

# **ESTUDO DO PERFIL DA EXPRESSÃO GÊNICA GLOBAL EM LEUCEMIAS LINFÓIDES AGUDAS DE LINHAGENS DE CÉLULAS B E T**

**Aluna: Diana Azevedo Queiroz**

**Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira**

Dissertação de Mestrado apresentada ao Instituto de Biologia da Universidade Estadual de Campinas para a obtenção do título de Mestre em Biologia Funcional e Molecular, na área de Bioquímica.

**Campinas - SP**

**2004**

## **Ficha catalográfica**

**Campinas, 06 de abril de 2004**

**BANCA EXAMINADORA**

Prof. Dr. Gonçalo Amarante Guimarães Pereira (orientador)

Deptº Genética e Evolução – Instituto de Biologia – Unicamp

Dr. Fernando Lopes Alberto

Instituto Fleury de Ensino e Pesquisa – São Paulo

Prof. Dr. Antônio Carlos Boschero

Deptº Fisiologia – Instituto de Biologia – Unicamp

Prof. Dr. José Camilo Novello

Deptº Bioquímica – Instituto de Biologia – Unicamp

Dr. Marcos Antônio de Oliveira

Laboratório Nacional de Luz Síncrotron – Campinas

Dedico este trabalho à minha família que sempre foi minha principal fonte de paz e maior razão de estar aqui. Com todo o apoio, confiança e amor que tive de vocês fica fácil transpor qualquer obstáculo imposto e alcançar a glória. Obrigada por serem a imagem do que eu venho tentando me tornar.

## **AGRADECIMENTOS**

Ao Gonçalo, que com esse jeito baiano de ser teve minha admiração e respeito crescente ao longo desses seis anos de orientação.

Ao Fernando Lopes e ao André Fattori, pela enorme ajuda na criação e desenvolvimento desse trabalho, eu não teria convencido meu orientador sozinha!

À Maricene Sabha, pela luta incansável atrás do sucesso com os microarranjos e pela grande amizade concretizada com sabor de vitória.

Às toscas, Ana, Raquel, Camila e Cinthia por todas as ajudas, broncas e discussões, eu sei que não foi fácil me convencerem de que eu não estava certa e me colocarem novamente no eixo da minha pesquisa. Ainda bem que também tive o amor de vocês fora do laboratório.

Aos maravilhosos homens da Bioinformática, tio Du, Marujo, Tchelo, Luciano, Diego, Danilo e Fernando por terem atendido aos meus inúmeros gritos de socorro e por terem sido sempre um ótimo colírio aos meus olhos nos dias de trabalho.

Aos colegas antigos de laboratório, Ana Paula, Carlinha, Odalys, Fernandinha, Marcos e Vitor Genu por todo o ensinamento e carinho dentro e fora do laboratório.

Aos meus novos amigos de trabalho, Julio, Zapa, Hugo, Zé, Dani, Diana, Leandra, Marcio, tenho certeza que vocês irão manter esse delicioso ambiente de trabalho, que bom que pude conhecê-los.

Aos queridíssimos Vitão, Anderson, Gisele, Alessandra, Gabriel, Carol e Sula, amigos eternos com quem tive a honra de trabalhar e o prazer de adorar.

Aos meus amigos irmãos, Lu, Tales, Flavia, Rubens, Paty e Tati que me deram o prazer de relaxar, contar longas estórias e dar várias gargalhadas.

À Andréia e Eliane, por suportarem minhas trapalhadas com tanto carinho e bom humor, é muito bom contar com a organização e amizade de vocês.

À Andréia e Marina, secretárias do Departamento de Bioquímica por serem tão atenciosas e por torcerem de coração pelas nossas conquistas.

À Maria Inês, onde quer que esteja, por ter sem nenhum benefício próprio acreditado cegamente em mim tornando possível que esse trabalho fosse realizado .

Ao CNPQ, pelo apoio financeiro concedido.

## **ÍNDICE**

### **1. INTRODUÇÃO**

|      |                                   |    |
|------|-----------------------------------|----|
| 1.2. | Câncer : Processo e epidemiologia | 1  |
| 1.3. | Projeto Genoma Humano do Câncer   | 4  |
| 1.4. | Leucemia Linfóide Aguda           | 5  |
| 1.5. | Microarranjos de DNA              | 10 |

### **2. OBJETIVOS**

14

### **3. MATERIAIS E MÉTODOS**

|        |  |    |
|--------|--|----|
| 3.1.   | Extração de RNA total                                    | 15 |
| 3.2.   | Preparação dos clones                                    | 17 |
| 3.3.   | Amplificação dos clones                                  | 19 |
| 3.4.   | Confecção dos microarranjos de DNA                       | 20 |
| 3.4.1. | Impressão dos clones                                     | 20 |
| 3.4.2. | Pós Processamento dos microarranjos                      | 21 |
| 3.4.3. | Marcação das sondas                                      | 21 |
| 3.4.4. | Purificação das sondas                                   | 22 |
| 3.4.5. | Hibridação   | 23 |
| 3.4.6. | Lavagem da lâmina  | 24 |
| 3.5.   | Re-seqüenciamento  | 25 |
| 3.6.   | Confirmação da expressão gênica por <i>northern Blot</i> | 26 |

### **4. RESULTADOS**

|      |   |    |
|------|---|----|
| 4.1. | Obtenção de RNA total   | 28 |
| 4.2. | Seleção e deposição dos clones                                | 29 |
| 4.3. | Confecção de microarranjos invertidos                         | 30 |
| 4.4. | Confecção de microarranjos heterólogos                        | 32 |
| 4.5. | Confecção de microarranjos com cDNA purificado                | 33 |
| 4.6. | Análise das triplicatas                                       | 35 |
| 4.7. | Seleção e Seqüenciamento dos clones para <i>Northern Blot</i> | 36 |

|                   |  |    |
|-------------------|--|----|
| 4.8.              | Confirmação da expressão gênica por <i>Northern Blot</i> | 37 |
| <b>5.</b>         | <b>DISCUSSÃO</b>   |    |
| 5.1.              | Padronização da técnica de microarranjos de DNA          | 38 |
| 5.2.              | Genes candidatos   | 40 |
| <b>6.</b>         | <b>CONCLUSÃO</b>   | 43 |
| <b>7.</b>         | <b>PERSPECTIVAS</b>                                      | 44 |
| <b>8.</b>         | <b>REFERÊNCIAS BIBLIOGRÁFICAS</b>                        | 45 |
| <b>APÊNDICE A</b> |  | 50 |
| <b>APÊNDICE B</b> |  | 57 |
| <b>APÊNDICE C</b> |  | 70 |

## **ABSTRACT**

The comprehension of data generated by many finished or in progress Genome Projects will provide essential informations to molecular biology research. The identification of differentially expressed genes in distinct stages of any given disease, for example, will permit an enormous advance in the determination of its physiopathology, identifying molecular markers for its prognostic and possible targets for the intervention of the process. Among human diseases, cancer consists in one of the main subjects of research in clinical and molecular biology, basically because involves a complex and poorly understood process of cellular reprogramming. Among the cancer subtypes, acute lymphoblastic leukemia is prevalent in Brazil, ranked in 8<sup>th</sup> position in number of clinical cases and decorrent deaths. Attempting to increase the understanding of cancer molecular bases, Human Cancer Genome Project (LICR/ FAPESP) had detached by generation of thousand of expressed sequences of normal and tumoral human tissues, increasing the possibility of discovery of novel genes related to carcinogenic processes. Thus, the present work employed the Human Cancer Genome Project fragments to investigate the global gene expression profile in acute lymphoblastic leukemia by DNA microarrays technique. It was identified the genes of IgK VLJ (imunoglobulina kappa light chain VLJ region) e TPT1 (Tumor Protein, translationally-controlled 1), which differential expression was confirmed by Northern blot assays.

## RESUMO

A compreensão das informações geradas pelos diversos Projetos Genoma já totalizados ou em andamento constitui-se em uma ferramenta essencial à pesquisa em biologia molecular. A identificação de genes diferencialmente expressos em estágios distintos de uma dada doença, por exemplo, constituirá um enorme avanço na determinação de sua fisiopatologia, fornecendo marcadores moleculares para seu prognóstico e possíveis alvos para a intervenção do processo. Dentre as moléstias humanas, o câncer constitui-se em um dos principais alvos de pesquisa em biologia molecular e clínica médica, basicamente porque envolve um quadro de reprogramação celular complexo e ainda pouco compreendido. Entre os vários tipos de cânceres, a leucemia linfóide aguda é bastante prevalente no Brasil, classificando-a como 8º lugar em número de casos e óbitos decorrentes. Buscando compreender as bases moleculares do câncer, o Projeto Genoma Humano do Câncer (LICR/ FAPESP) se destacou por ter gerado milhares de seqüências codantes do DNA de tecidos humanos normais e tumorais, possibilitando a descoberta de novos genes relacionados aos processos carcinogênicos. Assim, o presente estudo utilizou fragmentos produzidos durante o Projeto Genoma Humano do Câncer para investigar o perfil de expressão gênica em leucemias linfóides agudas através da técnica de microarranjos de DNA. Foram identificados como diferencialmente expressos os genes IgK VLJ (“*immunoglobulin kappa light chain VLJ region*”) e TPT1 (“*Tumor Protein , translationally-controlled 1*”), cuja expressão foi confirmada por ensaios de Northern blot.

## **1. INTRODUÇÃO**

### **1.1. Câncer: processo e epidemiologia**

Os cânceres compreendem um conjunto de mais de cem doenças caracterizadas pelo crescimento desordenado de células que invadem tecidos e órgãos e que podem espalhar-se para outras regiões do corpo, em um processo denominado metástase (Cecil, 1986)

Ao acúmulo localizado de células que se multiplicam vagarosamente e se assemelham ao tecido original é dado o nome de tumores benignos, enquanto o acúmulo de células atípicas ao tecido recebe o nome de tumor ou neoplasia maligna. Os tumores malignos são classificados inicialmente pelo local onde ocorrem, pela velocidade com que se multiplicam e pela capacidade de invadirem outros órgãos.(Altman & Schwartz, 1983)

Carcinogênese é o processo de formação das células cancerosas e apresenta três fases distintas de acordo com o seu grau de desenvolvimento. Na etapa de iniciação, as células normais sofrem efeito de agentes carcinogênicos que provocam sua regressão a um estado desdiferenciado. Em seguida, na etapa denominada promoção, as células sofrem malignização através da ação de oncopromotores. A última fase, conhecida como progressão, caracteriza-se pela multiplicação descontrolada e irreversível das células afetadas, e o câncer é considerado instalado e em evolução. O desenvolvimento de todas essas etapas depende das características individuais, da intensidade de estímulos carcinogênicos, da presença de agentes externos ou internos do tecido afetado (localização primária) (Hoffbrand & Pettit, 1993).

Dentre os principais fatores apontados como agentes carcinogênicos, cerca de 80 a 90% são considerados extrínsecos, ou seja, relacionados às condições a que o indivíduo está submetido, como o ambiente ocupacional (indústrias, poluição), os hábitos alimentares e práticas rotineiras (tabagismo, alcoolismo, sedentarismo, entre outros). Os fatores intrínsecos, ou seja, aqueles ligados à predisposição genética e ao perfil imunológico, contribuem com apenas 10% do total de casos de câncer (Altman & Schwartz, 1983).

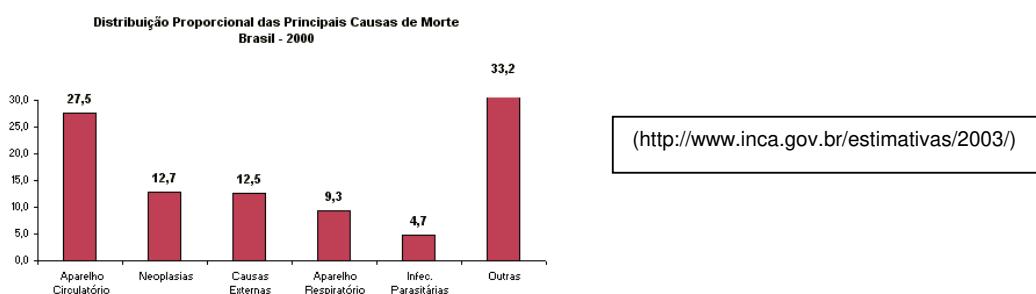
A epidemiologia do câncer permite o estudo da distribuição das várias formas dessa doença na população, levando-se em conta a localização primária, o perfil histopatológico e a extensão anatômica dos tumores.

Segundo o Instituto Nacional do Câncer (INCA), órgão vinculado ao Ministério da Saúde, a mortalidade decorrente de neoplasias malignas constitui-se atualmente na segunda causa de óbitos no Brasil, e este índice apresenta uma tendência crescente e proporcional ao avanço da faixa etária (<http://www.inca.gov.br/>).

Dentre os tipos de cânceres de alta prevalência, as leucemias têm despertado grande interesse na comunidade médica e científica mundial, devido às suas peculiaridades epidemiológicas. Segundo dados internacionais fornecidos pela Agência Internacional de Pesquisas de Câncer (IARC), em 2000, as leucemias ocupavam o 10º lugar em número de casos (2,5%) e 7º lugar em número de mortes (3,9%) por câncer em todo o mundo. Em relação à distribuição regional dos casos, o Brasil apresentava-se como o 8º país com maior número de óbitos por leucemia, na América do Sul 45% dos casos de morte por câncer ocorrem no Brasil (<http://www-dep.iarc.fr/globocan/globocan.html>).

Atualmente, a taxa de incidência do câncer infantil tem crescido em torno de 1% ao ano, porém com o progresso da ciência e tecnologia, tem sido inversamente proporcional ao crescimento da taxa de mortalidade e estima-se que a cura global esteja em torno de 85%. No Brasil, o câncer é a terceira causa de morte por doença entre um e 14 anos ( Rodrigues & Camargo, 2003).

A epidemiologia dos cânceres em crianças e adolescentes aponta a leucemia linfóide aguda como uma das cinco principais causas de morte infantil por câncer no mundo. Esta estatística justifica o imenso interesse em identificar os agentes carcinogênicos responsáveis pelo desenvolvimento e progressão das leucemias em especial.



Estimativas para o ano 2003 de número de óbitos por câncer, no Brasil, por região.  
[\(http://www.inca.gov.br/estimativas/2003/\)](http://www.inca.gov.br/estimativas/2003/)

| Localização Primária        | Norte        | Nordeste      | Centro-Oeste | Sudeste       | Sul           |
|-----------------------------|--------------|---------------|--------------|---------------|---------------|
| Pele não Melanoma           | 55           | 170           | 55           | 410           | 185           |
| Mama Feminina               | 205          | 1.280         | 450          | 5.570         | 1.830         |
| Traquéia, Brônquio e Pulmão | 565          | 1.840         | 910          | 8.450         | 4.465         |
| Estômago                    | 575          | 1.540         | 590          | 6.130         | 2.310         |
| Próstata                    | 260          | 1.480         | 510          | 4.370         | 1.610         |
| Colo do Útero               | 320          | 880           | 320          | 1.730         | 860           |
| Côlon e Reto                | 180          | 710           | 380          | 4.900         | 1.800         |
| Esôfago                     | 105          | 590           | 260          | 2.980         | 1.660         |
| Leucemias                   | 235          | 850           | 290          | 2.320         | 910           |
| Cavidade Oral               | 115          | 430           | 160          | 1.870         | 670           |
| Pele Melanoma               | 20           | 120           | 55           | 600           | 330           |
| Outras Localizações         | 1.745        | 8.670         | 3.155        | 29.660        | 11.265        |
| <b>TOTAL</b>                | <b>4.380</b> | <b>18.560</b> | <b>7.135</b> | <b>68.990</b> | <b>27.895</b> |

## **1.2 Projeto Genoma Humano do Câncer**

O Projeto Genoma Humano do Câncer (HCGP) foi uma iniciativa lançada em 1999 pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) juntamente com o Instituto Ludwig de Pesquisas do Câncer (LICR), com o objetivo de identificar genes expressos em cânceres prevalentes no Brasil. Até então uma grande quantidade de seqüências expressas (ESTs) advindas de outros projetos genomas tinham sido depositadas em bancos de dados, mas 91% do total dessas seqüências estavam concentradas nas extremidades desses genes (65% representavam a extremidade 3' do cDNA e 26%, a extremidade 5'). Baseado nesse fato o projeto foi delineado para utilizar uma nova metodologia, denominada ORESTES (*Open Reading Frame Expressed Sequence Tags*), fundamentada na amplificação de mRNAs com o uso de oligonucleotídeos inespecíficos em condições de baixa estringência, a qual gera preferencialmente a extensão da região central da parte codante dos genes expressos (Dias Neto *et al.*, 2000). O consórcio de seqüenciamento montado contou com uma equipe de 25 laboratórios de seqüenciamento, que gerou um banco de dados de mais de 1.200.000 ORESTES, obtidos a partir de 24 tecidos diferentes entre tumorais e normais (Dias Neto *et al.*, 2000).

A combinação das seqüências 3' e 5' dos Projetos Genomas Humanos com os ESTs internos do HCGP podem gerar transcritos inteiros facilitando o trabalho de anotação desses genes (Camargo *et al.*, 2001). Esta iniciativa tem possibilitado descobrir novos genes relevantes aos processos carcinogênicos, catalogar genes altamente expressos em vários cânceres e selecionar grupos de genes para análises de tumores específicos por metodologias de análise de expressão gênica

em larga escala como, por exemplo, microarranjos de DNA (Strausberg *et al*, 2002).

Nesse intuito, diversos pesquisadores participantes do “The Human Cancer Genome Project”, do “Cancer Genome Anatomy Project Annotation Consortium” e do consórcio de seqüenciamento estabelecido pelo Projeto Genoma Humano do Câncer se reuniram, juntaram informações de ESTs advindos dos vários projetos e compilaram 1250 genes humanos expressos em processos malignos ao qual deram o nome de genes relacionados a câncer (CR - “Cancer Related”). As seqüências foram depositadas em bancos de dados públicos gerando dois milhões de EST's relativos a tumores humanos e seu correspondente tecido normal. Suas análises foram a base dos dados extraídos para a montagem desse estudo e podem ser consultados no **Apêndice A** (Camargo et al, 2003).

## 1.2. Leucemia Linfóide Aguda

Leucemias são cânceres que acometem os glóbulos brancos sanguíneos (leucócitos). Os fatores responsáveis pelas alterações do processo de divisão e diferenciação dos leucócitos são pouco conhecidos e a principal característica desse processo leucêmico é o acúmulo de células jovens (blásticas) anormais na medula óssea, prejudicando ou impedindo a produção de leucócitos e das demais células sanguíneas, como hemárias e plaquetas, tendo como consequências da falência medular, o surgimento de infecções, anemia e hemorragias constantes (Hoffbrand & Pettit, 1993).

As leucemias são inicialmente classificadas quanto às células predominantemente envolvidas (linfóide ou mielóide) e quanto ao tempo de

sobrevida sem tratamento (aguda ou crônica). A distinção entre leucemias agudas linfóides e mielóides é crítica para o sucesso de um tratamento. A primeira classificação foi baseada em análises citoquímicas que divide as leucemias em dois subgrupos principais: precursores mielóides (leucemia mielóide aguda ou crônica) e precursores linfóides (leucemia linfóide aguda ou crônica). A confirmação molecular desses subgrupos deu-se com o desenvolvimento de anticorpos que reconhecem moléculas de superfície específicas às células linfóides ou mielóides (Harris *et al*, 1997).

Leucemias linfóides agudas (LLA) são neoplasias agressivas caracterizadas pela presença de precursores linfóides (linfoblastos) na medula óssea e/ou no sangue periférico. Segundo a classificação FAB (*French-American-British*), as LLAs podem ser diagnosticadas pela presença de mais de 30% de células blásticas na medula óssea, enquanto a WHO -*World Health Organization* considera 20% suficiente para o diagnóstico dessa doença. LLAs são o tipo de leucemia mais comum na infância (80% dos casos), decrescendo para um índice de 20% em adultos (Schoch *et al.*, 2002; Lai *et al*, 2000).

Atualmente as leucemias agudas, tanto mielóides quanto linfóides, têm sido associadas a translocações cromossômicas específicas, como por exemplo, a translocação t(12;21)(p13;q22) que ocorre em 25% dos pacientes com LLA (Golub, 1999).

As anormalidades cromossômicas, tanto de número (hipoploidias, hiperploidias e pseudoploidias) como de estrutura (translocações, deleções parciais ou monossomias parciais), são encontradas em 80% das crianças com LLAs e em 70% dos casos em adultos. As anormalidades cromossômicas estão

diretamente associadas a determinados tipos da doença e podem fornecer um grande apoio ao prognóstico eficaz das diferentes LLAs (Lafage-Pochitaloff & Charrin, 2003).

Estudos de expressão gênica têm sido usados como alternativa para se mapear translocações cromossômicas. Em LLA infantil de células B foram identificadas 6 anormalidades cromossômicas diferentes e esse resultado foi analisado estatisticamente, comparado com dados de outras técnicas comumente utilizadas e verificou-se uma acurácia de 96% no diagnóstico da doença (Staudt, 2003).

Cromossomo Philadelphia (Ph), a primeira aberração cromossônica especificamente associada a uma doença humana, é um marcador característico de leucemia mielóide crônica, embora também tenha sido descrito em algumas linhagens de leucemias agudas. Ele é produto da translocação (9;22)(q34;q11), o que acarreta molecularmente a fusão do gene BCR (cromossomo 9) e do gene ABL (translocado do cromossomo 22), elevando a atividade da tirosina quinase. Isso causa uma desordem na leucemogênese, desencadeando um quadro patogênico. Leucemia linfóide aguda positiva para cromossomo Philadelphia (LLA Ph+) ocorre em 10 a 40% dos casos em adultos e 2 a 5% em crianças e está associada a um pior prognóstico da doença (*Drexler et al*, 1999).

O prognóstico inicial de leucemias agudas é usualmente baseado em análises morfológicas e citoquímicas segundo a classificação da WHO, e normalmente complementados com dados do cariótipo e imunofenótipo. No entanto, análises citométricas com uso de anticorpos para receptores específicos

de células T ou B permitem até 98% de precisão na distinção entre origem mielóide ou linfóide de leucemias agudas (Bain, 1999; Lai *et al*, 2000).

A linhagem linfóide é composta por três tipos celulares: B, T e NK. As células B desenvolvem-se no fígado fetal ou medula óssea adulta e são responsáveis pelo reconhecimento de抗ígenos intactos, pela codificação de um receptor de superfície específico e pela liberação de anticorpos (imunoglobulinas). As células T desenvolvem-se no timo, possuem receptores específicos e são responsáveis pela liberação de citocinas, por destruir células infectadas e coordenar a resposta imune. As células NK – natural killer (exterminadora natural) se diferenciam na medula óssea e são capazes de reconhecer e destruir determinadas células tumorais e infectadas por vírus.

O diagnóstico de leucemias linfóides agudas possibilita a diferenciação em LLA de células T, quando mais de 20% dos blastos forem positivos para os抗ígenos pan-T (CD7 e CD3), e em LLA de linhagem B, quando mais de 20% das células leucêmicas apresentarem expressão completa de imunoglobulina de superfície (SmIg) (Foon & Todd, 1986).

As leucemias linfóides agudas de células T representam 15% das LLA e estão normalmente associadas a translocações específicas envolvendo genes receptores de células T, uma rara translocação não associada a esses receptores vem sendo descrita e pode estar envolvida em rearranjos complexos (Douet-Guilbert *et al*, 2003).

Muitos estudos têm sido realizados no sentido de se verificar a relação entre LLAs e linfomas. Linfomas são um tipo de câncer que se origina de linfócitos presentes em tecidos característicos por seu acúmulo e se desenvolve no sistema

linfático. Apesar da leucemia também ser desse grupo de células a grande diferença é que esta se origina na medula óssea, o grande centro de produção e formação dessas células. O sistema FAB classifica citologicamente diferentes subtipos de LLA em L1, L2 e L3. O tipo L1 é caracterizado pela presença de células leucêmicas regulares e homogêneas enquanto no tipo L2 são encontradas células mais heterogêneas. O tipo L3 inclui o Linfoma de Burkitt, um tipo de leucemia linfóide aguda de células B maduras com características distintas, associado ao vírus Epstein Barr, colocando-o como uma entidade única de genótipo e imunofenótipo (Emanuel et al, 1984).

A precisão do diagnóstico de LLA tipo 3 tem uma importância crítica para a sobrevida dos pacientes, pois este tipo de leucemia não responde satisfatoriamente aos tratamentos que são normalmente bem-sucedidos em outros tipos de LLA (Lai et al, 2000).

O subtipo L1 é mais freqüente em crianças (76 a 89%) do que em adultos (31 a 43%), o subtipo L2 é mais freqüente em adultos (49 a 60%) que em crianças (14 a 22%) e o subtipo L3, raramente encontrado, é mais freqüente em crianças do que em adultos (Plasschaert et al , 2004).

Com a adoção de técnicas para as comparações entre as leucemias linfóides agudas, como a imunocitoquímica e a citometria de fluxo com anticorpos monoclonais, tornou-se possível um melhor conhecimento dos mecanismos de regulação da hematopoiese e formação de células leucêmicas, permitindo relacionar esses processos ao papel das diversas moléculas de adesão celular, como a família CD44, encontradas em vários processos e estágios de instalação de quadros leucêmicos (Cavalcanti Jr, 2001).

O principal desafio no tratamento dos cânceres e em especial das leucemias tem sido estabelecer métodos eficientes de identificação, classificação e portanto prognóstico eficaz. Estes fatores são de central importância ao tratamento dos pacientes afetados porque baseiam-se primariamente nas particularidades morfológicas do tumor que, apesar de apresentarem uma histopatologia semelhante, normalmente progridem diferentemente e assim devem responder de formas distintas às terapias (Lai *et al*, 2000).

Em uma abordagem alternativa, a utilização de microarranjos de anticorpos permitiu o estudo de mais de 60抗ígenos em leucócitos ou células leucêmicas. Nesta técnica, uma suspensão de células é aplicada sobre o arranjo, permitindo a ligação apenas entre as células que expressem o antígeno correspondente a determinado anticorpo, sendo possível não só identificar linhagens e estágios de maturação como descobrir novos marcadores de prognósticos. Antígenos de superfície foram identificados em resíduos com valores mínimos de desenvolvimento das doenças (Belov *et al*, 2001).

### **1.3. Microarranjos de DNA**

Para compreender o enorme potencial dos bancos de dados gerados pelos diversos projetos genoma, será de grande interesse obter um acesso rápido aos padrões de expressão dos genes e ESTs (*Expressed Sequence Tags*) depositados, tanto em um nível transcricional quanto tradicional. Essa progressão simultânea de genomas de organismos procariotos e eucariotos permite uma caracterização gênica em larga escala baseada em índices de similaridade entre seqüências (Blumental, 2001). Para acompanhar o dinamismo do

seqüenciamento de genomas, técnicas baseadas em análises de níveis de expressão, como os microarranjos de DNA sobressaem-se como uma ferramenta importante para identificar os perfis transcripcionais de tecidos e condições de interesse (Brown *et al*, 2000).

A técnica de microarranjos de DNA consiste na deposição de fragmentos de DNA em lâminas de vidro em altas densidades através de um equipamento automatizado. Paralelamente, duas populações de mRNA obtidas em condições experimentais distintas são submetidas a reações de transcrição reversa na presença de citosinas ou uracilas contendo diferentes fluorocromos fotossensíveis, Cy3 (emissão de fluorescência verde) e Cy5 (emissão de fluorescência vermelha), de forma a marcar diferencialmente cada população de cDNA gerada (Hedge *et al*, 2000). Estas sondas são então misturadas e hibridadas contra as seqüências complementares depositadas na lâmina, permitindo inferir a abundância relativa de transcritos em cada um dos tratamentos realizados (Schena *et al*, 1996).

Cada sonda é excitada separadamente por uma fonte de luz e os sinais são armazenados em canais de imagens distintos para posterior quantificação. Após o processamento das imagens é possível aplicar ferramentas computacionais para extrair informações sobre o padrão de expressão de cada gene sob as condições experimentais amostradas, a partir da razão entre a emissão de fluorescência verde ou vermelha observada em cada posição na lâmina (Kerr & Churchill, 2001).

Microarranjos de DNA podem revelar a alteração na expressão gênica decorrente da inibição ou superexpressão de alguns fatores regulatórios, e mesmo que os passos intermediários pós-transcripcionais e pré-traducionais não possam ser visualizados, é possível uma análise quantitativa dessas alterações. Assim, as

transformações oncogênicas que envolvem translocações e/ou amplificações de genes que codificam fatores de regulação, ou mesmo a perda desses genes por deleções cromossomais, são processos que podem ser identificados através da análise de expressão seguida por técnicas mais sensíveis (Lee *et al*, 2000). Esses eventos podem ser experimentalmente modelados e o uso de microarranjos de DNA podem ser usados na compreensão das consequências destes eventos na regulação celular (Staudt, 2002).

O conhecimento da expressão gênica fornece um atalho conveniente para a descoberta de dados efetivos que levem à compreensão da diversidade clínica e biológica das células normais e malignas, sendo possível um agrupamento em categorias de diagnósticos e prognósticos. A aplicação de microarranjos de DNA em busca de anormalidades cromossômicas pode substituir múltiplos testes de diagnóstico, aumentar a acurácia de prognóstico alcançando o sucesso na terapia. (Staudt, 2003).

Neste contexto, a análise global da expressão gênica através de microarranjos de DNA tem se mostrado uma abordagem altamente eficiente na busca por marcadores moleculares específicos a cada subtipo tumoral, pois permite o monitoramento simultâneo da expressão de milhares de genes em um número limitado de experimentos (Armstrong *et al*, 2002) (Golub, 1999).

Considerando a quantidade de informações que podem ser extraídas dos microarranjos de DNA, a utilização das bibliotecas produzidas durante o Projeto Genoma Humano do Câncer (LICR/ FAPESP) irá facilitar a busca por alterações moleculares estabelecidas na leucemia linfóide aguda, uma vez que permite que sejam utilizados fragmentos de cDNA obtidos a partir de mensagens produzidas

em tecidos humanos normais e tumorais. Assim, a produção de experimentos que avaliem simultaneamente a expressão de centenas de genes candidatos a estarem relacionados com o câncer, relacionando-os com as mensagens derivadas de linhagens de células de leucemia linfóide aguda, fornecerá pistas importantes para a compreensão das bases moleculares responsáveis pelas modificações subcelulares ocorridas neste tipo de câncer.

