DEFINIÇÃO *IN SILICO* DO TRANSCRIPTOMA DE DIFERENTES TIPOS DE TUMOR

MAARTEN RUDOLPH LEERKES

Tese de doutorado apresentada à Fundação Antônio Prudente para obtenção do Grau de Doutor em Ciências.
Área de concentração: Oncologia

Orientador: Dr. Sandro José de Souza

São Paulo
2004

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A.C. Camargo

Leerkes, Maarten Rudolph

Definição *in silico* do transcriptoma de diferentes tipos de tumor / Maarten Rudolph Leerkes -- São Paulo, 2004.

87p.

Tese(doutorado)–Fundação Antônio Prudente.

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

Orientador: Sandro José de Souza.

Descritores:1. EXPRESSÃO GÊNICA. 2.CÂNCER DA MAMA. 3.CÂNCER COLORRETAL. 4. GENES DO CÂNCER. 5. TRANSCRIPTOMA.

DEDICATÓRIA

Dedico este trabalho ao meu filho, Lucas,

à minha esposa, Aline e aos meus pais.

AGRADECIMENTOS

Ao **Dr. Sandro José de Souza**, pelo apoio e pela amizade e por sempre ter estimulado o senso crítico.

Aos colegas do Laboratório de Biologia computacional, **Helena, Daniel, Italo, Fabio Brito, Fabio Passetti, Adriana, Paulão, Silvia, Ana Claudia, Pedro, André Zaiats, Elisson, Artur, Leonardo, André Leme, Rodrigo, Jorge, Natanja, Noboru, Maria, Elza e Rogério** pelo convívio e pela amizade.

Ao **Prof. Dr. Ricardo Renzo Brentani**, pelo privilégio de poder fazer um trabalho deste no Instituto Ludwig de Pesquisa sobre o Câncer.

Ao **Dr. Andrew John George Simpson**, por ter me recomendado ao Dr. Sandro José de Souza.

Ao **Dr. Humberto Torloni**, pelos ótimos conselhos.

À **Dra. Otávia Luísa Damas de Caballero**, por ter validado os genes candidatos de mama.

Ao **Dr. Fernando Augusto Soares**, pelas amostras de mama.

À **Dra. Maria Mitzi Brentani**, pela colaboração.

À **Dra. Anamaria Aranha Camargo**, e aos colegas no seu laboratório, pela ajuda na 'bancada'.

A **Ricardo Pereira de Moura e Elisângela Monteiro** pelos seqüenciamentos.

Aos **professores** do programa de pós-graduação do Hôspital do Câncer A.C. Camargo.

A **Anamaria e Márcia** da secretaria de pós-graduação, pela ajuda.

Aos funcionários da biblioteca **Suely, Rosinéia, Adriana, Francyne e Alessander.**

Aos funcionários da administração do Instituto Ludwig de Pesquisa sobre o Câncer, pela ajuda com os relatórios financeiros e outros trabalhos financeiros.

À **Teresa**, uma grande amiga que me deu um grande apoio.

A todas as outras pessoas que me apoiaram com amizade e bom humor.

À **FAPESP** pela bolsa de doutorado concedida.

RESUMO

Leerkes, MR. **Definição *in silico* do transcriptoma de diferentes tipos de tumor**. São Paulo; 2004. [Tese de Doutorado-Fundação Antônio Prudente].

Nesta tese são apresentados dois estudos *in silico* de conjuntos de dados de expressão. Ambos os estudos originaram-se de conjuntos de dados de expressão feitas de "Open Reading Frame ESTs" ou ORESTES. Estes ORESTES foram gerados no projeto genoma humano de câncer da FAPESP/LICR. A ênfase da primeira parte da tese está no estudo de expressão diferencial de genes em tumor de mama. Na segunda parte da tese a ênfase da tese está no estudo da expressão de novos genes em câncer colorretal. A combinação do uso de ORESTES e a informação disponível dos bancos de dados de UniGene e SAGE caracterizaram o transcriptoma de células normais e tumorais de mama. Neste estudo, identificamos 154 genes como candidatos de genes que são super-expressos em células tumorais de mama. Entre estes, achamos 28 genes que foram anteriormente validados por outros como sendo super-expressos em tumor de mama ou em outros tumores. Onze genes candidatos foram testados, usando RT-PCR, sendo que 9 deles realmente eram super-expressos em tumor de mama. Além disso, 99 foram validados *in silico* por dados de SAGE. Dos 55 genes não confirmados por SAGE, 42 tinham o seu 'cluster' composto somente por ESTs sem anotação funcional, ou seja, eram genes com função desconhecida. Avaliando o tamanho de 'cluster', estes 42 genes provavelmente foram expressos em baixos níveis em tecido mamário. Estes resultados levaram ao aprofundamento do estudo sobre os genes de baixa abundância

entre os quais, provavelmente, se encontra a maior parte dos genes novos. Como a maioria destes genes novos foram expressos em níveis baixos, tornou-se difícil identificá-los por metodologias de expressão gênica como 'SAGE' (Serial Analysis of Gene Expression) ou abordagens convencionais como 'ESTs' (Expressed Sequence Tags). Neste trabalho demonstramos que a metodologia ORESTES pode contribuir para a descoberta de novos genes. Em câncer colorretal observa-se um tipo específico de instabilidade genômica, caracterizado por alterações no tamanho das unidades simples de seqüência repetitiva ou microsatélites. Muitos dos transcritos de baixa abundância podem ser essenciais para determinar fenótipos celulares normais e patológicos e podem ser responsáveis pelas diferenças fundamentais poucos compreendidas entre os diferentes fenótipos de câncer colorretal. Neste trabalho, usamos 'Open Reading Frame ESTs' ou 'ORESTES' para identificar novos genes expressos especificamente em diferentes fenótipos de câncer colorretal. Realizamos uma análise *in silico* dos transcriptomas de dois tipos diferentes de câncer colorretal. Comparamos estes transcriptomas tumorais com o trancriptoma normal vindo de tecido colorretal. Identificamos transcritos únicos expressos em níveis baixos em tumores com estabilidade em microsatélites, e transcritos únicos expressos em níveis baixos em tumores com instabilidade em microsatélites. Estes transcritos poderão ser específicos para tecido colorretal e, além disso, podem determinar uma fase específica da tumorigênese. As análises aqui apresentadas poderão contribuir para o entendimento das diferenças fundamentais em características clínicas, patológicas e moleculares dos cânceres coloretais com estabilidade (MSS) e instabilidade (MSI) de microsatélites. Com a abordagem computacional apresentada, observamos que a metodologia ORESTES pode ser complementar a outras tecnologias de larga escala

de expressão gênica (SAGE, bibliotecas normalizadas de ESTs) na identificação de genes novos com importantes papéis em tumorigênese.

SUMMARY

Leerkes, MR. **Definição *in silico* do transcriptoma de diferentes tipos de tumor** [*In silico* definition of the transcriptome of different types of tumour]. São Paulo; 2004. [Tese de Doutorado-Fundação Antônio Prudente].

In this thesis, two *in silico* studies of expression datasets are presented. Both studies start with datasets of ORESTES or "Open Reading Frame Expressed Sequence Tags". These ORESTES were produced within the human cancer genome project of the FAPESP / LICR. The emphasis of the first part of the thesis is on the differential expression of genes in breast tumors. The second part of the thesis emphasizes the study of novel genes in colorectal cancer. The combination of the use of ORESTES and the publicly available information in the databases of UniGene and SAGE lead to the characterization of the transcriptome of normal and tumour breast cells. In this study, we identified 154 genes as candidate up-regulated genes in breast tumour cells. Among these, 28 genes have been shown by others to be over-expressed in breast or other tumours. Using RT-PCR, we tested 11 candidate genes and found that 9 were indeed overexpressed in breast tumour cells. Furthermore, 99 genes were validated *in silico* by SAGE data. Of the 55 genes that were not confirmed by SAGE, 42 have their corresponding cluster composed solely by ESTs. These 42 clusters have no functional annotation and the function of these genes is unknown. The transcripts of the genes that are represented by these 42 clusters are likely to be expressed at low levels in breast tissue. These results led to a more profound study of low abundance genes among which probably most of the novel

genes can be found. As the majority of novel genes are expressed at low levels, difficulties are encountered in identifying them with gene expression techniques like SAGE (Serial Analysis of Gene Expression) or EST (Expressed Sequence Tag) libraries. In this work we show the contribution of the ORESTES methodology in identifying novel genes. In colorectal cancer, a specific type of genetic instability characterized by length alterations within simple repeated sequences, termed microsatellite instability (MSI) is seen in the majority of hereditary nonpolyposis colorectal cancers (HNPCCs) and in a subset of sporadic cancers. Many of the low abundance transcripts could distinguish between different phases of tumorigenesis. The analyses presented in this work could contribute to the understanding of fundamental clinical, pathological and molecular differences between colorectal cancers with stability in microsatellites and colorectal cancers with instability in microsatellites. With the described computational approach, we observed that the ORESTES methodology could be complementary to other large scale gene expression technologies (SAGE, normalized EST libraries) in the identification of novel genes with important roles in tumorigenesis.

