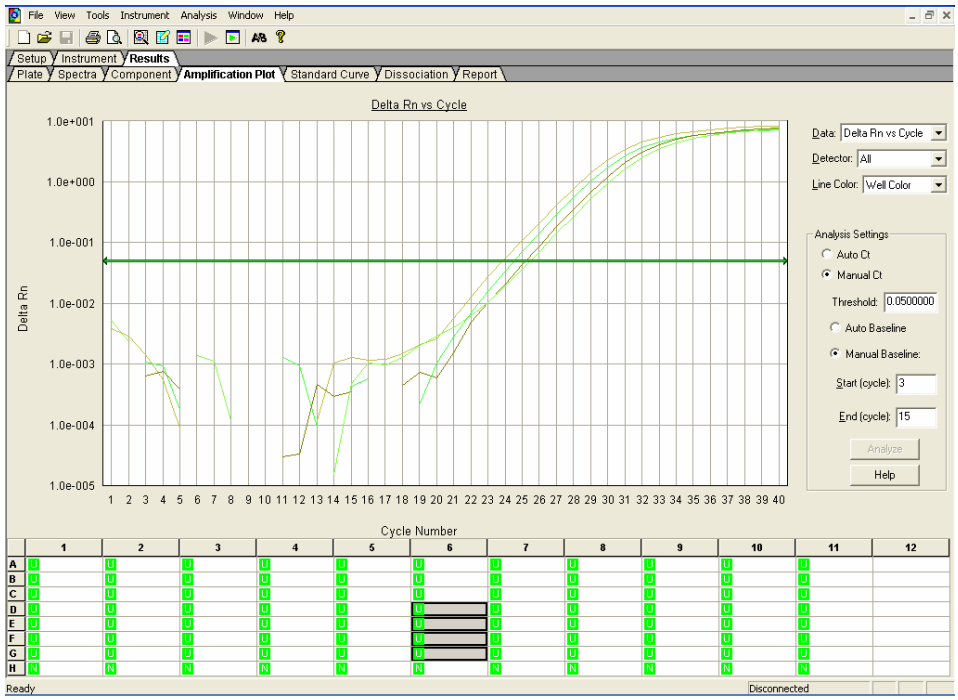


(b)

Figura 17- Curvas de amplificação do gene *GAPDH* para amostras normais (a) e tumorais (b). A variação entre as amostras pode ser observada através da variação dos CTs (intersecção entre a linha verde horizontal e a respectiva curva de amplificação).

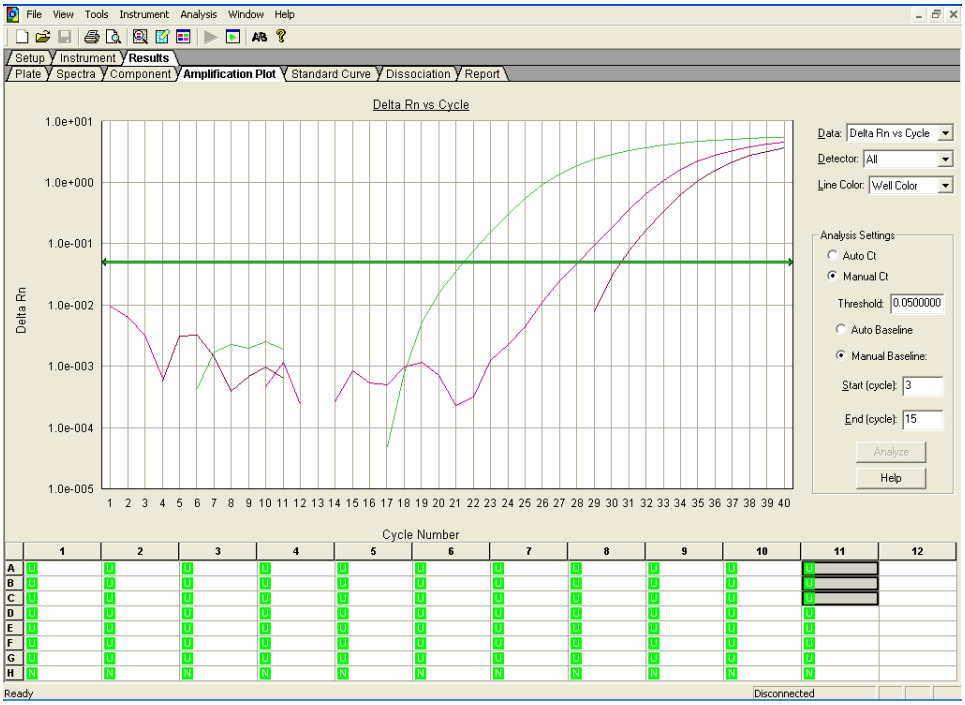


(a)

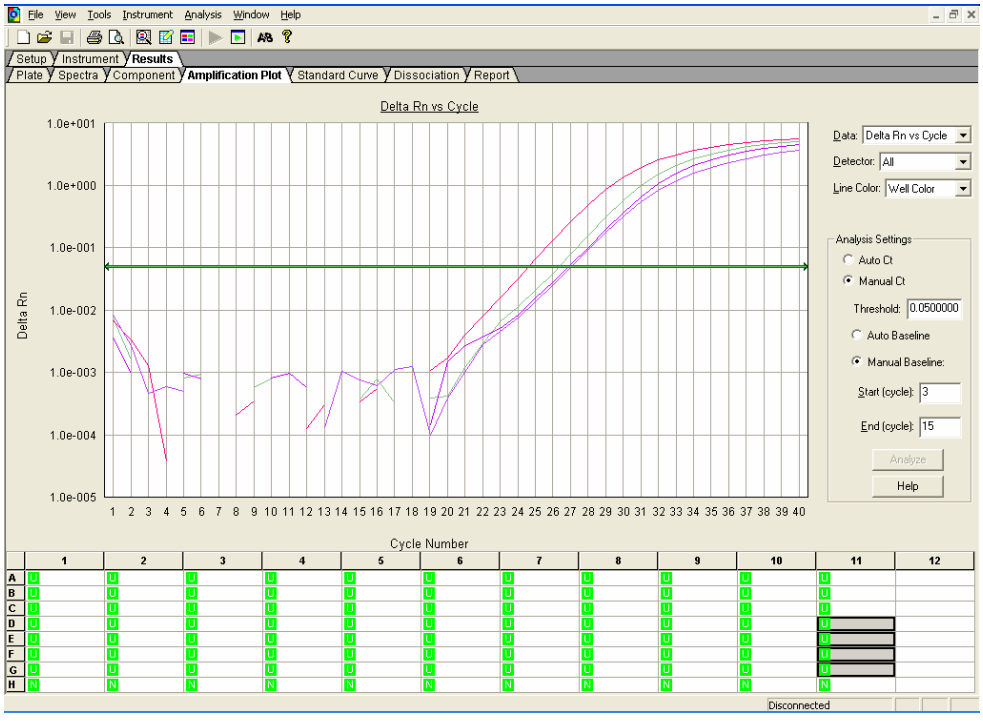


(b)

Figura 18- Curvas de amplificação do gene *IGHG3* para amostras normais (a) e tumorais (b). A variação entre as amostras, observada através da variação do CT, é mais discreta do que a variação de genes habitualmente utilizados como normalizadores como *GAPDH* (Figura 17).

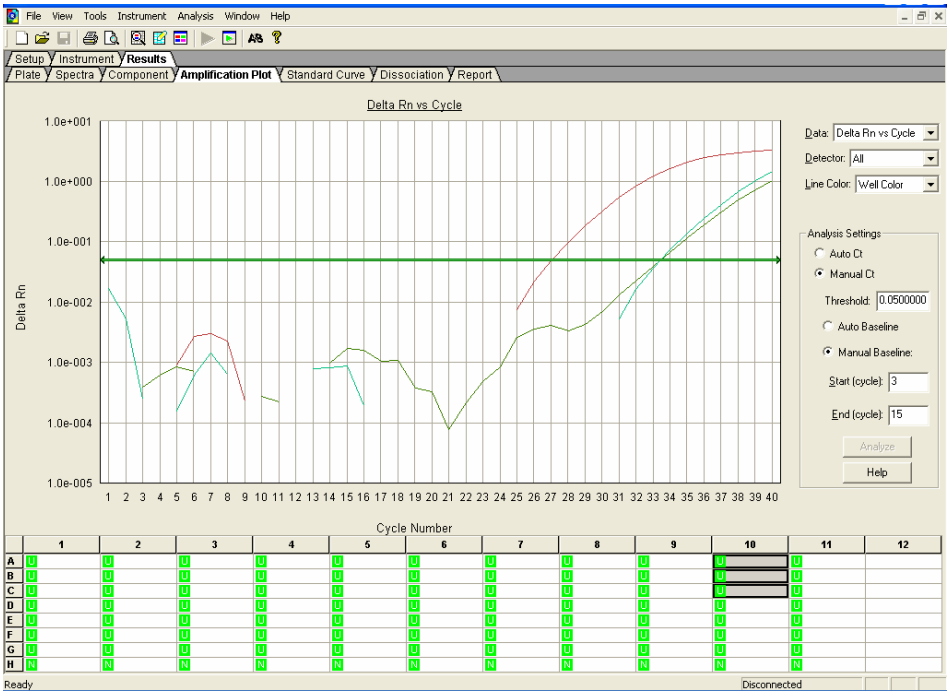


(a)

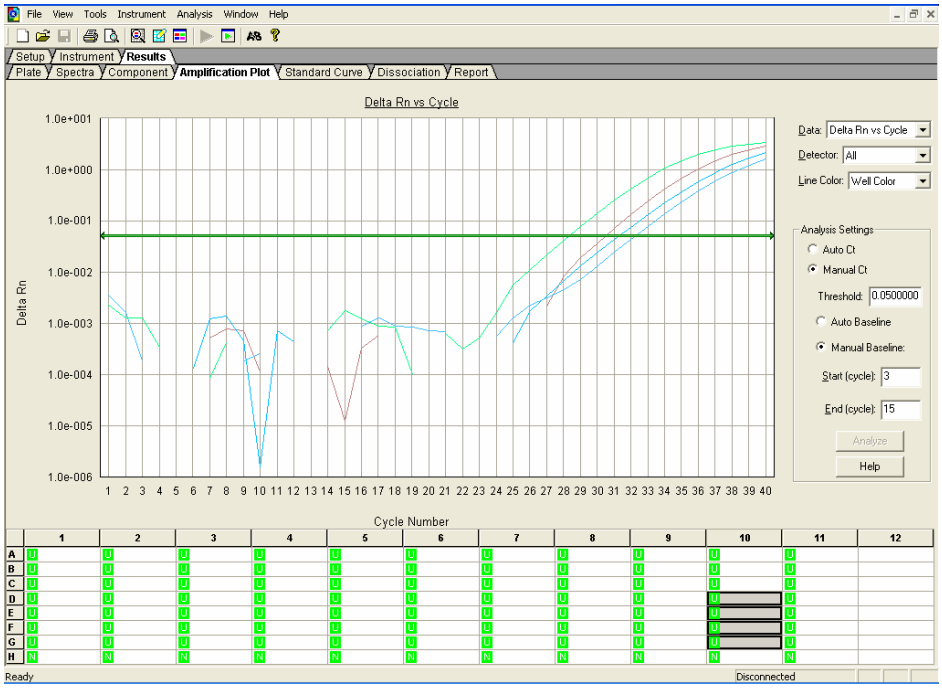


(b)

Figura 19- Curvas de amplificação do gene *B2M* para amostras normais (a) e tumorais (b). O gene *B2M* teve a maior estabilidade entre os genes de referência "padrão" que foram testados (*B2M, GAPDH e ACTB*).



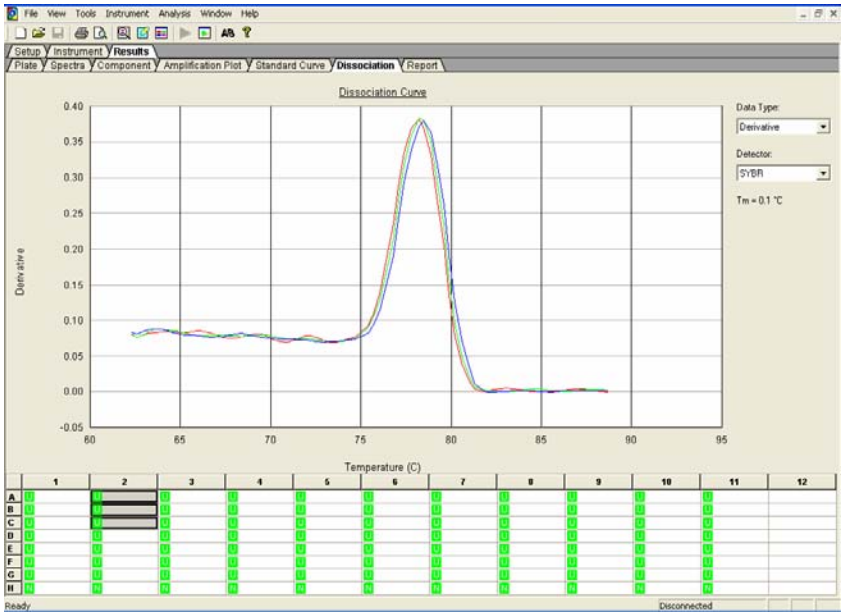
(a)



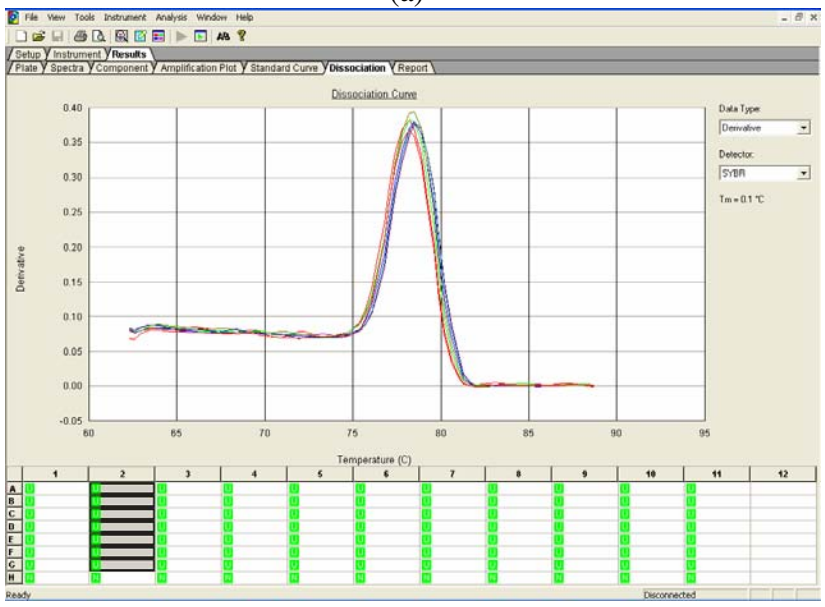
(b)

Figura 20- Curvas de amplificação do gene *ACTB* para amostras normais (a) e tumorais (b).