## **2. OBJETIVOS**

O objetivo do presente trabalho foi avaliar, utilizando a técnica de microarranjos de DNA, a possível expressão diferencial de genes previamente relacionados a processos carcinogênicos em linhagens leucêmicas de células B e T.

### **3. MATERIAIS E MÉTODOS**

#### **3.1. Extração de RNA total**

O RNA total usado como controle foi extraído diretamente das tonsilas normais de pacientes sadios, isolado através do protocolo baseado em *TRIZOL* (*Life Technologies*). Para isso, o tecido foi macerado na presença de nitrogênio líquido, transferido para tubos *eppendorfs* de 1,5 ml e imediatamente adicionado *TRIZOL* para garantir degradação mínima das moléculas de RNA. O volume do reagente utilizado manteve a proporção de 1 ml para cada 50-100mg de tecido. Após a homogeneização, o material foi centrifugado a 14000g por 10 minutos para sedimentar o material insolúvel, como restos celulares, proteínas e gordura. O sobrenadante foi transferido para novo tubo *eppendorf* de 1,5 ml e incubado a 30ºC por 5 minutos para permitir a completa dissociação das nucleoproteínas. Em seguida adicionou-se 200µl de clorofórmio, incubando-se novamente a 30º C por 3 minutos e centrifugando-se por 15 minutos a 12000 g. Após esta etapa, a solução divide-se em duas fases, uma orgânica e uma aquosa, esta última contendo as moléculas de RNA. Transferiu-se esta fase aquosa para novo tubo *eppendorf* de 1,5 ml e adicionou-se 500µl de álcool isopropílico para precipitar o RNA. Após incubar o material a 30ºC durante 5 minutos, centrifugou-se a 12000g por 10 minutos e, ao fim desse passo, pôde ser observado nos tubos um *pellet* referente ao RNA total precipitado. Descartou-se o sobrenadante e o RNA foi lavado com 1 ml de etanol 75º C. As amostras foram novamente centrifugadas, desta vez a 7500 g durante 5 minutos apenas para sedimentar o RNA, descartando-se em seguida o sobrenadante. Após deixar secar brevemente, o material foi ressuspêndido em

água DEPC (dietilpirocarbonato), incubando-se a 50ºC por cerca de 10 minutos para garantir a completa dissolução do RNA. O material foi armazenado a –80ºC.

O protocolo foi conduzido separadamente para o material extraído em cada tonsila de cada paciente (5 materiais diferentes) e a qualidade do RNA total foi avaliada em gel denaturante de agarose. O controle foi obtido através da junção do RNA de três tonsilas formando um *pool* como controle a fim de minimizar a variabilidade individual.

O RNA total das linhagens leucêmicas em estudo foi disponibilizado pelo médico André Fattori do Hemocentro da Unicamp. Células leucêmicas são capazes de formar colônias *in vitro* e neste aspecto reside a facilidade de obtenção desse material. As células foram cultivadas numa densidade de aproximadamente  $1 \times 10^6$  células/ml em meio de cultura IMDM (Invitrogen) suplementado com 10% soro bovino fetal (Invitrogen), 100U/ml de penicilina e 100ug/ml de estreptomicina (Invitrogen). As culturas foram incubadas a 37°C, numa atmosfera úmida com 5% de CO<sub>2</sub>. Após a incubação as células foram colhidas, centrifugadas a 1500 RPM por 10 minutos à temperatura ambiente e ressuspensas em 1 ml de *TRIZOL* seguindo então o mesmo protocolo descrito anteriormente. As linhagens (cultura de células) usadas nesse projeto foram:

- Células de Daudi – Derivadas de um indivíduo de 16 anos com linfoma de Burkitt (leucemia linfóide aguda de células B maduras) por Klein & Klein em 1967.
- Células de Raji – Extraídas de um linfoma de Burkitt da maxila esquerda de um paciente de 11 anos por Pulvertaft em 1963.

- Células de Jurkat – Coletadas de sangue periférico de um indivíduo de 14 anos, portador de leucemia linfóide aguda de células T - por Schneider e colaboradores.
- Células CCRF-CEM – Obtida de sangue periférico de uma paciente de 4 anos com leucemia linfóide aguda de células T derivadas de Foley e colaboradores em 1964.

As descrições dessas linhagens podem ser encontradas no site da coleção de células da Universidade Federal do Rio de Janeiro (<http://www.nce.ufrj.br/bcrj>) e no site da ATCC – American Type Culture Collection (<http://www.atcc.org>).

### **3.2. Preparação dos clones**

Inicialmente, o protocolo adotado consistiu da amplificação dos insertos por PCR diretamente das culturas bacterianas armazenadas a  $-70^{\circ}\text{C}$ . Neste caso, alíquotas de  $3\mu\text{l}$  foram empregadas para as reações. Posteriormente, devido à necessidade de uma purificação desse material para a obtenção de resultados melhores, a amplificação começou a ser feita a partir de plasmídeos tratados com proteinase K. Para isso, as amostras armazenadas em culturas permanentes foram novamente inoculadas em placas de inóculo de 96 poços (marca *Sorenson*) contendo 1ml de meio LB com ampicilina 50mg/ml, *overnight* a  $37^{\circ}\text{C}$  com agitação de 300 rpm e utilizadas na preparação de DNA por lise alcalina.

A mini-preparação de DNA foi feita em placas de 96 poços e seguiu o seguinte protocolo: centrifugação das placas de inóculo a 4000 rpm por 8 minutos para sedimentação das células; descarte do sobrenadante com secagem sobre

papel absorvente por aproximadamente 1 minuto; lavagem em solução 1; adição de 80 $\mu$ l de solução 1, agitação com selo por 2 minutos em vórtex na velocidade máxima; em uma microplaca ELISA de fundo U de 96 poços (*Greiner*) foi adicionado 3 $\mu$ l de RNase (10mg/ml) em cada poço; transferência de 60 $\mu$ l da placa de inóculo para a microplacas de fundo U; adição de 60 $\mu$ l da solução 2; inversão das placas seladas; descanso por 5 minutos; centrifugação rápida para a solução se desprender do adesivo; adição de 60 $\mu$ l da solução 3, inversão das placas seladas; centrifugação rápida para a solução desgrudar do adesivo; incubação em estufa a 90 $^{\circ}$ C por exatos 30 minutos; gelo por 7 minutos; centrifugação a 4000 rpm, 20 $^{\circ}$ C, 10 minutos; montagem de uma microplaca fundo V com o filtro Millipore MAGV N22, transferência de 110 $\mu$ l da microplaca fundo U para o filtro/microplaca fundo V; centrifugação a 4000 rpm, 20 $^{\circ}$ C, 8 minutos; remoção do filtro; adição de 75  $\mu$ l de isopropanol, selagem (selo resistente a álcool) e inversão por 20 vezes; centrifugação a 4000 rpm, 20 $^{\circ}$ C, 45 minutos; descarte do sobrenadante e secagem da placa sobre papel absorvente; adição de 200ul de etanol 70% gelado; centrifugação a 4000 rpm, 20 $^{\circ}$ C, 10 minutos; descarte, secagem por no mínimo 1 hora; acréscimo de 60 $\mu$ l de H<sub>2</sub>O milli-Q autoclavada e armazenagem a -20 $^{\circ}$  C.

Os plasmídeos foram então diluídos (seguindo uma proporção de 1:50) em solução tampão de proteinase K, incubados por 15 minutos a 55 $^{\circ}$  C e 15 minutos a 80 $^{\circ}$  C e então armazenados a -20 $^{\circ}$  C para posterior amplificação dos insertos.

As soluções utilizadas foram:

|                                |   |
|--------------------------------|---|
| Solução 1                      | Glicose 20%; EDTA 0,5M pH 8,0; Tris HCl 1M pH 7,4 |
| Solução 2                      | NaOH 0,2M; SDS 1%                                 |
| Solução 3                      | Acetato de Potássio 3M pH5,2; Ac. Acético glacial |
| Solução Tampão de Proteinase K | Tris 10mM. EDTA 1mM, Proteinase K 50 µg/ml        |

### 3.3. Amplificação dos clones

As reações de PCR para a amplificação dos clones, conduzidas em placas de 96 poços, seguiram o seguinte protocolo (medidas referentes a uma reação): 18,2 µl água Milli-Q, 3,0µl Tampão de PCR 10X (Invitrogen), 0,8 µl MgCl<sub>2</sub> 50mM, 3,0 µl dNTP 1,25mM, 0,8 µl oligonucleotídeo M13 forward 5pmol/µl, 0,8 µl oligonucleotídeo M13 reverse 5pmol/µl, 0,4 µl Taq polimerase (Invitrogen) e 3 µl de DNA. O termociclador *GeneAmp® PCR System 9700 (Applied Byosystem)* foi ajustado para realizar o seguinte programa: 94° C – 2 minutos e posteriores 40 ciclos [94°C – 20 segundos; 52°C – 1 minuto; 72°C – 2 minutos].

As seqüências dos oligonucleotídeos utilizados para a amplificação dos insertos por PCR foram :

M13 PUC23 Forward: 5'- AGCGGATAACAATTTCACACAGG -3'

M13 PUC23 Reverse: 5'- CCCAGTCACGACGTTGTAAAACG -3'

Os mesmos foram utilizados no re-seqüenciamento desses fragmentos e na marcação da sonda para *Northern blot*.

### **3.4. Confecção dos microarranjos**

#### **3.4.1. Impressão dos clones**

O equipamento utilizado para construção dos microarranjos foi *Flexys® Robot* (*Perkin Elmer*) e, para digitalização, o *scanner GeneTAC™ Microarray Analyzer* (*Perkin Elmer*).

O protocolo adotado pelo nosso Laboratório de Genética e Evolução (LGE) prevê a utilização de lâminas de vidro tratadas com poli-L-lisina para a impressão dos clones de DNA. As lâminas tratadas com poli-L-lisina possuem uma superfície hidrofóbica e positivamente carregada, características que auxiliam a posicionar e reter as moléculas de DNA no exato local onde foram depositadas. A produção das lâminas tratadas com poli-L-lisina está sendo realizada conjuntamente com a Maricene Sabha, coordenadora da equipe de *Microarrays* do LGE.

Previamente à impressão dos clones, as placas de 96 poços contendo os produtos de PCR foram rearranjadas em placas de 384 poços, disposição que atende aos requisitos do equipamento *Flexys® Robot* (*Perkin Elmer*). Para isso, 3 µl do produto de PCR foram adicionados a 3 µl de 3xSSC e a transposição desse material atendeu a um mapa de consolidação que permite correlacionar as posições de cada clone em ambos os tipos de placas. O equipamento realizou a transferência do material das placas 384 poços para as lâminas de vidro automaticamente, em ambiente mantido a 20°C e umidade do ar 37%.

### **3.4.2. Pós-processamento dos microarranjos**

Após a impressão das lâminas, os microarranjos passaram por uma etapa de pós-processamento para a fixação definitiva do DNA à superfície da lâmina. O protocolo obedeceu a seguinte ordem: (1) reidratação do DNA (câmara úmida contendo 100 ml de 1xSSC, durante 2 hs); (2) secagem (bloco invertido a 85-90°C, durante 2 hs); (3) fixação (*cross-linking* por UV a 65 mJ); (4) fixação (banho de formalina, durante 1 minuto); (5) lavagem (imersão rápida, 5-10 vezes, em banho de água destilada); (6) secagem (centrifugação a 700 rpm, durante 7 minutos); (7) bloqueio da lâmina (imersão em solução de ácido succínico e ácido bórico durante 30 segundos, seguida por agitação no *shaker* orbital por 25 minutos); (8) desnaturação do DNA (banho em água destilada a 95°C, durante 5 minutos); (9) lavagem (imersão rápida, 5 vezes, em etanol 95%); (10) secagem (centrifugação a 700 rpm, durante 7 minutos); (11) armazenagem dos microarranjos em caixas de lâminas a temperatura ambiente.

### **3.4.3. Marcação das sondas**

Previamente à marcação das sondas, as amostras de RNA total foram quantificadas, de acordo com a fórmula:  $[RNA] = A_{260nm} \times 40\mu g/ml \times \text{fator de diluição}$ . Assim, buscou-se determinar o volume necessário de cada amostra para atingir a concentração ideal para a síntese das sondas, aproximadamente 50  $\mu g$  de RNA total.

O protocolo adotado foi o seguinte (medidas expressas em sua concentração final): (1) incubação do RNA total com oligo dT (70°C, durante 2 minutos); (2)

preparação das amostras para a reação de transcrição reversa, adicionando-se: 31,5 µl de uma solução contendo 1x PCR *buffer*, 2,5mM MgCl<sub>2</sub>, 10mM DTT, 500µM A/G/CTP + 40 µM TTP; 3,6 µl de Cy3 ou Cy5 e, por fim, 300 unidades da enzima transcriptase reversa (*Superscript II / GIBCO*) a cada amostra; (3) incubar à temperatura ambiente durante 10 minutos; (4) incubar a 39ºC durante 120 minutos, adicionando mais 75 U da enzima na metade deste período; (5) finalizar a reação através da desnaturação da enzima (incubar a 70ºC, durante 15 minutos); (6) adicionar 2 unidades de RNase H para degradar o RNA molde (incubar a 37ºC, durante 30 minutos); (7) misturar as amostras marcadas com os fluoróforos Cy3 e Cy5 em um único tubo (refazendo cerca de 180 µl).

Neste ponto, o material está na forma de cDNA marcado (cada sonda sintetizada foi incorporada com um fluoróforo distinto) e pronto para o próximo passo: a purificação das sondas para a eliminação do excesso de reagentes e material não incorporado.

#### **3.4.4. Purificação das sondas**

As etapas de purificação são conduzidas em 3 tipos distintos de colunas: por exclusão de tamanho, troca iônica e para concentração da amostra.

Antes de ser utilizada, a primeira coluna (*CENTRI-SEP / Princeton Separations*) foi hidratada por no mínimo 30 minutos em água DEPC. Só então as sondas foram adicionadas (cada 180 µl requerem 3 colunas distintas, pois o volume máximo de cada coluna CENTRI-SEP é de 60 µl). As colunas foram

centrifugadas a 750 g durante 2 minutos, e o material eluído (aproximadamente 60 µl) foi utilizado na coluna seguinte.

Na coluna de troca iônica (*QIAquick PCR Purification kit / QIAGEN*), ao material obtido no passo anterior foram adicionados 300 µl de tampão PB (*binding buffer*) (*QIAGEN*) e esta solução foi inserida na coluna, centrifugando-se em seguida a 14000 g durante 1 minuto e descartando-se o filtrado. Em seguida, adicionou-se 750µl de tampão PE (*wash buffer*) (*QIAGEN*), centrifugou-se por 1 minuto a 14000 g e novamente descartou-se o filtrado. A centrifugação foi repetida (14000 g por 1 minuto) e adiciona-se então 100 µl de tampão EB (*elution buffer*) (*QIAGEN*). Esta solução foi incubada a temperatura ambiente por 2 minutos e após este período, centrifugou-se a 14000 g por 2 minutos. O material eluído (aproximadamente 300 µl) foi utilizado na última coluna.

A este volume foram adicionados 170 µl de tampão TE pH 7.4 e esta solução foi aplicada em uma coluna *Microcon YM-30 (AMICON)* para a concentração das amostras purificadas. Centrifugou-se a coluna por 12 minutos a 10000 g e em seguida esta coluna foi transferida invertida para novo *tubo eppendorf de 1,5 ml*. Uma nova centrifugação, desta vez a 3000 g por 3 minutos, recuperou a amostra purificada e concentrada, em um volume final de aproximadamente 15 µl.

### **3.4.5. Hibridação**

A amostra obtida nas etapas anteriores foi incubada a 95ºC por 5 minutos para permitir sua completa desnaturação. Em seguida preparou-se a solução de hibridação, composta pelos seguintes reagentes (medidas expressas em

concentração final): DNA de esperma de salmão ( $1\mu\text{g}/\mu\text{l}$ ), 20x SSC, 10% SDS e o volume total das sondas desnaturadas, completando um volume final de  $28\ \mu\text{l}$  (suficiente para cobrir a área estampada de  $24 \times 40\ \text{mm}$ ).

Finalmente, a sonda foi aplicada sobre a lâmina e coberta por uma lamínula para garantir o contato completo entre as sondas e o material depositado. A hibridação prosseguiu a  $65^\circ\text{C}$  por um período de 16-17 horas, em um forno de hibridação preservado da luz (para proteção dos fluoróforos).

#### **3.4.6. Lavagem da lâmina**

Previamente ao processamento digital dos microarranjos foram necessárias 3 etapas de lavagens das lâminas para eliminar sinais inespecíficos (*background*). As soluções de lavagem foram, respectivamente,  $0,1\%\text{SDS} + 1\times\text{SSC}$  (solução 1),  $0,2\times\text{SSC}$  (solução 2) e  $0,05\times\text{SSC}$  (solução 3). Os banhos foram realizados imergindo-se a lâmina em cada uma das soluções a temperatura ambiente e agitando-se levemente por 2 minutos. Após os 3 banhos, a lâmina foi centrifugada a  $700\ \text{g}$  durante 7 minutos para garantir a secagem completa de sua superfície.

Neste ponto o microarranjo está pronto para ser analisado, podendo alternativamente ser armazenado a  $4^\circ\text{C}$  por semanas antes da fase de digitalização. Os microarranjos produzidos no presente trabalho foram imediatamente digitalizados e armazenados em caixa para lâminas à temperatura ambiente.

### **3.5. Re-Seqüenciamento**

Os clones selecionados após a análise dos microarranjos foram novamente seqüenciados para a confirmação das suas identidades utilizando-se o seqüenciador automático *ABI PRISM™ Model 377 (Perkin Elmer)* (Instituto de Biologia – Departamento de Genética e Evolução, Unicamp). O protocolo adotado para o seqüenciamento foi o seguinte (medidas referentes a uma reação): 1 µl de *ABI PRISM® BigDye™ Terminator v3.0 (Applied Biosystems)*, 1 µl oligonucleotídeo M13 5pmol/µl (forward ou reverse), 7 µl água Milli-Q e 1 µl de DNA. Os oligonucleotídeos utilizados foram os mesmos adotados no protocolo de amplificação dos clones da biblioteca. O termociclador *GeneAmp® PCR System 9700 (Applied Biosystem)* foi ajustado para realizar o seguinte programa: 96ºC – 2 minutos e posteriores 40 ciclos [96ºC – 20 segundos; 55ºC – 15 segundos; 60ºC – 4 minutos].

Após a amplificação, conduzida em placas de 96 poços, este material foi purificado segundo o seguinte protocolo (medidas referentes a uma reação): (1) precipitação do DNA (adição de 2µl de acetato de amônia 7,5M e 50 µl álcool etílico 100% e agitação para misturar, deixando em seguida 20 minutos em repouso sem exposição à luz a 4ºC); (2) sedimentação do DNA precipitado (centrifugação a 4000 rpm durante 40 minutos, na temperatura de 4ºC, descartando-se então o sobrenadante); (3) lavagem (adição de 200 µl etanol 70%); (4) sedimentação do DNA (centrifugação a 4000g durante 15 minutos, na temperatura de 4ºC, descartando-se novamente o sobrenadante); (5) secagem (centrifugação com a placa invertida sobre papel absorvente a 300 g por 10

segundos). Após secas, as amostras foram submetidas à denaturação, aplicação em gel e seqüenciamento automático.

As seqüências obtidas foram submetidas ao *software* de avaliação de qualidade e trimagem de vetor e contaminação *Phred* (EWING et al, 2002), a um alinhamento local para busca de similaridade com o próprio banco de dados do Projeto Genoma Humano do Câncer (BLASTn) e com bancos de dados de proteínas (BLASTx) atualizado em fevereiro de 2004. Com base nas identidades e no perfil de expressão observado nos microarranjos, alguns genes foram selecionados para confirmação de expressão por *Northern blot*.

### **3.6. Confirmação da expressão gênica por *northern blot***

O protocolo adotado para *Northern blot* foi o seguinte: 10 µg de cada uma das amostras de RNA total (Controle, CCRF-CEM, Jurkat, Daudi e Raji) foi aplicada em um gel denaturante de agarose e transferida para uma membrana de nylon *Hybond-N+* (*Amershan Pharmacia Biotech*). A membrana foi incubada a 80°C durante 2 horas para a fixação do RNA. Em seguida, foi preparada uma solução de pré-hibridização contendo os seguintes reagentes: DNA de esperma de salmão (10µg/µl) previamente denaturado (95°C por 10 minutos e gelo por 10 minutos), Denhardt's (100x), SSC 20x, SDS10%, formamida e água deionizada. O volume final desta solução (25 µl) foi adicionado à garrafa de hibridação contendo a membrana e este conjunto foi então incubado por 6 horas a 40°C em forno de hibridação *Big SHOT Hybridization Oven* (*Boekel Scientifics*). A pré-hibridização é necessária para diminuir a reatividade da membrana e assim minimizar sinais

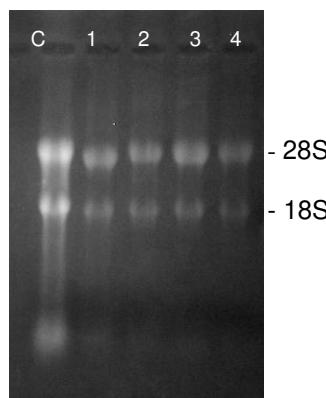
inespecíficos (*background*). Após este período, adicionou-se a sonda marcada por  $\alpha$ -ATP<sup>32</sup>, preparada segundo o seguinte protocolo: 5  $\mu$ l DNA desnaturado (produto de PCR obtido a partir de regiões dos diferentes genes testados) + 0,5mM dNTP (exceto ATP) + 1 $\mu$ g/ $\mu$ l oligo degenerado (PdN<sub>6</sub>) + 10x tampão React 2 + água desionizada + *Klenow* polimerase +  $\alpha$ -ATP<sup>32</sup>, totalizando um volume final de 20  $\mu$ l. Esta solução foi incubada a 37°C durante 30 minutos para permitir a polimerização da fita marcada de DNA. Em seguida, a sonda foi purificada com a coluna *Microspin* (*Amershan Biosciences*) para eliminar o excesso de nucleotídeos não incorporados e desnaturada (incubar a 95°C por 10 minutos e gelo por 10 minutos). Finalmente, a sonda foi adicionada à garrafa de hibridação e incubada a 42°C *overnight*.

Antes da etapa de exposição ao filme fotográfico, a membrana passou por uma fase de lavagem para eliminar a radiação em excesso. Esta lavagem consiste na substituição da solução de hibridação (que foi descartada em local apropriado) por soluções salinas em concentração serial: solução 1 (2xSSC + 0,1%SDS, 10 minutos a 25°C), solução 2 (1xSSC + 0,1%SDS, 15 minutos a 65°C) e solução 3 (0,1xSSC + 0,1%SDS, 15 minutos a 65°C). Após as lavagens, a membrana foi testada para a radioatividade por dosímetro para confirmar a eliminação do excesso de radiação e então colocada em um cassete com filme fotográfico e mantida 2 dias a -80°C. A fase de revelação foi conduzida em uma câmara escura, consistindo nas etapas de revelação (o filme é imerso e levemente agitado por 3 minutos em solução reveladora), lavagem (30 segundos imerso em água) e fixação (5 minutos imerso e levemente agitado em solução fixadora).

## 4. RESULTADOS

### 4.1. Obtenção de RNA total

A qualidade das amostras de RNA total obtidas a cada linhagem leucêmica e em tonsilas normais foram avaliadas através de gel denaturante de agarose, conforme a **Figura 1** e quantificadas por espectrofotometria, conforme a **Tabela 1**, afim de determinar sua integridade, concentração e pureza.



**Figura 1.** Eletroforese em gel denaturante de agarose corado por brometo de etídio para visualização de qualidade das amostras de RNA total extraídas de células leucêmicas e normais. C: Controle; 1: CCRF-CEM; 2: Jurkat; 3: Daudi; 4: Raji. Foram aplicadas 2 µl de cada amostra.

**Tabela 1.** Valores de RNA total, obtidos por espectrofotometria. A pureza é calculada pela razão das medidas de absorbância a um comprimento de onda de 260nm e 280nm ( $A_{260}/A_{280}$ ) e deve permanecer em torno de 2,0. A concentração está expressa em µg/µl

| Amostra  | Pureza | Concentração |
|----------|--------|--------------|
| CCRF-CEM | 2,13   | 7,16         |
| JURKAT   | 2,42   | 6,20         |
| DAUDI    | 2,67   | 6,86         |
| RAJI     | 2,14   | 6,38         |
| NORMAL   | 2,63   | 4,42         |
| NORMAL   | 2,62   | 5,50         |
| NORMAL   | 2,52   | 6,04         |

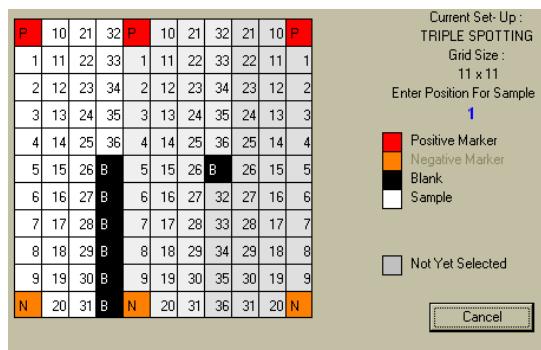
A partir desses dados, os volumes foram calculados para atender aos 50µg necessários para a marcação com fluoróforos Cy3 e Cy5 na produção dos microarranjos. Esse material também foi usado para a validação dos genes por *Northern blot*.

#### **4.2. Seleção e deposição dos clones:**

Os ORESTES gerados pelo HCGP foram analisados de acordo com registros de envolvimento dos genes a que pertencem em diferentes tipos de câncer, tendo sido a partir daí 1127 genes selecionados (**Apêndice A**). A seqüência completa desses genes foi comparada com o banco de dados de ORESTES usando a ferramenta BlastN. Dos “hits” resultantes, os que constavam dos grupos CM2 e CM4, cujos clones físicos encontram-se armazenados no nosso Laboratório de Genômica e Expressão, foram selecionados, rearranjados em 12 placas de 96 poços e utilizados para a deposição nos microarranjos. No **Apêndice B** encontra-se a lista dos genes e dos 1152 clones ORESTES usados neste trabalho.

A deposição dos clones nas lâminas de vidro tratadas com poli-lisina seguiu um desenho composto por 32 grades, pois o equipamento automatizado trabalha com 32 pinos. No presente trabalho, foram desenhadas 32 grades de 11x11 compostas por 36 clones distintos em triplicata distribuídos em posições equivalentes em cada grade, totalizando 1152 no microarranjo todo.

O desenho dos microarranjos pode ser visualizado na **figura 2**.

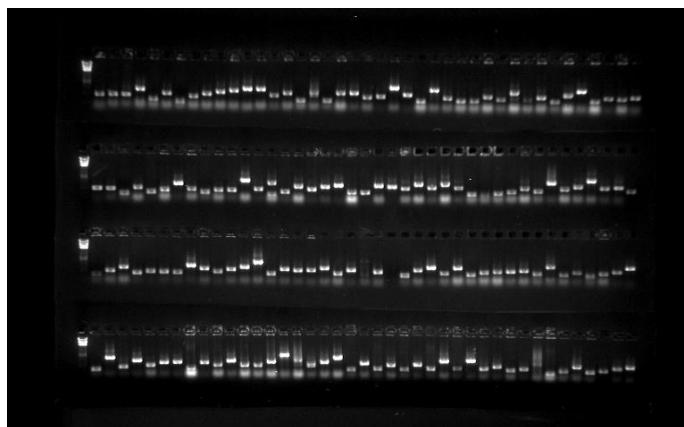


**Figura 2.** Desenho esquemático da disposição dos clones nas lâminas de vidro. Grid de 11x11. Cada microarranjo conta com 32 grids iguais a esse tendo 1.152 clones distintos em triplicata. P= Actina (controle positivo), N= DMSO (controle negativo) e B= blank, sem deposição de material.

#### **4.3. Confecção de microarranjos invertidos**

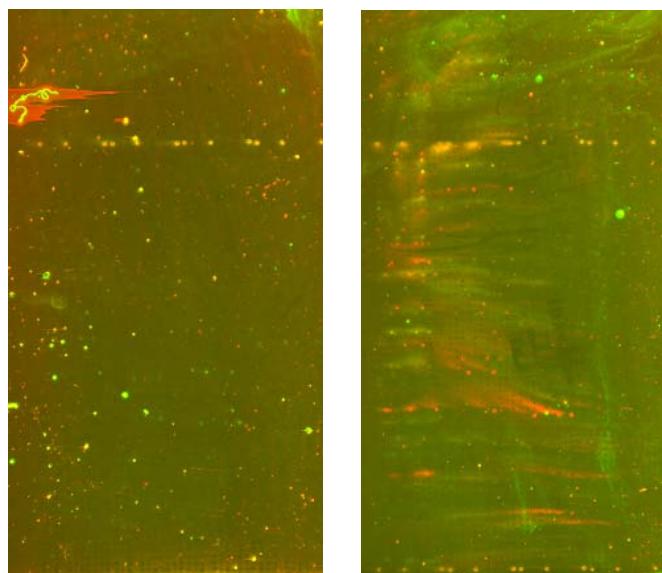
O presente trabalho contou com uma série de testes até que alcançássemos condições ótimas para o uso da técnica. Obtivemos resultados com dados relevantes ao estudo utilizando dois protocolos de preparação do material a ser amplificado. Inicialmente os clones não passaram por nenhum tipo de purificação.

Os fragmentos selecionados foram rearranjados em 12 placas de 96 poços e então amplificados diretamente por PCR. A qualidade das amostras pode ser verificada na **figura 3**.