ÍNDICE

1 INTRODUÇÃO..... 1

1.1 Expressão Diferencial..... 4

1.2 Genes de Baixa Abundância..... 6

2 OBJETIVOS..... 9

3 ARTIGOS CIENTÍFICOS..... 10

3.1 *In silico* comparison of the transcriptome derived from purified normal breast cells and breast tumor cell lines reveals candidate up-regulated genes in breast tumor cells..... 10

3.2 An *in silico* comparison of the SAGE and ORESTES transcriptomes shows little overlap in novel tissue-specific colorectal cancer genes..... 39

4 DISCUSSÃO..... 64

5. CONCLUSÕES..... 71

6 REFERÊNCIAS BIBLIOGRÁFICAS..... 73

1 INTRODUÇÃO

Genes que exercem um papel importante em câncer afetam as funções normais de vários processos celulares como proliferação, interações célula-célula e célula-matriz, reparo de DNA, invasão e motilidade celular, angiogênese, apoptose entre outros. A genética da expressão ou o estudo da expressão dos genes é uma abordagem atualmente usada para a identificação desses genes. Classicamente, a genética do câncer tem ignorado o potencial de estudar eventos ao longo do processo de expressão gênica e tem focado nas mutações no genoma [103]. Por métodos clássicos, porém, muitos poucos genes mutados que afetam os processos celulares têm sido identificados em cânceres humanos. Apesar de várias tentativas não foi possível estabelecer uma correlação direta entre mutação e câncer [129-131].

Além disso, hoje já é bem estabelecido que em câncer existem muito mais genes alterados na expressão do que genes mutados. Genes não-mutados em tumores que possuem alterações no nível de expressão são componentes chave para os problemas da genética do câncer, não só pelas suas contribuições no entendimento das bases moleculares do câncer como também para os seus potenciais papéis no desenvolvimento de quimioterápicos. Um evento de alteração de expressão em um gene individual pode ser causado tanto por uma mutação no promotor do gene como por mudanças na regulação do mesmo.

Genes, cuja regulação foi perturbada, são um alvo de terapia tão importante quanto os genes mutados [102-103,142] e nos últimos 50 anos, uma considerável parte da pesquisa de câncer tem sido dedicada à análise de genes que são expressos

diferencialmente em células tumorais em comparação à sua contraparte normal, embora estas pesquisas tenham sido feitas em pequena escala. A modificação na expressão pode levar à um fenótipo anormal numa quantidade significativa de cânceres [61, 145]. Esta modificação na expressão pode facilitar a iniciação ou progressão de um neoplasma, como fazem os oncogenes, ou inibi-la, como fazem os genes supressores de tumor [60].

Apesar do fato de centenas de estudos isolados terem observado diferenças em expressão de somente um gene ou poucos genes, nenhum estudo amplo de expressão gênica em células cancerosas foi feito até meados dos anos 90, e não se sabe quantos genes foram expressos diferencialmente em células tumorais em comparação às células normais. Portanto, era desconhecido se as diferenças entre essas células ocorriam devido à uma reação ao micro-ambiente do tumor ou se era uma expressão diferencial autônoma da célula, podendo ser específica do tipo celular e não específica para o tumor [93].

Avanços tecnológicos na última década possibilitaram a análise simultânea dos padrões de expressão de milhares de genes [14-15,122,135]. Três exemplos de estratégias empregados para gerar dados de expressão são "Expressed Sequence Tags" (ESTs) [119], "Serial Analysis of Gene Expression" (SAGE) [136] e "Open Reading frame Expressed Sequence Tags" (ORESTES) [34]. Estas estratégias contribuiram com um grande volume de dados nos bancos públicos de seqüências e vêm contribuindo significativamente para um melhor entendimento de vários fenômenos biológicos, inclusive o aparecimento e desenvolvimento de várias doenças. Algumas iniciativas têm surgido, sendo que o objetivo principal é correlacionar características genéticas e epigenéticas de um determinado tumor com o seu desenvolvimento e progressão.

Talvez a mais conhecida destas iniciativas seja o Cancer Genome Anatomy Project (CGAP), iniciado pelo 'National Cancer Institute' nos EUA [86,101,105,123,125-126]. Este projeto visa, entre outras coisas, gerar milhões de "expressed sequence tags" (ESTs) de tecidos normal e tumoral, formando assim um catálogo genético de um determinado tecido. ESTs são seqüências curtas de "single pass" (de uma única passagem de seqüenciamento) derivadas do cDNA [2,88]. Este tipo de seqüência é responsável pelo crescimento exponencial de dados observado no Genbank [7]. Há, inclusive, um banco especializado em ESTs, chamado dbEST, hoje com 18,140,083 seqüências (05/09/2003). Dentre estas, 5,413,050 correspondem a ESTs humanas (05/09/2003). Vários centros de estudo no mundo todo têm organizado as ESTs de um mesmo organismo em "clusters". Em princípio, um cluster corresponderia a apenas um transcrito celular. A mais conhecida destas iniciativas é o banco UniGene [11] que, para a espécie humana, contém hoje 108,094 clusters (05/09/2003). Outros bancos incluem o "Human Gene Index" da TIGR [99] e o STACK do SANBI [47,84]. Esta indexação das ESTs em clusters só é possível devido ao fato de serem geradas, em sua maioria, a partir da extremidade 3' dos transcritos. Uma limitação do processo de indexação é a baixa freqüência de ESTs que atingem a região codificadora de um determinado transcrito. Esta limitação levou ao seqüenciamento da outra extremidade dos clones de cDNA, gerando o que se convencionou chamar de 5'ESTs. Uma forma alternativa de se gerar ESTs é a metodologia "ORESTES". Esta técnica baseia-se em uma reação de PCR feita em baixa estringência usando cDNA como "template" [33]. Dias-Neto et al.[34] demonstraram que seqüências geradas a partir de Orestes localizam-se primariamente no centro dos transcritos, ou seja, na região codificadora. Tal fato torna estas seqüências complementares àquelas presentes nos

bancos públicos, além de contribuirem para a completa caracterização do transcriptoma humano (o transcriptoma seria a totalidade dos genes expressos em uma determinada espécie). Dentro deste contexto, a geração de seqüências consenso para os clusters (contigs) ganha uma importância ainda maior. Acredita-se hoje que seja possível definir o transcriptoma humano somente com dados de ESTs. Adicionamos a isto o completo seqüenciamento do genoma humano, onde o papel das ESTs na definição dos genes é primordial. O sequenciamento completo do cromossomo 22 humano mostrou que programas de predição gênica como Genscan [17] apresentam uma eficiência muito baixa. É cada vez mais aceito que a forma mais confiável de se identificar um gene no DNA genômico seja através do uso de ESTs [42, 47].

A metodologia "ORESTES" foi usada no Projeto Genoma Humano do Câncer (HCGP, co-financiado pelo Instituto Ludwig de Pesquisa sobre o Câncer e a FAPESP) e gerou aproximadamente 1,200,000 ESTs de diferentes tipos de tumor e de tecido normal. O nosso laboratório, como responsável pela coordenação de Bioinformática do HCGP, vem desenvolvendo uma série de projetos e ferramentas para a análise dos dados gerados dentro do HCGP. O uso de ORESTES e ferramentas próprias de bioínformática foram responsáveis por diferenciar de outros estudos [110], que também relataram a expressão diferencial entre tecidos normais e tumorais.

1.1 EXPRESSÃO DIFERENCIAL

Uma forma de pesquisar os genes que perturbam a regulação de células cancerosas é através da medição dos padrões de expressão gênica desses genes. Neste contexto, o estudo da totalidade dos transcritos, ou o transcriptoma, de um determinado fenótipo através da biologia computacional torna-se indispensável ao entendimento da regulação da expressão gênica. A clonagem diferencial de genes [66] e a hibridização subtraente ("subtractive hybridization") [112] são metodologias bem-sucedidas e visam obter padrões de expressão diferencial ao nível do RNA. Além destas metodologias outras foram utilizadas, dentre elas citamos a exposição diferencial ("differential display" [71-73], análise de representação diferencial de cDNA ("representational difference analysis of cDNA") [49], e análise serial de expressão gênica ("serial analysis of gene expression, SAGE"), [13, 134-136].