A qualidade das reações também foi verificada através da observação das curvas de dissociação de cada gene, confome exemplos a seguir:



(a)



(b)

Figura 21- Curvas de dissociação do gene *JUN* para amostras normais (a,b) e tumorais (b). O pico das curvas correspondem à temperatura (representada no eixo x) de dissociação do fragmento amplificado (*amplicon*). A morfologia e a coincidência das curvas em torno da temperatura esperada para o *amplicon* (neste caso cerca de 78°C) indica ausência de amplificação de produtos inespecíficos.

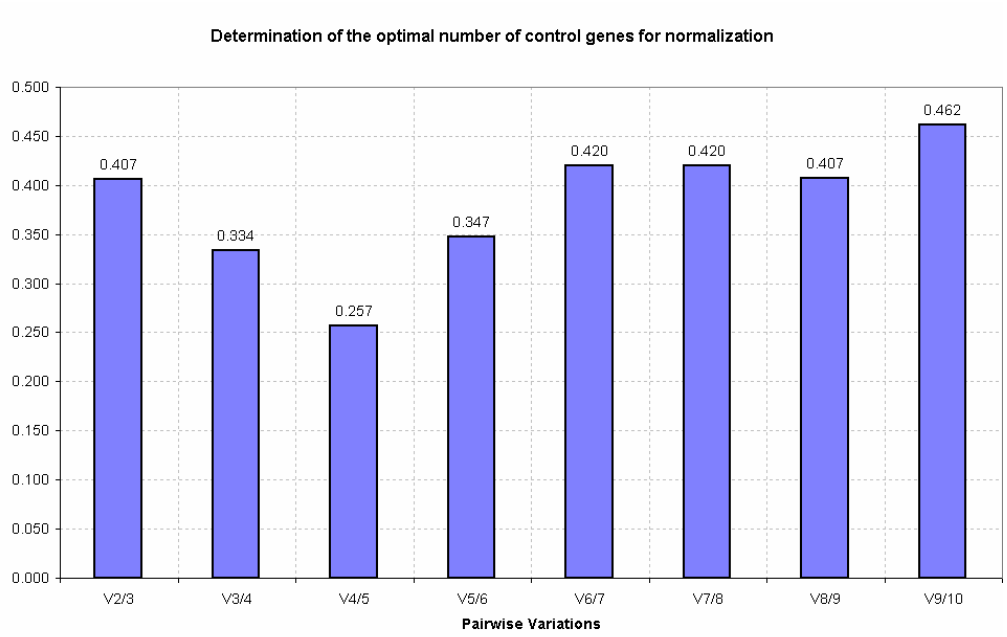


Figura 22- Variação par-a-par para o conjunto de genes avaliados. A menor variação encontra-se em $V_{4/5}$, indicando que o número ótimo de genes de controle para as amostras estudadas é quatro.

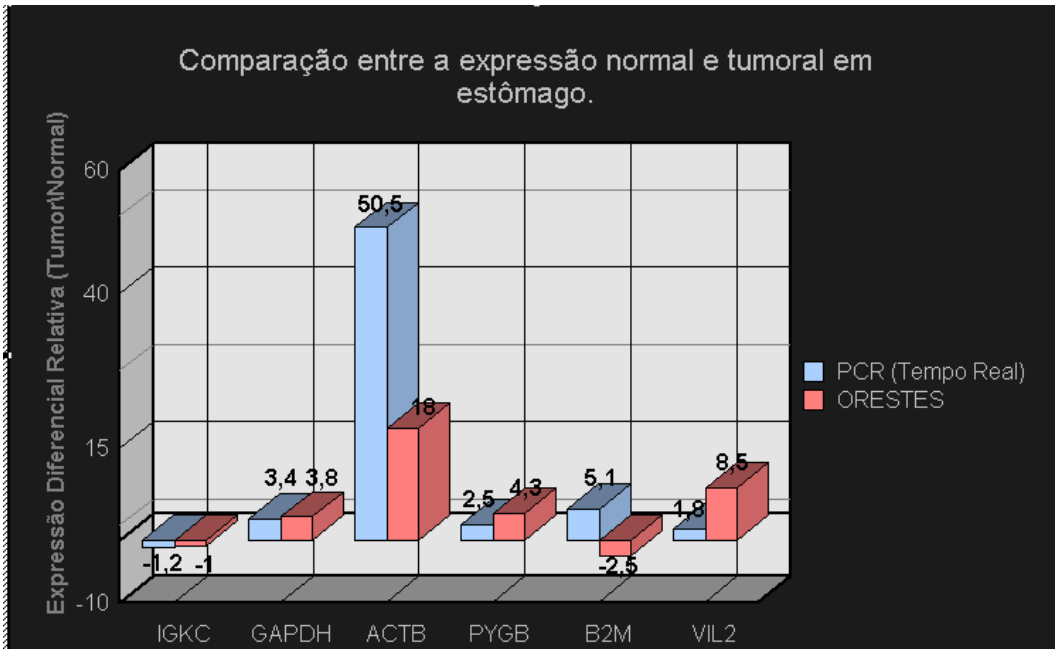


Figura 23- Expressão relativa (normal *vs* tumor) em estômago para a análise ORESTES e para a análise por RT-PCR em Tempo Real. Os genes utilizados como controle foram os quatro genes mais estáveis para as amostras utilizadas: *PFN1, IGHG3, ACTG1* e *JUN*.

5- DISCUSSÃO

A Bionformática se tornou, além de apoio, uma necessidade no entendimento de fenômenos biológicos considerando o volume de informações a serem interpretadas e a própria dinâmica e complexidade intrínseca a tais fenômenos. A capacidade computacional vem crescendo exponencialmente (SCHALLER, 1997), porém o potencial da informação gerada cresce a taxas ainda maiores. Neste contexto, novas ferramentas e abordagens para relacionar as informações relevantes ou novas formas de visualizar um conjunto de dados são extremanente importantes na descoberta científica. Este trabalho fez uso intensivo de ferramentas computacionais no apoio e entendimento do fenômeno neoplásico por meio da avaliação de dados gerados pelo Projeto Genoma Humano do Câncer. Foi apresentada uma metodologia pioneira que, a partir de mais de um milhão de fragmentos "desconhecidos", permitiu classificação de genes particularmente interessantes. Outras metodologias poderiam ser aplicadas, e o critério de "genes interessantes" variará de acordo com a ótica estruturada nos sucessivos passos da análise. O que caracteriza a presente abordagem é a abrangência dos fatores e critérios considerados, aliada à flexibilidade de ajustes permitidos, garantindo à ótica humana o poder analítico de ferramentas computacionais de processamento em larga escala.

A qualidade dos dados foi considerada desde a primeira etapa: ao invés de utilizar apenas as seqüências depositadas em bases públicas para montagem, este trabalho contou com o apoio do LBI para fornecimento dos arquivos de qualidade (phd) das seqüências. A primeira montagem foi baseada em critérios de qualidade bastante rígidos e o resultado foi conservador. A segunda montagem, realizada com parâmetros comumente usados, configurou-se na melhor relação custo/benífício, ratificando o fato de que os parâmetros-padrão são orientados a projetos de seqüenciamento em larga escala. Conforme apresentado (Tabela 3), as tarefas iniciais - nas quais o volume de dados ainda é grande - exigem esforço computacional considerável. No final da montagem, o número total de seqüências consenso (Tabela 8) era em média oito vezes menor do que o número original de fragmentos ORESTES (Tabela 1).

Na etapa de BLAST apenas as seqüências consenso foram consideradas. Como o objetivo deste trabalho envolvia principalmente a váriável "nível de expressão gênica", assumimos que as seqüências que não se agruparam em consensos representariam genes

distintos daqueles atribuídos às seqüências consenso (caso contrário estas seqüências teriam sido incluídas em algum consenso), com nível de expressão unitário. Assim sendo, a perturbação causada pela falta das seqüências "solitárias" não foi considerada como sendo relevante para a determinção do nível de expressão. Se as seqüências fossem incluídas, o número de falsos positivos para genes com expressão diferencial iria aumentar consideravelmente por causa da assimetria de tamanho das bibliotecas normal e tumoral de cada tecido, e poderia comprometer a identificação dos sinais biológicos relevantes. Porém, esta suposição tem reflexos discutidos posteriormente na análise global de expressão também utilizada como objeto de estudo deste trabalho.

Apesar de largamente utilizado, o BLAST é suscetível a resultados falsos positivos em parte por seu caráter heurístico. Em virtude desta característica este trabalho não utilizou o melhor "*blast hit*" (abordagem tradicional) para mapeamento de uma seqüência desconhecida. Para cada seqüência, foram utilizados todos os mapeamentos que obedecessem a um *e-value* mínimo. Além disso, as bases de dados públicas possuem grande quantidade de informação redundante e algumas vezes com qualidade heterogênea. Nomenclaturas distintas para o mesmo gene ou proteína também são comuns. Dadas estas limitações, todos os *blast hits* que obedeciam ao critério de *e-value* anteriormente citado foram utilizados ao invés do uso do melhor *hit* (ENGELHARDT et al, 2005). Isto gerou redundância adicional que teve que ser considerada posteriormente durante o mapeamento proteína-gene.

Para avaliação estatística do fator diferencial de expressão gênica foi desenvolvido o programa LyM. Seu desenvolvimento foi motivado por dois fatores: a ferramenta disponível (DGED) não permitia a entrada de dados (atualmente isto é possível apenas para bibliotecas SAGE) e a lógica da ferramenta é focada na escolha arbitrária de fatores e não na confiabilidade, o que a torna extremamente limitada. O desenvolvimento de LyM inverteu este conceito, mudando o foco para a confiabilidade até encontrar o melhor fator que respeite este limite. Os gráficos apresentados (Figura 11) mostram que as duas ferramentas convergem para o mesmo resultado, porém para o DGED são necessárias múltiplas interações manuais. Nos demais gráficos (Figuras 7-10), os pontos acima e abaixo do valor "um" demonstram a informação diferencial não capturada pelo DGED numa

primeira interação (fator arbitrário igual a dois), ou seja, demonstra fatores diferenciais subestimados para vários genes. Para exemplificar, pode-se considerar um gene que para LyM possui fator diferencial de 50 com p = 0,05. Este mesmo gene para DGED possui p=0,00 para o fator difencial arbitrário de dois ou, ainda, p=0,01 para o fator diferencial arbitrário de oito. Na convergência: p=0,05 para o fator diferencial arbitrário de 50. Convém lembrar que a escolha de cada um dos fatores é um processo manual no DGED. A ferramenta LyM se mostrou bastante rápida em decorrência de sua busca binária. Além disso, o "curto-circuito" nas operações de multiplicação por zero e a reutilização de valores calculados tiveram papel de destaque na melhoria do desempenho.

No mapamemento das proteínas com seus respectivos genes, foi utilizado outro filtro para palavras-chave, eliminando seqüências ribossomiais, seqüências desconhecidas e seqüências desconhecidas oriundas de projetos de seqüenciamento em larga escala, pelos motivos de qualidade discutidos anteriormente. Esta abordagem diminuiu o número de genes presentes no estudo, porém a avaliação de elementos pouco conhecidos que podem estar envolvidos com o câncer foi considerada por meio das seqüências hipotéticas que foram utilizadas.

Após o mapeamento, a literatura de câncer foi utilizada na classificação para permitir que genes conhecidos no processo neoplásico pudessem ser considerados. Convém lembrar que a expressão gênica é um processo complexo e não se pode atribuir todas as diferenças de expressão encontradas ao fenômeno neoplásico. Outras dinâmicas clínicas e fisiopatológicas estão envolvidas com alterações de expressão gênica. Os pesos para a classificação foram ajustados para que o fator de expressão gênica fosse o principal componente da pontuação de um gene e não a literatura. Evitamos a tendência de acreditar que a presença de um gene na literatura está relacionada a expressão diferencial considerável, visto que grande parte destes resultados foram evidenciados por técnicas distintas – e muitas vezes pouco comparáveis – de análise. Neste caso, a identificação da expressão diferencial de um gene já descrito na literatura acaba contribuindo para evidenciar ainda mais a sua relevância biológica. A influência de outros trabalhos é fundamental, dado que todas as publicações científicas têm viés explícito em suas respectivas seções de referência.

A identificação dos genes não redundantes entre os tecidos com vias relacionadas ao câncer (HAHN e WEINBERG, 2002) revelou que aproximadamente metade deles estava com fator de expressão diferencial (em módulo) superior a cinco em várias partes destas vias. Nas demais, os genes não possuíam expressão diferencial significativa ou simplesmente não foram localizados, provavelmente pela eliminação das seqüências que não se agruparam na etapa de montagem dos fragmentos ORESTES.

Para validação biológica, amostras de seis pacientes foram selecionadas e dez genes representados em tecido normal e tumoral de estômago (*ACTB, ACTG1, B2M, GAPDH, IGKC, IGHG3, JUN, PFN1, PYGB* e *VIL2*) foram submetidos à técnica de PCR quantitativo em tempo real. Esta técnica é robusta para avaliação de expressão gênica e a confiabilidade de seus resultados depende tanto da padronização do ensaio (desenho de iniciadores, padronização de reações, distribuição das amostras na placa) quanto da intepretação dos dados (análise de curvas de amplificação e dissociação, definição da quantidade de genes de referência normalizadores, análise de CTs e amostras). Neste trabalho, optamos por utilizar o algoritmo do geNorm para seleção de vários genes de referência em detrimento do uso de um único normalizador. O geNorm realiza uma análise consistente baseada na média geométrica dos genes normalizadores e, em nosso caso, evidenciou que os melhores normalidores não eram os genes tradicionalmente utilizados. Este é um dado importante que reforça a idéia de que um único normalizador "padrão" produz um resultado com menor confiabilidade. Dos seis genes resultantes para avaliação, cinco deles (*ACTB, GAPDH, IGKC, PYGB* e *VIL2*) apresentaram conformidade qualitativa com os dados previstos pela análise ORESTES (Figura 23).

6- CONCLUSÕES

1. A ferramenta LyM mostrou possuir compatibildade com os resultados da ferramenta DGED com a vantagem de ser orientada à confiabilidade, evitando multiplas interações manuais. Além disso, a busca binária mostrou-se um diferencial importante. Sem o desenvolvimento da ferramenta, a avaliação dos níveis de expressão gênica seria inviável neste trabalho.

2. Foram identificados 9261 genes diferencialmente expressos em tumores de mama, cólon, cabeça e pescoço, pulmão, sistema nervoso central, próstata, estômago, testículo e útero utilizando a análise de dados ORESTES.

3. A análise de algumas vias relacionadas ao fenômeno neoplásico revelou diferenças de expressão com fator (em módulo) maior do que cinco em 52% dos genes que participavam destas vias.

4. Foram selecionados dez genes para confirmação dos achados ORESTES com os dados biológicos. Destes, quatro genes foram utilizados como normalizadores e cinco tiveram concordância qualitativa com a previsão computacional.