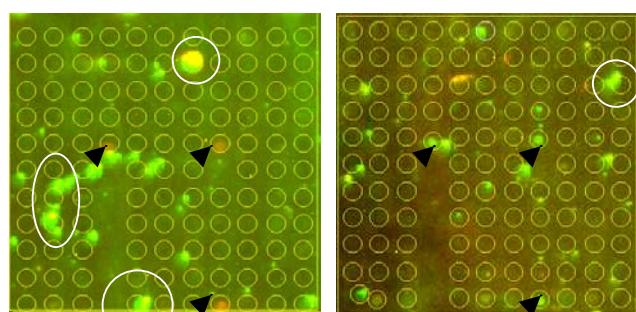
**Figura 3.** Eletroforese em gel de agarose para visualização de qualidade e quantidade dos produtos de PCR amplificados diretamente de colônia. As amostras aplicadas foram escolhidas de forma aleatória. Foram aplicados 5 $\mu$ l de cada amostra. Em cada primeiro *lane* foi aplicado uma marcador de peso molecular, confirmando o tamanho dos insertos de 500-1000pb.

Para a síntese das sondas marcadas por cada fluoróforo, foram utilizadas as amostras de RNA total obtidas das linhagens CCRF e Raji, respectivamente leucemias de células T e B. Estes dois microarranjos foram feitos invertidos para que pudéssemos confirmar a eficiência de marcação das sondas pelos fluoróforos e podem ser visualizados na **figura 4**.



**Figura 4.** Microarranjos produzidos no presente estudo, sem purificação dos fragmentos depositados. A marcação com os fluoróforos foi invertida, CCRF inicialmente com Cy3 e Raji com Cy5 e posteriormente ao inverso.

Apesar de um alto *background* (sinais inespecíficos), foi possível uma análise detalhada dos resultados desses microarranjos e dados positivos e importantes para a padronização dessa técnica foram obtidos como demonstrado na **figura 5**. Foi possível verificarmos a coerência das triplicatas e a eficiência da marcação pelos fluoróforos. Entretanto, é evidente que a existência do *background* interfere substancialmente na análise dos resultados.



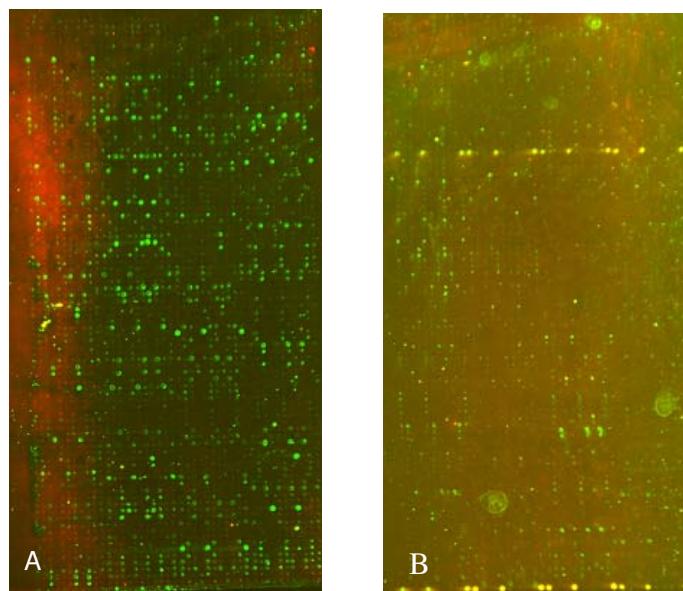
**Figura 5.** As setas demonstram a qualidade das triplicatas e a inversão dos resultados. Ao fundo, o evidente *background* com os círculos brancos exemplificando pontos inespecíficos.

#### **4.4. Confecção de microarranjos heterólogos**

Buscando obter uma maior qualidade dos microarranjos, conduzimos um teste para verificar a qualidade dos fragmentos da biblioteca depositados nas lâminas e também a qualidade do RNA total utilizado para a síntese das sondas. Para isso, utilizamos como controle experimental o material produzido a partir de coração de rato (*Rattus norvegicus*), que apresentou sinais bastante nítidos e consistentes em todos os microarranjos produzidos (Ana Carolina Deckmann, projeto de mestrado, suporte Capes).

O primeiro teste foi conduzido para verificar a qualidade das amostras de RNA total de leucemia, utilizando para isto a lâmina contendo ESTs de coração de rato. Foi realizada a hibridação heteróloga com 50ug de RNA total de CCRF marcado com Cy3. Um teste paralelo foi realizado para verificar a qualidade da biblioteca de ESTs de câncer, utilizando para isto uma das lâminas produzidas no presente estudo, e hibridada com Daudi marcado por Cy3 e RNA total de coração de rato marcado por Cy5. Esses testes podem ser visualizados na **figura 6**.

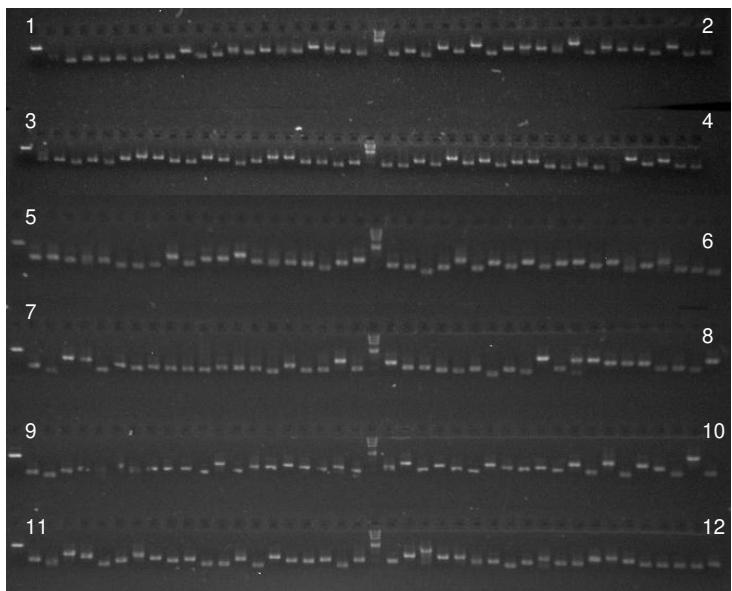
Com o resultado desses testes pudemos verificar a baixa qualidade do material impresso nos microarranjos produzidos no presente trabalho. Iniciou-se uma nova etapa de testes buscando melhorar a qualidade dos produtos de PCR impressos no microarranjo até que obtivéssemos um protocolo funcional para os padrões do presente estudo.



**Figura 6.** Em **A**: Lâmina de biblioteca de rato hibridada com linhagem leucêmica de humano (CCRF - Cy3). Em **B**: Lâmina de clones de câncer hibridada com linhagem leucêmica (Daudi – Cy3) e RNA de coração de rato (Cy5).

#### 4.5. Confecção de microarranjos com cDNA purificado

A fim de obtermos um resultado melhor, foram aplicadas diversas técnicas de purificação buscando melhorias na qualidade do material a ser amplificado. Os clones selecionados e rearranjados em placas de 96 poços foram submetidos à mini-preparações de DNA por lise alcalina e os plasmídeos advindos dessa técnica foram posteriormente tratados com proteinase K e então amplificados em PCR. A qualidade dos produtos de PCR pode ser visualizada na **figura 7**.

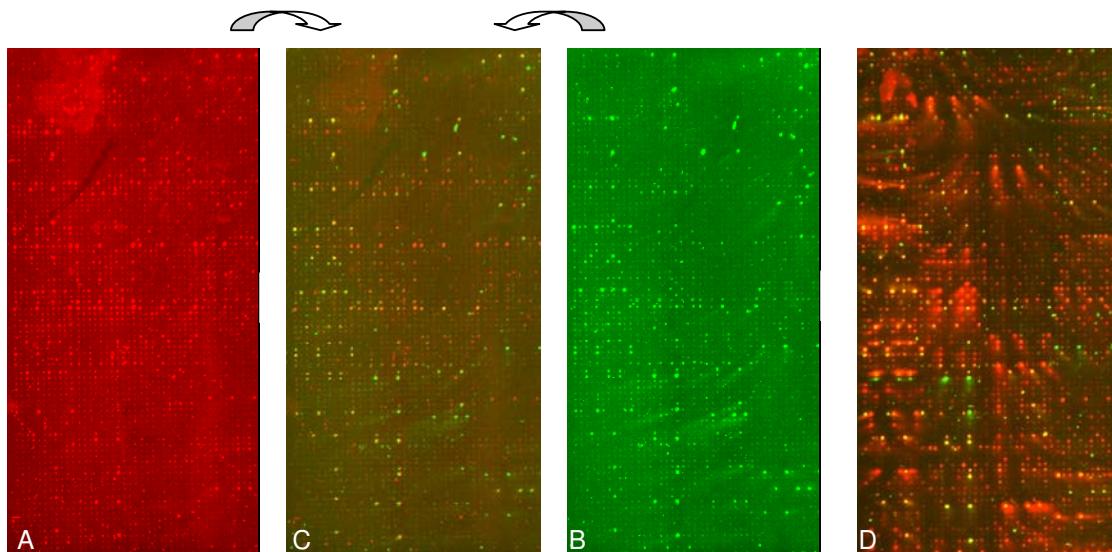


**Figura 7.** Eletroforese em gel de agarose para visualização de qualidade e quantidade dos produtos de PCR dos plasmídeos amplificados. A numeração de 1 a 12 refere-se ao número das placas de 96 poços. As 20 amostras de cada placa foram escolhidas de forma aleatória. Foram aplicados 5 $\mu$ l de cada amostra. Entre cada placa foi aplicado 6 $\mu$ l de marcador de peso molecular, confirmando o tamanho dos insertos de 500-1000pb.

A digitalização das amostras marcadas com cada fluoróforo é armazenada em canais de imagem distintos para que os softwares de processamento de imagem sejam capazes de quantificar a expressão gênica em função da intensidade do sinal emitido por cada fluoróforo. Cy3 emite fluorescência verde e Cy5 emite vermelha, a composição das imagens gera uma coloração amarela quando presente a emissão dos dois fluoróforos.

Devido ao mecanismo de digitalização, é possível a sobreposição de imagens como pode ser visualizado na **figura 8** onde o microarranjo hibridado com linhagem leucêmica de células T (CCRF) foi elaborado em lâminas separadas, ambas marcadas com Cy3. Os dados de expressão podem ser triangulados, analisados e comparados com outros resultados não sobrepostos. Outra marcação realizada pode também ser visualizada na **figura 8**, onde a linhagem leucêmica de células B – Daudi, foi marcada com o fluoróforo Cy5 e o controle

normal com Cy3. A quantificação dos sinais foi realizada através do software *Gene TAC Biochip Analyser (Perkin Elmer)*.



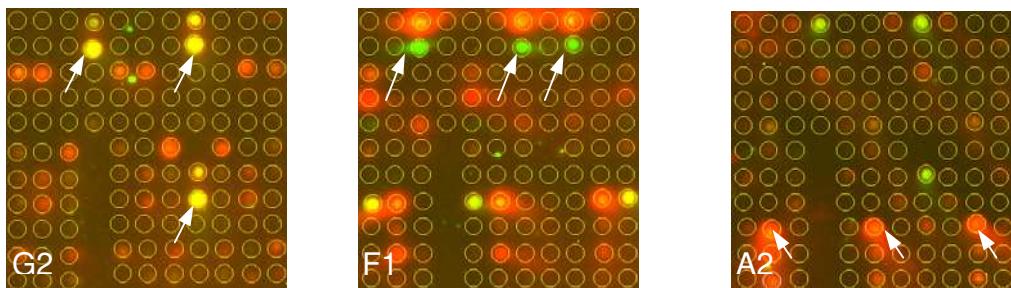
**Figura 8.** Microarranjos produzidos com purificação dos clones depositados. Em **A**: Marcação da linhagem leucêmica de células T (CCRF) com fluoróforo Cy3. Em **B**: a condição normal (controle) marcada com Cy5. Em **C**: a imagem sobreposta com os dois canais. Em **D**: Imagem composta de um microarranjo hibridado usando a condição normal (controle) marcado com Cy3 e a linhagem leucêmica de células B (Daudi) marcadas com Cy5. Os três primeiros demonstram a possibilidade de sobreposição das imagens captadas de cada um dos canais do fluoróforo.

#### 4.6. Análise das triplicatas

Fundamental ao início de qualquer análise advinda da técnica de microarranjos de DNA, é a reproduzibilidade do sinal das triplicatas, pois além de demonstrarem a coerência dos resultados a serem analisados, permitem-nos a confiabilidade da técnica empregada. E nesse sentido, esse trabalho vem demonstrar a possibilidade de empregar-se a técnica de microarranjos em nosso laboratório como usual em diversos tipos de organismos.

A trajetória desse presente trabalho contou positivamente com a total confiança aos resultados apresentados como pode ser visualizado na **figura 9** e

conferido na **tabela 2** que constam com exemplos de alguns dos clones de microarranjos produzidos e a sua reproduutibilidade.



**Figura 9.** Qualidade e reproduutibilidade dos sinais emitidos pelas réplicas no microarranjo marcados o controle (Normal) com o fluoróforo Cy3 e a linhagem de células B (Daudi) com Cy5. O mesmo padrão de qualidade foi obtido em todos os outros microarranjos descritos nesse trabalho. As setas indicam alguns exemplos de triplicatas encontradas nesses grids.

#### 4.7. Seleção e seqüenciamento dos clones para *northern blot*

A seleção dos clones para confirmação por *Northern blot* baseou-se inicialmente em 3 critérios: 1) Análise visual – triplicatas com forte intensidade do fluoróforo; 2) Análise estatística – triplicata com valores de razão ( $Cy5 * \text{Fator de normalização} / Cy3 > 2.0$  ou  $< 0.5$ ); 3) Análise da seqüência – clones confirmados através de BLASTn com banco de dados local, e com qualidade alta através de BLASTx com  $phrap > 20$  e  $e-value < 10^{-4}$ .

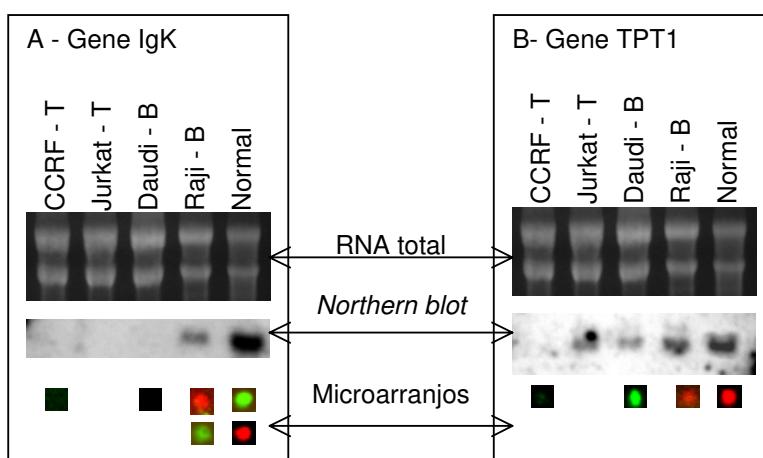
Foi refeito o seqüenciamento de 128 amostras (~11% das 1152 sequencias iniciais) e o resultado obtido pode ser visualizado no **Apêndice C**. Os dados foram apresentados a partir de seu endereçamento nos microarranjos, identidade por BLASTn e seus valores de qualidade e pela média das razões de intensidade dos fluoróforos. O BLASTn é uma ferramenta de comparação de seqüências com bancos públicos de dados de nucleotídeos identificando produtos de genes

codificados pela determinada seqüência e seus valores de qualidade são fornecidos na forma de “valor esperado” (*e-value*) entendendo-se que quanto menor o valor, maior a chance da similaridade da seqüência analisada não ser ao acaso. O cálculo dos valores obtidos nas triplicatas foi baseado na média das razões dos três pontos considerando-se 1 desvio-padrão de eliminação do ponto.

#### 4.8. Confirmação da expressão gênica por *northern blot*

Os microarranjos produzidos contaram apenas com um gene interno constitutivo positivo (actina), mas não possuíram genes balizadores para controle do processo. Sendo assim, a confirmação da expressão dos genes selecionados foi utilizada como balizadora da técnica de microarranjos.

Os resultados do perfil de expressão de alguns genes obtidos por *Northern blot* podem ser visualizados **figura 10**.



**Figura 10.** A figura mostra o RNA total aplicado em gel denaturante, o filme exposto com o determinado gene marcado radioativamente (*Northern blot*) e as respectivas imagens pontuais obtidas nos microarranjos. Gene em **A**: Cadeia leve kappa da Imunoglobulina região VLJ. Gene em **B**: proteína tumoral TPT1. Não foram realizados microarranjos utilizando RNA de cultura de célula de Jurkat.

## **5. DISCUSSÃO**

Os resultados obtidos pelo presente trabalho foram analisados a partir dos esforços de padronização da técnica de microarranjos e aplicabilidade no problema biológico da leucemia linfóide aguda encontrando genes diferencialmente expressos em culturas de células B e T.

### **5.1. Padronização da técnica de microarranjos de DNA**

A formação de um grupo de trabalho buscando a padronização e implementação da técnica de microarranjos de DNA foi essencial perante as diversas dificuldades encontradas. Esse grupo formado deveria padronizar uma mesma técnica, porém havia adotado bibliotecas confeccionadas de diferentes formas, adquiridas de diferentes organismos e para o estudo de diferentes problemas biológicos. Composto por esse estudo de leucemia linfóide aguda advindo de EST's de humanos, por um estudo de hipertrofia ventricular em rato com biblioteca clonada em fago (projeto de mestrado, suporte Capes), por um estudo do fungo causador da vassoura-de-bruxa com biblioteca genômica (projeto de doutoramento, suporte Capes) e posteriormente pelo estudo de eucalipto com biblioteca de cDNA (projeto Genolyptus, suporte FAPESP), esse grupo conseguiu alcançar protocolos funcionais a todos os projetos.

Diversos testes foram necessários até que os microarranjos fossem padronizados para cada um desses projetos, pois contaram com específicas alterações a cada um. O motivo dessa diferença de resultados ainda não tem uma explicação definitiva. Sabemos apenas que os resultados se apresentam de forma bem diferente dependendo de cada uma dessas variáveis (organismo, biblioteca,

quantidade e qualidade do material a ser depositado e do RNA a ser marcado pelos fluoróforos).

O primeiro microarranjo de alta qualidade que obtivemos foi o de coração de rato e serviu como base para a padronização dos seguintes, pois alguns genes constitutivos da biblioteca em fago passaram a ser usados em diversos testes realizados (não mostrados) como controle do processo até a padronização de cada um dos estudos. Além disso, o protocolo de preparação da biblioteca de coração de rato utiliza um tratamento com proteinase K, que digere preferencialmente complexos protéicos associados ao DNA, que pode ser adaptado a insertos clonados em plasmídeos purificando o material depositado. Esse procedimento foi eficaz no presente estudo, porém não se mostrou eficiente no material de vassoura-de-bruxa e eucalipto, que tiveram que adotar protocolos com lâminas de aminosilano e purificação por colunas específicas.

Os microarranjos produzidos no presente trabalho contaram com uma padronização da técnica com o menor custo possível uma vez que utilizou as próprias lâminas confeccionadas pelo grupo (poli-L-lisina) e não precisou de colunas de purificação para o material a ser depositado. Os resultados obtidos, mesmo nos microarranjos de menor qualidade, apresentam grande coerência, o que pode ser observado através da reprodutibilidade das triplicatas e das inversões de marcação. Além disso, obtivemos clara correlação na expressão obtida para o microarranjos e o *Northern blot* em alguns genes selecionados. Esses resultados sugerem que os dados obtidos podem ser ao menos preliminarmente analisados para identificar correlações entre genes específicos e os processos biológicos analisados.

## **5.2. Genes candidatos**

Esse presente estudo verificou com clareza a correlação da expressão obtida pelas metodologias utilizadas como também com a comparação desses resultados com os dados de literatura. Alguns genes analisados são discutidos a seguir:

**IgK VLJ** (“*Immunoglobulin kappa light chain VLJ region*”): os genes de imunoglobulina são formados através de recombinações entre cadeias leves e pesadas durante a diferenciação dos linfócitos. Esse jogo de recombinação é responsável pela origem dos diversos anticorpos existentes no homem. (Bentley & Rabitt, 1981). Emanuel e colaboradores em 1984 descreveram a existência das cadeias leves kappa em linhagens de linfomas de Burkitt e tumores de células B como sendo preferenciais. Porém, Halusca e colaboradores em 1987, evidenciaram que a recombinação em linhagem leucêmica Daudi envolvem cadeias pesadas de imunoglobulina. Esses achados nos confirmam os dados obtidos pelos microarranjos e Northern blot confeccionados durante esse estudo.

**TPT1** (“*Tumor Protein , translationally-controlled 1*”): localizado no cromossomo 13q12-q14, é um fator de liberação de histamina dependente de imunoglobulina E (MacDonald et al, 1995). Encontrado em líquidos biológicos de pacientes alérgicos, se tornou em grande candidato ao estudo de asma (MacDonald et al, 1999). Foi encontrado em vários tipos de células tumorais com exceção de tumores de fígado e carcinoma renal. No presente estudo se apresentou expresso em todos os tecidos testados inclusive no tecido normal, menos em linhagem leucêmica de células T (CCRF-CEM) e portanto deve ter sua expressão confirmada por novos testes.

**MAGE A4:** pertencente à família dos genes MAGE A, situados no cromossomo Xq28, são antígenos tumores-específicos que podem ser reconhecidos por linfócitos T, sendo de grande importância em imunoterapia (Rogner *et al*, 1995). De Plaen e colaboradores em 1994 identificaram 12 genes dessa família e encontraram através de RT-PCR, uma forte expressão do gene MAGE A4 em vários tumores de diversos tipos histológicos, mas silencioso em tecido normal, com exceção de testículo e placenta. No presente estudo pudemos verificar uma alta expressão deste gene em linhagem leucêmica de células B (Daudi) e nenhuma em tecido normal, porém esse dado deve ser confirmado por *Northern blot*.

**CASP4:** encontrado com diversos sinônimos (ICE, TX e CED3), esse gene pertence à família das caspases, genes responsáveis pela regulação positiva da apoptose. As caspases encontram-se naturalmente como proenzimas inativas e são ativadas por elas mesmas (Munday *et al.*, 1995). Quando induzido, esse gene induz a apoptose celular que serve como um balanço mitótico da regulação de tecido animais (Faucheu *et al.*, 1995). A expressão desse gene está descrita no banco de dados de proteínas Swiss-Prot (<http://ca.expasy.org/cgi-bin/niceprot.pl?P49662>) onde altos níveis de expressão são encontrados em baço e pulmão, moderado em coração e fígado, baixo em músculo esquelético, rim e testículos e nenhuma em cérebro. No presente estudo, verificamos níveis de expressão presente em todas as nossas amostras que serão confirmados em *Northern blot* e então relacionados ao processo biológico em estudo.

**ERBB3:** esse oncogene mapeado no cromossomo 12q13é um receptor de fator de crescimento epidermal pertencente à subfamília de receptores de tirosina

quinase e sua elevada expressão foi descrita em algumas linhagens de tumores de mama sugerindo sua importante participação em malignizações humanas (Kraus *et al*, 1989). Carraway e colaboradores em 1994 verificaram a existência desses transcritos em vários tecidos humanos. O presente estudo irá confirmar os dados obtidos de maior expressão em linhagens leucêmicas de células B do que em tecidos normais e linhagens de células T.

## **6. CONCLUSÕES**

Essencial para assegurar a qualidade e eficiência da técnica de microarranjos de DNA desenvolvida no Laboratório de Genômica e Expressão do Departamento de Genética e Evolução da Universidade de Campinas, o presente trabalho padronizou e consolidou alterações eficazes dos protocolos de preparo das bibliotecas de diferentes organismos e os resultados obtidos garantiram a alta qualidade obtida pela aplicação da técnica.

Por se tratar de uma biblioteca conhecida e de alta qualidade e de uma seleção realizada anteriormente, o estudo iniciou-se bem fundamentado e os resultados se mostraram pertinentes ao perfil esperado.

As leucemias linfóides agudas são um grupo de doenças tardivamente detectadas e que possuem etiologia pouco conhecido. Nesse sentido, o presente trabalho vem através de seus resultados agregar diferenças específicas entre dois tipos de leucemias linfóides agudas provenientes da diferenciação de células B e de células T.

A adequação dessa técnica como um estudo do perfil de expressão gênica global foi comprovada através da identificação de genes diferencialmente expressos que obtiveram sua expressão adequada ao esperado como verificado através do gene IgK VLJ (*"imunoglobulina kappa light chain VLJ region"*) e do gene TPT1 (*"Tumor Protein , translationally-controlled 1"*) e de outros que ainda serão confirmados por *Northern blot* como por exemplo os genes MAGE A4 (família dos genes MAGE), CASP4 (família das caspases) e ErbB3 (receptor de fator de crescimento epidermal).

## **7. PERSPECTIVAS**

Uma vez padronizada e consolidada a técnica de microarranjos de DNA, inicia-se um processo amplo da enorme aplicabilidade desse conhecimento adquirido e instalado em nosso laboratório.

O presente estudo é o primeiro passo da utilização dessa poderosa ferramenta de estudo global de processos complexos da instalação de doenças humanas, podendo estender-se a outros quadros clínicos.

O estudo do perfil de expressão global de genes relacionados às leucemias linfóides agudas é de grande interesse para a compreensão do processo de formação dos quadros de câncer, da diferenciação celular que ocorre na hematopoiese, das diferenças de progressão dos quadros clínicos, mas principalmente, torna-se uma grande ferramenta para se encontrar marcadores moleculares específicos a cada subtipo tumoral possibilitando diagnósticos precoces e consequentemente a aplicação de tratamentos específicos e adequados.

Projetos futuros deveriam quantificar a expressão de todos os genes diferencialmente expressos através de PCR quantitativo em tempo real, ampliar os estudos a outros tipos de leucemias, assim como encontrar alvos de marcadores moleculares e possíveis estudos de quadros clínicos.

## **8. REFERÊNCIAS BIBLIOGRÁFICAS**

- Altman A.J. & Schwartz, A.D. (1983). *Malignant diseases of infancy, childhood and adolescence*, 2ed, Saunders, 605pp.
- Armstrong, S.A., Stauton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub,T.R.& Korsmeyer, S.J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genetics* , 30: 41-47.
- Bain, B.J. (1999). The relationship between the myelodysplastic syndromes and the myeloproliferative disorders. *Leuk Lymphoma*, 34: 443-9.
- Belov, L., de la Vega, O., dos Remedios, C.G., Mulligan, S.P.& Christopherson, R.I. (2001). Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Research*, 61: 4483-4489.
- Bentley, D. L.; Rabbits, T. H. (1981). Human V(kappa) immunoglobulin gene number: implications for the origin of antibody diversity. *Cell* 24: 613-623.
- Blumenthal, H.T. (2001). Milestone or genomania? The relevance of the Human Genome Project to biological aging and the age-related diseases. *J. Gerontology*, 56A, 9:M529-M537.
- Brown, M.P.S.; Grundy, W. N.; Lin, D. Cristianini, N.; Sugnet, C.W.; Furey, T.S.; Ares Jr., M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS* 97:262-267.
- Camargo, A. A. *et al.* (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *PNAS*, 98: 12103-12108.
- Camargo, A.A., Souza, S.J., Brentani, R.R. & Simpson, A.J.G. (2001). Human gene Discovery trhough experimental definition of transcribed regions of the human genome. *Current Opinion in Chemical Biology*, 6: 13-16.