Em adição, foram desenvolvidas modificações e melhoramentos nas técnicas conservando-se a simplicidade do procedimento original. Uma outra metodologia, como o vetor de cDNA [107] ou de oligonucleotídeos [74] ("microarray"), foi usada para comparar a expressão de milhares de genes em uma variedade de tecidos e estados patológicos. Esta metodologia foi limitada à análise de transcritos previamente identificados.

A análise serial de expressão gênica ("serial analysis of gene expression, SAGE") propiciou uma abordagem rápida e ampla para o esclarecimento de padrões de expressão gênica sem depender da disponibilidade prévia de informação de transcritos [136]. A metodologia de SAGE é baseada na criação de etiquetas ("tags") únicas de transcritos individuais e na concatenação desta etiqueta, formando uma molécula longa de DNA. O seqüenciamento rápido de clones de concatâmeros revela as etiquetas individuais e possibilita a identificação de transcritos celulares. Assim, a

técnica de SAGE possibilita a construção de um perfil amplo de expressão com cada mRNA que é representado por etiquetas de cDNA de 10 pares de bases, obtidas de uma localização precisa da extremidade 3' do mRNA. Ao contrário das metodologias de hibridização ("subtractive hybridization") [112] e a exposição diferencial ("differential display") [71-73], uma vantagem de SAGE é que ela resulta na quantificação dos níveis de expressão de cada gene. Desta forma torna-se possível fazer comparações entre um estado normal e tumoral de um determinado tecido, estabelecendo-se um painel de expressão diferencial. Um obstáculo técnico em usar a metodologia SAGE foi a obtenção de RNA de células tumorais em espécimes heterogêneos, pois contêm diversos tipos celulares. Atualmente, existem metodologias que resolvem este problema como, por exemplo, a dissecção de tecidos fixados ou congelados [54] ou classificação de células por características de superfície ("cell sorting") [44].

Nesta tese demonstramos o uso das ORESTES para facilitar o processo de encontrar genes diferencialmente expressos. Um dos tecidos 'modelo' utilizado foi a mama, onde identificamos 154 genes como candidatos de super-expressão em células tumorais de mama. Outros grupos observaram anteriormente que entre estes, 28 genes são de fato super-expressos em tumores de mama ou em outros tumores. Testamos 11 genes candidatos , usando RT-PCR, e achamos que 9 deles são de fato super-expressos em células mamários tumorais.

1.2 GENES DE BAIXA ABUNDÂNCIA

Os atuais bancos de dados de genes expressos existentes incluem praticamente todos os genes humanos abundantemente expressos, e parte dos genes de médio e baixo nível de expressão. Os genes que ainda não foram descobertos são genes expressos em baixos níveis ou são expressos especificamente em somente alguns tipos de células, estágios de desenvolvimento ou condições de crescimento. Como a maioria dos genes é expresso em baixos níveis, torna-se difícil a identificação destes genes através de tecnologias mais comumente usadas, como SAGE (Serial Analysis of Gene Expression) [136] ou abordagens de ESTs (Expressed Sequence Tags) comuns [2,88]. Ao aumentar o tamanho dos bancos de dados de expressão, a velocidade com que novos genes são descobertos, diminui, consideravelmente. Porém, um amplo estudo dos genes de baixa abundância, realizado por nós, poderá levar à descoberta de novos genes, identificando fatores regulatórios chave responsáveis por fenótipos diferenciados, progressão de desenvolvimento ou regulação de crescimento celular.

Em câncer colorretal, muitos dos transcritos de baixa abundância podem ser essenciais para determinar fenótipos celulares normais e patológicos, e podem ser responsáveis pelas diferenças fundamentais poucos compreendidas entre os diferentes fenótipos deste tipo de câncer. Em um conjunto de cânceres esporádicos e na maioria dos cânceres coloretais hereditários que não formam polipóses (HNPCC, ou Hereditary Non Polyposis Colorectal Cancer), é observado um tipo específico de instabilidade genômica, caracterizado por alterações no tamanho dos microsatélites (unidades simples de seqüência repetitiva) [1,52,59,96,128].

Nesta tese demonstramos o uso das ORESTES para facilitar o processo de busca de marcadores tumorais específicos para estadiamento de tumores. Foi

realizada uma análise *in silico* (computacional) dos transcriptomas de dois diferentes tipos específicos de câncer colorretal. Também foram realizadas comparações entre esses tipos e o transcriptoma normal de uma linhagem celular derivado de tecido colorretal. Através do uso da metodologia ORESTES, foram identificados 650 transcritos únicos de baixa abundância em tumores com estabilidade em microsatélites. Para tumores com instabilidade em microsatélites, 1,223 transcritos únicos de baixa abundância foram encontradas. No transcriptoma normal, identificamos 1,433 transcritos únicos expressos em baixa abundância. A análise destes transcritos de baixa abundância indica que eles podem ser específicos para tecido colorretal e, além disso, podem determinar uma fase da tumorigênese. Um aprofundamento nas análises aqui apresentadas poderá contribuir para o entendimento das fundamentais diferenças em características clínicas, patológicas e moleculares dos cânceres coloretais com estabilidade (MSS ou RER-) e instabilidade (MSI ou RER+) de microsatélites. Com a abordagem computacional apresentada, achamos que a metodologia ORESTES pode ser complementar a outras tecnologias de larga escala de expressão gênica (SAGE, bibliotecas normalizadas de ESTs) na identificação de genes com importantes papéis em tumorigênese.

2 OBJETIVOS

1 Demonstrar, através de métodos computacionais, uma metodologia alternativa para identificar potenciais marcadores tumorais que podem contribuír a uma melhor compreensão do processo tumorigênico.

2 Identificar, com uma abordagem computacional, genes de baixa abundância expressos em câncer colorretal com instabilidade e estabilidade de microsatélites.

3 ARTIGOS CIENTÍFICOS

3.1 *In silico* comparison of the transcriptome derived from purified normal breast cells and breast tumor cell lines reveals candidate up-regulated genes in breast tumor cells

Maarten R. Leerkes, Otavia L. Caballero, Alan Mackay, Humberto Torloni, Michael J. O'Hare, Andrew J.G. Simpson, Sandro J. de Souza

Genomics 2002; 79(2):257-65.

ABSTRACT

Genes that are differentially expressed in tumor tissues are potential diagnostic markers and drug targets. The DNA sequence information available in the public databases can be used to identify transcripts differentially expressed in cancer. We report here the combined use of the ORESTES sequences generated in the FAPESP/LICR Human Cancer Genome Project and information available in the UniGene and SAGE databases to characterize the transcriptome of normal and breast tumor cells. We have identified 154 genes as candidates for being overexpressed in breast tumor cells. Among these, 28 genes have been shown by others to be overexpressed in breast or other tumors. Using RT-PCR we tested 11 candidate genes and found that 9 were indeed overexpressed in breast tumor.

Keywords: Breast Cancer, Cancer Genes, Tumor, Gene Expression, Up-Regulation

INTRODUCTION

The huge amount of sequence data in the public databases poses a new challenge to biologists. In the post-genomic era, interpretation of the data and the development of models that account for specific biological phenomena is a priority. Within this context, a very important task is the characterization of the transcriptome of tissues in normal and pathological situations and the identification of genes that are differentially expressed. Several databases with different degrees of completeness and annotation cover most of the human transcriptome. Expressed Sequence Tags (EST) and Serial Analysis of Gene Expression (SAGE) are the two major sources of expression data in the public domain. Analysis of these cDNA data has proved to be an effective method of identifying and characterizing genes expressed in a variety of human tissues [50,132]. In a similar way, techniques have been used in combination for mining a variety of databases to define candidate genes to be differentially expressed in several pathological situations [108,110]. It is also important to mention that initiatives like CGAP [124] and SAGEmap [65] are adopting a more integrated approach for expression profiling in cancer.

Breast cancer is among the major causes of death from cancer in women [97]. In regions like United Kingdom and Switzerland, among others, breast cancer is the leading cause of death from cancer in women [20,64]. Although progress has been made towards a better understanding of breast tumorigenesis and markers have been identified (for example, estrogen receptor has been used as a predictive and prognostic marker [16]), the need to devise new approaches for identifying new

prognostic and predictive markers is clear [75]. Furthermore, since early detection is the most effective way of reducing mortality, it is crucial, again, that new markers be identified and more important, a correlation between them and several parameters, such as response to treatment and mortality, be established.