7- REFERÊNCIAS BIBLIOGRÁFICAS

ALIZADEH, A. A. e EISEN, M. B. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 403 503-511, 2000.

ALTSCHUL, S. F., GISH, W., MILLER, W. et al. Basic local alignment search tool. J Mol Biol. 215(3): 403-410, 1990.

ASHBURNER, M., BALL, C. A, e BLAKE, J. A. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. 25(1): 25-29, 2000.

AUDIC, S. e CLAVERIE, J.-M. The significance of digital gene expression profiles. Genome Res. 7 986-995, 1997.

BACHMANN, P. Die Analytische Zahlentheorie. Analytische Zahlentheorie. 2, 1894.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J. et al. GenBank. Nucleic Acids Res. 33(Database): D34-D38, 2005.

BOLSTAD, W. M. Introduction to Bayesian Statistics. John Wiley. ISBN 0-471-27020-2, 2004.

BONALUME, N. R. Brazilian scientists team up for cancer genome project. Nature. 398(6727): 450, 1999.

BREDEL, M., BREDEL, C., JURIC, D. et al. Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. Cancer Res. 65(19): 8679-8689, 2005.

BUTTE, A. J., TAMAYO, P., SLONIM, D. et al. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc.Natl.Acad.Sci.U.S.A. 97(22): 12182-12186, 2000.

CAMARGO, A. A. et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc.Natl.Acad.Sci.U.S.A. 98(21): 11837-11838, 2001.

CHEN, H., CENTOLA, M., e ALTSCHUL, S. F. et al. Caracterization of gene expression in resting and activated mast cells. Proc.Natl.Acad.Sci.U.S.A. 188 1657-1668, 1998.

CHOMCZYNSKI, P. e SACCHI, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 162(1): 156-159, 1987.

DIAS, N. E., GARCIA, C. R., e ALMEIDA, S. V. et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc.Natl.Acad.Sci.U.S.A. 97 3491-3496, 2000.

DOOLITTLE, R. F. et al. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science. 221 275-277, 1983.

DORKELD, F., BERNHEIM, A., DESSEN, P. et al. A database on cytogenetics in haematology and oncology. Nucleic Acids Res. 27(1): 353-354, 1999.

ENGELHARDT, B. E, JORDAN, M. I., MURATORE, K. E. et al. Protein molecular function prediction by bayesian phylogenomics. PLoS Comput Biol. 1(5): e45, 2005.

EWING, B. e GREEN P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3): 186-194, 1998.

FUTREAL, P. A., COIN, L., MARSHALL, M. et al. A census of human cancer genes. Nat Rev Cancer. 4(3): 177-183, 2004.

GIBAS, C. e JAMBECK, P. Developing Bioinformatics Computer Skills. O'Reilly press. ISBN 1-565-92664-1, 2001.

GINZINGER, D. G. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol. 30(6): 503-512, 2002.

GOLUB, T. R. e SLONIM, D. K. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286 531-537, 1999.

HAHN, W. C. e WEINBERG, R. A. Modelling the molecular circuitry of cancer. Nat Rev Cancer. 2(5): 331-341, 2002.

HANAHAN, D. e WEINBERG, R. A. The hallmarks of cancer. Cell. 100(1): 57-70, 2000.

HARRIS, H. Cell fusion and the analysis of malignancy. Proc.R.Soc.Lond.B Biol.Sci. 179 1-20, 1971.

HINTON, G. e SEJNOWSKI, T. J. Unsupervised Learning and Map Formation: Foundations of Neural Computation. MIT Press. 1999.

HONTS, J. E. Evolving Strategies for the Incorporation of Bioinformatics Within the Undergraduate Cell Biology Curriculum. Cell Biol Educ. 2 233-243, 2003.

HUANG, X. e MADAN, A. CAP3: A DNA sequence assembly program. Genome Res. 9(9): 868-877, 1999.

HUERTA, M., DOWNING, G., SETO, B. et al. Nih Working Definition Of Bioinformatics And Computational Biology. NIH WORKING DEFINITION. 2000.

IHAKA, R. e GENTLEMAN, R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 5 299-314, 1996.

JOHNSON, S. C. Hierarchical clustering schemes. Psychometrika. 32(3): 241-254, 1967.

KLEIHUES, P. e SOBIN, L. H. World Health Organization classification of tumors. 88(12): 2887-2887, 2000.

KNUTH, D. The Art of Computer Programming. Addison-Wesley. ISBN 0-201-89685-0, 1969.

KOHONEN, T. D. Self-organized formation of topologically correct feature maps. Biological Cybernetics. 43(1): 59-69, 1982.

KUMAR, R., GURURAJ, A. E., VADLAMUDI, R. K. et al. The clinical relevance of steroid hormone receptor corepressors. Clin Cancer Res. 11(8): 2822-2831, 2005.

LAL, A., LASH, A. E., e ALTSCHUL, S. F. et al. A Public Database for Gene Expression in Human Cancers. Cancer Res. 59 5403-5407, 1999.

LASH, A. E., TOLSTOSHEV, C. M., WAGNER, L. et al. SAGEmap: A Public Gene Expression. Genome Res. 10(7): 1051-1060, 2000.

LAURELL, H., BOUISSON, M., e BERTHELEMY, P. et al. Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation with gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples. World J Gastroenterol. 12(21): 3344-3351, 2006.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions and reversals. Sov.Phys.Dokl. 6 707-710, 1966.

LEVINSON, A. D., OPPERMANN, H., LEVINTOW, L. et al. Evidence that the transforming gene of avian sarcoma virus encodes a protein kinase associated with a phosphoprotein. Cell. 15 561-572, 1978.

LEWIN, B. Gene Expression. John Wiley & Sons Inc. ISBN 0-471-01976-3, 1980.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1 281-297, 1967.

MONTELIONE, G. T. e ANDERSON, S. Structural genomics: keystone for a Human Proteome Project. Nature Structural Biology. 6 11-12, 1999.

MUGGERUD, A. A., JOHNSEN, H., e BARNES, D. A. et al. Evaluation of MetriGenix custom 4D arrays applied for detection of breast cancer subtypes. BMC Cancer. 6(59): 2006.

NEEDLEMAN, S. B. e WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3): 443-53, 1970.

NILSSON, J. A. e CLEVELAND, J. L. Myc pathways provoking cell suicide and cancer. Oncogene. 22(56): 9007-9021, 2003.

NOWELL, P. C. The clonal evolution of tumor cell populations. Science. 194 23-28, 1976.

NOWELL, P. C. e HUNGERFORD, D. A. A minute chromosome in human chronic granulocytic leukemia. Science. 132 1488-1501, 1960.

OGATA, H, GOTO, S, SATO, K et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27(1): 29-34, 1999.

PEARSON K. On the Influence of Past Experience on Future Expectation. Philosophical Magazine. 13 365-378, 1907.

PERES, T. S., MACHADO, T. F., COSTA, F. F. et al. LyM: a new tool to reach the best fold in gene expression comparison. X-Meeting 2005 (Anals). 2005.

POP, M. e KOSACK, D. Using the TIGR assembler in shotgun sequencing projects. Methods Mol Biol. 255 279-294, 2004.

PORTER, D. A., KROP, I. E., e NASSER, S. et al. A SAGE (Serial Analysis of Gene Expression) View of Breast Tumor Progression. Cancer Res. 61 5697-5702, 2001.

REINHOLD, W. C., KOUROS-MEHR, H., e KOHN, K. W. et al. Apoptotic susceptibility of cancer cells selected for camptothecin resistance: gene expression profiling, functional analysis, and molecular interaction mapping. Cancer Res. 63(5): 1000-1011, 2003.

RIDLEY M. Nature via Nurture: Genes, Experience, & What Makes Us Human. Harper Collins. ISBN 0-06-000678-1, 2003.

SAEED, A. I., SHAROV, V., e WHITE, J. et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 34(2): 374-378, 2003.

SAKAMURO, D. e PRENDERGAST, G. C. New Myc-interacting proteins: a second Myc network emerges. Oncogene. 18(19): 2942-2954, 1999.

SCHALLER, R. R. Moore's law: past, present and future. IEEE Spectrum. 34(6): 52-59, 1997.

SCHENA, M., SHALON, D., DAVIS, R. W. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270(5235): 467-470, 1995.

SHANNON C.E. e WEAVER W. The Mathematical Theory of Communication. University of Illinois Press. ISBN 0-252-72548-4, 1949.

SOGAYAR, M. C., CAMARGO, A. A., e BETTONI, F. et al. A transcript finishing initiative for closing gaps in the human transcriptome. Genome Res. 14(7): 1413-1423, 2004.

TAMAYO, P., SLONIM, D., e MESIROV, J. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc.Natl.Acad.Sci.U.S.A. 96(6): 2907-2912, 1999.

THYKJAER T, WORKMAN C, KRUHOFFER M et al. Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res. 61(6): 2492-2499, 2001.

VAN'T VEER, L. J., DAL, H., e VAN DE VIJVER, M. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 15 530-526, 2002.

VANDESOMPELE, J., DE PRETER, K., PATTYN, F. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3(7): RESEARCH0034.1-RESEARCH0034.11, 2002.

VINTSYUK T.K. Recognition of spoken words by the dynamic programming method. Kibernetika. 4(1): 81-88, 1968.

VON HANSEMANN, D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. Virchow's Arch.Path.Anat. 119 299, 1890.

WADA, A., YAMAMOTO, M., e SOEDA, E. Automatic DNA sequencer: computer-programmed microchemical manipulator for the Maxam-Gilbert sequencing method. Rev Sci Instrum. 54(11): 1569-1572, 1983.

WATERFIELD, M. D. et al. Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. Nature. 304 35-39, 1983.

WELSH, J. B., ZARRINKAR, P. P., e SAPINOSO, L. M. et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc.Natl.Acad.Sci.U.S.A. 98(3): 1176-1181, 2001.

WESTON, A. D. e HOOD, L. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. J Proteome Res. 3(2): 179-196, 2004.

WHEELER, D. L., BARRETT, T., e BENSON, D. A. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33(Database): D39-D45, 2005.

WHITE, R. J. RNA polymerase III transcription — a battleground for tumor suppressors and oncogenes. Eur J Cancer. 40(1): 21-27, 2004.

WIBOM, C., PETTERSSON, F., e SJOSTROM, M. Protein expression in experimental malignant glioma varies over time and is altered by radiotherapy treatment. Br J Cancer. 94(12): 1853-1863, 2006.

WORLD HEALTH ORGANIZATION AND INTERNATIONAL UNION AGAINST CANCER, 2005. Global Action Against Cancer - Updated Version. WHO Press. 2005.

YEUNG, K. Y. e RUZZO, W. L. Principal component analysis for clustering gene expression data. Bioinformatics. 17(9): 763-774, 2001.

8- ANEXOS

ANEXO 1

Termo de Informação e Consentimento

CENTRO DE HEMATOLOGIA E HEMOTERAPIA DA UNICAMP
TERMO DE INFORMAÇÃO E CONSENTIMENTO

Para obter um maior conhecimento clínico e científico do câncer, o Corpo Clínico deste Centro (médicos e pesquisadores) desenvolve pesquisa clínica científica. Através dessa pesquisa é possível conhecer melhor os mecanismos da doença e, portanto, oferecer novas possibilidades de diagnóstico e tratamento.

O presente trabalho envolve o entendimento dos genes diferencialemnte expressos em vários tumores com o objetivo de selecionar potenciais candidatos à terapêutica, alvo de drogas, diagnóstico e seguimento, cujo pesquisador responsável é Tarcísio de Souza Peres sob orientação do Prof. Dr. Fernando Lopes Alberto. Estes estudos são realizados em fragmentos de tumores removidos em cirurgias, ou em material biológico colhido.

Você está sendo admitido(a) neste Centro para estabelecimento de diagnóstico e/ou tratamento de alguma forma de tumor . Para fins de diagnóstico, fator prognóstico e/ou como parte de seu tratamento, há a necessidade da remoção do tumor e/ou material biológico retirado, para exames clínicos laboratoriais, necessários para um diagnóstico definitivo. O restante do tumor ou material biológico retirado é não utilizado será congelado e armazenado para novos exames se necessários. Caso contrário será descartado, conforme Legislação Sanitária regulamentar sobre o assunto.

A obtenção e o estudo dos referidos fragmentos de tumor e material biológico não implicarão em riscos adicionais no seu tratamento ou extensão do mesmo. O material biológico armazenado e utilizado neste estudo poderá ser envolvido em futuras pesquisas apenas após nova submissão e aprovação do Comitê de Ética em Pesquisa da FCM/UNICAMP.

O fragmento de tumor e/ou material biológico será identificado no laboratório por um código formado por números e letras e, portanto, sua privacidade e identidade serão sempre preservadas. A eventual inclusão dos resultados em publicação científica será feita de modo a manter o anonimato do paciente.

Concordando com o uso do material para os fins acima descritos, é necessário esclarecê-lo(a) que não existem quaisquer benefícios ou direitos financeiros a receber sobre os eventuais resultados decorrentes da pesquisa. Se você não concordar em doar o material para pesquisa, sua decisão não influenciará, de nenhum modo, no seu tratamento.

Caso você ainda tenha questões a fazer sobre este Termo de Consentimento ou alguma dúvida que não tenha sido esclarecida, por gentileza, entre em contato com a Coordenadoria do Hemocentro da UNICAMP pelo telefone: (0XX19)3 788-8734.

Você receberá uma cópia deste documento e o original será arquivado em seu prontuário. Somente assine este Termo se consentir.

D E C L A R A Ç Ã O
Declaro estar ciente das informações ora prestadas, tendo lido atentamente e concordando com todo o teor.
Campinas (SP), ...de ..... de .....

.....
Responsável ou Paciente
Nome: .....
RG : .....
RGH : .....