- Carraway K.L. 3rd, Sliwkowski M.X., Akita R., Platko J.V., Guy P.M., Nuijens A., Diamanti A.J., Vandlen R.L., Cantley L.C., Cerione R.A. (1994). The erbB3 gene product is a receptor for heregulin *J Biol Chem.*; 269(19):14303-6.
- Cavalcanti Jr., G.B., Savino, W., Maia, R.C., Dobbin, J.A., Carriço M.K., Harab, H.C. & Oliveira, M.S.P. (2001). Correlação entre a expressão celular do CD44 e formas tumorais das leucemias linfoblásticas. *Rev. Bras. Cancerologia*. 43: 1-11.
- Cecil, R. L. (1986). *Tratado de medicina interna*, 16ed, Guanabara.
- De Plaen, E.; Arden, K.; Traversari, C.; Gaforio, J. J.; Szikora, J.-P.; De Smet, C.; Brasseur, F.; van der Bruggen, P.; Lethe, B.; Lurquin, C.; Brasseur, R.; Chomez, P.; De Backer, O.; Cavenee, W.; Boon, T. (1994). Structure, chromosomal localization, and expression of 12 genes of the MAGE family. *Immunogenetics* 40: 360-369.
- Dias Neto, E., et al. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *PNAS*, 97: 3491-3496.
- Douet-Guilbert, N., Morel, F., Le Bris, M., Herry, A., Le Calvez, G., Marion, V., Berthou, C. & De Braekeleer, M. (2003). t(4;11)(q21;p15), including one complex translocation t(1;4;11)(p32;q21;p15), in adult T-cell acute lymphoblastic leukemia. *Leukemia Research*, 27: 965-967.
- Drexler, H.G., MacLeod, A.F. & Uphoff, C.C. (1999). Leukemia cell lines: in vitro models for the study of Philadelphia chromosome-positive leukemia. *Leukemia research*, 23: 207-15.
- Emanuel, B. S.; Selden, J. R.; Chaganti, R. S. K.; Jhanwar, S.; Nowell, P. C.; Croce, C. M. (1984). The 2p breakpoint of a 2;8 translocation in Burkitt lymphoma interrupts the V(kappa) locus. *Proc. Nat. Acad. Sci.* 81: 2444-2446.
- Ewing, B., Green, P., Gordon, D. (2002). Phred, Phrap and Consed –Autofinish.
- Faucheu C., Diu A., Chan A.W., Blanchet A.M., Miossec C., Herve F., Collard-Dutilleul V., Gu Y, Aldape R.A., Lippke J.A., Rocher, C., Su, M.S., Livingston, D.J., Hercend, T. & Lalane, J.L. (1995). A novel human protease similar to the

- interleukin-1 beta converting enzyme induces apoptosis in transfected cells. *EMBO J.*, 14: 1914-1922.
- Foon, K. A & Todd, E.D. (1986). Immunological classification of leukemias and lymphomas. *Blood*, 68: 1-31.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. & Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 286: 531-537.
- Harris, N.L., Jaffe, E.S., Diebold, J., Flandrin, G., Muller-Hermelink, H.K., Vardiman, J., Lister, T.A. & Bloomfield, C.D. (2000). World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Comitte meeting-Airlie House , Virginia, November 1997. *Histopathology*. 36: 69-86.
- Harris, N.L., Jaffe, E.S., Diebold, J., Flandrin, G., Muller-Hermelink, H.K., Vardiman, J., Lister, T.A. & Bloomfield, C.D. (1999). World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Comitte meeting-Airlie House , Virginia, November 1997. *J. Clin. Oncol.* 17: 3835-49.
- Hegde, P.; Qi R.; Abernathy, K.; Gay, C.; Dharap, S.; Gaspard, R.; Hughes, J.E.; Snesrud, E.; Lee, N. & Quackenbush, J. (2000). A concise guide to DNA microarray analysis. *Biotechniques*, 29(3):548-562.
- Hoffbrand, A. V. & Pettit, J. E. (1993) .Acute Leukaemias. *Essencial Haematology*, Blackwell Science, 3ed, pp 209.
- INCA – MINISTÉRIO DA SAÚDE. Estimativa da Incidência e Mortalidade por Câncer no Brasil. *Instituto Nacional do Câncer*. 2003
- Kerr, M.K. & Churchill, G.A. (2001). Statistical design and the analysis of gene expression microarray data. *Genet Res.*, 77(2):123-8.
- Kraus M.H., Issing W., Miki T., Popescu N.C., Aaronson S.A. (1989). Isolation and characterization of ERBB3, a third member of the ERBB/epidermal growth factor

- receptor family: evidence for overexpression in a subset of human mammary tumors. *Proc Natl Acad Sci U S A*. 1989 Dec;86(23):9193-7
- Lafage-Pochitaloff, M. & Charrin, C. (2003). Anomalies cytogénétiques dans les leucémies aiguës lymphoblastiques. *Pathologie Biologie*, 51: 329-336.
- Lai, R., Hirsch-Ginsberg, C.F. & Bueso-Ramos, C. (2000). Pathologic diagnosis of acute lymphocytic leukemia. *Hemat. Oncol. Clin. North Amer.* 14: 1209-1235.
- Lee, M.T.; Kuo, F.C.; Whitmore, G.A. & Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive DNA hybridizations. *PNAS*, 97 (18): 9834-9839.
- MacDonald, S. M.; Rafnar, T.; Langdon, J.; Lichtenstein, L. M. (1995). Molecular identification of an IgE-dependent histamine-releasing factor. *Science* 269: 688-690.
- MacDonald, S. M.; Paznekas, W. A.; Jabs, E. W. (1999). Chromosomal localization of tumor protein, translationally-controlled 1 (TPT1) encoding the human histamine releasing factor (HRF) to 13q12-q14. *Cytogenet. Cell Genet.* 84: 128-129.
- Munday N.A, Vaillancourt J.P., Ali A, Casano F.J., Miller D.K., Molineaux S.M., Yamin T.T., Yu V.L., Nicholson D.W. (1995). Molecular cloning and pro-apoptotic activity of ICErelIII and ICErelIII, members of the ICE/CED-3 family of cysteine proteases. *Biol Chem.*, 270(26):15870-6.
- NCBI, NIH. National Center for Biotechnology Information. *NCBI/NIH*. 2002.
- Plasschaert, SLA., Hamps, WA., Vellenga, E., Vries, EGE., & de Bont, ESJM. (2004). Prognosis in chilhood and acute lymfoblastic leukaemia: a question of maturation? *Cancer Treatment Reviews*, 30: 37-51.
- Rodrigues, K. M. & Camargo, B. (2003). Diagnóstico precoce do câncer infantil: Responsabilidade de todos. *Rev. Assoc. Med. Bras.*, 49 (1): 29-34.

- Rogner, U. C.; Wilke, K.; Steck, E.; Korn, B.; Poustka, A. (1995). The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28. *Genomics*, 29: 725-731.
- Schena, M., Shalon, D., Davies, R.W., Brown, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 6: 639-645.
- Schoch, C., Kohlmann, A., Schittger, S., Brors, B., Dugas, M., Mergenthaler, S., Kern, W., Hiddermann, W., Eils, R. & Haferlach, T. (2002). Acute myeloid leukemias with reciprocal specific gene expression profiles. *PNAS*. 99: 10008 – 10013.
- Staudt, L.M. (2002). Gene expression profiling of lymphoid malignancies. *Annu. Rev. Med.* 53: 303-318.
- Staudt, L.M. (2003). Molecular diagnosis of the hematologic cancers. *New England Journal of Medicine*, 348 (18): 1777-85.
- Strausberg, R.L., Camargo, AA., Riggins, G.L., Schaefer, C.F., Souza, S.J., Grouse, L.H., Lal, A., Buetow, K.H., Boon, K., Greenhut, S.F. & Simpson, A.J.G.(2002). An international database and integrated analysis tools for the study of cancer gene expression. *The Pharmacogenomics Journal*, 2: 156-164.

## **Apêndice A**

# The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags

Helena Brentani<sup>a</sup>, Otávia L. Caballero<sup>a</sup>, Anamaria A. Camargo<sup>a</sup>, Aline M. da Silva<sup>b</sup>, Wilson Araújo da Silva, Jr.<sup>c</sup>, Emmanuel Dias Neto<sup>d</sup>, Marco Grivet<sup>e</sup>, Arthur Gruber<sup>f</sup>, Pedro Edson Moreira Guimaraes<sup>d</sup>, Winston Hide<sup>g</sup>, Christian Iseli<sup>h</sup>, C. Victor Jongeneel<sup>h</sup>, Janet Kelso<sup>g</sup>, Maria Aparecida Nagai<sup>i</sup>, Elida Paula Benquique Ojopi<sup>d</sup>, Elisson C. Osorio<sup>a</sup>, Eduardo M. R. Reis<sup>b</sup>, Gregory J. Riggins<sup>j</sup>, Andrew John George Simpson<sup>a,k</sup>, Sandro de Souza<sup>a</sup>, Brian J. Stevenson<sup>h</sup>, Robert L. Strausberg<sup>l</sup>, Eloiza H. Tajara<sup>m</sup>, Sergio Verjovski-Almeida<sup>b</sup>, The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium\*, and The Human Cancer Genome Project Sequencing Consortium†

<sup>a</sup>Laboratorio de Genética Molecular do Cancer, Departamento de Radiologia, Universidade de São Paulo, Travessa da Rua Dr. Ovídeo Pires de Campos S/N, 4º andar, 05403-010, São Paulo, SP, Brazil; <sup>b</sup>Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; <sup>m</sup>Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 15054, São José do Rio Preto, SP, Brazil; <sup>a</sup>Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente 109 4º andar, 01509-010, São Paulo, SP, Brazil; <sup>f</sup>Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, Avenida Professor Orlando Marques de Paiva 87, 05508-000, São Paulo, SP, Brazil; <sup>c</sup>Fundação Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Rua Tenente Catão Roxo 2501, 14051-140, Ribeirão Preto, SP, Brazil; <sup>d</sup>Laboratory of Neurosciences (LIM-27), Instituto de Psiquiatria, Faculdade de Medicina, Universidade de São Paulo, Rua Dr. Ovídeo de Campos, S/N, 05403-010, São Paulo, SP, Brazil; <sup>l</sup>National Cancer Institute, Bethesda, MD 20892; <sup>j</sup>Duke University Medical Center, Durham, NC 27710; <sup>g</sup>South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa; <sup>h</sup>Office of Information Technology, Ludwig Institute for Cancer Research, CH-1066 Épalinges, Switzerland; and <sup>e</sup>Centro de Estudo de Telecomunicações-PUC, Rua Marquês de São Vicente, 225, 22453-900, Rio de Janeiro, RJ, Brazil

Contributed by Walter Bodmer, June 12, 2003

**Whereas genome sequencing defines the genetic potential of an organism, transcript sequencing defines the utilization of this potential and links the genome with most areas of biology. To exploit the information within the human genome in the fight against cancer, we have deposited some two million expressed sequence tags (ESTs) from human tumors and their corresponding normal tissues in the public databases. The data currently define ≈23,500 genes, of which only ≈1,250 are still represented only by ESTs. Examination of the EST coverage of known cancer-related (CR) genes reveals that <1% do not have corresponding ESTs, indicating that the representation of genes associated with commonly studied tumors is high. The careful recording of the origin of all ESTs we have produced has enabled detailed definition of where the genes they represent are expressed in the human body. More than 100,000 ESTs are available for seven tissues, indicating a surprising variability of gene usage that has led to the discovery of a significant number of genes with restricted expression, and that may thus be therapeutically useful. The ESTs also reveal novel nonsynonymous germline variants (although the one-pass nature of the data necessitates careful validation) and many alternatively spliced transcripts. Although widely exploited by the scientific community, vindicating our totally open source policy, the EST data generated still provide extensive information that remains to be systematically explored, and that may further facilitate progress toward both the understanding and treatment of human cancers.**

**H**uman cancer results from the accrual of genetic mutations or epigenetic changes in the genomes of individual somatic cells. These exert their effect via alterations in the structure and abundance of individual mRNA molecules that, in turn, alter crucial protein-mediated cellular functions. One step in the path toward building a comprehensive molecular portrait of human cancer is the definition of the genes actively expressed in specific tumors and corresponding normal tissues. It is within these sets of genes that we must search for cancer-defining mutations and epigenetic changes and delineate the extent of the molecular milieu within which we will deepen our understanding of cancer. Very short sequence tags of transcripts and hybridization techniques such as microarrays identify many of the genes expressed in tumors and are widely used for measuring the relative levels of gene expression (1, 2). However,

longer transcript sequences such as ESTs permit genes expressed in individual cells and tissues to be identified in a completely unambiguous manner, provide additional data on transcript and gene variants, and represent a key source for the search for as yet incompletely characterized genes.

In two large projects, extensive EST sequencing of human tumor tissues has been undertaken: the Cancer Genome Anatomy Project (CGAP) (3) and the Fundação de Amparo à Pesquisa do Estado de São Paulo/Ludwig Institute for Cancer

Abbreviations: EST, expressed sequence tag; CGAP, Cancer Genome Anatomy Project; HCGP, Human Cancer Genome Project; SNP, single-nucleotide polymorphism; CR, cancer-related.

\*The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium: Marcio Luis Ascencio<sup>o</sup>, Mário Henrique Bengtson<sup>o</sup>, Fabiana Bettoni<sup>p</sup>, Walter F. Bodmer<sup>q</sup>, Marcelo R. S. Briones<sup>q</sup>, Luiz Paulo Camargo<sup>o</sup>, Webster Cavenee<sup>l</sup>, Janete M. Cerutti<sup>u</sup>, Luís Eduardo Coelho Andrade<sup>v</sup>, Paulo César Costa dos Santos<sup>o</sup>, Maria Cristina Ramos Costa<sup>o</sup>, Israel Tojal da Silva<sup>w</sup>, Marcos Roberto H. Estécio<sup>o</sup>, Karine Sa Ferreira<sup>w</sup>, Frank B. Furnari<sup>t</sup>, Milton Faria, Jr.<sup>s</sup>, Pedro A. F. Galante<sup>p</sup>, Gustavo S. Guimaraes<sup>o</sup>, Adriano Jesus Holanda<sup>w</sup>, Edna Teruko Kimura<sup>z</sup>, Maarten R. Leerkes<sup>p</sup>, Xin Lu<sup>a</sup>, Rui M. B. Maciel<sup>u</sup>, Elizabeth A. L. Martins<sup>bb</sup>, Katlin Brauer Massirer<sup>o</sup>, Analy S. A. Melo<sup>o</sup>, Carlos Alberto Mestriner<sup>cc</sup>, Elisabete Cristina Miracca<sup>o</sup>, Leandro Lorenci Miranda<sup>o</sup>, Francisco G. Nobrega<sup>dd</sup>, Paulo S. Oliveira<sup>p</sup>, Apuá C. M. Paquola<sup>ee</sup>, José Rodrigo C. Pandolfi<sup>cc</sup>, Maria Inês de Moura Campos Pardin<sup>ff</sup>, Fabio Passetti<sup>p</sup>, John Quackenbush<sup>gg</sup>, Beatriz Schnabel<sup>l</sup>, Mari Cleide Sogayar<sup>o</sup>, Jorge E. Souza<sup>p</sup>, Sandro R. Valentini<sup>cc</sup>, and Andre C. Zaiats<sup>p</sup>.

†The Human Cancer Genome Project Sequencing Consortium: Elisabete Jorge Amaral<sup>x</sup>, Liliane A. T. Arnaldi<sup>u</sup>, Amélia Goés de Araújo<sup>w</sup>, Simone Aparecida de Bessa<sup>n</sup>, David C. Bicknell<sup>p</sup>, Maria Eugenia Ribeiro de Camarao<sup>o</sup>, Dirce Maria Carraro<sup>p</sup>, Helaine Carrer<sup>hh</sup>, Alex F. Carvalho<sup>p</sup>, Christian Colin<sup>o</sup>, Fernando Costa<sup>ii</sup>, Cyntia Curcio<sup>o</sup>, Ismael Dale Cotrim Guerreiro da Silva<sup>y</sup>, Neusa Pereira da Silva<sup>z</sup>, Márcia Dellamano<sup>o</sup>, Hamza El-Dorry<sup>kk</sup>, Enilza Maria Espreafico<sup>o</sup>, Ari José Scattone Ferreira<sup>kk</sup>, Cristiane Ayres Ferreira<sup>w</sup>, Maria Angela H. Z. Fortes<sup>mm</sup>, Angelita Habr Gama<sup>nn</sup>, Daniel Giannella-Neto<sup>mm</sup>, Maria Lúcia C. G. Giannella<sup>mm</sup>, Ricardo R. Giorgi<sup>mm</sup>, Gustavo Henrique Goldman<sup>oo</sup>, Maria Helena S. Goldman<sup>pp</sup>, Christine Hackel<sup>l</sup>, Paulo Lee Ho<sup>bb</sup>, Elza Myiuki Kimura<sup>qq</sup>, Luiz Paulo Kowalski<sup>rr</sup>, Jose E. Krieger<sup>ss</sup>, Luciana C. C. Leite<sup>bb</sup>, Ademar Lopes<sup>rr</sup>, Ana Mercedes S. C. Luna<sup>mm</sup>, Alan Mackay<sup>tt</sup>, Suely Kazue Nagashashi Mari<sup>rr</sup>, Adriana Aparecida Marques<sup>w</sup>, Waleska K. Martins<sup>p</sup>, André Montagnini<sup>rr</sup>, Mario Mourão Neto<sup>rr</sup>, Ana Lucia T. O. Nascimento<sup>bb</sup>, A. Munro Neville<sup>uu</sup>, Marina P. Nobrega<sup>dd</sup>, Mike J. O'Hare<sup>tt</sup>, Audrey Yumi Otsuka<sup>w</sup>, Anna Izabel Ruas de Melo<sup>p</sup>, Maria Luisa Paço-Larson<sup>ww</sup>, Gonçalo Guimarães Pereira<sup>ii</sup>, Neusa Pereira da Silva<sup>y</sup>, João Bosco Pesquero<sup>oi</sup>, Juliana Gilbert Pessoal<sup>jj</sup>, Paula Rahal<sup>x</sup>, Claudia Aparecida Rainho<sup>xx</sup>, Vanderlei Rodrigues<sup>yy</sup>, Silvia Regina Rogatto<sup>xx</sup>, Camila Malta Romanó<sup>zz</sup>, Janaína Gusmão Romeiro<sup>x</sup>, Benedito Mauro Rossi<sup>rr</sup>, Monica Rusticci<sup>ll</sup>, Renata Guerra de Sá<sup>vv</sup>, Simone Cristina Sant' Anna<sup>qq</sup>, Míriam L. Sarmazo<sup>x</sup>, Teresa Cristina de Lima e Silva<sup>y</sup>, Fernando Augusto Soares<sup>rr</sup>, Maria de Fátima Sonati<sup>qq</sup>, Josane de Freitas Sousa<sup>ll</sup>, Diana Queiroz<sup>y</sup>, Valéria Valente<sup>ww</sup>, André Luiz Vettore<sup>p</sup>, Fabiola Elizabeth Villanova<sup>vv</sup>, Marco Antonio Zago<sup>w</sup>, and Heloisa Zalberg<sup>p</sup>.

Research–Human Cancer Genome Project (HCGP) (4, 5). CGAP, launched in 1997 by the National Cancer Institute, has used single-pass sequencing from the 5' and/or 3' extremities of cDNA clones for sequence generation (6). The HCGP project adopted an alternative EST-based strategy, termed ORESTES, which generates sequences biased toward the central coding regions of transcripts (7). The data gathered by these two projects are thus complementary and have been combined into an International Database of Cancer Gene Expression (5), available at <http://cgap.nci.nih.gov>. They also constitute the basis of the Human Cancer Index at TIGR ([www.tigr.org/tdb/tgi/hcgi](http://www.tigr.org/tdb/tgi/hcgi)).

These sequencing initiatives are unique in providing a very large disease-oriented transcriptional database that contains an unprecedented amount of information on expressed human genes. To maximize the benefit of these data, we have made them

all publicly available immediately on generation. We here provide a description of these data, as well as their utility for the identification of human genes, the tissue specificity of their expression, and the structure of some of their transcript variants. We find that, although the data have been widely used by the scientific community, much additional information remains untapped.

## Materials and Methods

**Transcript Sequencing.** ORESTES sequences were generated as previously described (7). Tumors and corresponding normal tissues were mostly obtained from the Hospital de Câncer A.C. Camargo, São Paulo, with the exception of purified breast tissue samples that were obtained from University College London (7). In addition, extensive use was made of both breast and colon cell lines. Identification of the source of mRNA for all libraries is available at the CGAP homepage. CGAP sequences were generated from the 3' and 5' extremities of both standard and normalized cDNA libraries from a variety of sources as described elsewhere (5).

**Transcript to Genome Mapping.** A comprehensive reconstruction of human transcripts based on genome data were undertaken based on two datasets: (i) a set of alignments between transcripts and genome regions, thoroughly filtered to eliminate the effects of pseudogenes, highly conserved gene families, repetitive elements and EST sequencing errors; (ii) a mapping onto the genome of all polyadenylation sites that could be extracted from the chromatograms of the EST sequencing projects, thus marking sites where polyadenylation has been experimentally documented (8). To be included as a spliced cluster, we stipulated that canonical splice sites must be present on at least one of the transcript sequences. ORF identification required that for the longest ORF predicted by ESTScan ([www.ch.embnet.org/software/ESTScan.html](http://www.ch.embnet.org/software/ESTScan.html)) for each transcript cluster, the ESTScan score divided by the ORF length had to be >1 (9), an empirical measure that covers >99% of human SwissProt entries. In addition, the ESTScan-predicted ORFs had to have at least three nucleotides 5' and 3', as a measure of having at least some 5' and 3' UTR.

**Generation of a Representative Set of CR Genes.** This manually curated compilation was drawn up during a week-long meeting of the Annotation Consortium in August of 2001.<sup>‡</sup> It is a nonredundant list comprising 1,127 human cancer-associated genes based on initial querying of GenBank ([www.ncbi.nlm.nih.gov/GenBank/index.html](http://www.ncbi.nlm.nih.gov/GenBank/index.html)), GenCard (<http://bioinfo.weizmann.ac.il/cards/index.html>), and Harvard University ([http://sbweb.med.harvard.edu/research/breast\\_cancer/currentlistofgenes.htm](http://sbweb.med.harvard.edu/research/breast_cancer/currentlistofgenes.htm)) with the words “cancer” and “tumor.” The list is available at <http://bit.fmrp.usp.br/jamborestes>.

**Detection of Single-Nucleotide Polymorphisms (SNPs).** The transcript sequences were aligned against one another by using FASTA and base quality values files generated by using PHRED (10, 11). A SNP was considered for further analysis only if indicated by reads from at least two different ORESTES libraries, at least two different CGAP reads, or one read from each source. All selected SNPs were compared with SNPs already deposited in dbSNP ([www.ncbi.nlm.nih.gov/SNP](http://www.ncbi.nlm.nih.gov/SNP)) by BLASTN. Only newly identified human SNPs were included in experimental validation, by PCR with DNA from a panel of 150 Brazilian individuals from three different ethnic backgrounds: 50 whites (mostly of Western and Southern European ancestry), 50 blacks (mulattoes

<sup>†</sup>Laboratorio de Genética Molecular do Cancer, and <sup>‡</sup>Department of Gastroenterology, Faculdade de Medicina, Universidade de São Paulo, Travessa da Rua Dr. Ovídeo Pires de Campos S/N, 4º andar, 05403-010, São Paulo, SP, Brazil; <sup>ee</sup>Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; <sup>bb</sup>Insti-  
to Butantan, Avenida Vital Brazil 1500, 05503-900, São Paulo, Brazil; <sup>vw</sup>Departamento de Biologia Celular e Molecular e de Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil; <sup>u</sup>Laboratory of Molecular Endocrinology, Department of Medicine, Federal University of São Paulo, Rua Pedro de Toledo 781, 12th Floor, 04023-039, São Paulo, SP, Brazil; <sup>o</sup>Chemistry Institute, University of São Paulo, 05513-970, São Paulo, SP, Brazil; <sup>yy</sup>Departamento de Bioquímica e Imunologia, and <sup>ll</sup>Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil; <sup>jj</sup>Department of Biophysics, and <sup>d</sup>Departamento de Microbiologia, Imunologia, e Parasitologia, Universidade Federal de São Paulo, Rua Botucatu, 862 3º andar, 04023-062, São Paulo, SP, Brazil; <sup>ff</sup>Laboratório de Biologia Molecular, Hemocentro, Faculdade de Medicina Universidade Estadual Paulista, 18618-000, Botucatu, SP, Brazil; <sup>de</sup>Departamento de Bioinformática–UNAERP, Universidade de Ribeirão Preto, 14096-380, Ribeirão Preto, SP, Brazil; <sup>z</sup>Departamento de Histologia e Embriologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Avenida Prof Lineu Prestes 1524, 05508-900, São Paulo, Brazil; <sup>ss</sup>Department of Medicine, Laboratory of Genetics and Molecular Cardiology, Heart Institute (InCor), Universidade de São Paulo, 05403-000, São Paulo, SP, Brazil; <sup>x</sup>Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 15054, São José do Rio Preto, SP, Brazil; <sup>qq</sup>Department of Clinical Pathology, School of Medical Sciences, State University of Campinas–UNICAMP, 13083-970 Campinas, SP, Brazil; <sup>pl</sup>Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente 109, 4º andar, 01509-010, São Paulo, SP, Brazil; <sup>rr</sup>Fundação Antonio Prudente, Hospital do Câncer, Rua Professor Antonio Prudente 211, 01509-900, São Paulo, SP, Brazil; <sup>zz</sup>Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, Avenida Professor Orlando Marques de Paiva 87, 05508-000, São Paulo, SP, Brazil; <sup>w</sup>Fundaçao Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Rua Tenente Catão Roxo 2501, 14051-140, Ribeirão Preto, SP, Brazil; <sup>dd</sup>Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraíba, 12244-000, São José dos Campos, SP, Brazil; <sup>hh</sup>Department of Biological Sciences, Escola Superior de Agricultura “Luiz de Queiroz,” Universidade de São Paulo, 13418-900, Piracicaba, SP, Brazil; <sup>v</sup>Departamento de Reumatologia, Universidade Federal de São Paulo, Rua Botucatu 740, 04023-062, São Paulo, SP, Brazil; <sup>kk</sup>Department of Biochemistry, Institute of Chemistry, University of São Paulo, Avenida Professor Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil; <sup>cc</sup>Departamento de Ciências Biológicas, Faculdade de Ciências Farmacêuticas de Araraquara, Universidade Estadual Paulista, 14801-902, São Paulo, Brazil; <sup>xx</sup>Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, 18618-970, Botucatu, SP, Brazil; <sup>vv</sup>Gynecology Department, Federal University of São Paulo, Rua Botucatu 740, 04023-062, São Paulo, SP, Brazil; <sup>oo</sup>Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Avenida da Caffe S/N, 14040-903, Ribeirão Preto, SP, Brazil; <sup>pp</sup>Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, Avenida Bandeirantes, 3900, 14040-901, Ribeirão Preto, SP, Brazil; <sup>mm</sup>Laboratory for Cellular and Molecular Endocrinology (LIM-25/HCFMUSP), School of Medicine, Universidade de São Paulo, Avenida Dr. Arnaldo, 455 no. 4305, 01246-903, São Paulo, SP, Brazil; <sup>yd</sup>Departamento de Genética e Evolução, State University of Campinas–UNICAMP, 13083-970, Campinas, SP, Brazil; <sup>ii</sup>National Cancer Institute, Bethesda, MD 20892; <sup>cc</sup>Cancer Research U.K. Cancer and Immunogenetics Laboratory, Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom; <sup>ld</sup>Ludwig Institute for Cancer Research, Department of Medicine, Center for Molecular Genetics, University of California at San Diego, La Jolla, CA 92093-0660; <sup>aa</sup>Ludwig Institute for Cancer Research, Imperial College School of Medicine, St. Mary’s Campus, Norfolk Place, London W2 1PG, United Kingdom; <sup>gg</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; <sup>tl</sup>Ludwig Institute for Cancer Research, Breast Cancer Laboratory, Department of Surgery, Royal Free and University College Medical School, 67-73 Riding House Street, London W1W 7EJ, United Kingdom; and <sup>uu</sup>Ludwig Institute for Cancer Research, Horatio House, 77-85 Fulham Palace Road, London W6 8JC, United Kingdom.

<sup>†</sup>To whom correspondence should be addressed. E-mail: asimpson@licr.org.

© 2003 by The National Academy of Sciences of the USA

<sup>‡</sup>The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium, Jamborestes, Aug. 20–25, 2001, São Paulo, Brazil.

**Table 1.** HCGP and CGAP transcript sequence generation and clustering

| Form of gene representation                            | Number of sequences, clusters, or genes |
|--|---|
| ORESTES submitted to GenBank                           | 823,121 sequences                       |
| CGAP EST submitted to GenBank                          | 1,214,358 sequences                     |
| TOTAL EST submitted to GenBank                         | 2,037,479 sequences                     |
| Total clusters   | 32,129 clusters                         |
| Total clusters with known genes                        | 22,152 clusters                         |
| Clusters without known genes                           | 9,977 clusters                          |
| Clusters without known genes but with coding potential | 1,285 clusters                          |
| Estimated total genes based on HCGP and CGAP data      | 23,437 genes                            |

excluded), and 50 Japanese. All groups reported no racial admixture in their four grandparents.

**Alternative Splicing.** Two approaches were used. In the first, exon skipping was detected by using J.EXPLORER (available for download from [www.sanbi.ac.za/exon-skipping](http://www.sanbi.ac.za/exon-skipping)), which reduces the complexity of the gene sequences to a set of possible splice junctions used to search for ESTs spanning the annotated exon-exon junctions (12). The second approach listed all variant exons as revealed by transcript to genome alignments by using only those cDNAs that span at least two exons. Variants are represented by a binary matrix where each row corresponds to a sequence, and each column corresponds to an exon (33).

## Results and Discussion

**Gene Identification.** Collectively, the Fundação de Amparo à Pesquisa do Estado de São Paulo/Ludwig Institute for Cancer Research–Human Cancer Genome Project and the Cancer Genome Anatomy Project have deposited over two million sequences from tumors and normal tissues in GenBank (Table 1). To give perspective, this number of sequences is greater than that required to complete the high-quality shotgun sequencing of ≈40 bacterial genomes. The two projects are the largest individual contributors to the public human EST database and are responsible for >40% of all publicly available human EST data available (dbEST release 122002, [www.ncbi.nlm.nih.gov/dbEST](http://www.ncbi.nlm.nih.gov/dbEST)).

The first, and most obvious, use of ESTs is to define genes within the human genome. The precise experimental definition of these structures is crucial not only for CR research but also for research undertaken in all aspects of human biology. To achieve this, we have organized our EST data by aligning them, together with all other publicly available transcript sequences, to the human genome assembly. This allows an accurate organization of sequences into clusters corresponding to individual genes even in cases of minimal overlap or high similarity between paralogs. The transcript to genome alignment also provides an arbiter of quality, in that ESTs and EST clusters that span splice sites (referred to as spliced ESTs and spliced clusters, respectively) but correspond to still incompletely defined genes can be unambiguously distinguished from EST sequences derived from contaminating DNA or immature mRNA molecules.

We identified transcript clusters that contain at least one EST derived from one of the two cancer transcript sequencing projects, and which we judge to have a high probability of being derived from authentic expressed genes due to the presence of a reportedly full-length cDNA sequence, evidence of splicing or the presence of a predicted ORF. A total of 22,152 of the clusters contain full-length cDNAs. We use the presence of such full-length cDNA-containing transcript clusters as the gold standard

**Table 2.** Novel genes and definition of genes with restricted expression patterns utilizing EST clusters containing sequences from the HCGP and CGAP databases

| Gene type                             | No. of genes |
|---------------------------------------|--------------|
| Positionally cloned genes             | 5            |
| Genes cloned on the basis of homology | 16           |
| Tissue restricted expression          |              |
| Brain                                 | 9            |
| Breast                                | 14           |
| Colon                                 | 11           |
| Prostate                              | 15           |
| Cancer/testis                         | 4            |
| Various                               | 13           |
| Total                                 | 87           |

Full details are provided in Tables 6 and 7.

for identification of human genes and refer to the genes they define as known genes. The known genes mapped by our ESTs comprise a subset of the total of the 29,332 known genes currently contained within the human genome, indicating that ≈75.5% of known human genes are expressed in the tumors and normal tissues we have studied. The 25% of genes that were not represented by our ESTs are defined by ESTs from other projects or full length cDNAs from various sources. We assume that these genes are either not expressed or are expressed at low levels in the cells and tissues that we have studied. Some 4,000 of the known genes to which our ESTs map have been defined since the publication of the draft human genome in 2001 (13, 14). In the original published draft, these genes were represented only by the EST clusters. Interestingly, this number is 75% of all novel genes defined in this period, demonstrating that our ESTs equally well represent previously known genes and those still being defined. The preexistence of the EST clusters led directly to the generation of full-length cDNAs of a number of CR genes. The ESTs served either to indicate the existence of genes within defined regions of the human genome or as evidence of previously unknown members of paralogous families (Table 2 and cited in detail in Table 6, which is published as supporting information on the PNAS web site).