Here we describe a comparative analysis of more than 50,000 ESTs generated from immunopurified normal breast cells and from different tumor breast cell lines. The strength of the approach used here lies in the combination of the following features: a) immunomagnetic methods were used to generate highly purified populations of normal breast cells (as done before by several groups [28,43,91]); b) transcripts were sequenced with the ORESTES methodology, which produces sequences that are biased towards less abundant messages [34] and c) differential expression was evaluated by using the ratio of tumor/normal sequences in UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) as well as SAGE data available in the SAGEmap database [65] (http://www.ncbi.nlm.nih.gov/SAGE/).

RESULTS

Generation of sequences

EST sequencing was performed using the ORESTES methodology and resulted in 21,437 and 37,890 sequences from normal and tumor samples respectively. We evaluate the degree of redundancy in the normal and tumor dataset

by using the software CAP3 Clustering the 21,437 sequences from the normal mammary set resulted in 2,465 clusters with more than one sequence and 8,836 singletons. We obtained 1,097 clusters with more than one sequence and 1,691 singletons when we applied the same procedure to the 37,890 sequences from the tumor dataset.

The program BLASTN was run sequentially for each ORESTES from both the 'normal' and 'tumor' datasets against a database of known human genes retrieved from UniGene. In the dataset from the normal sample, 8,290 ORESTES sequences matched 1,084 "full-length" cDNAs. Similarly, 13,673 distinct sequences from the tumor dataset matched 1,547 "full-length" cDNAs. The remaining sequences were searched against a database containing the consensus sequence for all UniGene clusters (the nucleotide version of the trEST database [92]). For the normal dataset 6,134 distinct ORESTES matched 1,656 distinct UniGene clusters. Similarly, 11,366 distinct sequences from the tumor dataset matched 1,960 distinct UniGene clusters. Merging normal full-length sequences with normal sequences retrieved from the nucleotide-trEST database resulted in a set of 2,740 known genes represented in the normal breast cell transcriptome. In the same way, we obtained a set of 3,489 known genes represented in the tumor cell transcriptome by merging tumor full-length sequences to tumor sequences retrieved from the nucleotide-trEST database.

Defining functional categories from normal and tumor genes

We defined nine functional categories based on the annotation given to the UniGene cluster. Thus, a collection of 2,740 annotations of UniGene clusters present in the normal dataset and a collection of 3,489 annotations of UniGene clusters present in

the tumor dataset were used for the definition of the functional classes. The results of this analysis are shown in Table 1. Interestingly, the categories "protein metabolism" ($P < 0.005$), "cellular motility / cell structure" ($P < 0.005$) and "metabolism" ($P < 0.0005$) are represented by more genes in the tumor than in the normal set. In contrast, the categories "cellular communication" ($P < 0.0001$) and "unknown function" ($P < 0.0001$) seem to be represented by more genes in normal than in tumor cells.

Comparisons of tumor and normal sequences in UniGene clusters

We evaluated the degree of overlap between the normal and tumor transcriptomes and found that a set of 749 genes were present in both the normal and tumor datasets. Since one of the features of the ORESTES methodology is a normalization effect, sequences generated using this method *per se* can not be used as a quantitative evaluation of the pattern of expression. They can, however, be used as a starting point to guide subsequent analysis. We focused our attention on the 2,740 genes present exclusively in the tumor dataset to define candidates for being up-regulated in breast cancer using a series of evaluations as shown in Figure 1.

We first compared the number of sequences from tumor samples in relation to the number of sequences from normal samples for each exclusive UniGene cluster present in the tumor-restricted dataset. A total of 634 clusters (present in the tumor dataset) had a higher number of tumor sequences and a higher number of breast tumor sequences. A list of all these clusters with their respective annotation is given in our web site (http://www.ludwig.org.br/biocomp/cluster.html). We used the information contained in the SAGE database to further validate these data. For the

analysis with SAGE, we chose libraries on the basis of tissue, tumor state and size. For normal mammary tissue, the library SAGE mammary epithelium that contains 49,167 SAGE tags was chosen. For breast tumor, the libraries SAGE_SciencePark_MCF7_Control_0h (61,079 SAGE tags) and SAGE_DCIS_2 (28,888 tags) were chosen.

The SAGE libraries from breast tumor were then used as a starting point. First, for each of the 634 genes obtained in the previous analysis (candidates for being overexpressed in the tumor sample) a list of tags was generated from the tumor SAGE libraries. Then, for each of these tags, its frequency of occurrence in the SAGE library from normal breast tissue was verified. The significance of the difference in the frequency of the tag between each of the two tumor libraries and the normal library was tested by a Chi-square test or a Fischer's exact test [39,80]. Of 634 genes with more tumor derived sequences in the respective UniGene clusters, 99 were confirmed as being overexpressed in the tumor transcriptome on the basis of a significant excess of the respective tag in at least one of the tumor SAGE library. Table 2 lists all 99 genes, their annotation, the results from the SAGE tag analysis, and the data from the comparisons of tumor and normal sequences in UniGene clusters. For the remaining list of 535 genes (634-99) with an excess of tumor sequences in their respective UniGene cluster, 55 were further selected based on a more stringent criterion of more than 80 % of tumor sequences in the UniGene cluster (Table 3).

Experimental Validation

In a literature search using the 99 candidate genes listed in Table 2, we found that for 18 of them there was evidence showing an over-expression in breast tumor. For other 10, there was evidence of them being over-expressed in other types of tumor. We undertook also several semi-quantitative RT-PCR experiments. Table 04 shows that for eleven candidate genes tested, nine had a significant higher expression in at least 50% of the informative tumor samples when compared with the normal counterpart. Figure 02 shows the results of a gel eletrophoresis of some candidates. As expected, there is a heterogeneity among all tumor samples. In some samples a gene was not detected probably due to its low expression and the sensitivity of the RT-PCR.

DISCUSSION

Three different expression data resources were used in our analysis. First, more than 50,000 human ESTs generated using the ORESTES methodology from normal and tumor breast cells were collected. The whole UniGene database and three SAGE libraries (one from normal breast, one from the breast tumor cell line MCF7 and one from bulk breast tumor) were also incorporated as a means of identifying candidates to be up-regulated in breast tumor cells. We identified using this strategy

a set of 154 candidate genes to be overexpressed in breast tumor cells. For a subset of 99 genes, we were able to confirm this overexpression by an analysis of the SAGE data available in the SAGEmap database. Functional characterization of this set of 99 genes shows that most of them are related to DNA stability and the DNA repair system. This is intriguing in the light of the putative involvement of BRCA1 and BRCA2 in double-strand break repair [9]. At least two genes identified here are reported to interact with BRCA1 [10,143]. Furthermore several reports have linked deficiency of the DNA repair mechanism with the development of breast tumors [69].

The conceptual consistency that is hidden within the set of 99 genes should be a reflection of the set of 55 genes that were not confirmed by SAGE. As this set was derived using a more stringent criterion (>80% tumor sequences in the UniGene cluster) they are also candidates to be differentially expressed in breast tumor cells. In evaluating these results it is striking that 42 out of 55 genes have their UniGene cluster composed only by ESTs with no annotation. The transcripts of these genes are likely to be expressed at low levels in breast tissue. A strong indication of the abundance of a particular transcript is its cluster size in the UniGene database [18]. In assessing the size of each of these clusters for both the set of 55 genes and the set of 99 genes, we observed that the median cluster size of the set of 55 clusters is 4 sequences per cluster, whereas for the set of 99 clusters, the median cluster size is 353 sequences per cluster. It should be kept in mind that the choice of SAGE libraries can lead to deviations in composition from the initial set of sequences. This in turn could lead to false assumptions as to whether the supposed differential expression should be attributed to tumorigenesis and not cell differentiation or even differences that are related to cell culture. One should therefore bear in mind that

confirmation by SAGE was a mean to substantiate our methodology with an already well established experimental technique. Although records of tissue origin in the UniGene database report a variety of breast-tissue sources, the choice of SAGE libraries remains an issue to be carefully considered in selecting from the larger set of 634 possible candidates for being up-regulated in mammary tumor tissues. Among the set of 55 genes, there are few known genes with an annotation. It is tempting to speculate on their potential role in carcinogenesis. For example, Rhombotin 2 (a member of the Rhombotin family) is an oncogene involved in interaction with retinoblastoma binding-proteins [81]. Mutations in the DNA mismatch repair gene hMSH2 are associated with an aberrant expression of Rhombotin 2. This is quite interesting within the context of this work since we have identified overexpression of several genes related to DNA stability and repair mechanisms.