Comitê de Ética (0xx19) 3788-8936

ANEXO 2

Código-fonte do programa LyM

```
#!/usr/bin/perl -w

use Math::Integral::Romberg 'integral';

################################################################################

################################################################################
# Fonte: "Sagemap: A Public Gene Expression Resource" (Genome Research)
################################################################################
#
# Dados:
#       - Y e Z : dois tipos de celulas;
#       - y e z : concentracoes desconhecidas de mRNA;
#       - A     : total de tags sequenciados to tipo celular Y
#       - B     : total de tags sequenciados to tipo celular Z
#       - a     : numero de tags correspondentes ao mRNA de interesse em A
#       - b     : numero de tags correspondentes ao mRNA de interesse em B
#
################################################################################
#
# Abordagem Bayesiana:
#
#                 x = y/(y+z)
#
#               f(x) -> [0,1]   ;  f(x) = x^c * (1-x)^c
# Aqui usaremos c = 3 (Lal 1999)
#
#                                      x^a * (1-x)^b
#               g(x) = f(x) * --------------------------
#                                [1 + (A/B-1) * x]^(a+b)
#
#  A concentracao y excede z por um fator de ao menos F quando x>=L,
#  onde L = F/(F+1)
#
#  Segue que:
#                              integral( g(x)dx, L , 1)
#             P (x>=L)  = --------------------------
#                              integral( g(x)dx, 0 , 1)
#
################################################################################
```

```
##########
#  g(x)  #
##########################################################
sub GdeX {
    my $x = shift;
    my $aux1 = $x ** $a; if ($aux1 == 0) {return $aux1;}
    my $aux2 = (1 - $x) ** $b;  if ($aux2 == 0) {return $aux2;}
    my $FdeX = ($x**3)*((1 - $x)**3);  if ($FdeX == 0) {return $FdeX;}
    my $numer = $FdeX * $aux1 * $aux2;
    my $denom = ( 1 +  $XdivYMinus1 * $x ) ** $aplusb;
    if ($denom == 0) { return -1;}
    return $numer/$denom;
}
##########################################################

########################
#  Calculo de P(x>=L)  #
##########################################################
sub P {
    my $F = shift;

    $L = $F/($F+1);
    if ( $odd_ratio > 1 ) {$x1 = $L; $x2 = 1;} else { $x1 = 0; $x2=$L;}
    $numerador = integral(\&GdeX, $x1, $x2,10^-20,10^-20,20,16);

    $x1 = 0;
    $x2 = 1;
    if ($F == 2*$Fc) {$denominador = integral(\&GdeX, $x1, $x2,10^-20,10^-20,20,16);}
    # P(x>=L) eh a divisao entre as integrais
    if ($denominador == 0) { $P = - 1; }
    else { $P = $numerador/$denominador; }
    if ( $P > $k) { $P = sprintf($dig, $P); } else { unless ($P == -1) {$P = 0;} }
    return $P;
}
##########################################################

#########################################################
# Busca binaria do melhor Fator F, ponderado por pmin #
#########################################################################

sub buscaF {


 my $Fr = 2*$Fc;  #right
 my $Fl = $Fc/2;  #left
 my $inner = 0;
 my $pr = 0;
 my $pl = 0;

 while (abs($pl-$pr) < $k) {

     $pr = P($Fr);
     if ($pr == $pmin) { return $Fr; }
     if ($pr == -1) { return "OUT";}
     $pl = P($Fl);
     if ($pl == $pmin) { return $Fl; }
     if ($pl == -1) { return "OUT";}
     $Fr = 2*$Fr;
     if ($Fr > $FlimSup) { $inner = 1; last; }
     $Fl = $Fl/2;

 }

 if ($inner) { return busca_bin ($Fc/2,$Fc*2,$pl,$pr) } else {
     if ($pl > $pr) { #odd_ratio > 1
             if ($pl == 1) {
                 while ($pr > $pmin) {
                     $pl = $pr;
                     $pr = P($Fr);
                     $Fr = 2*$Fr;
                     if ($Fr > $FlimSup) { last; }
                 }
                 return busca_bin ($Fr/4, $Fr/2,$pl,$pr);
             } else {
                 while ($pl < $pmin) {
                     $pr = $pl;
                     $pl = P($Fl);
                     $Fl = $Fl/2;
                     if ($Fl < $FlimInf) { last; }
                 }
                 return busca_bin ($Fl*2, $Fl*4,$pl,$pr);
             }
     } else {
             if ($pr == 1) {
```

```
                    while ($pl > $pmin) {
                        $pr = $pl;
                        $pl = P($Fl);
                        $Fl = $Fl/2;
                        if ($Fl < $FlimInf) { last; }
                    }
                    return busca_bin ($Fl*2, $Fl*4,$pl,$pr);
                } else {
                    while ($pr < $pmin) {
                        $pl = $pr;
                        $pr = P($Fr);
                        $Fr = 2*$Fr;
                        if ($Fr > $FlimSup) { last; }
                    }
                    return busca_bin ($Fr/4, $Fr/2,$pl,$pr);
                }
            }
        }
    }

    sub busca_bin {
        my $inicio = shift;
        my $fim = shift;
        my $pl = shift;
        my $pr = shift;
        my $meio = ($inicio + $fim) /2 ;
        $meio = substr($meio,0,6);
        my $pm = P($meio);
        $pm = sprintf($dig,$pm);

        if (( $pm == $pmin ) || ($meio == $inicio) || ($meio == $fim)) {
            if ($pm == -1) { return -1;} else {return $meio;}
            }
        my @aux = sort ($pl,$pm,$pr);
        if ( $pm < $pmin ) {
            if ($pl == $aux[0]) { return  busca_bin ($fim, $meio, $pr, $pm);
            } else { return  busca_bin ($inicio, $meio, $pl, $pm);}
        } else {
            if ($pr == $aux[2]) {  return  busca_bin ($inicio, $meio, $pl, $pm);
            } else {   return  busca_bin ($meio, $fim, $pm, $pr);}
        }

}

###############
# OTHER TASKS #
###############

sub read_in {  #use: read_in(file_in,\@h)
    my $inf = shift;
    my $msg_fold = shift;
    my @temp = ();
    open (IN, "<$inf") || return -1 ;
    while (<IN>) {
        @temp = split(/\t/,$_);
        if ( $temp[4] eq "FLAG WITH HS" ) {
                push(@$msg_fold, $_);
        }
        if ($temp[2] eq "NORMAL") { next; }
        if ($temp[2] eq "") { last; }
    }
    close (IN);
    return 0;
}

sub read_cache {  #use: read_cache(cache_file,\%h)
    my $cf = shift;
    my $msg_fold = shift;
    my @temp = ();
    open (CACHE, "<$cf") || return -1 ;
    while (<CACHE>) {
        chomp;
        @temp = split(/\t/,$_);
        $$msg_fold{$temp[0]} = $temp[1];
    }
    close (CACHE);
    return 0;
}

sub write_cache {  #use: write_cache(cache_file,\%h)
    my $cf = shift;
    my $msg_fold = shift;
    my $k = "";
```

```
    open (CACHE, ">$cf") || return 1 ;
    foreach  $k (keys (%$msg_fold)) {
        print CACHE "$k\t$msg_fold{$k}\n";
    }
    close (CACHE);
    return 0;
}

sub print_result {   #use: print_result(cache_file,\@hash values to print)
    my $rf = shift;
    my $r = shift;
    open (OUT, ">$rf") || return 1 ;
    print OUT @$r;
    close (OUT);
    return 0;
}


#############
#PARAMETERS
####################
$dig = "%.2f";
$k=0.01;
$pmin = 0.95 ;
$FlimSup = 1024;  #up: +1024
$FlimInf = 0.000976563;  #down -1024

########################################
# DESIGNED BY: Tarcísio de Souza Peres  #
# Faculty of Medical Sciences           #
# State University of Campinas - Brazil #
########################################
```

ANEXO 3

Os 100 primeiros genes (de cada tecido) da classificação final.

Tabela 11- Os 100 primeiros genes (ordenados por pontuação), para o tecido "mama".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
8821	INPP4B	inositol polyphosphate-4-phosphatase, type II	363,96	0	2477	1013
226	ALDOA	aldolase A, fructose-bisphosphate	127,2	0	362	224
3611	ILK	integrin-linked kinase	124,4	0	265	168
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	120,8	0	115	76
55522	ILK-2	integrin-linked kinase-2	110,4	0	265	168
3065	HDAC1	histone deacetylase 1	100,4	0	136	88
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	92,725	0	20	15,75
7534	YWHAZ	tyrosine 3-monooxygenase/tryptophan	90,8	0	143	96
4938	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa	88	0	125	80
71	ACTG1	actin, gamma 1	85,6	0	105	72
4539	MT-ND4L	mitochondrially encoded NADH 4L	84	0	122	80
4519	MT-CYB	mitochondrially encoded cytochrome b	80,4	2	201	68
824	CAPN2	calpain 2, (m/II) large subunit	79,2	728	4	-64
10217	CTDSPL	CTD (carboxy-terminal domain, RNA polymerase	78,8	0	84	56
4763	NF1	neurofibromin 1	78,7	0	9	9
146556	MGC45438	hypothetical protein MGC45438	78,6	0	96	62
23637	RABGAP1	RAB GTPase activating protein 1	78	0	88	60
25821	MTO1	mitochondrial translation optimization 1 homolog	78	0	90	60
4881	NPR1	natriuretic peptide receptor A/guanylate cyclase	78	0	90	60
25	ABL1	v-abl Abelson murine leukemia viral oncogene	77,864	2	3	2,88
7402	UTRN	utrophin	77,6	340	0	-52
5250	SLC25A3	solute carrier family 25	76,8	367	0	-56
5546	PRCC	papillary renal cell carcinoma	76,364	0	7	7,88
4513	MT-CO2	mitochondrially encoded cytochrome c oxidase	75,6	8	413	52
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	74,369	4	0	1,23
1019	CDK4	cyclin-dependent kinase 4	73,4	0	22	18
5058	PAK1	p21/Cdc42/Rac1-activated kinase 1	73,4	0	71	48
5966	REL	v-rel reticuloendotheliosis viral oncogene	72,564	2	0	1,88
595	CCND1	cyclin D1	71,75	82	0	-12,5
8148	TAF15	TAF15 RNA polymerase II, TATA box binding	67,2	0	32	24
10263	CDK2AP2	CDK2-associated protein 2	66,3	0	27	21
3329	HSPD1	heat shock 60kDa protein 1 (chaperonin)	65,8	6	221	36

8800	PEX11A	peroxisomal biogenesis factor 11A	53,3	0	12	11
9135	RABEP1	rabaptin, RAB GTPase binding effector protein	53,3	0	12	11
7170	TPM3	tropomyosin 3	52,55	0	8	8,5
4926	NUMA1	nuclear mitotic apparatus protein 1	52,364	0	7	7,88
1345	COX6C	cytochrome c oxidase subunit VIc	52,25	0	6	7,5
23085	RAB6IP2	RAB6 interacting protein 2	51,65	0	3	5,5
7076	TIMP1	TIMP metallopeptidase inhibitor 1	51,6	0	13	12
5934	RBL2	retinoblastoma-like 2 (p130)	51,6	1	21	12
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	51,575	2	10	5,25
4830	NME1	non-metastatic cells 1, protein (NM23A)	51,3	0	12	11
4683	NBN	nibrin	50,7	0	9	9
2033	EP300	E1A binding protein p300	50,425	0	2	4,75
5878	RAB5C	RAB5C, member RAS oncogene family	50,4	0	38	28
2072	ERCC4	excision repair cross-complementing rodent	50,375	7	0	-1,25
4605	MYBL2	v-myb myeloblastosis viral oncogene homolog	50,2	0	16	14
867	CBL	Cas-Br-M (murine) ecotropic retroviral	48,564	2	0	1,88
5894	RAF1	v-raf-1 murine leukemia viral oncogene homolog 1	48,425	0	2	4,75
3815	KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral	48,425	0	2	4,75
5395	PMS2	PMS2 postmeiotic segregation increased 2	48,425	0	2	4,75
2120	ETV6	ets variant gene 6 (TEL oncogene)	48	24	21	-1
329	BIRC2	baculoviral IAP repeat-containing 2	47,8	0	4	6
8881	CDC16	CDC16 cell division cycle 16 homolog	47,65	0	3	5,5
2132	EXT2	exostoses (multiple) 2	47,65	0	3	5,5
328	APEX1	APEX nuclease (multifunctional DNA repair)	47,6	0	13	12
7251	TSG101	tumor susceptibility gene 101	47,6	0	13	12
7158	TP53BP1	tumor protein p53 binding protein, 1	46,6	0	30	22
5914	RARA	retinoic acid receptor, alpha	46,564	2	0	1,88
84925	DIRC2	disrupted in renal carcinoma 2	46,5	4	17	5
5899	RALB	v-ral simian leukemia viral oncogene homolog B	46,425	0	2	4,75
9950	GOLGA5	golgi autoantigen, golgin subfamily a	46,25	0	6	7,5
2130	EWSR1	Ewing sarcoma breakpoint region 1	45,768	2	2	2,56
5034	P4HB	procollagen-proline, 2-oxoglutarate	45,5	2	69	25
5265	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1)	45	0	41	30
4478	MSN	moesin	44,6	5	5	2
6625	SNRP70	small nuclear ribonucleoprotein 70kDa	44,6	0	30	22

Tabela 12- Os 100 primeiros genes (ordenados por pontuação), para o tecido "cólon".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
7633	ZNF79	zinc finger protein 79	232,8	0	1011	576
284349	ZNF283	zinc finger protein 283	232,8	0	1011	576
84967	LSM10	LSM10, U7 small nuclear RNA associated	151,2	0	531	304
57120	GOPC	golgi associated PDZ and coiled-coil motif	132	0	402	240
928	CD9	CD9 molecule	129,2	3	767	184
5284	PIGR	polymeric immunoglobulin receptor	108	0	276	160
23640	HSPBP1	hsp70-interacting protein	107,6	0	260	152
330	BIRC3	baculoviral IAP repeat-containing 3	98,2	0	17	14
6629	SNRPB2	small nuclear ribonucleoprotein polypeptide	97,2	2	414	124
63967	CLSPN	claspin homolog (Xenopus laevis)	95,6	0	193	112
3301	DNAJA1	DnaJ (Hsp40) homolog, subfamily A, member 1	92	3	414	100
2762	GMDS	GDP-mannose 4,6-dehydratase	91,2	0	177	104
1366	CLDN7	claudin 7	84,8	6	531	76
25832	NBPF14	neuroblastoma breakpoint family, member 14	84	0	131	80
284565	NBPF15	neuroblastoma breakpoint family, member 15	84	0	136	80
5663	PSEN1	presenilin 1 (Alzheimer disease 3)	83,6	0	86	52
3105	HLA-A	major histocompatibility complex, class I, A	82,8	909	10	-56
10541	ANP32B	acidic (leucine-rich) nuclear phosphoprotein	78,8	2	178	56
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	78,364	8	63	7,88
1999	ELF3	E74-like factor 3	77,4	8	514	58
25962	KIAA1429	KIAA1429	76,8	0	89	56
10099	TSPAN3	tetraspanin 3	75,6	0	83	52
10181	RBM5	RNA binding motif protein 5	75,2	0	54	34
5908	RAP1B	RAP1B, member of RAS oncogene family	75,15	4	47	10,5
25	ABL1	v-abl Abelson murine leukemia viral oncogene	74,534	2	0	1,78
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	72,2	11	157	14
4610	MYCL1	v-myc myelocytomatosis viral oncogene homolog	68,425	0	2	4,75
672	BRCA1	breast cancer 1, early onset	68,425	0	2	4,75
4763	NF1	neurofibromin 1 (neurofibromatosis, von	68,425	0	2	4,75