To investigate the extent to which our ESTs represent known CR genes, we compiled a list of 1,127 human genes known or presumed to play a role in the process of transformation to malignancy, which we refer to as CR (cancer-related) genes (see <http://bit.fmrp.usp.br>). The CR set contains extensively studied CR genes such as *TP53*, *RBI*, *BRCA1*, *CDKN2*, and *ERBB2*, as well as members of paralogous gene families that function in critical signal-transducing pathways, such as cadherins, integrins, and mitogen-activated protein kinases. Although incomplete, we believe the CR set is representative of well characterized genes relevant to the development of human cancer. Most importantly, many of these have been cloned on the basis of strategies, such as positional cloning, that do not depend on transcript abundance. Of the genes in the list, we found that 1,009 (89%) have at least one corresponding ORESTES sequence; 1,099 genes (97%) have at least one CGAP sequence; and 1,102 genes (97%) have EST sequences derived from at least one of the two projects. Of the 25 genes for which we have not generated ESTs, 18 have no EST coverage at all, indicating that their overall expression is at very low levels in the human body.

The 9,977 clusters composed only of ESTs represent an unknown number of additional genes for which a full-length cDNA is not yet available (Table 1). A particular difficulty in relating these clusters to genes lies in the very high level of heterogeneity at their 3' ends (8). Thus, noncoding, spliced clusters may be extensions of known genes; those that are not

might represent multiple different forms of as-yet-undefined genes. To avoid the complications of the 3' heterogeneity of transcripts, and also to exclude clusters that represent noncoding transcripts, one approach is to use only clusters that contain an identifiable ORF (9). Although we identified 9,977 EST clusters with splicing, only 1,285 are apparently coding sequences. We take these as each representing an individual human gene in the present discussion, although this one-to-one correlation remains to be demonstrated by full-length cDNA sequencing. Nevertheless, on the basis of these criteria, the two million or so EST sequences that we have generated translate into 23,437 genes, of which 1,285 are currently not defined by full-length cDNAs. Taking our data to include 75.5% of all genes, on the basis of the coverage of known genes described above, this value would predict a total of 31,042 genes in the human genome, in line with other recent estimates (13–16). It should be noted, however, that this value is likely to steadily increase as more small, nonspliced, and nonprotein-coding genes are discovered (16–18). These analyses indicate, however, that about the same number of genes confirmed by full-length cDNA sequencing since the publication of the draft genome sequence could be added to the genome immediately on the basis of the EST clusters with predicted ORFs to which we have contributed. These data are likely to contain a rich variety of novel genes relevant to cancer. For example, we found in September 2001 that among clusters of this kind, there were 19 that apparently encode novel paralogs of genes in the CR list. Of these, five have already been deposited by others in the GenBank database: CAMK1 (accession no. BC032726), MBD2 (accession no. AY038022), TLN2 (accession no. NM\_015059), BCHE (accession no. BC028713), and CNK (accession no. AK054808). The remainder will be described in detail elsewhere as the determination of the full-length sequences is completed.

We have previously shown that the number of ESTs corresponding to a gene is an indication of its level of expression (19). Thus, genes more frequently expressed in tumors than in normal tissues, for example, can be identified from the EST data. Several examples of genes with differential gene expression identified from our ESTs are included in Table 2 and are listed in detail in Table 7, which is published as supporting information on the PNAS web site. It should be pointed out, however, that in general we have rather less data from normal than tumor-derived tissues, which limits the power of this approach at present. Moreover, both normal and tumor samples can be highly heterogeneous, with normal tissues in particular being mixtures of different cell types. Thus, all conclusions drawn from analysis of our sequencing data require confirmation both in subsequent RT-PCR experiments and ultimately by histochemical analysis at the protein level. Nevertheless, EST cluster size appears to be a useful general indicator of at least overall gene expression and comparison of the number of CGAP and ORESTES sequences with serial analysis of gene expression (SAGE) tags for the same genes are positively correlated ( $r = 0.6$ ), despite the very different sets of tissues that have been accessed by the two databases. The average cluster size for all of the known human genes for which we have generated ESTs to date is 606, whereas the spliced EST clusters with predicted ORFs contain an average of only 19 ESTs. Thus, the latter represent genes much less frequently expressed than the average in the tissues we have studied. The average cluster size of the novel known genes added to the databases since the publication of the draft genome is 55, indicating that with time genes with lower levels of expression are being defined.

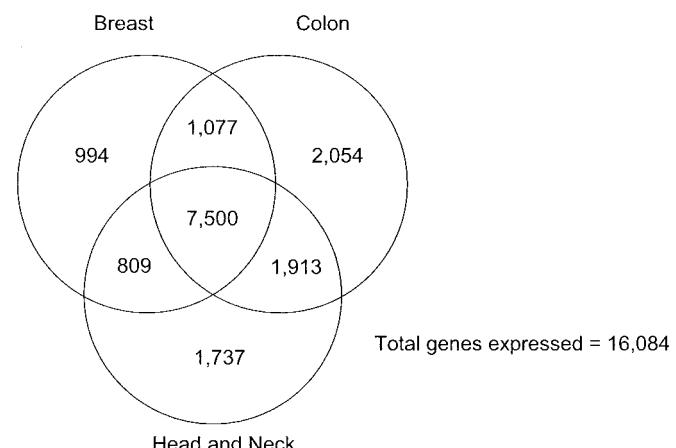
**Tissue Specificity of Gene Expression.** The genomes of all cells in the human body are copied from that in the fertilized egg, although they are continually modified over the lifetime of the individual due to the relentless acquisition of somatic mutations. Although

**Table 3. Documentation of gene expression in individual human tissues and estimation of tissue transcriptome complexity**

| Tissue    | ESTs    | Total clusters with known genes | Clusters with splicing and coding potential | Estimated total tissue transcriptome |
|-----------|---------|---------------------------------|---|--------------------------------------|
| Brain     | 185,193 | 12,746                          | 282   | 13,028                               |
| Breast    | 137,867 | 10,153                          | 227   | 10,380                               |
| Colon     | 186,870 | 12,218                          | 326   | 12,544                               |
| Head/neck | 186,298 | 11,698                          | 261   | 11,959                               |
| Kidney    | 115,096 | 12,368                          | 341   | 12,709                               |
| Lung      | 166,469 | 13,076                          | 314   | 13,390                               |
| Ovary     | 53,070  | 7,472                           | 137   | 7,609                                |
| Prostate  | 85,910  | 10,004                          | 224   | 10,228                               |
| Uterus    | 125,079 | 11,614                          | 244   | 11,858                               |
| Others    | 795,627 | 19,818                          | 1,079                                       | 20,897                               |

it is these somatic alterations that are ultimately responsible for human cancer, their relative importance will depend on the transcriptional programming of the cell in which they occur. Thus, the careful documentation of the tissue specificity of gene expression is vital to understanding the genetic basis of cancer. For this purpose, the availability of details of the tissue origin of our ESTs represents a powerful resource. All information on the tissue origins of our ESTs is available on the CGAP web site (<http://cgap.nci.nih.gov>). For each of seven human tissues (brain, head and neck, colon, lung, breast, uterus, and kidney), we have generated >100,000 EST sequences providing, in each case, a deep survey of gene expression. This value is comparable to the number of serial analysis of gene expression tags typically generated in individual experiments for the detection of differential gene expression and far in excess of any other EST compilation for individual tissues (20).

On the basis of ESTs that correspond to known genes and spliced EST clusters with predicted ORFs, we have evidence for the expression of between 10,000 and 13,500 genes for the seven tissue types that we have explored in depth, with lung having the highest number of expressed genes so far (Table 3). This indicates that not >57% of all of the genes defined by our ESTs are expressed in any one tissue type. It would thus appear that gene utility is quite variable between tissues. For example, if we take the 16,084 genes collectively expressed in breast, colon, and head and neck, there is evidence for expression for only 7,500 (47%) in all three tissues, whereas 4,785 (30%) are expressed in only one of the three tissues (Fig. 1). These numbers are arrived at despite the undoubtedly impurity of the tissue



**Fig. 1.** Coexpression of genes among breast, colon, and head/neck.

**Table 4. Pairwise comparison of gene expression in human tissues**

| Tissue      | Brain  | Breast | Colon  | Head/neck | Kidney | Lung   | Uterus |
|-------------|--------|--------|--------|-----------|--------|--------|--------|
| Total genes | 13,028 | 10,380 | 12,544 | 11,959    | 12,709 | 13,390 | 11,858 |
| Brain       | X      | 8,484  | 9,745  | 9,370     | 9,797  | 10,337 | 9,502  |
| Breast      |        | X      | 8,577  | 8,309     | 8,434  | 8,715  | 8,354  |
| Colon       |        |        | X      | 9,413     | 9,806  | 10,176 | 9,610  |
| Head/neck   |        |        |        | X         | 9,316  | 9,771  | 9,155  |
| Kidney      |        |        |        |           | X      | 10,299 | 9,621  |
| Lung        |        |        |        |           |        | X      | 9,922  |
| Uterus      |        |        |        |           |        |        | X      |

The numbers in each box refer to the number of genes expressed in common by the two tissues that head the rows and columns that intersect at that box. The number of genes that each of these tissues does not express in common with the other can be determined by subtracting the value in the box from the total number of genes expressed by each of the tissues, as listed in Table 3.

samples, in particular those derived from bulk tumors. Indeed, such impurity is readily detected by comparing ESTs present in libraries derived from bulk tumors but not corresponding cell lines, as can be easily accomplished by using the tools at the CGAP home page (<http://cgap.nci.nih.gov>). In the case of colon, for example, genes represented by ESTs from tumors but not cell lines include many examples (see Table 8, which is published as supporting information on the PNAS web site) that would not be expected to be present in epithelial cells. Given this tissue contamination, we might thus presume tissue specificity to be even higher than that measured, because the same contaminating lymphoid, fibroblast, endothelial, and smooth muscle cells are likely to be contaminants common to tumors from most locations.

On the other hand, we wondered whether the apparent high level of differential gene expression between tissues and between bulk tumors and cell lines might be exaggerated due to a still-insufficient level of sequence coverage. To address this possibility, we used data from the recently completed generation of several million transcript tags by Massively Parallel Signature Sequencing (MPSS) (21) from two human cell lines, one from breast and one from colon. In the MPSS experiment, the depth of sequencing guarantees the identification of all transcripts present at the level of at least one copy per cell (22). We found that in the MPSS experiment, ≈10,000 and 15,000 genes were expressed in the breast and colon cells, respectively, of which only 8,500 were expressed in both cell lines (21). In addition, inspection of the MPSS data also revealed that the great majority of the ESTs that we hypothesized were not derived from tumor epithelial cells and were also not detected in the colon cell line sequenced by MPSS, and those that were, were at very low levels (Table 8). Thus on both counts, the levels of EST coverage that we have achieved appear to provide a view of gene expression not significantly altered by very much deeper levels of transcript tagging.

A pairwise analysis of the tissues for which we have generated >100,000 ESTs indicates a consistency of shared and specific gene expression, with ≈70% of genes being expressed in common by any given pair (Table 4). These findings are consistent with the structure and function of human tissues being defined by the usage of highly variable permutations of genes, with almost all being expressed in more than one tissue. This would contrast, for example, with a model that requires a fixed set of ubiquitously expressed housekeeping genes with a substantial number of entirely tissue specific genes for individual tissues.

The EST data have proved extremely robust when used for the identification of genes with defined patterns of tissue specificity. Table 2 indicates examples of published studies that have identified CR genes with particular expression profiles and where the clusters representing the genes contained ESTs from our projects. The details of these studies are listed in Table 7. Of particular interest has been the identification of genes that are

restricted to organs such as the breast and prostate that could serve as therapeutic targets for cancers in these organs. In addition, growing interest is being focused on genes restricted in expression to normal testis and tumors (CT-antigens) that are important potential targets of therapeutic vaccines. The relatively sparse use of the data for tissue types other than breast and prostate indicates there are likely to be many more genes with restricted expression still to be identified by using our data.

**Identification of Gene and Transcript Variants.** In addition to gene definition and identification of the tissues in which a gene is expressed, ESTs also provide information concerning the existence of germline variants and alternatively spliced transcript forms. The former can have a significant impact on an individual's predisposition to cancer. Nevertheless, because such variants take the form of SNPs for the most part, the relatively high sequencing error rate (≈1%) of the single-pass transcript sequences we have generated requires that potential variants have to be carefully evaluated. Nevertheless, many studies have reported large-scale analyses of SNPs that have incorporated the ESTs we have generated (23, 24).

By using the original chromatograms, we identified a total of 237 SNPs for genes in the CR list, which are already listed in dbSNP. Of these, 47 were identified only with ORESTES data, 72 only with CGAP data, and 118 with both. In addition, a total of 210 further putative novel SNPs were identified by using the ORESTES data, and 295 were identified by using the CGAP data. Other studies have indicated that a high percentage of SNPs based on conventional ESTs, such as those generated in the CGAP project, can be subsequently verified (24). We suspected that, because ORESTES is a PCR-based methodology, a lower percentage of potential variants may be real. To test this, we examined experimentally 20 of the nonsynonymous SNPs identified with the ORESTES data alone and found that three (15%) were present in a bank of 100 normal human DNA sequences (Table 5). Although this percentage is (as expected) low, the ORESTES-based variants tend to be of particular

**Table 5. Validated new human SNPs based on ORESTES clusters**

| Gene   | Accession<br>name | no. | Variation   | Frequency of the variant allele<br>in 150 Brazilian individuals |        |          |
|--------|-------------------|-----|-------------|---|--------|----------|
|        |                   |     |             | Whites  | Blacks | Japanese |
| TM4SF3 | NM_004616         |     | G/C<br>G73A | 38% C   | 8% C   | 27% C    |
| CGM2   | X98311            |     | T/A F1201   | 39% A   | 63% A  | 62% A    |
| KISS1  | U43527            |     | A/G<br>Q36R | 8% G  | —      | —        |

—, Not yet performed.

interest, because they most frequently lie within the coding regions, thus potentially generating altered protein molecules. On the basis of our validation rate, we predict there are likely to be >30 true novel nonsynonymous SNPs identifiable by the ORESTES data alone within the CR list.

Alternative splicing generates variability at the transcriptome level and is conceivably of direct relevance to the generation of the malignant phenotype (25–27). Extensive utilization has been made of the ESTs we have generated for the identification of alternatively spliced genes within the human genome (12, 28–32). We explored the degree of variability due to alternative exon usage in the set of 1,124 CR genes by using two different approaches. The first approach, which detects transcripts displaying exon skipping through the use of a software called J-EXPLORER (12), predicts possible splice junctions and then searches the EST databases to identify transcripts spanning the annotated exon-exon junction. The second approach lists all variants based on transcript to genome mapping and is thus less stringent (33). The former found evidence of alternative splicing for 21.3% of all CR genes, with an average of 1.4 variants per gene, and the latter found that 47.5% of all genes in the list undergo alternative splicing with a total of 3,179 variants, and an average of 3.17 variants per gene. A total of 210 genes were found by both approaches as having more than one splicing variant. We sought to validate exon-skipping events identified in three of the genes: the skipping of exon 7 in CD 53 (NM\_00560), the skipping of exon 26 in PTPN13 (NM\_006264), and the insertion of an additional exon between denominated exons 1 and 2 in NM23A (NM\_000269). We had detected each of these only in ESTs derived from tumors, thus suggesting that the truncated transcripts might be tumor specific. However, by using a strategy that enhanced the amplification of the transcript with the skipped exon with an RT-PCR primer that spanned the novel splice site formed, we were able, in each case, to detect the alternative transcript in both normal and tumor derived samples. Thus, although carefully documented alternative splicing events can

almost always be readily validated, tumor-specific alternative splicing may be rare.

## Conclusion

The data we have generated have contributed to the identification of genes within the human genome, the definition of the tissue specificity of their expression, and the discovery of SNPs and alternatively spliced transcripts. It has been our intent that this should speed progress in the diagnosis, treatment, and understanding of human cancer. The number and variety of published studies in which the data we have generated appear to have played a significant role attest to the value of the open-source approach. Undoubtedly, vastly more progress has been made than if we had opted to concentrate on in-house utilization of our findings. Additionally, the thousands of high-quality EST clusters that are both spliced and contain predicted ORFs, the hundreds of genes that appear to have strongly tissue-restricted expression, as well as the thousands of potential SNPs and alternative splice forms that our data define, suggest that much more remains to be discovered within the existing publicly available data. We aim to continue generating sequence-based CR data with the aim of further improving the accurate transcriptional description of human tumors and the normal tissues from which they arise. We believe this will not only provide an important underlying support for broad-based cancer research but will also serve as the source for novel discoveries and hypotheses that will ultimately lead to improved cancer care.

We are indebted Lloyd J. Old, Scientific Director of the Ludwig Institute for Cancer Research, for support and stimulation; to J. Fernando Perez, Scientific Director of the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); and to Richard Klausner, Director of the National Cancer Institute, without whom these studies would not have been possible. We are also greatly indebted to Dr. Jucara Parra for dedication to the successful execution of these studies through her role as Project Manager for the FAPESP/Ludwig Institute for Cancer Research-Human Cancer Genome Project.

- Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., et al. (1999) *Cancer Res.* **59**, 5403–5407.
- DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996) *Nat. Genet.* **14**, 457–460.
- Strausberg, R. L., Dahl, C. A. & Klausner, R. D. (1997) *Nat. Genet.* **15**, Spec. No., 415–416.
- Bonalume, N. R. (1999) *Nature* **398**, 450.
- Strausberg, R. L., Camargo, A. A., Riggins, G. J., Schaefer, C. F., De Souza, S. J., Grouse, L. H., Lal, A., Buetow, K. H., Boon, K., Greenhut, S. F., et al. (2002) *Pharmacogenomics J.* **2**, 156–164.
- Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. (2000) *Trends Genet.* **16**, 103–106.
- Dias, N. E., Garcia, C. R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., et al. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 3491–3496.
- Iseli, C., Stevenson, B. J., De Souza, S. J., Samaia, H. B., Camargo, A. A., Buetow, K. H., Strausberg, R. L., Simpson, A. J., Bucher, P. & Jongeneel, C. V. (2002) *Genome Res.* **12**, 1068–1074.
- Iseli, C., Jongeneel, C. V. & Bucher, P. (1999) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
- Ewing, B. & Green, P. (1998) *Genome Res.* **8**, 186–194.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8**, 175–185.
- Hide, W. A., Babenko, V. N., van Heusden, P. A., Seoighe, C. & Kelso, J. F. (2001) *Genome Res.* **11**, 1848–1853.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) *Nature* **409**, 860–921.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001) *Science* **291**, 1304–1351.
- Roest, C. H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., et al. (2000) *Nat. Genet.* **25**, 235–238.
- Ewing, B. & Green, P. (2000) *Nat. Genet.* **25**, 232–234.
- Reymond, A., Camargo, A. A., Deutsch, S., Stevenson, B. J., Parmigiani, R. B., UCLA, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., et al. (2002) *Genomics* **79**, 824–832.
- Eddy, S. R. (2002) *Cell* **109**, 137–140.
- Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., et al. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 12103–12108.
- Boon, K., Osorio, E. C., Greenhut, S. F., Schaefer, C. F., Shoemaker, J., Polyak, K., Morin, P. J., Buetow, K. H., Strausberg, R. L., De Souza, S. J., et al. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 11287–11292.
- Jongeneel, C. V., Iseli, C., Stevenson, B. J., Riggins, G. J., Lal, A., Mackay, A., Harris, R. A., O'Hare, M. J., Neville, A. M., Simpson, A. J., et al. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 4702–4705.
- Brenner, S., Johnson, M., Brigham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., et al. (2000) *Nat. Biotechnol.* **18**, 630–634.
- Clifford, R., Edmonson, M., Hu, Y., Nguyen, C., Scherbier, T. & Buetow, K. H. (2000) *Genome Res.* **10**, 1259–1265.
- Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W. & Lee, C. J. (2000) *Nat. Genet.* **26**, 233–236.
- Caballero, O. L., De Souza, S. J., Brentani, R. R. & Simpson, A. J. (2001) *Dis. Markers* **17**, 67–75.
- Chun, S. Y., Bae, O. S. & Kim, J. B. (2000) *J. Korean Med. Sci.* **15**, 696–700.
- Sanchez, L. M., Hajos, S. E., Basilio, F. M., Mongini, C. & Alvarez, E. (2001) *Oncol. Rep.* **8**, 145–151.
- Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. (2002) *Nat. Genet.* **30**, 29–30.
- Kan, Z., States, D. & Gish, W. (2002) *Genome Res.* **12**, 1837–1845.
- Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001) *Nucleic Acids Res.* **29**, 2850–2859.
- Xie, H., Zhu, W. Y., Wasserman, A., Grebinskiy, V., Olson, A. & Mintz, L. (2002) *Genomics* **80**, 326–330.
- Xu, Q., Modrek, B. & Lee, C. (2002) *Nucleic Acids Res.* **30**, 3754–3766.
- Sakabe, N. J., Souza, J. E. S., Galante, P. F. A., Oliveira, P. S. L., Passetti, F., Brentani, H., Osorio, E. C., Zaiats, A., Leerkes, M. R., Kitajima, J. P., et al. (2003) *Proc. Fr. Acad. Sci.*, in press.

## **Apêndice B**

| <u>GeneCard</u> | <u>LocusLink</u> | <u>Gene Id</u> | <u>Annotation</u>   | <u>Reads CM2/CM4</u>      | <u>Reads CM2/CM4</u>      |
|-----------------|------------------|----------------|---|---------------------------|---------------------------|
| AB000410        | 4968             | OGG1           | 8-oxoguanine DNA glycosylase  | CM4-ST0189-051099-021-g08 |                           |
| AB002058        | 9127             | P2RXL1         | purinergic receptor P2X-like 1, orphan receptor   | CM2-HT0750-120700-269-g05 |                           |
| AB003698        | 8317             | CDC7L1         | CDC7 (cell division cycle 7, <i>S. cerevisiae</i> , homolog)-like 1                           | CM4-UT0075-021000-341-a07 |                           |
| AB011420        | 9263             | STK17A         | serine/threonine kinase 17a (apoptosis-inducing)  | CM4-UT0075-160900-319-g10 | CM4-UT0075-021000-349-b01 |
| AB011421        | 9262             | STK17B         | serine/threonine kinase 17b (apoptosis-inducing)  | CM4-NT0213-151200-607-e09 |                           |
| AB038162        | 7031             | TFF1           | trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)                    | CM4-CT0572-111100-429-c12 | CM2-CT0781-260301-702-d07 |
| AB040450        | 51771            | TOK-1          | cdk inhibitor p21 binding protein   | CM4-CN0096-191200-642-f03 |                           |
| AB040538        | 4928             | NUP98          | nucleoporin 98kD  | CM4-CT0276-201099-026-d01 | CM4-GN0329-010201-833-a11 |
| AF000231        | 8766             | RAB11A         | RAB11A, member RAS oncogene family  | CM4-CT0309-291199-051-h07 | CM4-CT0309-291199-051-f10 |
| AF003837        | 182              | JAG1           | jagged 1 (Alagille syndrome)  | CM4-CT0629-110101-669-c03 |                           |
| AF004162        | 10397            | NDRG1          | N-myc downstream regulated  | CM2-NN0116-100700-254-b02 | CM2-BN0303-190700-271-f03 |
| AF005774        | 8837             | CFLAR          | CASP8 and FADD-like apoptosis regulator   | CM4-ST0181-221099-028-b05 | CM4-MT0248-190101-818-h05 |
| AF010238        | 7428             | VHL            | von Hippel-Lindau syndrome  | CM4-ST0276-181299-064-d01 |                           |
| AF010309        | 9540             | PIG3           | quinone oxidoreductase homolog  | CM2-CT0662-020301-696-g12 | CM2-CT0662-020301-695-b02 |
| AF010312        | 9516             | PIG7           | LPS-induced TNF-alpha factor  | CM2-BN0303-110700-266-b09 |                           |
| AF010316        | 9536             | PTGES          | prostaglandin E synthase  | CM2-CI0179-181100-528-f12 | CM4-ST0276-181299-064-d01 |
| AF011905        | 5810             | RAD1           | RAD1 ( <i>S. pombe</i> ) homolog  | CM4-ST0276-181299-064-d01 |                           |
| AF015956        | 1616             | DAXX           | death-associated protein 6  | CM4-CT0308-291199-051-b04 | CM2-MT0158-291100-570-b11 |
| AF022212        | 395              | ARHGAP6        | Rho GTPase activating protein 6   | CM4-PT0015-071299-057-f03 |                           |
| AF023612        | 10907            | DIM1           | similar to <i>S. pombe</i> dim1+  | CM2-CT0482-300800-350-b07 |                           |
| AF027150        | 8487             | SIP1           | survival of motor neuron protein interacting protein 1  | CM2-CI0135-021100-475-g09 |                           |
| AF034795        | 1216             | CMAR           | cell matrix adhesion regulator  | CM4-CT0045-180200-512-e01 | CM2-CT0606-191200-643-c10 |
| AF035625        | 6794             | STK11          | serine/threonine kinase 11 (Peutz-Jeghers syndrome)   | CM4-HT0243-081199-037-d11 |                           |
| AF052124        | 6696             | SPP1           | secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1) | CM2-HT0947-070900-360-a12 | CM2-HT0945-230900-381-f12 |
| AF054183        | 5901             | RAN            | RAN, member RAS oncogene family   | CM4-CN0040-110100-073-c12 | CM2-CI0180-221100-541-c02 |
| AF059663        | 5002             | SLC22A1L       | solute carrier family 22 (organic cation transporter), member 1-like                          | CM2-CI0179-191200-635-a05 |                           |
| AF064243        | 6453             | ITSN1          | intersectin 1 (SH3 domain protein)  | CM2-CI0179-211200-650-d12 |                           |
| AF069747        | 9139             | CBFA2T2        | core-binding factor, runt domain, alpha subunit 2; translocated to, 2                         | CM4-ST0276-181299-064-d01 |                           |
| AF070539        | 8079             | MLF2           | myeloid leukemia factor 2   | CM4-FT0103-220600-210-g09 | CM4-IT0045-091200-582-h11 |
| AF087853        | 4616             | GADD45B        | growth arrest and DNA-damage-inducible, beta  | CM2-CT0782-090301-700-a01 | CM2-CT0782-190301-703-c05 |
| AF089868        | 4162             | MCAM           | melanoma cell adhesion molecule   | CM2-AN0077-210800-334-c10 |                           |
| AF092905        | 5500             | PPP1CB         | protein phosphatase 1, catalytic subunit, beta isoform  | CM2-CI0180-221100-541-d05 | CM2-GN0283-100101-683-d12 |
| AF096300        | 9448             | MAP4K4         | mitogen-activated protein kinase kinase kinase kinase 4                                       | CM2-AN0078-090800-320-g09 | CM2-NN0118-190700-270-f07 |
| AF097732        | 8915             | BCL10          | B-cell CLL/lymphoma 10  | CM2-CI0179-181100-528-f12 |                           |
| AF097996        | 5531             | PPP4C          | protein phosphatase 4 (formerly X), catalytic subunit   | CM4-LT0027-310100-505-d01 |                           |
| AF104863        | 4602             | MYB            | v-myb avian myeloblastosis viral oncogene homolog   | CM4-MT0024-230600-204-d11 | CM4-MT0024-230600-204-d05 |
| AF116604        | 4216             | MAP3K4         | mitogen-activated protein kinase kinase kinase 4  | CM4-UT0044-230900-310-h08 |                           |
| AF116610        | 7145             | TNS            | tensin  | CM4-FN0195-251100-468-e06 | CM4-CN0062-120101-799-g09 |