Close inspection of the literature regarding the list of 99 candidate genes validated by SAGE data revealed that for 18 of them there is evidence that they are over-expressed in breast tumors. For other 10, there is evidence of over-expression in tumors from other tissues. Literature references for these experimental studies are included in Table 2. Experiments using RT-PCR and primers specific to eleven candidate genes showed that for nine of them a significant up-regulation is observed in at least 50% of the tumor samples (Table 04 and Figure 02). Together, the literature inspection and the experimental validation show that the bioinformatics approach used here is robust and has the potential to identify tumor markers and to contribute to a better understanding of the tumorigenic process. This approach can be easily adapted and applied to any tumor type for which sufficient transcript sequences are available.

MATERIALS AND METHODS

Cells and Tissues

Normal breast luminal epithelial cells were sorted from primary cultures of mammoplasty tissue as described before [28]. Tumor sample was derived from a pool of 24 breast cancer cell lines (BT20, BT474, Cal51, Cama-1, Du4475, GI101, MCF-7, MDA134, MDA-MB-157, MDA-MB-175, MDA-MB-330, MDA-MB-361, MDA-MB-415, MDA-MB-435, MDA-MB-453, MDA-MB-468, PMC-42, SKBR-3, SKBR-5, SKBR-7, T47D, ZR75.1, ZR75.30, 734b).

For the RT-PCR experiments, breast cancer tissues, along with non-cancerous breast tissues from the same patients, were excised during surgery after obtaining informed preoperative consent from the patients. Immediately after surgery, tissue samples were frozen in liquid nitrogen and stored at -80°C until RNA extraction or OCT embedding. Seven samples were selected for this study. Each corresponding normal tissue was confirmed histopathologically to be free of cancer cells. The seven tumors were classified as ductal infiltrative carcinoma (5), adenocarcinoma (1) and granular cell myoblastoma (1).

ORESTES sequencing

Total RNA was extracted from the immunosorted cells and the pool of breast cancer cell lines. Qualified samples, with no detectable DNA, then were processed

for isolation of poly(A)+ RNA (MiniMacs; Miltenyi Biotec, Auburn, CA). To produce cDNA templates, samples of 10-100 ng of the purified mRNA were heated at 65°C for 5 min and then subjected to reverse transcription at 37°C for 60 min in the presence of 200 units of mouse murine leukemia virus reverse transcriptase and 15 pmol of a randomly selected primer in a final volume of 20 µl. After cDNA synthesis, one microliter of a 1:5 dilution of the single-stranded cDNA then was amplified by PCR by using the same or a single, alternative primer (for more details see Dias-Neto et al. [34]).

Data processing and analysis

All ORESTES sequences were processed as described [34] and submitted to GenBank. Sequences from UniGene and SAGE were retrieved from http://www.ncbi.nlm.nih.gov/UniGene and http://www.ncbi.nlm.nih.gov/SAGEmap respectively. All sequences were loaded into a relational database. Blastn searches were run under default parameters with an e-value cuttoff of 10^{-30}.

Statistical significance of the difference in the frequency of the SAGE tags was evaluated by the chi-square test or a Fischer's exact test when the number of tags in the normal sample was equal to zero. A database of known human genes was constructed by retrieving the largest cDNA from the Unigene cluster containing a known gene [18].

Semi-quantitative Reverse Transcription (RT)-PCR

Two micrograms of total RNA from normal and tumor tissue samples was reverse-transcribed for single-stranded cDNAs using oligo$(dT)_{12-18}$ primer and

Superscript II (Life Technologies, Inc.). Each cDNA mixture was diluted for subsequent PCR amplification. Primers specific for amplifying 195 bp from *B-actin* (5'CACTGTGTTGGCGTACAGGT 3'and 5'TCATCACCATTGGCAATGAG 3') were used to control for the amounts of cDNA generated from each sample. Amplification was carried out by using 1 unit of AmpliTaqGold (Applied Biosystems) in a final volume of 25 µl containing 1.5 mM $MgCl_2$, 200µM of each dNTP and 15 pmol of each primer. TaqGold was activated by incubation at 94°C for 10 min, and the reactions were cycled 25 to 31 times at 95°C for 1 min, 60°C for 1 min, and 72°C for 1 min, followed by a final extension at 72°C for 7 min. The number of PCR cycles was optimized in each case to ensure that product intensity fell within the linear phase of amplification. PCR products were visualized on 2% agarose gels stained with ethidium bromide, or alternatively on silver stained polyacrylamide gels. The samples were also analyzed by DHPLC (Transgenomic) to quantify the amount of DNA present after amplification. Each primer pair corresponding to the candidate genes resulted in amplification of a fragment of approximately 270 bp.

ACKNOWLEDGMENTS

The authors are indebted to Ricardo R. Brentani, Helena P.B. Samaia and Fabio Passetti for critically reading this manuscript. We are also indebted to Fernando Soares from Hospital AC Camargo for assistance in the macrodissection of

tumor samples. ML was supported by a doctoral fellowship from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). The authors would also like to express their gratitude to the sequencing network from the FAPESP/LICR Human Cancer Genome Project.

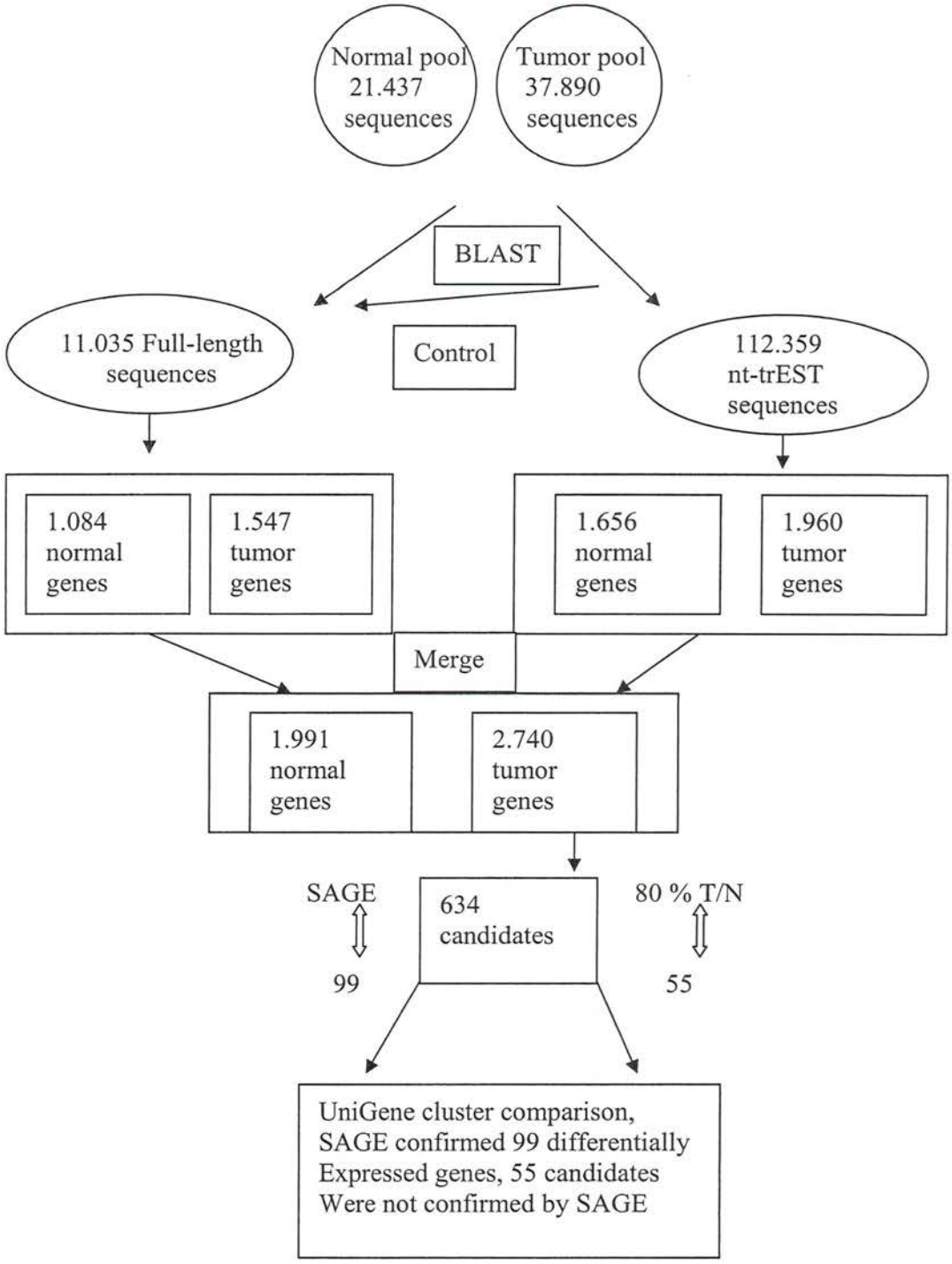

Figure 1 - Flow chart of the approach used in this work.