6678	SPARC	secreted protein, acidic, cysteine-rich	49,1	0	23	17
2073	ERCC5	excision repair cross-complementing rodent	48,525	16	3	-1,75
1942	EFNA1	ephrin-A1	48,5	0	19	15
2956	MSH6	mutS homolog 6 (E. coli)	48,425	0	2	4,75
4478	MSN	moesin	48,425	0	2	4,75
2178	FANCE	Fanconi anemia, complementation group E	48,425	0	2	4,75
5527	PPP2R5C	protein phosphatase 2, regulatory subunit B	48,3	196	3	-21
5747	PTK2	PTK2 protein tyrosine kinase 2	48,2	0	18	14
4609	MYC	v-myc myelocytomatosis viral oncogene homolog	48,05	4	11	3,5
991	CDC20	CDC20 cell division cycle 20 homolog	48,025	40	0	-6,75
8881	CDC16	CDC16 cell division cycle 16 homolog	47,95	0	5	6,5
329	BIRC2	baculoviral IAP repeat-containing 2	47,8	0	4	6

Tabela 13- Os 100 primeiros genes (ordenados por pontuação), para o tecido "cabeça e pescoço".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
7038	TG	thyroglobulin	212	290	325	500
1938	EEF2	eukaryotic translation elongation factor 2	103,4	12	3976	71,57
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	94,1	54	2	-37
51185	CRBN	cereblon	87,6	156	4	-92
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	78,35	0	8	4,5
3265	HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene	78,275	0	5	4,25
4893	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	78,125	0	2	3,75
25	ABL1	v-abl Abelson murine leukemia viral oncogene	78,125	0	2	3,75
29966	STRN3	striatin, calmodulin binding protein 3	76,2	58	0	-54
221178	SPATA13	spermatogenesis associated 13	76,2	58	0	-54
338872	C1QTNF9	C1q and tumor necrosis factor related protein	76,2	58	0	-54
1277	COL1A1	collagen, type I, alpha 1	76	0	160	20
7076	TIMP1	TIMP metallopeptidase inhibitor 1	73,8	40	0	-36
648	BMI1	B lymphoma Mo-MLV insertion region (mouse)	71,75	17	0	-12,5
4763	NF1	neurofibromin 1	68,5	0	13	5
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	68,5	0	11	5
5925	RB1	retinoblastoma 1 (including osteosarcoma)	68,425	0	10	4,75
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	68,275	0	6	4,25
602	BCL3	B-cell CLL/lymphoma 3	66,125	0	3	3,75
1974	EIF4A2	eukaryotic translation initiation factor 4A	63,6	0	77	12
2130	EWSR1	Ewing sarcoma breakpoint region 1	62,48	2	3	1,6
2120	ETV6	ets variant gene 6 (TEL oncogene)	61,95	0	25	6,5
54904	WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1	60,125	0	2	3,75
5018	OXA1L	oxidase (cytochrome c) assembly 1-like	60	53	0	-50

51128	SAR1B	SAR1 gene homolog B (S. cerevisiae)	59,4	52	0	-48
7150	TOP1	topoisomerase (DNA) I	58,425	0	9	4,75
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	58,325	0	39	7,75
4869	NPM1	nucleophosmin (nucleolar phosphoprotein B23	58,15	235	69	-20,5
11272	PRR4	proline rich 4 (lacrimal)	57,2	0	202	24
3326	HSP90AB1	heat shock protein 90kDa alpha (cytosolic)	56	0	161	20
7849	PAX8	paired box gene 8	55,7	55	10	-19
145165	FAM10A4	family with sequence similarity 10, member A4	55,6	0	82	12
3932	LCK	lymphocyte-specific protein tyrosine kinase	54,456	2	2	1,52
4620	MYH2	myosin, heavy polypeptide 2, skeletal muscle	54,45	153	24	-31,5
5695	PSMB7	proteasome (prosome, macropain) subunit, beta	54,3	0	289	31
9045	RPL14	ribosomal protein L14	54,3	0	292	31
7913	DEK	DEK oncogene (DNA binding)	54,125	0	2	3,75
2175	FANCA	Fanconi anemia, complementation group A	54,125	0	2	3,75
6418	SET	SET translocation (myeloid leukemia-associated)	53,95	0	25	6,5
7175	TPR	translocated promoter region	53,95	0	23	6,5
6760	SS18	synovial sarcoma translocation, chromosome 18	53,575	0	14	5,25
10801	SEPT9	septin 9	52,364	0	40	7,88
2073	ERCC5	excision repair cross-complementing rodent	52,125	0	3	3,75
1019	CDK4	cyclin-dependent kinase 4	51,8	0	20	6
5292	PIM1	pim-1 oncogene	51,725	0	18	5,75
2033	EP300	E1A binding protein p300	50,35	0	8	4,5
4297	MLL	myeloid/lymphoid or mixed-lineage leukemia	50,35	0	8	4,5
8148	TAF15	TAF15 RNA polymerase II, TATA box binding	50,25	0	34	7,5
5914	RARA	retinoic acid receptor, alpha	50,125	0	2	3,75
5727	PTCH	patched homolog (Drosophila)	50,125	0	2	3,75
6391	SDHC	succinate dehydrogenase complex, subunit C,	50,125	0	3	3,75
2335	FN1	fibronectin 1	50	0	170	20
22872	SEC31L1	SEC31-like 1 (S. cerevisiae)	49,92	0	1664	71,57
3726	JUNB	jun B proto-oncogene	49,8	0	19	6

Tabela 14- Os 100 primeiros genes (ordenados por pontuação), para o tecido “pulmão”.

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
3500	IGHG1	immunoglobulin heavy constant gamma 1	87,4	4	532	78
5211	PFKL	phosphofructokinase, liver	86,4	0	206	88
7841	GCS1	glucosidase I	79,8	290	0	-66
3502	IGHG3	immunoglobulin heavy constant gamma 3 (G3m	79,6	0	111	52
81887	LAS1L	LAS1-like (S. cerevisiae)	78	0	135	60
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	77,75	3	6	2,5
56907	SPIRE1	spire homolog 1 (Drosophila)	76,8	0	122	56
3537	IGLC1	immunoglobulin lambda constant 1 (Mcg marker)	76,8	0	127	56
1277	COL1A1	collagen, type I, alpha 1	76,6	2	92	22
28831	IGLJ3	immunoglobulin lambda joining 3	75,6	0	115	52
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	71,3	0	16	11
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	68,35	1	0	-4,5
3507	IGHM	immunoglobulin heavy constant mu	61,6	9	551	42
79058	ASPSCR1	alveolar soft part sarcoma chromosome region	60,164	0	1	3,88
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	58,739	148	16	-9,13
28796	IGLV3-21	immunoglobulin lambda variable 3-21	58,2	0	97	44
3493	IGHA1	immunoglobulin heavy constant alpha 1	58,2	0	97	44
3493	IGHA1	immunoglobulin heavy constant alpha 1	58,2	0	97	44
3159	HMGA1	high mobility group AT-hook 1	57,3	0	17	11
2719	GPC3	glypican 3	56,55	0	11	8,5
9098	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	55,95	0	6	6,5
92715	WDR85	WD repeat domain 85	55,8	0	77	36
7169	TPM2	tropomyosin 2 (beta)	55,2	0	69	34
3727	JUND	jun D proto-oncogene	54,525	8	0	-1,75
10627	MRCL3	myosin regulatory light chain MRCL3	54,45	4	212	31,5
2120	ETV6	ets variant gene 6 (TEL oncogene)	54,414	11	4	-1,38
7913	DEK	DEK oncogene (DNA binding)	54,35	0	2	4,5
5781	PTPN11	protein tyrosine phosphatase, non-receptor	54,35	0	2	4,5
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	52,1	0	7	7
994	CDC25B	cell division cycle 25B	51,65	0	4	5,5

Tabela 15- Os 100 primeiros genes (ordenados por pontuação), para o tecido "estômago".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
51131	PHF11	PHD finger protein 11	112,8	0	626	176
3838	KPNA2	karyopherin alpha 2	101,2	0	350	104
5222	PGA5	pepsinogen 5, group I (pepsinogen A)	98,4	367	0	-128
5937	RBMS1	RNA binding motif, single stranded interacting	97,6	0	397	112
3020	H3F3A	H3 histone, family 3A	96	353	0	-120
9836	LCMT2	leucine carboxyl methyltransferase 2	91,2	0	351	104
28959	LR8	LR8 protein	87	264	0	-90
11138	TBC1D8	TBC1 domain family, member 8	84	0	269	80
10388	SYCP2	synaptonemal complex protein 2	84	0	267	80
7167	TPI1	triosephosphate isomerase 1	82,4	0	229	68
5630	PRPH	peripherin	80,4	0	226	68
55249	YY1AP1	YY1 associated protein 1	78,8	0	194	56
7175	TPR	translocated promoter region	73,15	0	23	10,5
3320	HSP90AA1	heat shock protein 90kDa alpha	73	0	93	30
4763	NF1	neurofibromin 1	68,275	0	2	4,25
1050	CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha	65	0	89	30
3376	IARS	isoleucine-tRNA synthetase	63,2	0	104	34
12	SERPINA3	serpin peptidase inhibitor, clade A	61,4	0	156	48
5355	PLP2	proteolipid protein 2	60	0	164	50
2923	PDIA3	protein disulfide isomerase family A, member 3	59,8	0	117	36
4508	MT-ATP6	mitochondrially encoded ATP synthase 6	58,8	3	389	46
3932	LCK	lymphocyte-specific protein tyrosine kinase	58,5	0	4	5
5159	PDGFRB	platelet-derived growth factor receptor, beta	58,275	0	2	4,25
10724	MGEA5	meningioma expressed antigen 5 (hyaluronidase)	57,8	2	167	26
23522	MYST4	MYST histone acetyltransferase	56	0	55	20
30968	STOML2	stomatin (EPB72)-like 2	55,8	0	116	36
79574	EPS8L3	EPS8-like 3	55,2	0	105	34
4259	MGST3	microsomal glutathione S-transferase 3	55,1	0	44	17
8513	LIPF	lipase, gastric	54,9	98	0	-33

Tabela 16- Os 100 primeiros genes (ordenados por pontuação), para o tecido "útero".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
6892	TAPBP	TAP binding protein (tapasin)	101,61	0	3187	318,7
2923	PDIA3	protein disulfide isomerase family A, member 3	100	0	260	120
5702	PSMC3	proteasome (prosome, macropain) 26S subunit	98	0	254	120
1522	CTSZ	cathepsin Z	98	0	248	120
2783	GNB2	guanine nucleotide binding protein (G protein)	84	0	165	80
330	BIRC3	baculoviral IAP repeat-containing 3	77,65	0	4	5,5
84364	ZNF289	zinc finger protein 289, ID1 regulated	75,6	0	105	52
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	71,6	0	17	12
2120	ETV6	ets variant gene 6 (TEL oncogene)	68,55	0	10	8,5
672	BRCA1	breast cancer 1, early onset	68,35	0	2	4,5
8837	CFLAR	CASP8 and FADD-like apoptosis regulator	68,3	0	59	31
7538	ZFP36	zinc finger protein 36, C3H type, homolog	65,4	0	93	48
4221	MEN1	multiple endocrine neoplasia I	63,6	0	18	12
27125	AFF4	AF4/FMR2 family, member 4	63,3	0	15	11
1635	DCTD	dCMP deaminase	59,8	0	67	36
7150	TOP1	topoisomerase (DNA) I	58,5	0	3	5
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	58,325	0	9	7,75
641	BLM	Bloom syndrome	58,164	0	1	3,88
3020	H3F3A	H3 histone, family 3A	56,4	187	0	-38
4508	MT-ATP6	mitochondrially encoded ATP synthase 6	55,8	0	69	36
329	BIRC2	baculoviral IAP repeat-containing 2	55	0	13	10
7913	DEK	DEK oncogene (DNA binding)	54,35	0	2	4,5
2175	FANCA	Fanconi anemia, complementation group A	53,687	2	2	2,29
6418	SET	SET translocation (myeloid leukemia-associated)	53,65	0	4	5,5
27020	NPTN	neuroplastin	53,6	0	17	12
56654	NPDC1	neural proliferation, differentiation and control, 1	53,3	0	15	11
9098	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	52,5	0	3	5

22794	CASC3	cancer susceptibility candidate 3	52,25	0	8	7,5
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	51,65	0	4	5,5
2033	EP300	E1A binding protein p300	50,5	0	3	5
7175	TPR	translocated promoter region	50,275	2	9	4,25
2073	ERCC5	excision repair cross-complementing rodent	48,399	3	0	1,33
4292	MLH1	mutL homolog 1, colon cancer, nonpolyposis type	48,35	0	2	4,5
613	BCR	breakpoint cluster region	48,35	0	2	4,5
2114	ETS2	v-ets erythroblastosis virus E26 oncogene	48,35	0	2	4,5
999	CDH1	cadherin 1, type 1, E-cadherin (epithelial)	48,35	0	2	4,5
208	AKT2	v-akt murine thymoma viral oncogene homolog 2	48,35	0	2	4,5
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	48,164	0	1	3,88
4288	MKI67	antigen identified by monoclonal antibody Ki-67	47,9	0	19	13
6774	STAT3	signal transducer and activator of transcription 3	47,6	0	17	12
11200	CHEK2	CHK2 checkpoint homolog (S. pombe)	47,6	0	18	12
60	ACTB	actin, beta	47,2	0	45	24
6391	SDHC	succinate dehydrogenase complex, subunit C	46,543	9	0	-1,81
466	ATF1	activating transcription factor 1	46,5	0	3	5
1387	CREBBP	CREB binding protein	46,35	0	2	4,5
5930	RBBP6	retinoblastoma binding protein 6	46,325	2	24	7,75
1277	COL1A1	collagen, type I, alpha 1	46,318	10	5	-1,06
6317	SERPINB3	serpin peptidase inhibitor, clade B	46,25	3	30	7,5
6390	SDHB	succinate dehydrogenase complex, subunit B, iron	45,95	0	6	6,5
3909	LAMA3	laminin, alpha 3	45,7	0	32	19
801	CALM1	calmodulin 1 (phosphorylase kinase, delta)	45,2	0	45	24
7057	THBS1	thrombospondin 1	44,85	0	12	9,5
1019	CDK4	cyclin-dependent kinase 4	44,564	4	4	1,88
4478	MSN	moesin	44,534	2	0	1,78
5879	RAC1	ras-related C3 botulinum toxin substrate 1	44,5	0	23	15