|                 |              |                |   |                           |                           |
|-----------------|--------------|----------------|---|---------------------------|---------------------------|
| <u>AF118124</u> | <u>4170</u>  | <u>MCL1</u>    | myeloid cell leukemia sequence 1 (BCL2-related)   | CM4-UT0075-021000-341-e06 | CM4-UT0075-160900-324-e11 |
| <u>AF119891</u> | <u>7879</u>  | <u>RAB7</u>    | RAB7, member RAS oncogene family  | CM2-FT0050-040700-252-f12 | CM4-ET0094-011100-398-d07 |
| <u>AF120103</u> | <u>6608</u>  | <u>SMOH</u>    | smoothened ( <i>Drosophila</i> ) homolog  | CM2-HT0968-141100-504-b01 |                           |
| <u>AF130122</u> | <u>5870</u>  | <u>RAB6A</u>   | RAB6A, member RAS oncogene family   | CM4-CT0106-240899-008-b05 | CM4-CT0106-240899-008-b04 |
| <u>AF136373</u> | <u>5879</u>  | <u>RAC1</u>    | ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1) | CM2-GN0294-030101-677-e02 | CM2-NN1153-311000-457-a05 |
| <u>AF147204</u> | <u>7852</u>  | <u>CXCR4</u>   | chemokine (C-X-C motif), receptor 4 (fusin)   | CM2-HB0003-020101-664-g04 |                           |
| <u>AF151378</u> | <u>51438</u> | <u>MAGEE1</u>  | melanoma antigen, family E, 1, cancer/testis specific                                   | CM2-BT0856-311000-465-b09 |                           |
| <u>AF152512</u> | <u>56110</u> | <u>PCDHGA5</u> | protocadherin gamma subfamily A, 5  | CM4-NT0290-150101-681-e06 | CM2-NT0185-071200-589-d04 |
| <u>AF152524</u> | <u>5098</u>  | <u>PCDHGC3</u> | protocadherin gamma subfamily C, 3  | CM4-NT0290-150101-681-e06 | CM2-NT0185-071200-589-d04 |
| <u>AF153979</u> | <u>10955</u> | <u>TDE1</u>    | tumor differentially expressed 1  | CM4-NN0001-040400-131-g10 | CM4-NT0286-310101-832-d08 |
| <u>AF175770</u> | <u>51684</u> | <u>SUFU</u>    | suppressor of fused   | CM4-NT0286-310101-831-g07 |                           |
| <u>AF177198</u> | <u>7094</u>  | <u>TLN1</u>    | talin 1   | CM4-MT0246-190101-810-e04 | CM4-MT0289-160201-761-h05 |
| <u>AF201944</u> | <u>26355</u> | <u>E2IG5</u>   | hypothetical protein, estradiol-induced   | CM4-ET0037-010600-186-d04 |                           |
| <u>AF208850</u> | <u>8073</u>  | <u>PTP4A2</u>  | protein tyrosine phosphatase type IVA, member 2   | CM4-UT0075-021000-341-g09 | CM4-NT0291-150101-663-f09 |
| <u>AF213459</u> | <u>2042</u>  | <u>EPHA3</u>   | EphA3   | CM4-FT0061-260600-218-c11 | CM4-FT0061-050700-227-f02 |
| <u>AF217975</u> | <u>11200</u> | <u>CHEK2</u>   | protein kinase Chk2   | CM4-MT0289-010201-842-g08 | CM2-CN0063-020101-661-e10 |
| <u>AF231512</u> | <u>9939</u>  | <u>RBM8A</u>   | RNA binding motif protein 8A  | CM2-CI0180-221100-541-e06 |                           |
| <u>AF237776</u> | <u>2872</u>  | <u>GPRK7</u>   | G protein-coupled receptor kinase 7   | CM4-HB0022-300101-830-e11 |                           |
| <u>AF262988</u> | <u>54386</u> | <u>RAP1</u>    | TRF2-interacting telomeric RAP1 protein   | CM2-GN0294-030101-677-g05 | CM4-DT0050-271299-068-c11 |
| <u>AF264750</u> | <u>58508</u> | <u>MLL3</u>    | myeloid/lymphoid or mixed-lineage leukemia 3  | CM4-GN0333-161200-630-g02 | CM2-UT0084-270900-398-g11 |
| <u>AF274943</u> | <u>1163</u>  | <u>CKS1</u>    | CDC28 protein kinase 1  | CM4-UT0042-281000-388-a02 |                           |
| <u>AF286028</u> | <u>6690</u>  | <u>SPINK1</u>  | serine protease inhibitor, Kazal type 1   | CM4-CT0629-160101-694-h07 | CM4-CT0629-150101-679-c09 |
| <u>AF329454</u> | <u>7113</u>  | <u>TMPRSS2</u> | transmembrane protease, serine 2  | CM2-FT0117-190700-274-c12 | CM2-FT0119-180700-280-e07 |
| <u>AJ224901</u> | <u>7750</u>  | <u>ZNF198</u>  | zinc finger protein 198   | CM2-HT0872-170800-332-d11 |                           |
| <u>AJ243211</u> | <u>1755</u>  | <u>DMBT1</u>   | deleted in malignant brain tumors 1   | CM4-CT0656-130201-852-d08 | CM2-HT0871-160800-328-b01 |
| <u>AJ243425</u> | <u>1958</u>  | <u>EGR1</u>    | early growth response 1   | CM4-MT0246-190101-810-b12 | CM2-HT0193-191099-022-e03 |
| <u>AJ251595</u> | <u>960</u>   | <u>CD44</u>    | CD44 antigen (homing function and Indian blood group system)                            | CM4-MT0248-190101-816-g11 |                           |
| <u>AJ278110</u> | <u>55510</u> | <u>HAGE</u>    | DEAD-box protein  | CM2-HT0184-061099-018-g11 |                           |
| <u>AK025474</u> | <u>5878</u>  | <u>RAB5C</u>   | RAB5C, member RAS oncogene family   | CM4-IT0045-051200-483-g03 |                           |
| <u>AL050259</u> | <u>5863</u>  | <u>RAB2L</u>   | RAB2, member RAS oncogene family-like   | CM4-CN0095-120101-673-c09 |                           |
| <u>AL121734</u> | <u>998</u>   | <u>CDC42</u>   | cell division cycle 42 (GTP-binding protein, 25kD)                                      | CM4-ET0037-010600-186-d11 | CM2-TN0138-070900-371-h02 |
| <u>AL137321</u> | <u>5862</u>  | <u>RAB2</u>    | RAB2, member RAS oncogene family  | CM4-NN1011-100300-109-c02 | CM2-MT0191-131200-591-b12 |
| <u>AY026895</u> | <u>6651</u>  | <u>SON</u>     | SON DNA binding protein   | CM2-CT0639-030101-682-a04 | CM4-GN0317-180101-688-f10 |
| <u>D00068</u>   | <u>4938</u>  | <u>OAS1</u>    | 2',5'-oligoadenylate synthetase 1 (40-46 kD)  | CM4-CT0106-240899-008-e02 | CM4-HN0021-241100-456-a03 |
| <u>D14134</u>   | <u>5888</u>  | <u>RAD51</u>   | RAD51 ( <i>S. cerevisiae</i> ) homolog ( <i>E. coli</i> RecA homolog)                   | CM4-ST0276-181299-064-b12 |                           |
| <u>D21089</u>   | <u>7508</u>  | <u>XPC</u>     | xeroderma pigmentosum, complementation group C  | CM4-NN1029-150800-543-h09 | CM4-NN1029-310300-530-c06 |
| <u>D21255</u>   | <u>1009</u>  | <u>CDH11</u>   | cadherin 11, type 2, OB-cadherin (osteoblast)   | CM2-HT0630-220300-125-f12 |                           |
| <u>D28124</u>   | <u>4681</u>  | <u>NBL1</u>    | neuroblastoma, suppression of tumorigenicity 1  | CM4-HT0869-190800-268-c01 |                           |
| <u>D29808</u>   | <u>7102</u>  | <u>TM4SF2</u>  | transmembrane 4 superfamily member 2  | CM4-HN0022-051200-472-e02 | CM4-HN0022-090101-453-g09 |

|        |              |                |  |                           |
|--------|--------------|----------------|--|---------------------------|
| D37965 | <u>5157</u>  | <u>PDGFR</u>   | platelet-derived growth factor receptor-like   | CM2-AN0077-220800-338-b10 |
| D38583 | <u>6282</u>  | <u>S100A11</u> | S100 calcium-binding protein A11 (calgizzarin)   | CM2-HT0392-240800-050-e07 |
| D43772 | <u>2886</u>  | <u>GRB7</u>    | growth factor receptor-bound protein 7   | CM2-HT0969-181100-509-d07 |
| D50406 | <u>8434</u>  | <u>RECK</u>    | reversion-inducing-cysteine-rich protein with kazal motifs   | CM2-GN0165-021100-487-b07 |
| D67031 | <u>120</u>   | <u>ADD3</u>    | adducin 3 (gamma)  | CM4-UT0046-050900-287-b08 |
| D83233 | <u>5468</u>  | <u>PAPRG</u>   | peroxisome proliferative activated receptor, gamma   | CM4-HT0652-150400-143-e11 |
| D88460 | <u>8976</u>  | <u>WASL</u>    | Wiskott-Aldrich syndrome-like  | CM2-CI0180-221100-540-g02 |
| D90278 | <u>1084</u>  | <u>CEACAM3</u> | carcinoembryonic antigen-related cell adhesion molecule 3  | CM4-CT0658-130201-859-a07 |
| J02958 | <u>4233</u>  | <u>MET</u>     | met proto-oncogene (hepatocyte growth factor receptor)   | CM2-CT0482-300800-349-g04 |
| J03191 | <u>5216</u>  | <u>PFN1</u>    | profilin 1   | CM4-BT0859-011100-402-h04 |
| J04810 | <u>4437</u>  | <u>MSH3</u>    | mutS (E. coli) homolog 3   | CM4-ST0189-051099-021-g08 |
| J05070 | <u>4318</u>  | <u>MMP9</u>    | matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase)                             | CM4-MT0289-010201-842-f06 |
| J05392 | <u>6382</u>  | <u>SDC1</u>    | syndecan 1   | CM4-GN0365-160101-685-f02 |
| J05500 | <u>6710</u>  | <u>SPTB</u>    | spectrin, beta, erythrocytic (includes spherocytosis, clinical type I)   | CM2-NN0210-281000-443-c10 |
| J05593 | <u>7077</u>  | <u>TIMP2</u>   | tissue inhibitor of metalloproteinase 2  | CM4-OT0161-300101-708-d11 |
| L02785 | <u>1811</u>  | <u>SLC26A3</u> | solute carrier family 26, member 3   | CM2-CT0782-080301-699-f05 |
| L04288 | <u>4820</u>  | <u>NKTR</u>    | natural killer-tumor recognition sequence  | CM4-CT0629-120101-671-h04 |
| L05624 | <u>5604</u>  | <u>MAP2K1</u>  | mitogen-activated protein kinase kinase 1  | CM4-UT0044-260900-570-H10 |
| L07648 | <u>4601</u>  | <u>MXI1</u>    | MAX-interacting protein 1  | CM2-FT0123-280700-305-G05 |
| L10333 | <u>6252</u>  | <u>RTN1</u>    | reticulon 1  | CM2-NN1032-190400-527-d12 |
| L11370 | <u>5097</u>  | <u>PCDH1</u>   | protocadherin 1 (cadherin-like 1)  | CM2-CI0135-021100-476-h06 |
| L12146 | <u>5523</u>  | <u>PPP2R3</u>  | protein phosphatase 2 (formerly 2A), regulatory subunit B'' (PP2B''), alpha isoform and (PP2B 130), beta isoform | CM4-CT0515-230900-321-f02 |
| L13288 | <u>7433</u>  | <u>VIPR1</u>   | vasoactive intestinal peptide receptor 1   | CM4-MT0246-180101-808-a12 |
| L13616 | <u>5747</u>  | <u>PTK2</u>    | PTK2 protein tyrosine kinase 2   | CM4-NN1011-260400-535-f08 |
| L14812 | <u>5933</u>  | <u>RBL1</u>    | retinoblastoma-like 1 (p107)   | CM4-ST0276-181299-064-d01 |
| L14848 | <u>4276</u>  | <u>MICA</u>    | MHC class I polypeptide-related sequence A   | CM2-MT0191-141200-590-g08 |
| L18877 | <u>4111</u>  | <u>MAGEA12</u> | melanoma antigen, family A, 12   | CM2-BT0856-311000-465-b09 |
| L20321 | <u>6787</u>  | <u>STK2</u>    | serine/threonine kinase 2  | CM4-ST0189-051099-021-g08 |
| L29277 | <u>6774</u>  | <u>STAT3</u>   | signal transducer and activator of transcription 3 (acute-phase response factor)                                 | CM4-CN0095-090101-670-h03 |
| L36033 | <u>6387</u>  | <u>SDF1</u>    | stromal cell-derived factor 1  | CM2-IT0039-211200-636-e07 |
| L36720 | <u>705</u>   | <u>BYSL</u>    | bystin-like  | CM2-CT0639-030101-680-c11 |
| L38929 | <u>5789</u>  | <u>PTPRD</u>   | protein tyrosine phosphatase, receptor type, D   | CM4-NN0011-170300-112-b07 |
| L38951 | <u>3837</u>  | <u>KPNB1</u>   | karyopherin (importin) beta 1  | CM4-CN0096-150101-678-e04 |
| L41142 | <u>6776</u>  | <u>STAT5A</u>  | signal transducer and activator of transcription 5A  | CM2-ET0193-261200-653-e04 |
| L42451 | <u>5164</u>  | <u>PDK2</u>    | pyruvate dehydrogenase kinase, isoenzyme 2   | CM4-NN1011-100300-110-f11 |
| L42542 | <u>10928</u> | <u>RALBP1</u>  | ralA binding protein 1   | CM4-MT0248-190101-818-c12 |
| L77886 | <u>5796</u>  | <u>PTPRK</u>   | protein tyrosine phosphatase, receptor type, K   | CM4-CN0095-090101-670-h07 |
| M12670 | <u>7076</u>  | <u>TIMP1</u>   | tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)                 | CM2-NT0194-301100-576-e07 |
| M12783 | <u>5155</u>  | <u>PDGFB</u>   | platelet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)                  | CM4-CT0574-101100-428-e08 |

|        |       |          |  |                           |
|--------|-------|----------|--|---------------------------|
| M13509 | 4312  | MMP1     | matrix metalloproteinase 1 (interstitial collagenase)  | CM2-HT0969-181100-509-a04 |
| M13975 | 5579  | PRKCB1   | protein kinase C, beta 1   | CM4-MT0287-160201-732-d04 |
| M15400 | 5925  | RB1      | retinoblastoma 1 (including osteosarcoma)  | CM2-GN0167-101100-489-e08 |
| M15796 | 5111  | PCNA     | proliferating cell nuclear antigen   | CM4-GN0057-180900-284-g01 |
| M16279 | 4267  | MIC2     | antigen identified by monoclonal antibodies 12E7, F21 and O13  | CM2-MT0158-301100-572-d09 |
| M18112 | 142   | ADPRT    | ADP-ribosyltransferase (NAD <sup>+</sup> ; poly (ADP-ribose) polymerase)   | CM4-CT0661-120201-872-d06 |
| M21574 | 5156  | PDGFRA   | platelet-derived growth factor receptor, alpha polypeptide   | CM4-UT0011-080900-306-f03 |
| M21616 | 5159  | PDGFRB   | platelet-derived growth factor receptor, beta polypeptide  | CM2-CI0135-021100-476-f07 |
| M24283 | 3383  | ICAM1    | intercellular adhesion molecule 1 (CD54), human rhinovirus receptor  | CM4-IT0042-131200-611-c12 |
| M24899 | 7067  | THRA     | thyroid hormone receptor, alpha (avian erythroblastic leukemia viral (v-erb-a) oncogene homolog)   | CM2-BT0856-311000-465-e09 |
| M26429 | 3107  | HLAC     | major histocompatibility complex, class I, C   | CM4-CT0572-011100-397-c01 |
| M26747 | 7068  | THRΒ     | thyroid hormone receptor, beta (avian erythroblastic leukemia viral (v-erb-a) oncogene homolog 2)  | CM4-CT0045-180200-512-h06 |
| M28209 | 5861  | RAB1     | RAB1, member RAS oncogene family   | CM2-CI0135-101100-484     |
| M28211 | 5867  | RAB4     | RAB4, member RAS oncogene family   | CM4-HN0021-241100-457-h02 |
| M28214 | 5865  | RAB3B    | RAB3B, member RAS oncogene family  | CM4-ST0276-181299-064-d01 |
| M28215 | 5868  | RAB5A    | RAB5A, member RAS oncogene family  | CM4-NN1032-280300-122-q07 |
| M28526 | 5175  | PECAM1   | platelet/endothelial cell adhesion molecule (CD31 antigen)   | CM4-MT0286-160201-729-c03 |
| M28650 | 2067  | ERCC1    | excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)            | CM2-FT0123-280700-304-G01 |
| M28696 | 2213  | FCGR2B   | Fc fragment of IgG, low affinity IIb, receptor for (CD32)  | CM4-CN0089-130201-726-c12 |
| M30142 | 1604  | DAF      | decay accelerating factor for complement (CD55, Cromer blood group system)   | CM4-CT0486-260900-573-H07 |
| M30938 | 7520  | XRCC5    | X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining; Ku autoantigen, 80kD)               | CM4-HN0020-221100-451-e01 |
| M31159 | 3486  | IGFBP3   | insulin-like growth factor binding protein 3   | CM2-KT0031-181200-629-e10 |
| M32315 | 7133  | TNFRSF1B | tumor necrosis factor receptor superfamily, member 1B  | CM2-TN0040-010800-242-c12 |
| M32977 | 7422  | VEGF     | vascular endothelial growth factor   | CM2-KT0030-211200-617-e08 |
| M33308 | 7414  | VCL      | vinculin   | CM2-NT0194-301100-567-e06 |
| M34668 | 5786  | PTPRA    | protein tyrosine phosphatase, receptor type, A   | CM4-NN1009-280300-123-b04 |
| M34840 | 55    | ACPP     | acid phosphatase, prostate   | CM4-CT0658-120201-857-b08 |
| M35416 | 5899  | RALB     | v-ras simian leukemia viral oncogene homolog B (ras related; GTP binding protein)  | CM4-MT0286-160201-729-b08 |
| M36951 | 5515  | PPP2CA   | protein phosphatase 2 (formerly 2A), catalytic subunit, alpha isoform  | CM4-UT0075-021000-341-h06 |
| M58286 | 7132  | TNFRSF1A | tumor necrosis factor receptor superfamily, member 1A  | CM2-CI0031-181000-420-c11 |
| M59465 | 7128  | TNFA1P3  | tumor necrosis factor, alpha-induced protein 3   | CM2-GN0166-101100-495-a02 |
| M59904 | 962   | CD48     | CD48 antigen (B-cell membrane protein)   | CM2-MT0158-051200-571-h01 |
| M60614 | 51352 | WIT-1    | Wilms tumor associated protein   | CM2-CT0315-211299-059-d03 |
| M60724 | 6198  | RPS6KB1  | ribosomal protein S6 kinase, 70kD, polypeptide 1   | CM4-CN0096-191200-642-g11 |
| M62399 | 5970  | RELA     | v-rei avian reticuloendotheliosis viral oncogene homolog A (nuclear factor of kappa light polypeptide gene enhancer in B-cells 3 (n&#651)) | CM4-MT0289-070201-743-b08 |
| M62626 | 3195  | HOX11    | homeo box 11 (T-cell lymphoma 3-associated breakpoint)   | CM2-CT0639-030101-681-d06 |
| M63960 | 5499  | PPP1CA   | protein phosphatase 1, catalytic subunit, alpha isoform  | CM4-FT0104-230600-215-a05 |
| M64572 | 5774  | PTPN3    | protein tyrosine phosphatase, non-receptor type 3  | CM4-CT0045-180200-512-b12 |
| M68941 | 5775  | PTPN4    | protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)  | CM2-CT0480-220800-340-b04 |

|           |      |                |   |                           |                           |
|-----------|------|----------------|---|---------------------------|---------------------------|
| M69066    | 4478 | <u>MSN</u>     | moesin  | CM4-CT0658-130201-859-g03 | CM2-HB0003-020101-665-h04 |
| M69225    | 667  | <u>BPAG1</u>   | bullous pemphigoid antigen 1 (230/240kD)  | CM4-UT0042-080900-307-b11 | CM4-ET0097-111100-426-a01 |
| M74524    | 7319 | <u>UBE2A</u>   | ubiquitin-conjugating enzyme E2A (RAD6 homolog)   | CM2-HT0631-220300-126-d04 |                           |
| M74525    | 7320 | <u>UBE2B</u>   | ubiquitin-conjugating enzyme E2B (RAD6 homolog)   | CM2-HT0631-220300-126-e04 | CM2-HT0631-220300-126-d04 |
| M74903    | 5777 | <u>PTPN6</u>   | protein tyrosine phosphatase, non-receptor type 6   | CM2-ET0193-211200-652-d03 | CM4-MT0237-180101-805-d12 |
| M77198    | 208  | <u>AKT2</u>    | v-akt murine thymoma viral oncogene homolog 2   | CM2-NN0116-100700-254-d11 | CM2-BN0303-190700-271-d05 |
| M77349    | 7045 | <u>TGFBI</u>   | transforming growth factor, beta-induced, 68kD  | CM4-CN0095-150101-683-c01 | CM4-CN0095-090101-670-a11 |
| M80783    | 7126 | <u>TNFAIP1</u> | tumor necrosis factor, alpha-induced protein 1 (endothelial)  | CM2-HT0671-010600-214-e08 | CM4-CT0656-120201-855-h07 |
| M81735    | 5424 | <u>POLD1</u>   | polymerase (DNA directed), delta 1, catalytic subunit (125kD)   | CM4-MT0218-120101-797-a12 |                           |
| M81830    | 6752 | <u>SSTR2</u>   | somatostatin receptor 2   | CM4-ET0097-111100-426-c12 |                           |
| M83221    | 5971 | <u>RELB</u>    | v-rei avian reticuloendotheliosis viral oncogene homolog B (nuclear factor of kappa light polypeptide gene enhancer in B-cells 3)                         | CM2-KT0030-141200-615-f08 |                           |
| M83738    | 5780 | <u>PTPN9</u>   | protein tyrosine phosphatase, non-receptor type 9   | CM2-CI0179-211200-650-e08 | CM2-NT0194-051200-565-e04 |
| M84757    | 6699 | <u>SPRR1B</u>  | small proline-rich protein 1B (cornifin)  | CM2-HT0670-290400-173-h09 | CM2-HT0870-170800-331-h10 |
| M86546    | 5087 | <u>PBX1</u>    | pre-B-cell leukemia transcription factor 1  | CM2-CT0662-080301-697-h11 |                           |
| M86699    | 7272 | <u>TTK</u>     | TTK protein kinase  | CM4-ST0189-051099-021-g08 |                           |
| M87339    | 5984 | <u>RFC4</u>    | replication factor C (activator 1) 4 (37kD)   | CM4-GN0332-211200-643-c12 |                           |
| M89914    | 4763 | <u>NF1</u>     | neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)   | CM4-CT0656-120201-853-a03 | CM4-HT0290-121199-042-c01 |
| M90102    | 4352 | <u>MPL</u>     | myeloproliferative leukemia virus oncogene  | CM4-ST0276-181299-064-d01 |                           |
| M92424    | 4193 | <u>MDM2</u>    | mouse double minute 2, human homolog of; p53-binding protein  | CM2-CT5001-270900-403-b04 |                           |
| M95678    | 5330 | <u>PLCD2</u>   | phospholipase C, beta 2   | CM2-MT0158-051200-571-g08 | CM2-CI0179-211200-639-c01 |
| M97935    | 6772 | <u>STAT1</u>   | signal transducer and activator of transcription 1, 91kD  | CM2-CT0478-220800-341-c08 | CM4-HB0026-211200-640-c09 |
| NM_000020 | 94   | <u>ACVR1L1</u> | activin A receptor type II-like 1   | CM2-GN0288-020101-669-g08 | CM4-CN0062-120101-793-c08 |
| NM_000037 | 286  | <u>ANK1</u>    | ankyrin 1, erythrocytic   | CM2-UT0112-141200-608-a08 |                           |
| NM_000107 | 1643 | <u>DDB2</u>    | damage-specific DNA binding protein 2 (48kD)  | CM4-ST0276-101299-059-c02 | CM4-ST0276-181299-064-h01 |
| NM_000118 | 2022 | <u>ENG</u>     | endoglin (Osler-Rendu-Weber syndrome 1)   | CM4-OT0161-220101-705-d03 | CM2-GN0294-030101-677-d09 |
| NM_000123 | 2073 | <u>ERCC5</u>   | excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum complementation group G (Cockayne syndrome)) | CM2-HB0003-020101-665-h12 |                           |
| NM_000177 | 2934 | <u>GSN</u>     | gelsolin (amyloidosis, Finnish type)  | CM4-MT0236-180101-803-d06 | CM2-NT0169-181100-534-e04 |
| NM_000210 | 3655 | <u>ITGA6</u>   | integrin, alpha 6   | CM2-CI0180-221100-532-e06 | CM2-CI0135-021100-492-b02 |
| NM_000211 | 3689 | <u>ITGB2</u>   | integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit)                                | CM4-MT0361-060201-735-e07 |                           |
| NM_000215 | 3718 | <u>JAK3</u>    | Janus kinase 3 (a protein tyrosine kinase, leukocyte)   | CM4-ST0189-051099-021-g08 |                           |
| NM_000234 | 3978 | <u>LIG1</u>    | ligase I, DNA, ATP-dependent  | CM4-CT0629-160101-687-c12 | CM4-CT0515-021000-346-f08 |
| NM_000380 | 7507 | <u>XPA</u>     | xeroderma pigmentosum, complementation group A  | CM4-HN0022-090101-453-a11 |                           |
| NM_000389 | 1026 | <u>CDKN1A</u>  | cyclin-dependent kinase inhibitor 1A (p21, Cip1)  | CM4-LT0027-310100-505-h12 | CM2-UM0045-060300-110-c11 |
| NM_000499 | 1543 | <u>CYP1A1</u>  | cytochrome P450, subfamily 1 (aromatic compound-inducible), polypeptide 1   | CM2-GN0167-101100-489-d12 |                           |
| NM_000572 | 3586 | <u>IL10</u>    | interleukin 10  | CM2-CI0179-181100-528-f12 | CM4-ST0276-181299-064-d01 |
| NM_000596 | 3484 | <u>IGFBP1</u>  | insulin-like growth factor binding protein 1  | CM4-MT0361-060201-735-b05 | CM2-CI0180-221100-540-c06 |
| NM_000611 | 966  | <u>CD59</u>    | CD59 antigen p18-20 (antigen identified by monoclonal antibodies 16.3A5, EJ16, EJ30, EL32 and G344)   | CM4-HN0023-221100-455-f11 | CM2-GN0288-020101-669-g11 |
| NM_000619 | 3458 | <u>IFNG</u>    | interferon, gamma   | CM2-CT5001-270900-404-d06 |                           |
| NM_000629 | 3454 | <u>IFNAR1</u>  | interferon (alpha, beta and omega) receptor 1   | CM2-MT0101-041000-293-C03 | CM4-HN0019-090101-469-g04 |

|           |      |               |  |                           |
|-----------|------|---------------|--|---------------------------|
| NM_000873 | 3384 | <u>ICAM2</u>  | intercellular adhesion molecule 2  | CM2-GN0220-231100-542-b09 |
| NM_000874 | 3455 | <u>IFNAR2</u> | interferon (alpha, beta and omega) receptor 2  | CM2-NT0192-051200-577-b07 |
| NM_000875 | 3480 | <u>IGF1R</u>  | insulin-like growth factor 1 receptor  | CM4-NN1011-100300-109-c01 |
| NM_000885 | 3676 | <u>JTGA4</u>  | integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)                                       | CM2-MT0100-310700-288-e05 |
| NM_000903 | 1728 | <u>DIA4</u>   | diaphorase (NADH/NADPH) (cytochrome b-5 reductase)   | CM2-HT0657-150400-156-e02 |
| NM_000965 | 5915 | <u>RARB</u>   | retinoic acid receptor, beta   | CM2-BT0857-311000-472-d08 |
| NM_001150 | 290  | <u>ANPEP</u>  | alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase CD13 n150) | CM4-LT0067-130100-079-c10 |
| NM_001187 | 574  | <u>BAGE</u>   | B melanoma antigen   | CM4-ST0276-181299-064-d01 |
| NM_001191 | 598  | <u>BCL2L1</u> | BCL2-like 1  | CM2-CI0179-201100-537-a09 |
| NM_001226 | 839  | <u>CASP6</u>  | caspase 6, apoptosis-related cysteine protease   | CM4-ST0276-181299-064-d01 |
| NM_001228 | 841  | <u>CASP8</u>  | caspase 8, apoptosis-related cysteine protease   | CM4-MT0248-190101-817-d07 |
| NM_001230 | 843  | <u>CASP10</u> | caspase 10, apoptosis-related cysteine protease  | CM4-ST0276-181299-064-d01 |
| NM_001251 | 968  | <u>CD68</u>   | CD68 antigen   | CM2-ET0126-181100-519-e12 |
| NM_001255 | 991  | <u>CDC20</u>  | CDC20 (cell division cycle 20, <i>S. cerevisiae</i> , homolog)   | CM2-NT0194-051200-565-a04 |
| NM_001320 | 1460 | <u>CSNK2B</u> | casein kinase 2, beta polypeptide  | CM4-OT0180-060201-741-h08 |
| NM_001348 | 1613 | <u>DAPK3</u>  | death-associated protein kinase 3  | CM2-PT0047-301299-063-e12 |
| NM_001394 | 1846 | <u>DUSP4</u>  | dual specificity phosphatase 4   | CM2-LT0061-180200-094-d08 |
| NM_001429 | 2033 | <u>EP300</u>  | E1A binding protein p300   | CM2-GN0294-020101-675-b11 |
| NM_001436 | 2091 | <u>FBL</u>    | fibrillarin  | CM2-HT0969-181100-509-a01 |
| NM_001619 | 156  | <u>ADRBK1</u> | adrenergic, beta, receptor kinase 1  | CM2-CT0606-211200-644-h06 |
| NM_001621 | 196  | <u>AHR</u>    | aryl hydrocarbon receptor  | CM2-HT0967-181100-507-b09 |
| NM_001631 | 248  | <u>ALPI</u>   | alkaline phosphatase, intestinal   | CM4-HB0025-150101-680-f08 |
| NM_001706 | 604  | <u>BCL6</u>   | B-cell CLL/lymphoma 6 (zinc finger protein 51)   | CM4-MT0289-010201-842-b10 |
| NM_001718 | 654  | <u>BMP6</u>   | bone morphogenetic protein 6   | CM2-CT0782-190301-703-c09 |
| NM_001753 | 857  | <u>CAV1</u>   | caveolin 1, caveolae protein, 22kD   | CM2-GN0221-291100-554-h02 |
| NM_001759 | 894  | <u>CCND2</u>  | cyclin D2  | CM4-MT0247-190101-815-f10 |
| NM_001760 | 896  | <u>CCND3</u>  | cyclin D3  | CM4-NT0213-061200-488-c05 |
| NM_001766 | 912  | <u>CD1D</u>   | CD1D antigen, d polypeptide  | CM4-GN0329-010201-833-e09 |
| NM_001771 | 933  | <u>CD22</u>   | CD22 antigen   | CM2-ET0016-170500-202-a04 |
| NM_001774 | 951  | <u>CD37</u>   | CD37 antigen   | CM4-CT0663-150101-677-f08 |
| NM_001777 | 961  | <u>CD47</u>   | CD47 antigen (Rh-related antigen, integrin-associated signal transducer)                                   | CM4-MT0246-180101-808-c07 |
| NM_001782 | 971  | <u>CD72</u>   | CD72 antigen   | CM4-MT0099-190700-287-g12 |
| NM_001786 | 983  | <u>CDC2</u>   | cell division cycle 2, G1 to S and G2 to M   | CM2-GN0294-020101-672-a03 |
| NM_001788 | 989  | <u>CDC10</u>  | CDC10 (cell division cycle 10, <i>S. cerevisiae</i> , homolog)   | CM4-MT0286-080201-835-c09 |
| NM_001792 | 1000 | <u>CDH2</u>   | cadherin 2, type 1, N-cadherin (neuronal)  | CM4-UT0042-281000-388-h07 |
| NM_001793 | 1001 | <u>CDH3</u>   | cadherin 3, type 1, P-cadherin (placental)   | CM2-CT0779-260301-711-e02 |
| NM_001794 | 1002 | <u>CDH4</u>   | cadherin 4, type 1, R-cadherin (retinal)   | CM4-CI0063-181000-359-f02 |
| NM_001795 | 1003 | <u>CDH5</u>   | cadherin 5, type 2, VE-cadherin (vascular epithelium)  | CM4-LT0025-161299-062-d07 |
| NM_001893 | 1453 | <u>CSNK1D</u> | casein kinase 1, delta   | CM4-BN0304-100700-273-B01 |
|           |      |               |  | CM2-FT0123-280700-304-C05 |
|           |      |               |  | CM4-NN1009-180300-115-f06 |
|           |      |               |  | CM4-BN0304-100700-273-B01 |
|           |      |               |  | CM4-HT0830-080700-234-g06 |
|           |      |               |  | CM2-BN0304-100700-273-B01 |
|           |      |               |  | CM2-GN0288-020101-668-a04 |
|           |      |               |  | CM4-ET0097-111100-426-b12 |
|           |      |               |  | CM2-RT0009-091200-606-e10 |