Table 1 - Functional categories for normal and tumor sequences

	Normal and tumor sets.			
	Normal (rel)	Normal (abs)	Tumor (rel)	Tumor (abs)
1. DNA metabolism	1.3 %	36	1.8 %	63
2. RNA metabolism	4.6 %	126	5.0 %	176
3. Protein metabolism (+)	4.8 %	131	6.0 %	208
4. Cellular motility / cell structure (+)	5.1 %	140	6.6 %	229
5. Cell cycle	2.6 %	72	2.7 %	93
6. Immune system / homeostasis	2.0 %	54	2.2 %	76
7. Cellular communication (-)	11.0 %	299	7.0 %	244
8. Metabolism (+)	7.6 %	208	9.9 %	345
9. Unknown function(-)	61.0 %	1674	58.8 %	2055
Total:	100.00 %	2740	100.00 %	3489

The categories "protein metabolism", "cellular motility / cell structure" and "metabolism" are represented by significantly more genes in the tumor than in the normal cells ($p<0.005$, $p<0.005$ and $p<0.0005$, respectively). In contrast, the category "cellular communication" is significantly less represented ($p<0.0001$) in tumor than in normal cells.

Table 2 - SAGE and UniGene data for 99 candidate genes.

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
AB018331	200|178	15|1	12|4	KIAA0788 protein
AB019568	3092|1838	40|2	7|0	eukaryotic translation EF1
AB023208	237|139	12|0	20|10	MLL septin-like fusion
AB028955	4|4	1|0	19|8	KIAA1032 protein
AF014807	95|95	4|0	14|2	phosphatidylinositol synthase, [51]
AF015186	344|234	11|2	8|2	splicing factor, arginine/serine-rich 2
AF015926	83|67	9|0	32|3	solute carrier family 9, [120]
AF035191	113|102	13|2	7|1	nuclear sperm protein
AF036613	358|351	10|6	18|2	general transcription factor II, pseudogene 1
AF037339	58|38	5|0	4|1	cleft lip and palate associated protein 1
AF038451	113|59	13|2	6|2	anterior gradient 2 homolog
AF041260	14|11	4|0	5|1	BCAS1 [31]
AF047042	249|209	17|1	10|0	citrate synthase

Table 2 - *continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
AF055008	326|216	13|1	19|6	granulin [77]
AF055584	77|43	1|0	33|16	sulfotransferase 1C [38]
AF077866	125|101	3|2	15|0	solute carrier family 7
AF089814	104|74	8|0	5|1	tumor suppressor related 1
AF187554	2321|2019	1122|283	7|1	glucose phosphate isomerase
AF229162	71|58	5|0	5|1	nuclear LIM interactor interacting factor
AF282618	48|19	2|0	6|1	serine carboxypeptidase 1
AJ007509	200|174	20|8	10|3	E1B associated protein 5
AJ011001	136|79	22|0	9|0	G protein coupled receptor 56 [21]
AK000260	132|69	10|0	9|2	clone FLJ13660 [4]
AK000393	51|45	2|1	4|1	hypothetical protein FLJ20386
AK000472	551|277	54|3	141|48	nucleophosmin
AK001296	6|6	2|0	58|36	serologically defined colon cancer antigen 3
AK001313	1666|1332	35|1	37|18	ribosomal protein, large, P0
AK022207	72|56	14|3	27|9	lone FLJ12145

Table 2 - *continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
AK022611	69|49	4|0	23|0	clone FLJ12549
AK022673	7|5	1|0	19|8	hypothetical protein FLJ10520
AK023107	76|49	4|1	8|2	clone FLJ13045
AK025105	1|1	1|0	23|9	clone FLJ21452
AK026142	48|41	5|0	19|0	hypothetical protein
AK026603	264|216	39|19	180|30	CD24 antigen [40]
AL080119	284|240	15|2	6|1	DKFZP564M2423
AL117612	82|82	10|1	51|5	DKFZp564B1264
AL137377	232|198	9|1	11|1	valosin containing protein
AL162004	296|187	17|5	15|1	KIAA1096 protein
AL162068	323|286	9|1	16|0	nucleosome assembly protein 1 like 1
AY007121	36|20	1|0	13|5	clone CDABP0014
BE314614	273|230	30|0	42|24	muscle specific gene

Table 2-*continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
BE617342	66|46	5|0	12|1	Weakly similar to KIAA0946 protein
BE732735	208|198	4|1	171|4	ubiquitin B [36]
BF032309	63|31	4|1	8|3	mitochondrial gene
BF311490	368|346	25|12	52|1	clone FLJ23538
D14697	164|106	6|0	164|75	farnesyl diphosphate synthase
D28480	231|167	8|0	17|4	minichromosome maintenance deficient (S. cerevisiae) 7
D38441	115|80	3|0	7|0	N acylaminoacyl peptide hydrolase
D83735	62|60	7|1	7|1	calponin 2
J03934	69|62	6|0	13|1	diaphorase [82]
J04102	133|115	3|0	4|1	E26 oncogene homolog 2
K00558	933|628	18|1	95|20	tubulin, alpha
L13434	49|46	1|0	16|0	unknown gene

Table 2-*continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
L36983	51|24	3|0	6|1	dynamin 2
L37033	88|76	3|0	10|0	FK506 binding protein 8 (38kD)
L38810	102|79	1|0	29|11	proteasome 26S subunit
L76191	115|92	3|0	14|5	interleukin 1 receptor associated kinase 1
M12623	451|342	23|1	193|1	high mobility protein 17 [46]
M14630	660|558	51|4	32|8	prothymosin, alpha (gene sequence 28) [76]
M25753	103|49	1|0	9|0	cyclin B1 [30]
M26326	570|251	18|0	366|31	keratin 18 [106]
M27539	631|429	131|7	89|46	major histocompatibility complex, class I, A
M29541	183|21	18|0	9|1	CEA 6 [138]
M31627	247|187	34|8	87|43	X box binding protein 1 [140]
M86737	109|98	4|0	6|2	structure specific recognition protein 1
M86849	121|119	7|1	23|12	connexin 26 [53]
NM_001614	189|158	15|5	7|0	actin, gamma 1

Table 2-*continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
NM_004690	191|136	27|0	42|24	LATS homolog 1
NM_016079	65|62	7|0	6|2	CGI149 protein
S57501	140|105	12|0	31|3	protein phosphatase 1
U09953	510|388	10|0	99|17	ribosomal protein L9
U29344	4|4	2|0	26|9	fatty acid synthase [90]
U30255	147|145	8|1	4|1	phosphogluconate dehydrogenase
U47077	55|53	3|1	8|1	protein kinase [146]
U53204	78|67	4|1	6|1	plectin 1
U66618	58|36	8|0	39|2	SWI/SNF related
U91985	55|28	1|0	34|10	DNA fragmentation factor
X02812	53|36	4|0	20|0	TGFß1
X03444	283|246	42|0	79|11	lamin A/C
X07077	67|32	2|1	65|12	proline 4 hydroxylase [83]
X16064	396|396	12|7	30|8	tumor protein
X17206	1616|560	30|0	159|55	ribosomal protein S2

Table 2 - *continued*

GenBank Acc. No.	Proportion tumor/normal In cluster	Proportion breast_tumor| breast_normal In cluster	Proportion tumor|normal SAGE tags	Gene Description (References)
X52022	145|133	9|1	32|3	collagen, type VI, alpha 3 [79]
X53416	17|13	1|0	14|1	filamin A, alpha
X53777	516|456	20|1	118|58	ribosomal protein L17
X56465	246|236	29|4	43|9	zinc finger protein 6 [41]
X64707	866|814	22|0	18|5	ribosomal protein L13
X73460	1095|336	23|0	215|138	ribosomal protein L3
X74801	362|189	12|0	5|1	chaperonin containing TCP1 [23]
X77588	51|49	3|0	7|2	N acetyltransferase
X80026	88|33	11|0	10|1	Auberger b antigen [111]
X87838	33|28	7|3	9|3	catenin, beta 1 [109]
X94910	89|60	18|0	4|1	ER lumenal protein [98]
Y07569	114|65	4|0	36|0	acidic protein rich in leucine [56]
Y13936	49|29	4|2	8|3	protein phosphatase 1G [58]
Z21507	317|197	31|0	15|4	eukaryotic EF1 delta
Z23090	374|334	18|0	103|18	heat shock 27kD
Z48950	432|397	15|8	11|4	H3.3B
Z82022	39|38	4|0	9|0	dolichyl phosphate

Numbers after 'Gene Description' correspond to the reference showing evidence of overexpression of the respective gene in tumors.