7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	44,357	8	2	-1,19
3838	KPNA2	karyopherin alpha 2	44,35	72	0	-14,5
3320	HSP90AA1	heat shock protein 90kDa alpha (cytosolic)	44,239	52	9	-4,13
4301	MLLT4	myeloid/lymphoid or mixed-lineage leukemia	44,164	0	1	3,88
29893	PSMC3IP	PSMC3 interacting protein	44	0	36	20
4683	NBN	nibrin	43,8	0	5	6
4820	NKTR	natural killer-tumor recognition sequence	43,8	0	5	6
9793	CKAP5	cytoskeleton associated protein 5	43,65	0	4	5,5
3437	IFIT3	interferon-induced protein with	43	0	56	30
8125	ANP32A	acidic (leucine-rich) nuclear phosphoprotein	42,6	0	39	22
3726	JUNB	jun B proto-oncogene	42,534	2	0	1,78
4926	NUMA1	nuclear mitotic apparatus protein 1	42,35	0	2	4,5
10801	SEPT9	septin 9	42,35	0	2	4,5
3106	HLA-B	major histocompatibility complex, class I, B	42	97	0	-20
29	ABR	active BCR-related gene	41,6	0	17	12
6119	RPA3	replication protein A3, 14kDa	41,6	0	17	12
3326	HSP90AB1	heat shock protein 90kDa alpha (cytosolic)	40,4	39	0	-8
26511	CHIC2	cysteine-rich hydrophobic domain 2	40,35	0	2	4,5
4302	MLLT6	myeloid/lymphoid or mixed-lineage leukemia	40,35	0	2	4,5
7994	MYST3	MYST histone acetyltransferase	40,35	0	2	4,5
23352	ZUBR1	zinc finger, UBR1 type 1	40,25	0	8	7,5
1974	EIF4A2	eukaryotic translation initiation factor 4A	40,239	3	13	4,13
3123	HLA-DRB1	major histocompatibility complex, class II	39,9	0	19	13
3659	IRF1	interferon regulatory factor 1	39,3	0	16	11
9110	MTMR4	myotubularin related protein 4	39	0	58	30
2131	EXT1	exostoses (multiple) 1	38,6	3	3	2
578	BAK1	BCL2-antagonist/killer 1	38,55	0	10	8,5
7170	TPM3	tropomyosin 3	38,534	3	2	1,78
3945	LDHB	lactate dehydrogenase B	38,5	0	24	15
6814	STXBP3	syntaxin binding protein 3	38,4	0	51	28
1643	DDB2	damage-specific DNA binding protein 2, 48kDa	38,399	3	0	1,33

Tabela 17- Os 100 primeiros genes (ordenados por pontuação), para o tecido "testículo".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
4538	MT-ND4	mitochondrially encoded NADH dehydrogenase 4	218,25	1455	2	727,5
3507	IGHM	immunoglobulin heavy constant mu	155,2	4	150	304
4893	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	141,6	0	21	112
3538	IGLC2	immunoglobulin lambda constant 2	122,4	0	36	208
1936	EEF1D	eukaryotic translation elongation factor 1	112,8	0	31	176
3502	IGHG3	immunoglobulin heavy constant gamma 3	107,2	0	26	144
51237	PACAP	proapoptotic caspase adaptor protein	102,8	0	25	136
28786	IGLV4-3	immunoglobulin lambda variable 4-3	97,2	0	23	124
3537	IGLC1	immunoglobulin lambda constant 1 (Mcg marker)	97,2	0	23	124
335	APOA1	apolipoprotein A-I	93,6	0	21	112
2678	GGT1	gamma-glutamyltransferase 1	81,05	0	13	63,5
2023	ENO1	enolase 1, (alpha)	69,4	0	10	48
4763	NF1	neurofibromin 1	67,711	4	0	2,37
238	ALK	anaplastic lymphoma kinase (Ki-1)	61,768	3	0	2,56
6647	SOD1	superoxide dismutase 1, soluble	60,2	0	9	44
3500	IGHG1	immunoglobulin heavy constant gamma 1	59,8	10	38	36
3126	HLA-DRB4	major histocompatibility complex, class II	58,2	0	9	44
5159	PDGFRB	platelet-derived growth factor receptor, beta	57,834	2	0	2,78
2120	ETV6	ets variant gene 6 (TEL oncogene)	57,834	2	0	2,78
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	57,768	3	0	2,56
92140	MTDH	metadherin	56,3	0	7	31
10607	TBL3	transducin (beta)-like 3	55,8	0	8	36
718	C3	complement component 3	55,8	3	16	36
2260	FGFR1	fibroblast growth factor receptor 1	54,564	8	0	1,88
400	ARL1	ADP-ribosylation factor-like 1	54,3	0	7	31

28299	IGKV1-5	immunoglobulin kappa variable 1-5	54,3	0	7	31
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	51,725	2	2	5,75
387	RHOA	ras homolog gene family, member A	50,3	0	5	21
7175	TPR	translocated promoter region	49,834	2	0	2,78
2033	EP300	E1A binding protein p300	49,834	2	0	2,78
1277	COL1A1	collagen, type I, alpha 1	49,834	2	0	2,78
5914	RARA	retinoic acid receptor, alpha	49,768	3	0	2,56
6418	SET	SET translocation (myeloid leukemia-associated)	49,711	4	0	2,37
2263	FGFR2	fibroblast growth factor receptor 2	48,05	1	0	-3,5
613	BCR	breakpoint cluster region	47,834	2	0	2,78
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	47,834	2	0	2,78
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	47,768	3	0	2,56
648	BMI1	B lymphoma Mo-MLV insertion region (mouse)	47,711	4	0	2,37
2130	EWSR1	Ewing sarcoma breakpoint region 1	45,639	6	0	2,13
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	44,48	12	0	1,6
5371	PML	promyelocytic leukemia	43,834	2	0	2,78
8837	CFLAR	CASP8 and FADD-like apoptosis regulator	43,834	2	0	2,78
8021	NUP214	nucleoporin 214kDa	43,711	4	0	2,37
6794	STK11	serine/threonine kinase 11	43,711	4	0	2,37
3320	HSP90AA1	heat shock protein 90kDa alpha (cytosolic)	43,687	5	0	2,29
1870	E2F2	E2F transcription factor 2	43,2	6	17	24
2521	FUS	fusion (involved in t(12;16) in malignant liposarcoma)	41,711	4	0	2,37
4926	NUMA1	nuclear mitotic apparatus protein 1	41,687	5	0	2,29
684	BST2	bone marrow stromal cell antigen 2	41	2	4	10
51119	SBDS	Shwachman-Bodian-Diamond syndrome	40,391	0	2	7,97
6421	SFPQ	splicing factor proline/glutamine-rich	39,834	2	0	2,78
7184	HSP90B1	heat shock protein 90kDa beta (Grp94), member	39,834	2	0	2,78
8028	MLLT10	myeloid/lymphoid or mixed-lineage leukemia	39,834	2	0	2,78
84168	ANTXR1	anthrax toxin receptor 1	39,834	2	0	2,78
5327	PLAT	plasminogen activator	39,8	0	6	26

Tabela 18- Os 100 primeiros genes (ordenados por pontuação), para o tecido "sistema nervoso central".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
80975	TMPRSS5	transmembrane protease, serine 5 (spinesin)	189,6	0	373	432
2120	ETV6	ets variant gene 6 (TEL oncogene)	124,8	0	45	56
1277	COL1A1	collagen, type I, alpha 1	115,6	0	43	52
389	RHOC	ras homolog gene family, member C	114,8	0	116	136
10514	MYBBP1A	MYB binding protein (P160) 1a	108	1	202	160
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	93,2	0	27	34
1278	COL1A2	collagen, type I, alpha 2	92,8	0	63	76
10	NAT2	N-acetyltransferase 2	92,4	0	73	88
283209	PGM2L1	phosphoglucomutase 2-like 1	91,2	0	88	104
348	APOE	apolipoprotein E	89,6	0	77	92
3872	KRT17	keratin 17	88,4	0	73	88
1290	COL5A2	collagen, type V, alpha 2	86,4	0	76	88
1801	DPH1	DPH1 homolog (S. cerevisiae)	85,4	950	0	-78
9867	PJA2	praja 2, RING-H2 motif	85,2	0	71	84
84790	TUBA6	tubulin, alpha 6	84,8	0	46	56
25	ABL1	v-abl Abelson murine leukemia viral oncogene	81,65	0	2	5,5
79902	NUP85	nucleoporin 85kDa	81,6	0	59	72
2969	GTF2I	general transcription factor II, i	80,8	0	47	56
5966	REL	v-rel reticuloendotheliosis viral oncogene	79,95	0	3	6,5
1191	CLU	clusterin	78,8	128	2346	56
3187	HNRPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	78	0	51	60
3188	HNRPH2	heterogeneous nuclear ribonucleoprotein H2	78	0	49	60
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	77,9	15	19	3
4155	MBP	myelin basic protein	76,8	4	138	56
10492	SYNCRIP	synaptotagmin binding, cytoplasmic RNA	76,8	0	46	56
2308	FOXO1A	forkhead box O1A (rhabdomyosarcoma)	75,9	0	9	13
8481	OFD1	oral-facial-digital syndrome 1	75,6	0	44	52
5908	RAP1B	RAP1B, member of RAS oncogene family	75,3	0	7	11
3265	HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene	74,534	3	0	1,78
673	BRAF	v-raf murine sarcoma viral oncogene homolog B1	71,539	5	12	5,13
4478	MSN	moesin	71,3	0	7	11
2316	FLNA	filamin A, alpha (actin binding protein 280)	71	0	33	40

Tabela 19- Os 100 primeiros genes (ordenados por pontuação), para o tecido "próstata".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
56851	C15orf24	chromosome 15 open reading frame 24	141,6	0	341	272
63934	ZNF667	zinc finger protein 667	86,4	0	106	88
29927	SEC61A1	Sec61 alpha 1 subunit (S. cerevisiae)	84	0	95	80
26580	BSCL2	Bernardinelli-Seip congenital lipodystrophy 2	78,6	0	75	62
29101	SSU72	SSU72 RNA polymerase II CTD phosphatase homolog	76,8	0	66	56
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	74,6	2	0	2
25	ABL1	v-abl Abelson murine leukemia viral oncogene	74,6	2	0	2
2067	ERCC1	excision repair cross-complementing rodent	68,3	0	35	31
7178	TPT1	tumor protein, translationally-controlled 1	66,4	12	345	38
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	64,6	2	0	2
79058	ASPSCR1	alveolar soft part sarcoma chromosome region	63,95	0	4	6,5
5930	RBBP6	retinoblastoma binding protein 6	62	0	21	20
134359	FLJ35779	hypothetical protein FLJ35779	59,4	0	56	48
71	ACTG1	actin, gamma 1	58,45	4	110	31,5
595	CCND1	cyclin D1	58,382	0	6	7,94
3508	IGHMBP2	immunoglobulin mu binding protein 2	58,2	0	51	44
23646	PLD3	phospholipase D family, member 3	57	0	47	40
378	ARF4	ADP-ribosylation factor 4	57	0	46	40
607	BCL9	B-cell CLL/lymphoma 9	56,6	2	0	2
3817	KLK2	kallikrein 2, prostatic	56,564	2	73	31,88
3725	JUN	v-jun sarcoma virus 17 oncogene homolog (avian)	56,4	0	31	28
255394	TCP11L2	t-complex 11 (mouse) like 2	55,2	0	39	34
2260	FGFR1	fibroblast growth factor receptor 1	54,6	2	0	2
81608	FIP1L1	FIP1 like 1 (S. cerevisiae)	54,2	0	13	14
1613	DAPK3	death-associated protein kinase 3	53,75	103	0	-12,5
2073	ERCC5	excision repair cross-complementing rodent	52,5	0	2	5
7170	TPM3	tropomyosin 3	52,382	0	6	7,94
9098	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	51,825	2	2	2,75
3845	KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene	51,8	0	3	6
6421	SFPQ	splicing factor proline/glutamine-rich	50,382	0	6	7,94
8208	CHAF1B	chromatin assembly factor 1, subunit B (p60)	49	0	8	10

9168	TMSB10	thymosin, beta 10	35,6	0	11	12
9322	TRIP10	thyroid hormone receptor interactor 10	35,4	0	18	18
3910	LAMA4	laminin, alpha 4	35,4	3	50	18
23593	HEBP2	heme binding protein 2	35,1	0	17	17
2316	FLNA	filamin A, alpha (actin binding protein 280)	34,925	208	9	-9,75

Tabela 16- Os 100 primeiros genes (ordenados por pontuação), para o tecido "útero".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
6892	TAPBP	TAP binding protein (tapasin)	101,61	0	3187	318,7
2923	PDIA3	protein disulfide isomerase family A, member 3	100	0	260	120
5702	PSMC3	proteasome (prosome, macropain) 26S subunit	98	0	254	120
1522	CTSZ	cathepsin Z	98	0	248	120
2783	GNB2	guanine nucleotide binding protein (G protein)	84	0	165	80
330	BIRC3	baculoviral IAP repeat-containing 3	77,65	0	4	5,5
84364	ZNF289	zinc finger protein 289, ID1 regulated	75,6	0	105	52
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	71,6	0	17	12
2120	ETV6	ets variant gene 6 (TEL oncogene)	68,55	0	10	8,5
672	BRCA1	breast cancer 1, early onset	68,35	0	2	4,5
8837	CFLAR	CASP8 and FADD-like apoptosis regulator	68,3	0	59	31
7538	ZFP36	zinc finger protein 36, C3H type, homolog	65,4	0	93	48
4221	MEN1	multiple endocrine neoplasia I	63,6	0	18	12
27125	AFF4	AF4/FMR2 family, member 4	63,3	0	15	11
1635	DCTD	dCMP deaminase	59,8	0	67	36
7150	TOP1	topoisomerase (DNA) I	58,5	0	3	5
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	58,325	0	9	7,75
641	BLM	Bloom syndrome	58,164	0	1	3,88
3020	H3F3A	H3 histone, family 3A	56,4	187	0	-38
4508	MT-ATP6	mitochondrially encoded ATP synthase 6	55,8	0	69	36
329	BIRC2	baculoviral IAP repeat-containing 2	55	0	13	10
7913	DEK	DEK oncogene (DNA binding)	54,35	0	2	4,5
2175	FANCA	Fanconi anemia, complementation group A	53,687	2	2	2,29
6418	SET	SET translocation (myeloid leukemia-associated)	53,65	0	4	5,5
27020	NPTN	neuroplastin	53,6	0	17	12
56654	NPDC1	neural proliferation, differentiation and control, 1	53,3	0	15	11
9098	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	52,5	0	3	5

22794	CASC3	cancer susceptibility candidate 3	52,25	0	8	7,5
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	51,65	0	4	5,5
2033	EP300	E1A binding protein p300	50,5	0	3	5
7175	TPR	translocated promoter region	50,275	2	9	4,25
2073	ERCC5	excision repair cross-complementing rodent	48,399	3	0	1,33
4292	MLH1	mutL homolog 1, colon cancer, nonpolyposis type	48,35	0	2	4,5
613	BCR	breakpoint cluster region	48,35	0	2	4,5
2114	ETS2	v-ets erythroblastosis virus E26 oncogene	48,35	0	2	4,5
999	CDH1	cadherin 1, type 1, E-cadherin (epithelial)	48,35	0	2	4,5
208	AKT2	v-akt murine thymoma viral oncogene homolog 2	48,35	0	2	4,5
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	48,164	0	1	3,88
4288	MKI67	antigen identified by monoclonal antibody Ki-67	47,9	0	19	13
6774	STAT3	signal transducer and activator of transcription 3	47,6	0	17	12
11200	CHEK2	CHK2 checkpoint homolog (S. pombe)	47,6	0	18	12
60	ACTB	actin, beta	47,2	0	45	24
6391	SDHC	succinate dehydrogenase complex, subunit C	46,543	9	0	-1,81
466	ATF1	activating transcription factor 1	46,5	0	3	5
1387	CREBBP	CREB binding protein	46,35	0	2	4,5
5930	RBBP6	retinoblastoma binding protein 6	46,325	2	24	7,75
1277	COL1A1	collagen, type I, alpha 1	46,318	10	5	-1,06
6317	SERPINB3	serpin peptidase inhibitor, clade B	46,25	3	30	7,5
6390	SDHB	succinate dehydrogenase complex, subunit B, iron	45,95	0	6	6,5
3909	LAMA3	laminin, alpha 3	45,7	0	32	19
801	CALM1	calmodulin 1 (phosphorylase kinase, delta)	45,2	0	45	24
7057	THBS1	thrombospondin 1	44,85	0	12	9,5
1019	CDK4	cyclin-dependent kinase 4	44,564	4	4	1,88
4478	MSN	moesin	44,534	2	0	1,78
5879	RAC1	ras-related C3 botulinum toxin substrate 1	44,5	0	23	15