|           |      |                |   |                           |                           |
|-----------|------|----------------|---|---------------------------|---------------------------|
| NM_001896 | 1459 | <u>CSNK2A2</u> | casein kinase 2, alpha prime polypeptide  | CM4-CT0658-130201-858-e03 | CM4-HT0136-150999-014-a08 |
| NM_001904 | 1499 | <u>CTNNB1</u>  | catenin (cadherin-associated protein), beta 1 (88kD)  | CM2-CN0063-030101-659-b04 | CM2-CN0063-030101-662-h07 |
| NM_001909 | 1509 | <u>CTSD</u>    | cathepsin D (lysosomal aspartyl protease)   | CM4-HT0532-180400-107-d07 | CM4-HT0532-080300-107-d07 |
| NM_001923 | 1642 | <u>DDB1</u>    | damage-specific DNA binding protein 1 (127kD)   | CM2-GN0165-021100-485-a12 | CM2-NN0116-100700-254-b06 |
| NM_001942 | 1828 | <u>DSG1</u>    | desmoglein 1  | CM2-HT0947-070900-362-c12 |                           |
| NM_001962 | 1946 | <u>EFNA5</u>   | ephrin-A5   | CM2-HT0949-140900-374-a10 |                           |
| NM_001968 | 1977 | <u>EIF4E</u>   | eukaryotic translation initiation factor 4E   | CM4-KN0014-130201-730-c02 | CM2-BT0691-150300-117-b10 |
| NM_001987 | 2120 | <u>ETV6</u>    | ets variant gene 6 (TEL oncogene)   | CM4-CT0574-041100-409-e12 | CM2-ET0193-211200-652-b11 |
| NM_001992 | 2149 | <u>F2R</u>     | coagulation factor II (thrombin) receptor   | CM4-ST0189-051099-021-g08 | CM4-ST0276-181299-064-d01 |
| NM_001993 | 2152 | <u>F3</u>      | coagulation factor III (thromboplastin, tissue factor)  | CM2-FN0053-040700-250-f04 | CM2-FT0050-220600-247-g10 |
| NM_002019 | 2321 | <u>FLT1</u>    | fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)                            | CM4-GN0333-161200-630-a03 | CM4-GN0361-150101-647-g02 |
| NM_002020 | 2324 | <u>FLT4</u>    | fms-related tyrosine kinase 4   | CM2-CI0135-021100-476-f07 |                           |
| NM_002035 | 2531 | <u>FVT1</u>    | follicular lymphoma variant translocation 1   | CM2-CI0179-221100-527-d06 | CM2-ET0126-181100-518-g08 |
| NM_002060 | 2701 | <u>GJA4</u>    | gap junction protein, alpha 4, 37kD (connexin 37)   | CM2-GN0294-020101-678-a02 |                           |
| NM_002074 | 2782 | <u>GNB1</u>    | guanine nucleotide binding protein (G protein), beta polypeptide 1  | CM4-HT0243-081199-037-b03 |                           |
| NM_002078 | 2803 | <u>GOLGA4</u>  | golgi autoantigen, golgin subfamily a, 4  | CM4-CT0629-160101-687-h09 | CM2-GN0056-230800-345-b10 |
| NM_002082 | 2870 | <u>GPRK6</u>   | G protein-coupled receptor kinase 6   | CM2-MT0157-221100-548-f03 | CM2-MT0157-271100-547-d08 |
| NM_002089 | 2920 | <u>GRO2</u>    | GRO2 oncogene   | CM2-GN0165-031100-494-b04 |                           |
| NM_002094 | 2935 | <u>GSPT1</u>   | G1 to S phase transition 1  | CM4-NT0007-290400-157-f11 | CM2-GN0288-020101-669-g03 |
| NM_002153 | 3294 | <u>HSD17B2</u> | hydroxysteroid (17-beta) dehydrogenase 2  | CM4-OT0161-300101-708-b06 | CM4-OT0161-220101-705-e12 |
| NM_002183 | 3563 | <u>IL3RA</u>   | interleukin 3 receptor, alpha (low affinity)  | CM2-ET0126-181100-518-b10 |                           |
| NM_002205 | 3678 | <u>ITGA5</u>   | integrin, alpha 5 (fibronectin receptor, alpha polypeptide)   | CM2-EN0014-310500-207-c12 | CM2-GN0288-020101-668-c10 |
| NM_002206 | 3679 | <u>ITGA7</u>   | integrin, alpha 7   | CM4-ET0095-011100-405-c07 | CM2-GN0054-230800-344-f03 |
| NM_002209 | 3683 | <u>ITGAL</u>   | integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)                               | CM4-MT0248-190101-817-a05 | CM4-UM0003-050400-133-c07 |
| NM_002214 | 3696 | <u>ITGB8</u>   | integrin, beta 8  | CM2-BN0303-190700-271-d12 |                           |
| NM_002229 | 3726 | <u>JUNB</u>    | jun B proto-oncogene  | CM4-CT0658-120201-857-h05 | CM4-MT0237-180101-804-g11 |
| NM_002230 | 3728 | <u>JUP</u>     | junction plakoglobin  | CM2-HT0671-010600-214-f08 | CM4-CT0656-120201-853-g07 |
| NM_002231 | 3732 | <u>KAI1</u>    | Kai 1 (suppression of tumorigenicity 6, prostate; CD82 antigen (H2 leukocyte antigen, antigen detected by monoclonal antibody IA4)) | CM4-ST0276-181299-064-d01 |                           |
| NM_002253 | 3791 | <u>KDR</u>     | kinase insert domain receptor (a type III receptor tyrosine kinase)   | CM2-GN0055-060900-356-e01 | CM2-NN1147-250900-394-c07 |
| NM_002314 | 3984 | <u>LIMK1</u>   | LIM domain kinase 1   | CM4-CT0658-160201-758-a10 | CM4-HT0486-010200-513-d08 |
| NM_002332 | 4035 | <u>LRP1</u>    | low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)  | CM4-NT0213-151200-607-f01 | CM2-NT0244-211200-641-g08 |
| NM_002341 | 4050 | <u>LTB</u>     | lymphotoxin beta (TNF superfamily, member 3)  | CM2-GN0295-030101-657-g02 |                           |
| NM_002357 | 4084 | <u>MAD</u>     | MAX dimerization protein  | CM2-CI0180-201100-539-g05 |                           |
| NM_002359 | 4097 | <u>MAFG</u>    | v-maf musculoaponeurotic fibrosarcoma (avian) oncogene family, protein G  | CM4-NN0011-160300-113-d01 |                           |
| NM_002360 | 7975 | <u>MAFK</u>    | v-maf musculoaponeurotic fibrosarcoma (avian) oncogene family, protein K  | CM2-CI0180-221100-540-g08 |                           |
| NM_002417 | 4288 | <u>MKI67</u>   | antigen identified by monoclonal antibody Ki-67   | CM4-MT0248-190101-817-h07 | CM2-ET0126-181100-520-d06 |
| NM_002447 | 4486 | <u>MST1R</u>   | macrophage stimulating 1 receptor (c-met-related tyrosine kinase)   | CM2-CT0662-080301-697-c09 | CM2-BT0857-021100-473-f05 |
| NM_002466 | 4605 | <u>MYBL2</u>   | v-myb avian myeloblastosis viral oncogene homolog-like 2  | CM4-IT0045-091200-582-b07 |                           |
| NM_002524 | 4893 | <u>NRAS</u>    | neuroblastoma RAS viral (v-ras) oncogene homolog  | CM4-HT0243-081199-037-g07 |                           |

|           |      |        |   |   |
|-----------|------|--------|---|---|
| NM_002584 | 5081 | PAX7   | paired box gene 7   | CM2-HT0968-111100-500-g07                           |
| NM_002634 | 5245 | PHB    | prohibitin  | CM4-UT0009-050900-567-f03 CM4-KT0035-151200-603-e09 |
| NM_002710 | 5501 | PPP1CC | protein phosphatase 1, catalytic subunit, gamma isoform   | CM4-HB0025-150101-680-d11 CM4-HB0025-150101-680-h08 |
| NM_002870 | 5872 | RAB13  | RAB13, member RAS oncogene family   | CM4-HT1152-161200-627-c04 CM4-HT1152-191200-639-b06 |
| NM_002886 | 5912 | RAP2B  | RAP2B, member of RAS oncogene family  | CM2-MT0157-221100-548-g04                           |
| NM_002956 | 6249 | RSN    | restin (Reed-Steinberg cell-expressed intermediate filament-associated protein)                             | CM4-HT0285-281099-033-a10                           |
| NM_002957 | 6256 | RXRA   | retinoid X receptor, alpha  | CM4-HN0020-181100-444-b04 CM2-MT0157-231100-546-a08 |
| NM_002958 | 6259 | RYK    | RYK receptor-like tyrosine kinase   | CM2-CT5001-270900-403-b02                           |
| NM_003120 | 6688 | SPI1   | spleen focus forming virus (SFFV) proviral integration oncogene spi1  | CM4-KN0014-130201-730-c03                           |
| NM_003147 | 6757 | SSX2   | synovial sarcoma, X breakpoint 2  | CM4-PT0034-180200-506-c07 CM2-CI0135-021100-490-b04 |
| NM_003246 | 7057 | THBS1  | thrombospondin 1  | CM4-CT0629-120101-671-a05 CM4-CT0629-230101-696-e08 |
| NM_003295 | 7178 | TPT1   | tumor protein, translationally-controlled 1   | CM4-MT0286-060201-738-e02 CM4-MT0286-050201-836-c09 |
| NM_003319 | 7273 | TTN    | titin   | CM2-HT0968-141100-504-f04 CM2-HT0969-181100-505-g06 |
| NM_003326 | 7292 | TNFSF4 | tumor necrosis factor (ligand) superfamily, member 4 (tax-transcriptionally activated glycoprotein 1, 34kD) | CM2-ET0127-291100-522-d06                           |
| NM_003391 | 7472 | WNT2   | wingless-type MMTV integration site family member 2   | CM4-CT0308-291199-051-d10                           |
| NM_003423 | 7594 | ZNF43  | zinc finger protein 43 (HTF6)   | CM4-HT0396-201299-500-a05 CM4-HT0509-060600-533-a07 |
| NM_003461 | 7791 | ZYX    | zyxin   | CM2-ET0126-201100-517-c02                           |
| NM_003472 | 7913 | DEK    | DEK oncogene (DNA binding)  | CM2-HT0945-150900-379-a04 CM2-MT0191-141200-590-b09 |
| NM_003686 | 9156 | EXO1   | exonuclease 1   | CM4-ST0181-231199-049-c06                           |
| NM_003718 | 8621 | CDC2L5 | cell division cycle 2-like 5 (cholinesterase-related cell division controller)                              | CM4-HT0194-191099-025-a04                           |
| NM_003757 | 8668 | EIF3S2 | eukaryotic translation initiation factor 3, subunit 2 (beta, 36kD)  | CM4-BT0326-221199-048-a10 CM2-FT0123-280700-305-C04 |
| NM_003792 | 8721 | GAGEB1 | endothelial differentiation-related factor 1  | CM4-HT0920-281000-387-f06                           |
| NM_003875 | 8833 | GMPS   | guanine monophosphate synthetase  | CM4-ST0276-181299-064-a04 CM2-BT0688-210300-122-d04 |
| NM_003929 | 8934 | RAB7L1 | RAB7, member RAS oncogene family-like 1   | CM4-ST0276-181299-064-d01                           |
| NM_003992 | 1198 | CLK3   | CDC-like kinase 3   | CM4-HT0869-190800-267-c12 CM2-HT0193-191099-022-d08 |
| NM_003993 | 1196 | CLK2   | CDC-like kinase 2   | CM4-HT0481-180200-510-a10                           |
| NM_004040 | 388  | ARHB   | ras homolog gene family, member B   | CM4-HT0136-150999-014-c06                           |
| NM_004068 | 1173 | AP2M1  | adaptor-related protein complex 2, mu 1 subunit   | CM2-KT0031-211200-627-c05 CM2-CI0180-201100-539-c03 |
| NM_004073 | 1263 | CNK    | cytokine-inducible kinase   | CM4-LT0057-140100-080-g12 CM2-CT0481-070900-359-b03 |
| NM_004090 | 1845 | DUSP3  | dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)                                     | CM2-CI0180-221100-532-h03                           |
| NM_004111 | 2237 | FEN1   | flap structure-specific endonuclease 1  | CM4-NT0290-150101-681-d12                           |
| NM_004302 | 91   | ACVR1B | activin A receptor, type IB   | CM2-CT0662-080301-697-b09                           |
| NM_004311 | 403  | ARL3   | ADP-ribosylation factor-like 3  | CM2-HT0668-010500-175-c03                           |
| NM_004327 | 613  | BCR    | breakpoint cluster region   | CM2-GN0294-020101-676-a12 CM4-CI0062-181000-370-b10 |
| NM_004329 | 657  | BMPR1A | bone morphogenetic protein receptor, type IA  | CM2-CI0179-201100-537-b11                           |
| NM_004333 | 673  | BRAF   | v-raf murine sarcoma viral oncogene homolog B1  | CM2-HT0949-070900-373-a04                           |
| NM_004394 | 1611 | DAP    | death-associated protein  | CM4-NT0290-090101-661-b06 CM4-MT0218-120101-797-d09 |
| NM_004401 | 1676 | DFFA   | DNA fragmentation factor, 45 kD, alpha polypeptide  | CM4-ST0189-051099-021-g08 CM4-ST0276-181299-064-d01 |
| NM_004408 | 1759 | DNM1   | dynamin 1   | CM2-MT0246-190101-811-e05 CM2-LT0066-030100-109-c04 |

|           |       |        |   |                             |                             |
|-----------|-------|--------|---|-----------------------------|-----------------------------|
| NM_004409 | 1760  | DMPK   | dystrophia myotonica-protein kinase   | CM4-C10629-150101-6/9-c02_1 | CM4-FN0193-181100-442-e04   |
| NM_004419 | 1847  | DUSP5  | dual specificity phosphatase 5  | CM4-N10290-150101-684-e11_1 | CM4-AN0081-210800-274-h01   |
| NM_004428 | 1942  | EFNA1  | ephrin-A1   | CM4-NT0213-051200-482-h03   | CM4-CT0278-221099-027-b07   |
| NM_004429 | 1947  | EFNB1  | ephrin-B1   | CM4-CT0659-130201-864-g01   | CM2-GN0167-141100-497-d11   |
| NM_004430 | 1960  | EGR3   | early growth response 3   | CM2-ET0127-291100-524-a09   |                             |
| NM_004431 | 1969  | EPHA2  | EphA2   | CM2-GN0283-100101-683-g12   | CM2-CI0031-051000-409-h03   |
| NM_004444 | 2050  | EPHB4  | EphB4   | CM4-CT0574-011100-404-g05   | CM2-CI0180-201100-539-a07   |
| NM_004450 | 2079  | ERH    | enhancer of rudimentary ( <i>Drosophila</i> ) homolog   | CM4-PT0047-060100-069-a12   |                             |
| NM_004505 | 9098  | USP6   | ubiquitin specific protease 6 (Tre-2 oncogene)  | CM2-GN0284-100101-684-a01   |                             |
| NM_004616 | 7103  | TM4SF3 | transmembrane 4 superfamily member 3  | CM4-NN0003-160300-111-c02   |                             |
| NM_004641 | 8028  | MLLT10 | myeloid/lymphoid or mixed-lineage leukemia (trithorax ( <i>Drosophila</i> ) homolog); translocated to, 10         | CM4-NT0291-150101-663-e05   |                             |
| NM_004725 | 9184  | BUB3   | BUB3 (budding uninhibited by benzimidazoles 3, yeast) homolog   | CM4-NT0214-061200-489-h08   | CM2-KT0030-141200-615-g03   |
| NM_004799 | 9372  | MADHIP | MAU (mothers against decapentaplegic, <i>Drosophila</i> ) homolog interacting protein, receptor activation anchor | CM2-CT0481-070900-359-g10   |                             |
| NM_004938 | 1612  | DAPK1  | death-associated protein kinase 1   | CM2-GN0294-020101-675-b11   | CM2-GN0295-020101-658-c11   |
| NM_005085 | 8021  | NUP214 | nucleoporin 214kD (CAIN)  | CM4-NT0286-310101-831-d05   | CM2-NT0170-211200-647-e01   |
| NM_005106 | 9940  | DLEC1  | deleted in lung and esophageal cancer 1   | CM4-CT0658-160201-758-g07   |                             |
| NM_005163 | 207   | AKT1   | v-akt murine thymoma viral oncogene homolog 1   | CM2-BN0303-120700-265-g07   | CM2-BN0303-120700-265-h12   |
| NM_005197 | 1112  | CHES1  | checkpoint suppressor 1   | CM4-CT0515-181000-340-d07   | CM4-MT0286-060201-738-g07   |
| NM_005228 | 1956  | EGFR   | epidermal growth factor receptor (avian erythroblastic leukemia viral (v-erb-b) oncogene homolog)                 | CM2-FN0110-110700-264-f05   | CM2-BN0302-100700-259-g10   |
| NM_005229 | 2002  | ELK1   | ELK1, member of ETS oncogene family   | CM2-CI0135-021100-476-e06   |                             |
| NM_005231 | 2017  | EMS1   | ems1 sequence (mammary tumor and squamous cell carcinoma-associated (p80/85 src substrate)                        | CM2-CT0606-201200-645-e05   | CM2-CI0179-211200-650-d06   |
| NM_005245 | 2195  | FAT    | FAT tumor suppressor ( <i>Drosophila</i> ) homolog  |                             | CM2-HT0870-220800-339-h03   |
| NM_005253 | 2355  | FOSL2  | FOS-like antigen 2  |                             | CM4-HT0137-220999-017-e07   |
| NM_005312 | 2889  | GRF2   | guanine nucleotide-releasing factor 2 (specific for crk proto-oncogene)   | CM4-ET0096-011100-400-a09   | CM2-BN0220-080500-181-g01   |
| NM_005343 | 3265  | HRAS   | v-Ha-ras Harvey rat sarcoma viral oncogene homolog  | CM4-NN1029-150800-543-d11   | CM4-HT0243-081199-037-g07   |
| NM_005356 | 3932  | LCK    | lymphocyte-specific protein tyrosine kinase   | CM2-CT0606-100101-646-g04   |                             |
| NM_005377 | 4611  | MYCL2  | v-myc avian myelocytomatisis viral oncogene homolog 2   | CM4-ST0276-181299-064-d01   |                             |
| NM_005433 | 7525  | YES1   | v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1  | CM4-ST0276-181299-064-d01   |                             |
| NM_005564 | 3934  | LCN2   | lipocalin 2 (oncogene 24p3)   | CM4-UT0043-050900-293-f09   | CM2-HT0551-010300-108-b06   |
| NM_005574 | 4005  | LMO2   | LIM domain only 2 (rhomboitin-like 1)   | CM4-FT0104-230600-212-e04   |                             |
| NM_005583 | 4066  | LYL1   | lymphoblastic leukemia derived sequence 1   | CM2-CT5002-270900-402-f10   |                             |
| NM_005637 | 6760  | SS18   | synovial sarcoma, translocated to X chromosome  | CM2-HT0283-191199-036-g04   | CM2-CI0135-021100-490-b04   |
| NM_005808 | 10217 | HYA22  | HYA22 protein   | CM4-CT0656-120201-855-d09   | CM4-C10629-110101-669-h03_1 |
| NM_005814 | 10223 | GPA33  | glycoprotein A33 (transmembrane)  | CM2-CT0639-030101-681-f04   |                             |
| NM_005851 | 10263 | DOC-1R | tumor suppressor deleted in oral cancer-related 1   | CM4-BT0325-151299-061-h04   | CM2-CT0780-260301-714-b12   |
| NM_005858 | 10270 | AKAP8  | A kinase (PRKA) anchor protein 8  | CM4-NN1157-011100-407-f03   |                             |
| NM_005883 | 10297 | APCL   | adenomatous polyposis coli like   | CM2-UT0108-211200-621-c08   |                             |
| NM_005902 | 4088  | MADH3  | MAD (mothers against decapentaplegic, <i>Drosophila</i> ) homolog 3   | CM4-CT0656-120201-854-d10   |                             |
| NM_005938 | 4303  | MLLT7  | myeloid/lymphoid or mixed-lineage leukemia (trithorax ( <i>Drosophila</i> ) homolog); translocated to, 7          | CM2-MT0157-221100-548-d04   |                             |

|           |       |            |  |  |                           |
|-----------|-------|------------|--|--|---------------------------|
| NM_005940 | 4320  | MMP11      | matrix metalloproteinase 11 (stromelysin 3)  |  | CM4-GN0329-010201-833-h08 |
| NM_006010 | 7873  | ARMET      | arginine-rich, mutated in early stage tumors   |  | CM4-CT0574-041100-410-c05 |
| NM_006020 | 8846  | ABH        | alkylation repair; alkB homolog  |  | CM4-CT0656-120201-854-f03 |
| NM_006218 | 5290  | PIK3CA     | phosphoinositide-3-kinase, catalytic, alpha polypeptide  |  | CM2-GN0055-060900-356-b08 |
| NM_006533 | 8190  | MIA        | melanoma inhibitory activity   |  | CM4-ET0237-220101-699-d04 |
| NM_006633 | 10788 | IQGAP2     | IQ motif containing GTPase activating protein 2  |  | CM4-HB0021-300101-825-f02 |
| NM_006738 | 3928  | LBC        | lymphoid blast crisis oncogene   |  | CM2-NT0185-061200-585-g02 |
| NM_006763 | 7832  | BTG2       | BTG family, member 2   |  | CM4-FN0193-181100-442-b09 |
| NM_006831 | 10978 | HEAB       | ATP/GTP-binding protein  |  | CM4-UT0042-281000-386-g11 |
| NM_006890 | 1087  | CEACAM7    | carcinoembryonic antigen-related cell adhesion molecule 7  |  | CM2-CN0063-030101-662-f10 |
| NM_007068 | 11144 | DMC1       | DMC1 (dosage suppressor of mck1, yeast homolog) meiosis-specific homologous recombination  |  | CM4-ST0276-181299-064-d01 |
| NM_007097 | 1212  | CLTB       | clathrin, light polypeptide (Lcb)  |  | CM4-CT0629-110101-669-f11 |
| NM_007295 | 672   | BRCA1      | breast cancer 1, early onset   |  | CM4-ST0276-181299-064-d01 |
| NM_012138 | 26574 | DED        | apoptosis antagonizing transcription factor  |  | CM4-ST0163-081099-023-d04 |
| NM_012141 | 26512 | DDX26      | deleted in cancer 1; RNA helicase HDB/DICE1  |  | CM4-UT0011-080900-306-c12 |
| NM_012153 | 26298 | EHF        | ets homologous factor  |  | CM2-CT0478-220800-341-f02 |
| NM_012257 | 26959 | HBP1       | HMG-box containing protein 1   |  | CM2-HT0949-070900-373-b05 |
| NM_012317 | 23641 | LDOC1      | leucine zipper, down-regulated in cancer 1   |  | CM2-NN1154-311000-463-a03 |
| NM_012484 | 3161  | HMMR       | hyaluronan-mediated motility receptor (RHAMM)  |  | CM2-GN0167-101100-489-b02 |
| NM_013314 | 29760 | BLNK       | B cell linker protein  |  | CM2-CT0662-080301-697-b05 |
| NM_014366 | 26354 | E2IG3      | putative nucleotide binding protein, estradiol-induced   |  | CM4-HT0194-091199-038-g05 |
| NM_014503 | 27340 | DRIM       | down-regulated in metastasis   |  | CM2-CI0180-221100-541-b04 |
| NM_014622 | 4013  | LUH11CH2_A | loss of heterozygosity, 11, chromosomal region 2, gene A   |  | CM2-GN0166-101100-495-g12 |
| NM_015646 | 5908  | RAP1B      | RAP1B, member of RAS oncogene family   |  | CM2-HT0194-081099-020-b05 |
| NM_016131 | 10890 | RAB10      | RAB10, member RAS oncogene family  |  | CM2-CN0063-020101-661-a07 |
| NM_016520 | 51759 | LOC51759   | hepatocellular carcinoma-associated antigen 59   |  | CM2-HB0003-030101-663-f10 |
| NM_016561 | 51283 | LOC51283   | apoptosis regulator  |  | CM2-CI0179-181100-528-f12 |
| NM_016733 | 3985  | LIMK2      | LIM domain kinase 2  |  | CM2-HT0969-181100-509-d03 |
| NM_017528 | 54743 | HASJ4442   | putative methyltransferase   |  | CM2-ST0189-051099-021-g08 |
| NM_021135 | 6196  | RPS6KA2    | ribosomal protein S6 kinase, 90kD, polypeptide 2   |  | CM4-HT0509-220300-120-e06 |
| NM_021874 | 994   | CDC25B     | cell division cycle 25B  |  | CM4-CN0091-060201-739-c06 |
| NM_022171 | 6988  | TCTA       | T-cell leukemia translocation altered gene   |  | CM2-MT0099-190700-285-G10 |
| NM_022443 | 4291  | MLF1       | myeloid leukemia factor 1  |  | CM2-HT0945-230900-381-g06 |
| NM_023028 | 2263  | FGFR2      | fibroblast growth factor receptor 2 (beta-1-expressed kinase, keratinocyte growth factor receptor, craniofacial dysostosis 1, Craniostenosis, Pfeiffer syndrome, Jackson-Weiss syndrome) |  | CM2-NT0169-141200-611-f09 |
| NM_024423 | 1825  | DSC3       | desmocollin 3  |  | CM4-ST0276-181299-064-d01 |
| NM_031966 | 891   | CCNB1      | cyclin B1  |  | CM4-ET0037-010600-186-e11 |
| NM_032043 | 83990 | BRIP1      | BRCA1-interacting protein 1; BRCA1-associated C-terminal helicase 1  |  | CM4-ST0276-181299-064-d01 |
| NM_032982 | 835   | CASP2      | caspase 2, apoptosis-related cysteine protease (neural precursor cell expressed, developmentally down-regulated 2)   |  | CM2-GN0220-221100-543-e07 |
| S38729    | 2547  | G22P1      | thyroid autoantigen 70kD (Ku antigen)  |  | CM4-HT0920-090900-299-a05 |
|           |       |            |  |  | CM4-HT0920-160900-302-c10 |