Table 3 - 55 Candidates genes (not confirmed by SAGE) with more than 80 % of tumor sequences in UniGene cluster.

UniGene Cluster	UniGene cluster size	UniGene Annotation
Hs.103156	16	ESTs
Hs.111314	12	ESTs
Hs.1149	15	LIM domain only 1 (rhombotin 1) [81]
Hs.124629	18	ESTs
Hs.125258	17	Homo sapiens cDNA FLJ13795 fis, clone THYRO1000107
Hs.128898	6	ESTs
Hs.147647	9	ESTs
Hs.152677	15	Homo sapiens cDNA FLJ20338 fis, clone HEP12179
Hs.153444	11	ESTs
Hs.159089	8	ESTs
Hs.159471	362	ZAP3 protein
Hs.165839	2	ESTs
Hs.167119	2	ESTs
Hs.175322	2	ESTs
Hs.175355	2	ESTs
Hs.182362	11	ESTs
Hs.190160	5	ESTs
Hs.190225	3	ESTs
Hs.190488	162	ESTs

Table 3 - *continued*

UniGene Cluster	UniGene cluster size	UniGene Annotation
Hs.197074	4	ESTs
Hs.198694	7	ESTs
Hs.200242	8	ESTs
Hs.200857	3	EST
Hs.201218	4	ESTs
Hs.201947	3	ESTs
Hs.202331	3	ESTs
Hs.204600	2	ESTs
Hs.205420	3	ESTs
Hs.207092	3	ESTs
Hs.208690	6	ESTs
Hs.211462	2	ESTs
Hs.222190	6	ESTs, Weakly similar to secretory carrier membrane protein.
Hs.22572	157	KIAA0580 protein
Hs.232174	3	ESTs
Hs.232284	3	ESTs
Hs.232473	2	ESTs
Hs.232838	15	ESTs
Hs.241196	3	ESTs

Table 3 - *continued*

UniGene Cluster	UniGene cluster size	UniGene Annotation
Hs.244201	11	ESTs, Weakly similar to putative ankyrin-repeat containing protein.
Hs.252414	2	ESTs
Hs.254784	5	ESTs, Weakly similar to KIAA1276 protein.
Hs.258016	2	ESTs
Hs.267222	68	ESTs
Hs.270289	3	ESTs
Hs.278629	3	ESTs
Hs.278709	1	EST, Weakly similar to X-linked retinopathy protein.
Hs.278710	1	EST
Hs.279064	3	ESTs
Hs.282013	2	ESTs
Hs.282050	11	ESTs
Hs.282345	3	ESTs
Hs.28338	160	KIAA1546 protein
Hs.283642	2	ESTs
Hs.290131	8	ESTs
Hs.47334	84	ESTs

	AF282618	AB018331	D83735	M31627	AK001313	AL137377	AJ007509	X53416	U47077	X74801	J04102
GCMN	1,32	60,33	5,67	13,50	798,50	ND	ND	ND	ND	102,79	ND
GCMT	12,45	43,93	15,80	25,20	704,50	ND	ND	ND	8,70	52,46	ND
Ductal1N	ND	14,38	9,36	NT	NT	ND	ND	NT	NT	NT	NT
Ductal1T	4,51	45,61	70,10	NT	NT	ND	ND	NT	NT	NT	NT
Ductal2N	ND	2,68	ND	NT	NT	ND	ND	NT	NT	NT	NT
Ductal2T	112,19	27,04	26,60	NT	NT	ND	ND	NT	NT	NT	NT
AdenoN	NT	NT	NT	180,96	563,70	ND	ND	ND	NT	NT	NT
AdenoT	NT	NT	NT	304,56	514,90	ND	ND	ND	NT	NT	NT
Ductal3N	ND	ND	ND	3,85	19,21	ND	ND	ND	ND	ND	ND
Ductal3T	ND	ND	ND	209,00	24,88	ND	ND	ND	ND	ND	ND
Ductal4N	18,36	9,10	64,93	20,10	494,98	12,48	42,28	6,94	1,07	51,75	ND
Ductal4T	14,87	42,55	74,79	128,00	482,53	12,12	ND	23,58	9,76	47,00	7,66
Ductal5N	48,65	112,96	180,03	423,28	818,54	2,46	12,03	13,00	16,65	93,50	14,17
Ductal5T	93,68	176,72	190,01	569,98	620,52	6,84	31,04	60,45	48,66	110,50	14,51

"GCM = Granular Cell Myoblastoma; NT = Not Tested; ND = Not Detected."
*The table lists the amount of PCR product in ng as measured by DHPLC analysis that result from the amplification of approximately 60 ng of poly(A+) RNA following 31 cycles.

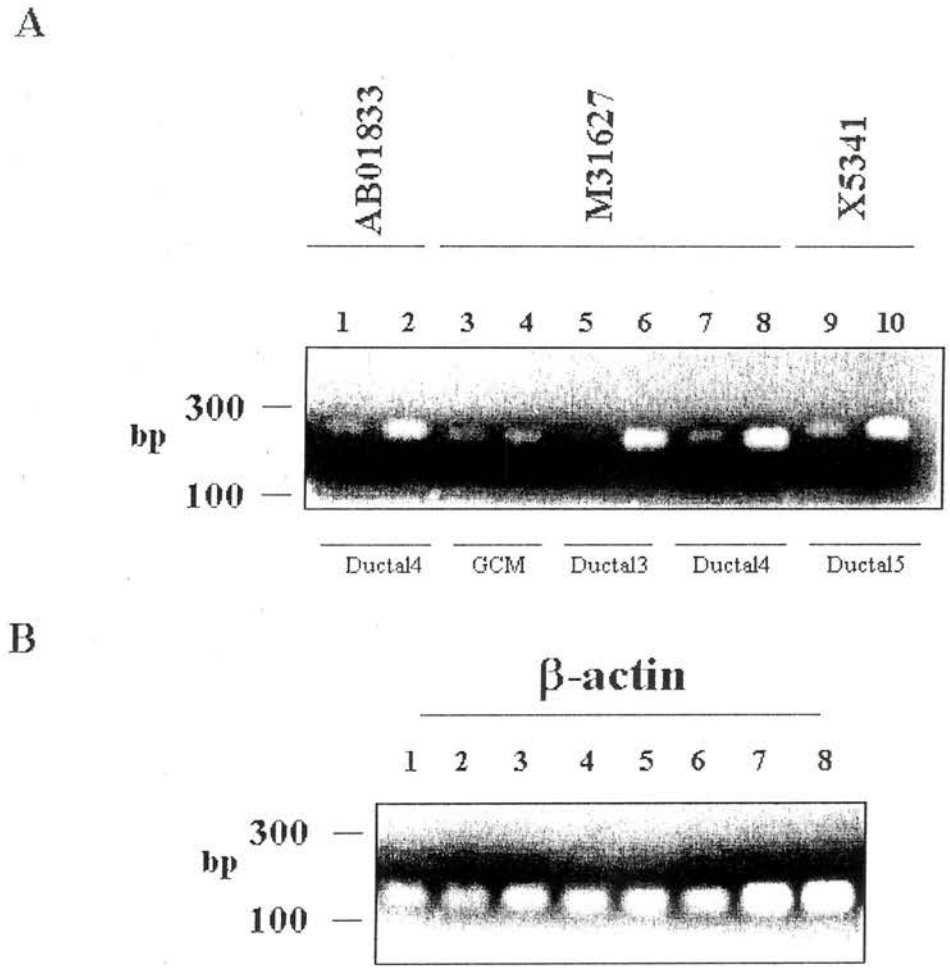

Figure 02 - (A) Electrophoresis of three candidate genes illustrating the RT-PCR analysis shown in Table 04 (B) B-actin was used as an internal control for normalization. Even numbers correspond to tumor samples (preceding odd numbers correspond to the respective normal samples)

3.2 An *in silico* study of ORESTES transcriptomes of colorectal cancer.

Leerkes MR[1], Bodmer WF[2], Simpson AJ[3], de Souza SJ[1#]

1 Ludwig Institute for Cancer Research, São Paulo Branch, São Paulo, Brazil.

2 Imperial Cancer Research Fund Cancer and Immunogenetics Laboratory and Wellcome Trust Centre for Human Genetics, University of Oxford, Headington OX3 9DS, United Kingdom.