7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	44,357	8	2	-1,19
3838	KPNA2	karyopherin alpha 2	44,35	72	0	-14,5
3320	HSP90AA1	heat shock protein 90kDa alpha (cytosolic)	44,239	52	9	-4,13
4301	MLLT4	myeloid/lymphoid or mixed-lineage leukemia	44,164	0	1	3,88
29893	PSMC3IP	PSMC3 interacting protein	44	0	36	20
4683	NBN	nibrin	43,8	0	5	6
4820	NKTR	natural killer-tumor recognition sequence	43,8	0	5	6
9793	CKAP5	cytoskeleton associated protein 5	43,65	0	4	5,5
3437	IFIT3	interferon-induced protein with	43	0	56	30
8125	ANP32A	acidic (leucine-rich) nuclear phosphoprotein	42,6	0	39	22
3726	JUNB	jun B proto-oncogene	42,534	2	0	1,78
4926	NUMA1	nuclear mitotic apparatus protein 1	42,35	0	2	4,5
10801	SEPT9	septin 9	42,35	0	2	4,5
3106	HLA-B	major histocompatibility complex, class I, B	42	97	0	-20
29	ABR	active BCR-related gene	41,6	0	17	12
6119	RPA3	replication protein A3, 14kDa	41,6	0	17	12
3326	HSP90AB1	heat shock protein 90kDa alpha (cytosolic)	40,4	39	0	-8
26511	CHIC2	cysteine-rich hydrophobic domain 2	40,35	0	2	4,5
4302	MLLT6	myeloid/lymphoid or mixed-lineage leukemia	40,35	0	2	4,5
7994	MYST3	MYST histone acetyltransferase	40,35	0	2	4,5
23352	ZUBR1	zinc finger, UBR1 type 1	40,25	0	8	7,5
1974	EIF4A2	eukaryotic translation initiation factor 4A	40,239	3	13	4,13
3123	HLA-DRB1	major histocompatibility complex, class II	39,9	0	19	13
3659	IRF1	interferon regulatory factor 1	39,3	0	16	11
9110	MTMR4	myotubularin related protein 4	39	0	58	30
2131	EXT1	exostoses (multiple) 1	38,6	3	3	2
578	BAK1	BCL2-antagonist/killer 1	38,55	0	10	8,5
7170	TPM3	tropomyosin 3	38,534	3	2	1,78
3945	LDHB	lactate dehydrogenase B	38,5	0	24	15
6814	STXBP3	syntaxin binding protein 3	38,4	0	51	28
1643	DDB2	damage-specific DNA binding protein 2, 48kDa	38,399	3	0	1,33

22931	RAB18	RAB18, member RAS oncogene family	38,35	0	2	4,5
5868	RAB5A	RAB5A, member RAS oncogene family	38,35	0	2	4,5
23011	RAB21	RAB21, member RAS oncogene family	38,35	0	2	4,5
2621	GAS6	growth arrest-specific 6	38,25	0	8	7,5
4258	MGST2	microsomal glutathione S-transferase 2	37,8	0	5	6
56898	BDH2	3-hydroxybutyrate dehydrogenase, type 2	37,8	0	47	26
26509	FER1L3	fer-1-like 3, myoferlin (C. elegans)	37,8	0	48	26
3059	HCLS1	hematopoietic cell-specific Lyn substrate 1	37,6	0	18	12
23157	SEPT6	septin 6	36,534	2	0	1,78
51517	NCKIPSD	NCK interacting protein with SH3 domain	36,534	2	0	1,78
219541	MED19	mediator of RNA polymerase II transcription,	36,534	2	0	1,78
79689	STEAP4	STEAP family member 4	36,534	2	0	1,78
6318	SERPINB4	serpin peptidase inhibitor, clade B	36,5	2	12	5
7037	TFRC	transferrin receptor (p90, CD71)	36,35	0	2	4,5

Tabela 17- Os 100 primeiros genes (ordenados por pontuação), para o tecido "testículo".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
4538	MT-ND4	mitochondrially encoded NADH dehydrogenase 4	218,25	1455	2	727,5
3507	IGHM	immunoglobulin heavy constant mu	155,2	4	150	304
4893	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	141,6	0	21	112
3538	IGLC2	immunoglobulin lambda constant 2	122,4	0	36	208
1936	EEF1D	eukaryotic translation elongation factor 1	112,8	0	31	176
3502	IGHG3	immunoglobulin heavy constant gamma 3	107,2	0	26	144
51237	PACAP	proapoptotic caspase adaptor protein	102,8	0	25	136
28786	IGLV4-3	immunoglobulin lambda variable 4-3	97,2	0	23	124
3537	IGLC1	immunoglobulin lambda constant 1 (Mcg marker)	97,2	0	23	124
335	APOA1	apolipoprotein A-I	93,6	0	21	112
2678	GGT1	gamma-glutamyltransferase 1	81,05	0	13	63,5
2023	ENO1	enolase 1, (alpha)	69,4	0	10	48
4763	NF1	neurofibromin 1	67,711	4	0	2,37
238	ALK	anaplastic lymphoma kinase (Ki-1)	61,768	3	0	2,56
6647	SOD1	superoxide dismutase 1, soluble	60,2	0	9	44
3500	IGHG1	immunoglobulin heavy constant gamma 1	59,8	10	38	36
3126	HLA-DRB4	major histocompatibility complex, class II	58,2	0	9	44
5159	PDGFRB	platelet-derived growth factor receptor, beta	57,834	2	0	2,78
2120	ETV6	ets variant gene 6 (TEL oncogene)	57,834	2	0	2,78
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	57,768	3	0	2,56
92140	MTDH	metadherin	56,3	0	7	31
10607	TBL3	transducin (beta)-like 3	55,8	0	8	36
718	C3	complement component 3	55,8	3	16	36
2260	FGFR1	fibroblast growth factor receptor 1	54,564	8	0	1,88
400	ARL1	ADP-ribosylation factor-like 1	54,3	0	7	31

28299	IGKV1-5	immunoglobulin kappa variable 1-5	54,3	0	7	31
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	51,725	2	2	5,75
387	RHOA	ras homolog gene family, member A	50,3	0	5	21
7175	TPR	translocated promoter region	49,834	2	0	2,78
2033	EP300	E1A binding protein p300	49,834	2	0	2,78
1277	COL1A1	collagen, type I, alpha 1	49,834	2	0	2,78
5914	RARA	retinoic acid receptor, alpha	49,768	3	0	2,56
6418	SET	SET translocation (myeloid leukemia-associated)	49,711	4	0	2,37
2263	FGFR2	fibroblast growth factor receptor 2	48,05	1	0	-3,5
613	BCR	breakpoint cluster region	47,834	2	0	2,78
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	47,834	2	0	2,78
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	47,768	3	0	2,56
648	BMI1	B lymphoma Mo-MLV insertion region (mouse)	47,711	4	0	2,37
2130	EWSR1	Ewing sarcoma breakpoint region 1	45,639	6	0	2,13
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	44,48	12	0	1,6
5371	PML	promyelocytic leukemia	43,834	2	0	2,78
8837	CFLAR	CASP8 and FADD-like apoptosis regulator	43,834	2	0	2,78
8021	NUP214	nucleoporin 214kDa	43,711	4	0	2,37
6794	STK11	serine/threonine kinase 11	43,711	4	0	2,37
3320	HSP90AA1	heat shock protein 90kDa alpha (cytosolic)	43,687	5	0	2,29
1870	E2F2	E2F transcription factor 2	43,2	6	17	24
2521	FUS	fusion (involved in t(12;16) in malignant liposarcoma)	41,711	4	0	2,37
4926	NUMA1	nuclear mitotic apparatus protein 1	41,687	5	0	2,29
684	BST2	bone marrow stromal cell antigen 2	41	2	4	10
51119	SBDS	Shwachman-Bodian-Diamond syndrome	40,391	0	2	7,97
6421	SFPQ	splicing factor proline/glutamine-rich	39,834	2	0	2,78
7184	HSP90B1	heat shock protein 90kDa beta (Grp94), member	39,834	2	0	2,78
8028	MLLT10	myeloid/lymphoid or mixed-lineage leukemia	39,834	2	0	2,78
84168	ANTXR1	anthrax toxin receptor 1	39,834	2	0	2,78
5327	PLAT	plasminogen activator	39,8	0	6	26

283651	C15orf21	chromosome 15 open reading frame 21	39,8	0	6	26
9793	CKAP5	cytoskeleton associated protein 5	39,768	3	0	2,56
10641	TUSC4	tumor suppressor candidate 4	39,687	5	0	2,29
1974	EIF4A2	eukaryotic translation initiation factor 4A	39,639	6	0	2,13
324	APC	adenomatosis polyposis coli	38	26	0	-1
9266	PSCD2	pleckstrin homology, Sec7 and coiled-coil	37,834	2	0	2,78
3421	IDH3G	isocitrate dehydrogenase 3 (NAD+) gamma	37,8	0	6	26
399818	LOC399818	similar to CG9643-PA	37,8	0	6	26
64983	MRPL32	mitochondrial ribosomal protein L32	37,8	0	6	26
117584	RFFL	ring finger and FYVE-like domain containing 1	37,8	0	6	26
6233	RPS27A	ribosomal protein S27a	36,725	0	4	15,75
4629	MYH11	myosin, heavy polypeptide 11, smooth muscle	36,564	8	0	1,88
427	ASAH1	N-acylsphingosine amidohydrolase (acid	36,3	0	5	21
29774	DKFZP434 P211	POM121-like protein	36,3	0	5	21
7037	TFRC	transferrin receptor (p90, CD71)	35,768	3	0	2,56
60528	ELAC2	elaC homolog 2 (E. coli)	35,711	4	0	2,37
7249	TSC2	tuberous sclerosis 2	35,711	4	0	2,37
54952	TRSPAP1	tRNA selenocysteine associated protein 1	35,55	17	32	18,5
404734	MASK-BP3	MASK-4E-BP3 alternate reading frame gene	34,725	0	4	15,75
196883	ADCY4	adenylate cyclase 4	34,725	0	4	15,75
3028	HADH2	hydroxyacyl-Coenzyme A dehydrogenase, type II	34,725	0	4	15,75
54882	ANKHD1	ankyrin repeat and KH domain containing 1	34,725	0	4	15,75
8861	LDB1	LIM domain binding 1	34,5	2	6	15
199775	C19orf9	chromosome 19 open reading frame 9	34,2	26	36	14
506	ATP5B	ATP synthase, H+ transporting, mitochondrial F1	33,75	2	5	12,5
3550	IK	IK cytokine, down-regulator of HLA II	33,711	4	0	2,37
116379	IL22RA2	interleukin 22 receptor, alpha 2	33,6	0	3	12
9352	TXNL1	thioredoxin-like 1	33,6	0	3	12
9744	CENTB1	centaurin, beta 1	33,6	0	3	12
7311	UBA52	ubiquitin A-52 residue ribosomal protein fusion	33,6	0	3	12
6154	RPL26	ribosomal protein L26	33,6	0	3	12

3514	IGKC	immunoglobulin kappa constant	33,3	5	7	11
967	CD63	CD63 molecule	32,309	31	0	-1,03
9733	SART3	squamous cell carcinoma antigen recognised by T	32,05	1	0	-3,5
1509	CTSD	cathepsin D (lysosomal aspartyl peptidase)	31,834	2	0	2,78
51614	ERGIC3	ERGIC and golgi 3	31,834	2	0	2,78
5932	RBBP8	retinoblastoma binding protein 8	31,768	3	0	2,56
9500	MAGED1	melanoma antigen family D, 1	31,768	3	0	2,56
10807	SDCCAG3	serologically defined colon cancer antigen 3	31,711	4	0	2,37
4678	NASP	nuclear autoantigenic sperm protein	31,711	4	0	2,37
7076	TIMP1	TIMP metallopeptidase inhibitor 1	31,65	7	4	5,5
1789	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta	30,391	0	2	7,97
51434	ANAPC7	anaphase promoting complex subunit 7	30,391	0	2	7,97
6388	SDF2	stromal cell-derived factor 2	29,834	2	0	2,78
55813	UTP6	UTP6, small subunit (SSU) processome component, homolog (yeast)	29,834	2	0	2,78

Tabela 18- Os 100 primeiros genes (ordenados por pontuação), para o tecido "sistema nervoso central".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
80975	TMPRSS5	transmembrane protease, serine 5 (spinesin)	189,6	0	373	432
2120	ETV6	ets variant gene 6 (TEL oncogene)	124,8	0	45	56
1277	COL1A1	collagen, type I, alpha 1	115,6	0	43	52
389	RHOC	ras homolog gene family, member C	114,8	0	116	136
10514	MYBBP1A	MYB binding protein (P160) 1a	108	1	202	160
4627	MYH9	myosin, heavy polypeptide 9, non-muscle	93,2	0	27	34
1278	COL1A2	collagen, type I, alpha 2	92,8	0	63	76
10	NAT2	N-acetyltransferase 2	92,4	0	73	88
283209	PGM2L1	phosphoglucomutase 2-like 1	91,2	0	88	104
348	APOE	apolipoprotein E	89,6	0	77	92
3872	KRT17	keratin 17	88,4	0	73	88
1290	COL5A2	collagen, type V, alpha 2	86,4	0	76	88
1801	DPH1	DPH1 homolog (S. cerevisiae)	85,4	950	0	-78
9867	PJA2	praja 2, RING-H2 motif	85,2	0	71	84
84790	TUBA6	tubulin, alpha 6	84,8	0	46	56
25	ABL1	v-abl Abelson murine leukemia viral oncogene	81,65	0	2	5,5
79902	NUP85	nucleoporin 85kDa	81,6	0	59	72
2969	GTF2I	general transcription factor II, i	80,8	0	47	56
5966	REL	v-rel reticuloendotheliosis viral oncogene	79,95	0	3	6,5
1191	CLU	clusterin	78,8	128	2346	56
3187	HNRPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	78	0	51	60
3188	HNRPH2	heterogeneous nuclear ribonucleoprotein H2	78	0	49	60
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	77,9	15	19	3
4155	MBP	myelin basic protein	76,8	4	138	56
10492	SYNCRIP	synaptotagmin binding, cytoplasmic RNA	76,8	0	46	56
2308	FOXO1A	forkhead box O1A (rhabdomyosarcoma)	75,9	0	9	13
8481	OFD1	oral-facial-digital syndrome 1	75,6	0	44	52
5908	RAP1B	RAP1B, member of RAS oncogene family	75,3	0	7	11
3265	HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene	74,534	3	0	1,78
673	BRAF	v-raf murine sarcoma viral oncogene homolog B1	71,539	5	12	5,13