|        |              |                 |   |                             |
|--------|--------------|-----------------|---|-----------------------------|
| S66431 | <u>5927</u>  | <u>RBBP2</u>    | retinoblastoma-binding protein 2  | CM4-CT0656-120201-854-b12   |
| S78271 | <u>8243</u>  | <u>SMC1L1</u>   | SMC1 (structural maintenance of chromosomes 1, yeast)-like 1                                      | CM4-CI0055-181000-360-h09   |
| S79942 | <u>1974</u>  | <u>E1F4A2</u>   | eukaryotic translation initiation factor 4A, isoform 2  | CM4-HT0486-010200-513-b07   |
| U01038 | <u>5347</u>  | <u>PLK</u>      | polo ( <i>Drosophila</i> )-like kinase  | CM4-CT0659-120201-866-g11   |
| U02570 | <u>392</u>   | <u>ARHGAP1</u>  | Rho GTPase activating protein 1   | CM4-AN0081-210800-273-f03   |
| U02680 | <u>5756</u>  | <u>PTK9</u>     | protein tyrosine kinase 9   | CM4-CT0478-230800-347-b08   |
| U03397 | <u>3604</u>  | <u>TNFRSF9</u>  | tumor necrosis factor receptor superfamily, member 9  | CM4-ST0276-181299-064-d01   |
| U03735 | <u>4102</u>  | <u>MAGEA3</u>   | melanoma antigen, family A, 3   | CM4-BT0856-311000-465-b09   |
| U03858 | <u>2323</u>  | <u>FLT3LG</u>   | fms-related tyrosine kinase 3 ligand  | CM4-HT0968-111100-512-b12   |
| U03911 | <u>4436</u>  | <u>MSH2</u>     | mutS ( <i>E. coli</i> ) homolog 2 (colon cancer, nonpolyposis type 1)                             | CM4-CT5003-270900-406-f03   |
| U04313 | <u>5268</u>  | <u>SERPINB5</u> | serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5                          | CM4-UT0042-090900-308-c10   |
| U06654 | <u>2315</u>  | <u>MLANA</u>    | melan-A   | CM4-ST0189-051099-021-g08   |
| U08839 | <u>5329</u>  | <u>PLAUR</u>    | plasminogen activator, urokinase receptor   | CM4-CT0779-190301-709-a06   |
| U10586 | <u>4110</u>  | <u>MAGEA11</u>  | melanoma antigen, family A, 11  | CM4-GN0332-161200-631-c06   |
| U10685 | <u>4109</u>  | <u>MAGEA10</u>  | melanoma antigen, family A, 10  | CM4-BT0856-311000-465-b09   |
| U10693 | <u>4107</u>  | <u>MAGEA8</u>   | melanoma antigen, family A, 8   | CM4-BT0856-311000-465-b09   |
| U10694 | <u>4108</u>  | <u>MAGEA9</u>   | melanoma antigen, family A, 9   | CM4-GN0332-161200-631-c06   |
| U12597 | <u>7186</u>  | <u>TRAF2</u>    | TNF receptor-associated factor 2  | CM4-CT0639-220101-695-f08   |
| U12779 | <u>9261</u>  | <u>MAPKAPK2</u> | mitogen-activated protein kinase-activated protein kinase 2                                       | CM4-UT0009-240800-279-d12   |
| U14588 | <u>5829</u>  | <u>PXN</u>      | paxillin  | CM4-GN0298-061200-496-g05   |
| U14658 | <u>5395</u>  | <u>PMS2</u>     | postmeiotic segregation increased ( <i>S. cerevisiae</i> ) 2                                      | CM4-HB0021-300101-825-e11   |
| U15131 | <u>6764</u>  | <u>ST5</u>      | suppression of tumorigenicity 5   | CM4-CT0486-050900-289-a02   |
| U16031 | <u>6778</u>  | <u>STAT6</u>    | signal transducer and activator of transcription 6, interleukin-4 induced                         | CM4-MT0289-010201-841-g04   |
| U16296 | <u>7074</u>  | <u>TIAM1</u>    | T-cell lymphoma invasion and metastasis 1   | CM4-NT0247-061200-583-h03   |
| U16307 | <u>11010</u> | <u>RTVP1</u>    | glioma pathogenesis-related protein   | CM4-FT0115-190700-283-h11   |
| U17714 | <u>6767</u>  | <u>ST13</u>     | suppression of tumorigenicity 13 (colon carcinoma) (Hsp70-interacting protein)                    | CM4-MT0286-070201-749-c07   |
| U18321 | <u>7818</u>  | <u>DAP3</u>     | death associated protein 3  | CM4-HT0551-170200-093-h03   |
| U20657 | <u>7375</u>  | <u>USP4</u>     | ubiquitin specific protease 4 (proto-oncogene)  | CM4-ST0163-081099-023-b09   |
| U24497 | <u>5310</u>  | <u>PKD1</u>     | polycystic kidney disease 1 (autosomal dominant)  | CM4-CI0031-181000-420-b12   |
| U28946 | <u>2956</u>  | <u>MSH6</u>     | mutS ( <i>E. coli</i> ) homolog 6   | CM4-HT0920-160900-302-a04   |
| U28977 | <u>837</u>   | <u>CASP4</u>    | caspase 4, apoptosis-related cysteine protease  | CM4-CN0045-120100-071-e07   |
| U29175 | <u>6597</u>  | <u>SMARCA4</u>  | SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 | CM4-CT0396-061299-047-b09   |
| U29656 | <u>4832</u>  | <u>NME3</u>     | non-metastatic cells 3, protein expressed in  | CM4-NN1157-041100-420-f03   |
| U33635 | <u>5754</u>  | <u>PTK7</u>     | PTK7 protein tyrosine kinase 7  | CM4-HT0869-190800-267-f03   |
| U34846 | <u>361</u>   | <u>AQP4</u>     | aquaporin 4   | CM4-C10663-150101-677-h08 ? |
| U35735 | <u>6563</u>  | <u>SLC14A1</u>  | solute carrier family 14 (urea transporter), member 1 (Kidd blood group)                          | CM4-ST0189-051099-021-g08   |
| U37408 | <u>1487</u>  | <u>CTBP1</u>    | C-terminal binding protein 1  | CM4-HT0509-240300-523-d08   |
| U38817 | <u>6827</u>  | <u>SUPT4H1</u>  | suppressor of Ty ( <i>S.cerevisiae</i> ) 4 homolog 1  | CM4-MT0361-060201-735-f11   |
| U40462 | <u>10320</u> | <u>ZNFN1A1</u>  | zinc finger protein, subfamily 1A, 1 (Ikars)  | CM4-GN0056-230800-345-b01   |

|        |       |                 |  |                           |
|--------|-------|-----------------|--|---------------------------|
| U40705 | 7013  | <u>TERF1</u>    | telomeric repeat binding factor (NIMA-interacting) 1   | CM4-ST0276-181299-064-d01 |
| U42412 | 5571  | <u>PRKAG1</u>   | protein kinase, AMP-activated, gamma 1 non-catalytic subunit   | CM4-HT0501-240300-519-b09 |
| U43148 | 5727  | <u>PTCH</u>     | patched (Drosophila) homolog   | CM4-CT0572-111100-439-a12 |
| U43368 | 7423  | <u>VEGFB</u>    | vascular endothelial growth factor B   | CM2-CT0779-190301-709-f04 |
| U43430 | 7531  | <u>YWHAE</u>    | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide  | CM4-NN1004-010400-129-f08 |
| U45879 | 329   | <u>BIRC2</u>    | baculoviral IAP repeat-containing 2  | CM2-BT0694-180300-120-f08 |
| U46691 | 6830  | <u>SUPT6H</u>   | suppressor of Ty (S.cerevisiae) 6 homolog  | CM4-NN0011-170300-112-e04 |
| U47077 | 5591  | <u>PRKDC</u>    | protein kinase, DNA-activated, catalytic polypeptide   | CM2-CT0662-020301-695-f06 |
| U47686 | 6777  | <u>STAT5B</u>   | signal transducer and activator of transcription 5B  | CM4-GN0289-120101-791-b11 |
| U47927 | 8078  | <u>USP5</u>     | ubiquitin specific protease 5 (isopeptidase T)   | CM2-ET0193-261200-653-e04 |
| U52682 | 3662  | <u>IRF4</u>     | interferon regulatory factor 4   | CM4-ST0189-051099-021-g08 |
| U52840 | 9037  | <u>SEMA5A</u>   | sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain (semanhorin) 5A | CM4-ST0181-231199-049-a02 |
| U53174 | 5883  | <u>RAD9</u>     | RAD9 (S. pombe) homolog  | CM4-MT0236-180101-802-d12 |
| U54617 | 5166  | <u>PDK4</u>     | pyruvate dehydrogenase kinase, isoenzyme 4   | CM4-FN0194-181100-443-c08 |
| U56386 | 6452  | <u>SH3BP2</u>   | SH3-domain binding protein 2   | CM2-MT0157-271100-547-c08 |
| U57650 | 3635  | <u>INPP5D</u>   | inositol polyphosphate-5-phosphatase, 145kD  | CM2-CT5001-270900-404-h11 |
| U58334 | 7159  | <u>TP53BP2</u>  | tumor protein p53-binding protein, 2   | CM4-GN0362-161200-621-d06 |
| U61262 | 4756  | <u>NEO1</u>     | neogenin (chicken) homolog 1   | CM4-NN0084-100600-195-a09 |
| U61835 | 4331  | <u>MNAT1</u>    | menage a trois 1 (CAK assembly factor)   | CM2-ST0133-240999-013-a04 |
| U63139 | 10111 | <u>RAD50</u>    | RAD50 (S. cerevisiae) homolog  | CM2-CI0179-191200-635-b01 |
| U65999 | 6455  | <u>SH3GL1</u>   | SH3-domain GRB2-like 1   | CM2-CT0782-090301-700-h11 |
| U66867 | 7329  | <u>UBE2I</u>    | ubiquitin-conjugating enzyme E2I (homologous to yeast UBC9)  | CM4-CT0376-271299-067-a06 |
| U67156 | 4217  | <u>MAP3K5</u>   | mitogen-activated protein kinase kinase kinase 5   | CM4-HN0021-241100-456-b03 |
| U67195 | 7078  | <u>TIMP3</u>    | tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)  | CM2-GN0294-020101-675-b11 |
| U69668 | 7175  | <u>TPR</u>      | translocated promoter region (to activated MET oncogene)   | CM4-IT0042-151200-605-b08 |
| U70439 | 10541 | <u>SSP29</u>    | acidic protein rich in leucines  | CM2-UT0086-270900-395-f07 |
| U74611 | 8718  | <u>TNFRSF12</u> | tumor necrosis factor receptor superfamily, member 12 (translocating chain-association membrane protein)                                   | CM2-GN0294-020101-672-d04 |
| U75330 | 4685  | <u>NCAM2</u>    | neural cell adhesion molecule 2  | CM4-BT0694-290300-133-d08 |
| U77845 | 10293 | <u>TRIP</u>     | TRAF interacting protein   | CM4-CT0662-160201-754-d12 |
| U77949 | 990   | <u>CDC6</u>     | CDC6 (cell division cycle 6, S. cerevisiae) homolog  | CM4-ST0276-181299-064-d01 |
| U81232 | 8764  | <u>TNFRSF14</u> | tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)   | CM4-HB0025-220101-697-c03 |
| U82671 | 1069  | <u>CETN2</u>    | centrin, EF-hand protein, 2  | CM4-FT0117-190700-274-f01 |
| U82828 | 472   | <u>ATM</u>      | ataxia telangiectasia mutated (includes complementation groups A, C and D)   | CM4-ST0276-181299-064-d01 |
| U82938 | 10572 | <u>SIVA</u>     | CD27-binding (Siva) protein  | CM2-NT0210-181200-625-c04 |
| U87954 | 5036  | <u>PA2G4</u>    | proliferation-associated 2G4, 38kD   | CM2-CT5003-051000-407-c10 |
| U88666 | 6733  | <u>SRPK2</u>    | SFRS protein kinase 2  | CM2-CI0179-191200-635-g02 |
| U92074 | 5890  | <u>RAD51L1</u>  | RAD51 (S. cerevisiae)-like 1   | CM2-MT0157-231100-545-g10 |
| U93236 | 4221  | <u>MEN1</u>     | multiple endocrine neoplasia I   | CM2-CI0031-051000-409-g11 |
| U95204 | 3690  | <u>ITGB3</u>    | integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)  | CM4-HT0193-061099-022-f01 |

|               |              |                |  |                           |                           |
|---------------|--------------|----------------|--|---------------------------|---------------------------|
| <u>U95299</u> | <u>4855</u>  | <u>NOTCH4</u>  | Notch (Drosophila) homolog 4   | CM2-GN0295-020101-655-d08 | CM2-GN0165-031100-494-c10 |
| <u>X03363</u> | <u>2064</u>  | <u>ERBB2</u>   | v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 (neurofibromatosis derived oncogene homolog) | CM4-GN0361-150101-648-h04 | CM4-CT0660-120201-871-b08 |
| <u>X06956</u> | <u>7277</u>  | <u>TUBA1</u>   | tubulin, alpha 1 (testis specific)   | CM4-HT0137-220999-017-a11 | CM2-HT0137-200999-010-e12 |
| <u>X07868</u> | <u>3481</u>  | <u>IGF2</u>    | insulin-like growth factor 2 (somatomedin A)   | CM2-GN0288-020101-670-b02 | CM2-CI0135-021100-490-b06 |
| <u>X52104</u> | <u>1655</u>  | <u>DDX5</u>    | DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 5 (RNA helicase, 68kD)  | CM4-ST0182-051099-021-d10 | CM4-MT0286-130201-721-h11 |
| <u>X52479</u> | <u>5578</u>  | <u>PRKCA</u>   | protein kinase C, alpha  | CM4-CT0275-040500-521-a11 |                           |
| <u>X53587</u> | <u>3691</u>  | <u>ITGB4</u>   | integrin, beta 4   | CM2-BN0303-120700-265-b05 | CM2-GN0221-291100-556-a08 |
| <u>X57025</u> | <u>3479</u>  | <u>IGF1</u>    | insulin-like growth factor 1 (somatomedin C)   | CM4-ST0189-051099-021-g08 |                           |
| <u>X59798</u> | <u>893</u>   | <u>CCND1</u>   | cyclin D1 (PRAD1: parathyroid adenomatosis 1)  | CM2-GN0166-101100-496-b07 |                           |
| <u>X66945</u> | <u>2260</u>  | <u>FGFR1</u>   | fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)                       | CM4-IT0045-080101-494-e02 | CM4-IT0045-051200-483-b06 |
| <u>X75042</u> | <u>5966</u>  | <u>REL</u>     | v-rel avian reticuloendotheliosis viral oncogene homolog   | CM4-ST0276-181299-064-d01 |                           |
| <u>X75308</u> | <u>4322</u>  | <u>MMP13</u>   | matrix metalloproteinase 13 (collagenase 3)  | CM2-HT0969-300101-501-c10 |                           |
| <u>X79568</u> | <u>26469</u> | <u>PTPN18</u>  | protein tyrosine phosphatase, non-receptor type 18 (brain-derived)   | CM4-NN1009-280300-123-g08 |                           |
| <u>X82206</u> | <u>10121</u> | <u>ACTR1A</u>  | ARP1 (actin-related protein 1, yeast) homolog A (centractin alpha)   | CM4-GN0328-300101-827-c10 | CM2-NT0170-201100-535-c03 |
| <u>X84709</u> | <u>8772</u>  | <u>FADD</u>    | Fas (TNFRSF6)-associated via death domain  | CM2-GN0283-100101-683-f10 |                           |
| <u>X84740</u> | <u>3980</u>  | <u>LIG3</u>    | ligase III, DNA, ATP-dependent   | CM4-IT0045-080101-494-f02 |                           |
| <u>X94612</u> | <u>5593</u>  | <u>PRKG2</u>   | protein kinase, cGMP-dependent, type II  | CM2-KT0031-141200-628-f05 |                           |
| <u>X94630</u> | <u>976</u>   | <u>CD97</u>    | CD97 antigen   | CM2-PT0048-060300-112-h11 | CM4-MT0289-070201-743-h07 |
| <u>X95152</u> | <u>675</u>   | <u>BRCA2</u>   | breast cancer 2, early onset   | CM4-ST0276-181299-064-c04 | CM2-HT0657-150400-156-a10 |
| <u>Y00062</u> | <u>5788</u>  | <u>PTPRC</u>   | protein tyrosine phosphatase, receptor type, C   | CM4-MT0248-190101-816-a03 |                           |
| <u>Y00093</u> | <u>3687</u>  | <u>ITGAX</u>   | integrin, alpha X (antigen CD11C (p150), alpha polypeptide)  | CM4-ST0189-051099-021-g08 |                           |
| <u>Y00815</u> | <u>5792</u>  | <u>PTPRF</u>   | protein tyrosine phosphatase, receptor type, F   | CM2-GN0288-020101-667-a07 | CM2-CT0662-020301-695-g05 |
| <u>Y07848</u> | <u>2130</u>  | <u>EWSR1</u>   | Ewing sarcoma breakpoint region 1  | CM2-CI0180-201100-539-e11 | CM2-RT0061-020101-671-e02 |
| <u>Y10256</u> | <u>9020</u>  | <u>MAP3K14</u> | mitogen-activated protein kinase kinase kinase 14  | CM2-GN0294-020101-672-h05 |                           |
| <u>Z11685</u> | <u>1656</u>  | <u>DDX6</u>    | DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 6 (RNA helicase, 54kD)  | CM4-UT0046-050900-297-b01 | CM2-CN0063-030101-662-a03 |
| <u>Z23022</u> | <u>595</u>   | <u>BCL1</u>    | B-cell CLL/lymphoma 1  | CM2-GN0166-101100-496-b07 |                           |
| <u>Z23090</u> | <u>3315</u>  | <u>HSPB1</u>   | heat shock 27kD protein 1  | CM2-RT0061-020101-671-g03 | CM4-HT0285-281099-033-a08 |

## **Apêndice C**

| Chip   | Resultados blastN   | Score | e-value | Media da razao |
|--------|---|-------|---------|----------------|
| A3.d8  | pdb 1O1M A Chain A, Deoxy Hemoglobin (A-Glyglygly-C:v1m,L29f,H58q B,D:v1m,V67w)   | 66    | 8e-10   | 31,675         |
| E4.a2  | pir S14725 casein kinase II (EC 2.7.1.-) beta chain - human   | 181   | 2e-44   | 11,535         |
| G3.a2  | ref NP_000215.1  cytokeratin 18 [Homo sapiens]  | 117   | 1e-41   | 10,84          |
| F1.d1  | sp Q16531 DDB1_HUMAN DNA damage binding protein 1 (Damage-specific DNA binding protein 1)   | 129   | 7e-29   | 10,04          |
| C1.k3  | ref NP_002256.2  karyopherin beta 1; nuclear factor p97; importin 90; importin beta-1 subunit [Homo sapiens]  | 244   | 7e-64   | 9,925          |
|        | ret NP_005007.2  poly(rC)-binding protein 2 isoform a; alpha-CP2; heterogenous nuclear ribonucleoprotein  |       |         |                |
| H1.i2  | E2[Homo sapiens]  | 83    | 3e-15   | 9,465          |
| B4.b4  | ref XP_059930.3  similar to Keratin, type II cytoskeletal 8 (Cytokeratin 8) (K8) (CK 8) [Homo sapiens]  | 76    | 1e-17   | 9,42           |
| D4.f2  | ref NP_006000.2  tubulin, alpha 3;brain-specific; hum-a-tub1; hum-a-tub2 [Homo sapiens]   | 104   | 3e-21   | 9,275          |
| D2.d1  | gb AAD00646.1  erbB3 binding protein EBP1 [Homo sapiens]  | 64    | 2e-09   | 9,085          |
|        | ret NP_002353.2  melanoma antigen, family A, 4; Melanoma-associated antigen 4; MAGE-4 antigen [Homo sapiens]  |       |         |                |
| B1.j3  |   | 174   | 9e-43   | 8,715          |
| G1.g3  | gb AAL48398.1  NADH dehydrogenase subunit 2 [Homo sapiens]  | 99    | 7e-20   | 8,625          |
|        | gb AAP36544.1  Homo sapiens tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide   |       |         |                |
| B1.c3  | sp P13010 KU86_HUMAN A1P-dependent DNA helicase II, 80 kDa subunit (Lupus Ku autoantigen protein p86) (CTC box binding factor 85 kDa subunit)(CTC85) (Nuclear factor IV) (DNA-repair protein XRCC5)               | 79    | 3e-20   | 8,275          |
| B4.i3  | sp Q13042 CC16_HUMAN Cell division cycle protein 16 homolog (CDC16Hs) (Anaphase promoting complex subunit 6) (APC6) (Cyclosome subunit 6)   | 195   | 5e-49   | 8,21           |
| A3.f3  | ref NP_057696.1  mitochondrial solute carrier protein [Homo sapiens] - HT015 protein  | 249   | 9e-68   | 7,99           |
| D3.g1  | pir G02317 transcription activator stat5A - human   | 107   | 2e-22   | 7,895          |
| E4.d2  |   | 137   | 3e-55   | 7,43           |
| B1.b4  | ref NP_002256.2  karyopherin beta 1; nuclear factor p97; importin 90; importin beta-1 subunit [Homo sapiens]  | 244   | 7e-64   | 7,305          |
| E4.c3  | gb AAG09037.1  EWS/ZSG fusion protein long B isoform [Homo sapiens]   | 62    | 3e-08   | 7,275          |
| C1.j1  | gb AAC28913.1  FBRL_HUMAN; 34 KD NUCLEOLAR SCLERODERMA ANTIGEN [Homo sapiens]   | 117   | 2e-25   | 7,255          |
|        | emb CAA40736.1  nuclear factor IV [Homo sapiens] /// gb AAH19027.1  A1P-dependant DNA helicase II   |       |         |                |
| A1.f3  | [Homo sapiens]  | 299   | 3e-80   | 7,23           |
|        | gb AAK07520.1  PNAS-18 [Homo sapiens] /// gb AAP36951.1  Homo sapiens CDC28 protein kinase regulatory subunit 1B synthetic construct]   |       |         |                |
| H1.f2  | gb AAK71522.1  moesin/anaplastic lymphoma kinase fusion protein [Homo sapiens]  | 72    | 1e-11   | 7,23           |
| E3.k2  | gb AAK17489.2  NADH dehydrogenase subunit 4 [Homo sapiens]  | 75    | 1e-12   | 6,92           |
|        | ret NP_203698.1  cell division cycle 2 protein isoform 2; cyclin-dependent kinase 1; p34 protein kinase; cell cycle controller CDC2 [Homo sapiens]  |       |         |                |
| F3.c3  | 54  | 2e-17 | 6,86    |                |
| D2.i3  | dbj BAA00016.1  F1 beta subunit [Homo sapiens]  | 75    | 4e-06   | 6,735          |
|        | ret NP_001677.2  A1P synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide [Homo sapiens]   |       |         |                |
| D4.i3  | (EC 3.6.3.14)   | 77    | 6e-13   | 6,37           |
| F1.j2  | ref NP_003899.2  eukaryotic translation initiation factor 2 beta; [Homo sapiens]  | 50    | 5e-05   | 6,365          |
| B2.b4  | ref NP_002256.2  karyopherin beta 1; nuclear factor p97; importin 90; importin beta-1 subunit [Homo sapiens]  | 242   | 4e-63   | 6,05           |
| E2.c3  | sp P15170 GSP1_HUMAN G1 to S phase transition protein 1 homolog (GTP-binding protein GST1-HS)   | 76    | 1e-19   | 6,03           |
| A3.e2  | ref NP_004387.1  DEAD (Asp-Glu-Ala-Asp) box polypeptide 5; [Homo sapiens]   | 133   | 5e-30   | 5,78           |
|        | ret NP_003470.1  protein tyrosine phosphatase type IVA, member 2 isoform 1; phosphatase of regenerating liver 2 [Homo sapiens]  |       |         |                |
| G4.h3  | 193   | 5e-48 | 5,67    |                |
| E2.a3  | ref NP_001407.1  eukaryotic translation initiation factor 4A, isoform 1 [Homo sapiens]  | 108   | 1e-22   | 5,64           |
| E2.i3  | ref NP_002256.2  karyopherin beta 1; nuclear factor p97; importin 90; importin beta-1 subunit [Homo sapiens]  | 244   | 1e-19   | 5,63           |
| A2.i2  | gb AAF26678.1  beta-actin [Rivulus marmoratus]  |       |         | 5,615          |
|        | ret NP_003470.1  protein tyrosine phosphatase type IVA, member 2 isoform 1; phosphatase of regenerating liver 2 [Homo sapiens]  |       |         |                |
| C4.a3  | 58  | 2e-07 | 5,29    |                |
|        | ret NP_003748.1  eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa; eukaryotic translation initiation factor 3, subunit 2(beta, 36kD); TGFbeta receptor-interacting protein 1[Homo sapiens]       |       |         |                |
| E3.g2  | 82  | 2e-17 | 5,28    |                |
| A4.i2  | gb AAF09482.1  E2IG3 [Homo sapiens]   | 84    | 3e-15   | 5,185          |
| D4.a3  | pir B27430 phosphoprotein phosphatase (EC 3.1.3.16) catalytic beta chain - pig (fragment) Human   | 189   | 7e-47   | 5,07           |
| E4.f2  | ref NP_005044.1  UV excision repair protein RAD23 homolog A [Homo sapiens]  | 83    | 2e-15   | 5,05           |
|        | pdb 1QGK A Chain A, Structure Of Importin Beta Bound To The Ibb Domain Of Importin Alpha /// Karyopherin beta 1 [Homo sapiens]  |       |         |                |
| G1.b4  | 145   | 7e-34 | 5,045   |                |
| H3.f2  | gb AAF09482.1  E2IG3 [Homo sapiens]   | 53    | 2e-06   | 5,015          |
|        | ret NP_004716.1  BUB3 budding uninhibited by benzimidazoles 3 homolog; mitotic checkpoint component; BUB3 (budding uninhibited by benzimidazoles 3, yeast) homolog; budding uninhibited by benomyl [Homo sapiens] |       |         |                |
| E1.c2  | 224   | 3e-57 | 4,73    |                |
| H3.h3  | gb AAH22817.2  PSMA4 protein (Proteasome alpha 4 subunit ) [Homo sapiens]   | 146   | 4e-34   | 4,49           |
|        | ret NP_003245.1  tissue inhibitor of metalloproteinase 1 precursor; Erythroid-potentiating activity (tissue inhibitor of metalloproteinases); erythroid potentiating activity [Homo sapiens]                      |       |         |                |
| C3.b10 | 142   | 2e-65 | 4,485   |                |
| B3.j2  | gb AAG35479.1  PRO0898 [Homo sapiens]   | 53    | 8e-06   | 4,425          |
| B2.k2  | sapiens]  | 102   | 2e-20   | 4,18           |
|        | ret NP_003470.1  eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa; eukaryotic translation initiation factor 3, subunit 2(beta, 36kD); TGFbeta receptor-interacting protein 1[Homo sapiens]       |       |         |                |
| F3.g2  | 46  | 8e-07 | 4,15    |                |
|        | ret NP_004387.1  DEAD (Asp-Glu-Ala-Asp) box polypeptide 5; DEAD box-5 [Homo sapiens] /// pir JC1087 RNA helicase, ATP-dependent - human   |       |         |                |
| A2.c4  | 239   | 3e-62 | 3,975   |                |

|                  |  |  |            |              |              |
|------------------|--|--|------------|--------------|--------------|
| H2.j2            | Ku70   |  | 70         | 7e-11        | 3,955        |
| G4.b3            | human  |  | 339        | 3e-92        | 3,895        |
| G2.c2            | sp Q15149 PLE1_HUMAN Plectin 1 (PLTN) (PCN) (Hemidesmosomal protein 1) (HD1)<br>emb CAA68679.1  tyrosine kinase [Homo sapiens] /// gb AAA35836.1  fibroblast growth factor receptor (FGFr)                           |  | 138        | 2e-34        | 3,7          |
| G2.f3            | transmembrane form   |  | 54         | 2e-06        | 3,62         |
| G3.e3            | gb AAP36879.1  Homo sapiens ADP-ribosylation factor 3  |  | 171        | 2e-41        | 3,48         |
| C4.k2            | pdb 1AWI A Chain A, Human Platelet Profilin Complexed With The L-Pro10 Peptide   |  | 82         | 2e-14        | 3,445        |
| C4.k3            | pir S35700 phosphoprotein phosphatase (EC 3.1.3.16) 1-gamma catalytic chain, splice form 2 - human   |  | 49         | 6e-05        | 3,42         |
| F1.a3            | ref NP_001005.1  40S ribosomal protein S10 [Homo sapiens]  |  | 97         | 2e-19        | 3,375        |
| D3.k3            | pir S35700 phosphoprotein phosphatase (EC 3.1.3.16) 1-gamma catalytic chain, splice form 2 - human   |  | 49         | 5e-05        | 3,29         |
| E4.c1            | ref NP_001216.1  caspase 4 isoform alpha precursor; apoptotic cysteine protease Mch1/TX [Homo sapiens]   |  | 117        | 1e-26        | 3,25         |
| C1.b4            | sp P07737 PRO1_HUMAN Profilin I  |  | 112        | 7e-24        | 3,205        |
| G4.h1            | gb AAP89457.1  ATP synthase F0 subunit 6 [Homo sapiens]  |  | 84         | 4e-15        | 3,07         |
| H4.d1            | sp P39189 ALU2_HUMAN Alu subfamily SB sequence contamination warning entry   |  | 87         | 2e-16        | 2,915        |
| A1.i2            | gb AAA62175.1  heat shock protein 27 [Homo sapiens]  |  | 104        | 3e-21        | 2,855        |
| C2.c2            | ref NP_009140.1  ribosomal protein L35; 60S ribosomal protein L35 [Homo sapiens]   |  | 0          | 0            | 2,535        |
| E4.i2            | ref NP_004209.1  RAB11B, member RAS oncogene family [Homo sapiens]   |  | 101        | 7e-21        | 2,525        |
| D2.k3            | pir S35700 phosphoprotein phosphatase (EC 3.1.3.16) 1-gamma catalytic chain, splice form 2 - human   |  | 49         | 6e-05        | 2,495        |
| E1.d3            | gb AAP36381.1  Homo sapiens profilin 1 [synthetic construct]   |  | 260        | 2e-68        | 2,165        |
| E1.g3            | gb AAP36381.1  Homo sapiens profilin 1 [synthetic construct]   |  | 141        | 1e-32        | 2,02         |
| <br><b>F3.a4</b> | <b>gb AAH22436.1  TPT1 protein [Homo sapiens]</b>  |  | <b>149</b> | <b>6e-63</b> | <b>0,955</b> |
| <b>C2.c4</b>     | <b>ref NP_000933.1  peptidylprolyl isomerase B (cyclophilin B) [Homo sapiens]</b>  |  | <b>140</b> | <b>3e-32</b> | <b>0,475</b> |
| <b>H1.a4</b>     | <b>gb AAD16405.1  cell cycle protein CDC20 [Homo sapiens]</b>  |  | <b>56</b>  | <b>6e-07</b> | <b>0,425</b> |
| <br><b>B4.a4</b> | <b>gb AAC50729.1  tissue inhibitor of metalloproteinases-2 [Homo sapiens] /// Prommp-2TIMP-2 Complex<br/>gb AAP36229.1  Homo sapiens matrix metalloproteinase 1 (interstitial collagenase) [synthetic construct]</b> |  | <b>100</b> | <b>4e-20</b> | <b>0,29</b>  |
| <b>C2.j1</b>     |  |  | <b>102</b> | <b>5e-25</b> | <b>0,265</b> |
| <b>H2.b4</b>     | <b>gb AAA62273.1  ORF2</b>   |  | <b>46</b>  | <b>0.001</b> | <b>0,245</b> |
| <b>F2.d1</b>     | <b>ref NP_001769.2  CD48 antigen (B-cell membrane protein) [Homo sapiens]</b>  |  | <b>106</b> | <b>5e-22</b> | <b>0,09</b>  |
| <br><b>F1.b3</b> | <b>gb AAA59664.1  MHC HLA-Bw65 chain /// emb CAC20461.1  MHC class I antigen [Homo sapiens]</b>  |  | <b>122</b> | <b>7e-27</b> | <b>0,04</b>  |
| <b>F2.b4</b>     | <b>dbj BAC01687.1  immunoglobulin kappa light chain VLJ region [Homo sapiens]</b>  |  | <b>232</b> | <b>5e-60</b> | <b>0,03</b>  |
| <b>E4.h1</b>     | <b>dbj BAC01674.1  immunoglobulin kappa light chain VLJ region [Homo sapiens]</b>  |  | <b>139</b> | <b>8e-32</b> | <b>0,02</b>  |