4478	MSN	moesin	71,3	0	7	11
2316	FLNA	filamin A, alpha (actin binding protein 280)	71	0	33	40
4926	NUMA1	nuclear mitotic apparatus protein 1	70,4	2	44	28
2534	FYN	FYN oncogene related to SRC, FGR, YES	69,3	0	7	11
4763	NF1	neurofibromin 1	67,939	2	2	3,13
7157	TP53	tumor protein p53 (Li-Fraumeni syndrome)	67,789	3	2	2,63
3105	HLA-A	major histocompatibility complex, class I, A	65,4	0	39	48
3569	IL6	interleukin 6 (interferon, beta 2)	65,4	0	40	48
7431	VIM	vimentin	65,4	7	185	48
7913	DEK	DEK oncogene (DNA binding)	65	0	6	10
5781	PTPN11	protein tyrosine phosphatase, non-receptor	64,25	0	4	7,5
3164	NR4A1	nuclear receptor subfamily 4, group A, member 1	64,2	0	37	44
83937	RASSF4	Ras association (RalGDS/AF-6) domain family 4	63,6	0	8	12
7150	TOP1	topoisomerase (DNA) I	61,95	0	3	6,5
7507	XPA	xeroderma pigmentosum, complementation group A	61,95	0	3	6,5
3932	LCK	lymphocyte-specific protein tyrosine kinase	61,65	0	2	5,5
7450	VWF	von Willebrand factor	61,2	0	28	34
4671	BIRC1	baculoviral IAP repeat-containing 1	60,5	0	11	15
302	ANXA2	annexin A2	60,2	0	35	44
607	BCL9	B-cell CLL/lymphoma 9	59,789	3	2	2,63
6191	RPS4X	ribosomal protein S4, X-linked	59,4	0	41	48
4650	MYO9B	myosin IXB	58,45	0	26	31,5
51768	TM7SF3	transmembrane 7 superfamily member 3	58,2	0	36	44
6185	RPN2	ribophorin II	58,2	3	88	44
2260	FGFR1	fibroblast growth factor receptor 1	58,125	2	3	3,75
10916	MAGED2	melanoma antigen family D, 2	58	0	15	20
5159	PDGFRB	platelet-derived growth factor receptor, beta	57,939	2	2	3,13
64098	PARVG	parvin, gamma	57,8	0	30	36
800	CALD1	caldesmon 1	57,8	0	29	36
572	BAD	BCL2-antagonist of cell death	57,65	0	2	5,5
27316	RBMX	RNA binding motif protein, X-linked	57,6	0	34	42
5644	PRSS1	protease, serine, 1 (trypsin 1)	57,2	0	27	34
1281	COL3A1	collagen, type III, alpha 1 (Ehlers-Danlos	57	0	33	40
4539	MT-ND4L	mitochondrially encoded NADH 4L	56,7	9	187	39

27242	TNFRSF21	tumor necrosis factor receptor superfamily	56,45	0	26	31,5
633	BGN	biglycan	56,45	0	26	31,5
55830	GLT8D1	glycosyltransferase 8 domain containing 1	56,4	0	31	38
10983	CCNI	cyclin I	56,4	0	31	38
1982	EIF4G2	eukaryotic translation initiation factor 4	56,3	0	16	21
29894	CPSF1	cleavage and polyadenylation specific factor 1	55,8	0	30	36
9219	MTA2	metastasis associated 1 family, member 2	55,8	2	57	36
23166	STAB1	stabilin 1	55,8	0	30	36
6616	SNAP25	synaptosomal-associated protein, 25kDa	55,1	204	0	-17
4540	MT-ND5	mitochondrially encoded NADH dehydrogenase 5	54,9	6	111	33
9741	LAPTM4A	lysosomal-associated protein transmembrane 4	54,764	0	12	15,88
2060	EPS15	epidermal growth factor receptor pathway	54,25	0	4	7,5
7175	TPR	translocated promoter region	53,95	0	3	6,5
11064	CEP110	centrosomal protein 110kDa	53,65	0	2	5,5
1495	CTNNA1	catenin (cadherin-associated protein), alpha	52,5	0	11	15
595	CCND1	cyclin D1	51,95	0	3	6,5
5395	PMS2	PMS2 postmeiotic segregation increased	51,95	0	3	6,5
1019	CDK4	cyclin-dependent kinase 4	51,65	0	2	5,5
9401	RECQL4	RecQ protein-like 4	51,65	0	2	5,5
4609	MYC	v-myc myelocytomatosis viral oncogene homolog	51,65	0	2	5,5
472	ATM	ataxia telangiectasia mutated	51,65	0	2	5,5
23085	RAB6IP2	RAB6 interacting protein 2	51,65	0	2	5,5
648	BMI1	B lymphoma Mo-MLV insertion region (mouse)	51,3	0	7	11
471	ATIC	5-aminoimidazole-4-carboxamide ribonucleotide	50,55	0	5	8,5
1508	CTSB	cathepsin B	50,3	0	16	21
10342	TFG	TRK-fused gene	50,25	0	4	7,5
4302	MLLT6	myeloid/lymphoid or mixed-lineage leukemia	50,25	0	4	7,5
51517	NCKIPSD	NCK interacting protein with SH3 domain	50,25	0	4	7,5
2033	EP300	E1A binding protein p300	50,164	3	5	3,88
2130	EWSR1	Ewing sarcoma breakpoint region 1	49,65	0	2	5,5
9112	MTA1	metastasis associated 1	49,639	2	0	2,13

2073	ERCC5	excision repair cross-complementing rodent	48,534	3	0	1,78
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	48,164	30	56	3,88
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	48,05	3	4	3,5
8021	NUP214	nucleoporin 214kDa	47,875	1	5	6,25
4916	NTRK3	neurotrophic tyrosine kinase, receptor, type 3	47,65	0	2	5,5

Tabela 19- Os 100 primeiros genes (ordenados por pontuação), para o tecido "próstata".

GeneID	*Símbolo*	*Descrição*	*Pontuação*	*Normal*	*Tumor*	*Fator*
56851	C15orf24	chromosome 15 open reading frame 24	141,6	0	341	272
63934	ZNF667	zinc finger protein 667	86,4	0	106	88
29927	SEC61A1	Sec61 alpha 1 subunit (S. cerevisiae)	84	0	95	80
26580	BSCL2	Bernardinelli-Seip congenital lipodystrophy 2	78,6	0	75	62
29101	SSU72	SSU72 RNA polymerase II CTD phosphatase homolog	76,8	0	66	56
604	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	74,6	2	0	2
25	ABL1	v-abl Abelson murine leukemia viral oncogene	74,6	2	0	2
2067	ERCC1	excision repair cross-complementing rodent	68,3	0	35	31
7178	TPT1	tumor protein, translationally-controlled 1	66,4	12	345	38
2064	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene	64,6	2	0	2
79058	ASPSCR1	alveolar soft part sarcoma chromosome region	63,95	0	4	6,5
5930	RBBP6	retinoblastoma binding protein 6	62	0	21	20
134359	FLJ35779	hypothetical protein FLJ35779	59,4	0	56	48
71	ACTG1	actin, gamma 1	58,45	4	110	31,5
595	CCND1	cyclin D1	58,382	0	6	7,94
3508	IGHMBP2	immunoglobulin mu binding protein 2	58,2	0	51	44
23646	PLD3	phospholipase D family, member 3	57	0	47	40
378	ARF4	ADP-ribosylation factor 4	57	0	46	40
607	BCL9	B-cell CLL/lymphoma 9	56,6	2	0	2
3817	KLK2	kallikrein 2, prostatic	56,564	2	73	31,88
3725	JUN	v-jun sarcoma virus 17 oncogene homolog (avian)	56,4	0	31	28
255394	TCP11L2	t-complex 11 (mouse) like 2	55,2	0	39	34
2260	FGFR1	fibroblast growth factor receptor 1	54,6	2	0	2
81608	FIP1L1	FIP1 like 1 (S. cerevisiae)	54,2	0	13	14
1613	DAPK3	death-associated protein kinase 3	53,75	103	0	-12,5
2073	ERCC5	excision repair cross-complementing rodent	52,5	0	2	5
7170	TPM3	tropomyosin 3	52,382	0	6	7,94
9098	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	51,825	2	2	2,75
3845	KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene	51,8	0	3	6
6421	SFPQ	splicing factor proline/glutamine-rich	50,382	0	6	7,94

8208	CHAF1B	chromatin assembly factor 1, subunit B (p60)	49	0	8	10
5573	PRKAR1A	protein kinase, cAMP-dependent, regulatory	48,35	2	6	4,5
5216	PFN1	profilin 1	47,8	2	58	26
8021	NUP214	nucleoporin 214kDa	47,764	100	6	-5,88
6418	SET	SET translocation (myeloid leukemia-associated)	46,6	2	0	2
1277	COL1A1	collagen, type I, alpha 1	46,393	9	0	-1,31
2130	EWSR1	Ewing sarcoma breakpoint region 1	46,164	3	7	3,88
6117	RPA1	replication protein A1, 70kDa	46	0	20	20
4926	NUMA1	nuclear mitotic apparatus protein 1	45,689	63	2	-5,63
7428	VHL	von Hippel-Lindau tumor suppressor	44,6	2	0	2
208	AKT2	v-akt murine thymoma viral oncogene homolog 2	44,6	2	0	2
10276	NET1	neuroepithelial cell transforming gene 1	44,6	0	23	22
2353	FOS	v-fos FBJ murine osteosarcoma viral oncogene	44,564	14	0	-1,88
2132	EXT2	exostoses (multiple) 2	44,5	0	2	5
5980	REV3L	REV3-like, catalytic subunit of DNA polymerase	44	0	21	20
22948	CCT5	chaperonin containing TCP1, subunit 5 (epsilon)	44	0	20	20
1662	DDX10	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10	43,95	0	4	6,5
10342	TFG	TRK-fused gene	43,8	0	3	6
3320	HSP90AA1	heat shock protein 90kDa alpha (cytosolic),	43,789	8	11	2,63
6596	SMARCA3	SWI/SNF related, matrix associated, actin	43,6	0	11	12
7060	THBS4	thrombospondin 4	42,725	0	16	15,75
1387	CREBBP	CREB binding protein	42,6	2	0	2
11159	RABL2A	RAB, member of RAS oncogene family-like 2A	41,8	0	3	6
57018	CCNL1	cyclin L1	41,6	0	11	12
3106	HLA-B	major histocompatibility complex, class I, B	41,4	0	18	18
5664	PSEN2	presenilin 2 (Alzheimer disease 4)	41,2	0	25	24
23600	AMACR	alpha-methylacyl-CoA racemase	40,7	2	64	29
7991	TUSC3	tumor suppressor candidate 3	40,6	2	0	2
4914	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	40,6	2	0	2
8028	MLLT10	myeloid/lymphoid or mixed-lineage leukemia	40,5	0	2	5
1213	CLTC	clathrin, heavy polypeptide (Hc)	40,5	3	10	5
4221	MEN1	multiple endocrine neoplasia I	40,5	0	2	5

8301	PICALM	phosphatidylinositol binding clathrin assembly	40,5	0	2	5
3927	LASP1	LIM and SH3 protein 1	40,5	0	2	5
1967	EIF2B1	eukaryotic translation initiation factor 2B	40,3	0	22	21
471	ATIC	5-aminoimidazole-4-carboxamide ribonucleotide	40,275	3	8	4,25
4302	MLLT6	myeloid/lymphoid or mixed-lineage leukemia	40,2	1	0	-4
8835	SOCS2	suppressor of cytokine signaling 2	40,2	0	13	14
1974	EIF4A2	eukaryotic translation initiation factor 4A	39,9	3	4	3
84168	ANTXR1	anthrax toxin receptor 1	39,75	19	0	-2,5
25800	SLC39A6	solute carrier family 39 (zinc transporter)	38,6	0	23	22
9470	EIF4E2	eukaryotic translation initiation factor 4E	38,4	0	31	28
3191	HNRPL	heterogeneous nuclear ribonucleoprotein L	38,4	4	97	28
1499	CTNNB1	catenin (cadherin-associated protein), beta 1	38	5	0	-1
3074	HEXB	hexosaminidase B (beta polypeptide)	37,8	0	29	26
7936	RDBP	RD RNA binding protein	37,8	0	28	26
140606	SELM	selenoprotein M	37,8	0	28	26
3303	HSPA1A	heat shock 70kDa protein 1A	37,5	0	27	25
3304	HSPA1B	heat shock 70kDa protein 1B	37,5	0	27	25
1361	CPB2	carboxypeptidase B2 (plasma, carboxypeptidase U)	37,4	3	50	18
3169	FOXA1	forkhead box A1	37,4	3	50	18
3500	IGHG1	immunoglobulin heavy constant gamma 1	37,3	0	9	11
6159	RPL29	ribosomal protein L29	37,2	0	25	24
57153	SLC44A2	solute carrier family 44, member 2	37,2	0	25	24
9547	CXCL14	chemokine (C-X-C motif) ligand 14	37,1	0	17	17
7392	USF2	upstream transcription factor 2, c-fos	36,725	0	16	15,75
7417	VDAC2	voltage-dependent anion channel 2	36,6	0	24	22
9793	CKAP5	cytoskeleton associated protein 5	36,6	2	0	2
4162	MCAM	melanoma cell adhesion molecule	36,6	2	0	2
4629	MYH11	myosin, heavy polypeptide 11, smooth muscle	36,552	207	81	-1,84
23551	RASD2	RASD family, member 2	36,339	7	0	-1,13
7799	PRDM2	PR domain containing 2, with ZNF domain	35,95	0	4	6,5
5225	PGC	progastricsin (pepsinogen C)	35,9	0	12	13
4869	NPM1	nucleophosmin	35,8	0	3	6
10916	MAGED2	melanoma antigen family D, 2	35,8	0	3	6

9168	TMSB10	thymosin, beta 10	35,6	0	11	12
9322	TRIP10	thyroid hormone receptor interactor 10	35,4	0	18	18
3910	LAMA4	laminin, alpha 4	35,4	3	50	18
23593	HEBP2	heme binding protein 2	35,1	0	17	17
2316	FLNA	filamin A, alpha (actin binding protein 280)	34,925	208	9	-9,75