3 Ludwig Institute for Cancer Research, 605 Third Avenue, New York, New York 10158, USA.

To whom correspondence should be addressed: Rua Prof. Antonio Prudente 109, 4º andar, 01509-010, São Paulo, Brazil

phone: +55-11-3207-4922

fax: +55-11-3207-7001

e-mail: sandro@compbio.ludwig.org.br

ABSTRACT

Although the rate of identification of novel genes from the human genome has dropped substantially in recent years, many genes have not yet been identified. Because most of these genes are expressed at low levels, they are difficult to identify by the most commonly used gene expression technologies like SAGE (Serial Analyses of Gene Expression) or conventional EST (Expressed Sequence Tag) approaches. In colorectal cancers, a specific type of genetic instability characterized by length alterations within simple repeated sequences, termed microsatellite instability (MSI), is seen in the majority of hereditary nonpolyposis colorectal cancers (HNPCCs) and in a subset of sporadic cancers. Many of the lower level transcripts may be essential for determining normal and pathological cell phenotypes, and may account for fundamental poorly understood differences between different colorectal cancer phenotypes. In this work we have used Open Reading Frame EST (ORESTES) data to address this issue. We performed an *in silico* analysis of the transcriptomes of two different types of colorectal cancer and compared them with the transcriptome of normal colorectal tissue. We found that the ORESTES methodology identified 650 unique low abundance transcripts in microsatellite stable tumors, and 1,223 unique low abundance transcripts in microsatellite unstable tumors. In the normal colon transcriptome, 1,433 unique transcripts were identified. These transcripts are likely to be tissue and tumorigenesis specific. In depth computational analyses are presented that may contribute to the understanding of the fundamental differences in clinical, pathologic, and molecular characteristics of

colorectal cancers with microsatellite instability (MSI or RER+) and microsatellite-stable (MSS or RER-) cancers. With the presented computational approach we found that the ORESTES methodology can be complementary to other high-throughput gene expression technologies in the identification of novel tissue-specific genes with important roles in tumorigenesis.

INTRODUCTION

Complex biological processes such as growth and differentiation can be characterized by patterns of differential gene expression. All cells in the human body contain the same genetic information and the selective expression of several groups of genes at various points in time leads to the development of differentiated cell types. These specific cell types are shaped during the dynamics of cell proliferation, cell specialization, cell-cell interactions and cell movements. As such, essential cellular processes are orchestrated through regulatory events that eventually lead to tissue specific gene expression patterns. In disease, the physiology of a cell is deranged through the differential expression of the genome as a result of disrupted transcriptional pathways [127]. In order to gain a more thorough understanding of these regulatory events that lead to tissue-specific and disease-specific gene expression, it would be desirable to provide an effective means of defining the expression pattern of all genes in terms of abundance levels for genes in a specific given cell type or pathological phenotype.

The currently most frequently used methodologies aimed at identifying transcribed genes or characterizing transcriptional profiles are Serial Analysis of Gene Expression (SAGE) [133,136] construction of cDNA libraries and micro-array technology. In the utilization of micro-arrays to study gene expression patterns, prior knowledge is required of the sequences of transcribed genes, making it unsuitable for gene discovery and the studying of expression patterns of novel genes. SAGE is a particularly powerful approach both to evaluate overall gene expression patterns as

well as to identify new transcripts of the human transcriptome [8,25-27,85,139]. Indeed, roughly half of the SAGE tags collected from currently available datasets do not match known expressed sequences [25,145], suggesting they may have originated from novel transcripts and represent novel genes. The other half of the SAGE tags that matches GenBank entries followed an expected distribution of number of genes expressed from different classes of abundance levels, as predicted by statistical models of gene expression distribution in eukaryotic cells [62]. Of these transcripts that match GenBank entries, twenty percent represent characterized mRNA sequences and eighty percent correspond to expressed sequences that are as yet uncharacterized [145]. As SAGE tag to gene assignment relies heavily on information contained in publicly avalaible transcript databases, these SAGE tags remain uncharacterized.

Tecnnologies widely used for normalization and subtraction [12,94,116] and strategies that combine normalization and subtraction in a single procedure [22] have also proven effective in representing low abundance mRNAs, but encounter the same problems as SAGE in relying heavily on publicly available transcript information. With a novel technique called massively parallel signature sequencing (MPSS) [14-15] the colon adenocarcinoma transcriptome of a cultured cell line (HCT116) has been characterized [55]. Using MPSS, Jongeneel et al.[55]. found that a colon adenocarcinoma cell line expresses around 15,000 genes, although uncertainty remains concerning the part of the transcriptome that covers the poorly expressed genes. This is due to the lack of knowledge regarding the number of signature tags that actually document novel transcripts, and it has not been possible to be more precise at this point. Furthermore, the cost and complexity of the technique make it unsuitable for high-throughput analysis.

Despite enormous efforts to identify the genes in the human genome, a considerable number of novel transcripts in human cells is still being identified [19,68,100,113,118], illustrating the ill-determined coverage of publicly available transcript databases [121]. Furthermore, as novel genes predicted by different groups are still largely non-overlapping [48], the notion remains that an increase in representation of low abundance transcripts will eventually lead to a more complete coverage of the human transcriptome.

Here, we present the use of a modified EST strategy in identifying novel genes *in silico*. The strategy is termed ORESTES (Open Reading Frame ESTs) and has an equally effective representation of both highly and poorly expressed genes [18,33]. The essential difference of this alternative EST strategy lies in the generation of short cDNA templates using arbitrarily selected, non-degenerate primers under low stringency conditions [34]. As a consequence, the ORESTES strategy compensates partially for transcript abundance [34]. Based on conservative estimations by Camargo et al.[18], a set of 700,000 ORESTES corresponded to a transcriptome coverage where more than 50% of the weakly and rarely expressed sequences were represented. An explanation for this can come from observations that genes in the low and rare expression classes are more likely to be tissue and stage specific than highly expressed genes [26,87].

These rare and low abundance transcripts may be essential for a more thorough understanding of pathological cell phenotypes, which is illustrated by the following. Of the few known cases in which small changes in low levels of expression have a profound effect on a disease phenotype, the adenomatous polyposis coli tumor suppressor gene (APC) is a striking example [141]. The low levels

of APC expression have been observed by studies of the relative expression level of the APC gene product in human colorectal cancer cell lines and cell lines of noncolorectal tissue origins [35,115]. These observations put forth even further the fundamental role that low abundance genes may play in the tumorigenesis of colorectal cancer.

A pathological phenotype of particular interest is the replication error or microsatellite mutator phenotype in subtypes of colorectal cancer. In this phenotype, a specific form of genetic instability is characterized by length alterations within simple repeated sequences, also termed microsatellite instability (MSI or RER+). MSI is observed in the majority of hereditary nonpolyposis colorectal cancers (HNPCCs), and in a subset of sporadic cancers [6,78,95]. The study of aberrant expression patterns in these phenotypes can bring to bear original new insights into the problem of microsatellite instability and its consequences in aberrant expression patterns in colorectal cancers

Here, we demonstrated by a computational approach that of the genes encoding the colorectal cancer ORESTES sequences that map exclusively to the genome and not to known transcripts, between 91 % and 94 % should represent genes exclusively expressed in typical colorectal ORESTES transcriptomes. As genes that pertain predominantly to the low and rare expression classes, they are likely to be tissue and tumorigenesis stage specific.

MATERIALS & METHODS

ORESTES generation: Template Preparation and DNA Sequencing.

Cell-lines were obtained from the Immunogenetics Laboratory of Dr. W.F. Bodmer and harvested at a defined number of generations with a defined confluence. Then, they were frozen in liquid nitrogen immediately after resection. They were allowed to partially thaw to 20°C and microdissected to enrich for tumor cells in the sample.

Total RNA was extracted with Trizol, and RNA degradation was evaluated by means of a Northern Blot by using a GAPDH cDNA probe. Those samples with intact mRNA were treated with DNaseI (10 units/50 µg of total RNA), and the absence of contaminating genomic DNA was confirmed by PCR using primers for the mitochondrial D loop and for the p53 gene. The amplified product was blotted onto nylon membranes and hybridized with [alpha-^{32}P]dCTP-labeled probes for the corresponding amplified sequences. Qualified samples, with no detectable DNA, were then processed for isolation of poly(A)$^+$ RNA (MiniMacs; Miltenyi Biotec, Auburn, CA). To produce cDNA templates, samples of 10-100 ng of the purified mRNA were heated at 65°C for 5 min and then subjected to reverse transcription at 37°C for 60 min in the presence of 200 units of mouse murine leukemia virus reverse transcriptase and 15 pmol of a randomly selected primer in a final volume of 20 µl. The criteria for primer selection were GC content of more than 50% and length of 18-25 nt.

No specific sequence constraints were imposed. Indeed, almost exclusively, the primers used originally had been designed for specific PCR amplification of DNA sequences in nonhuman genomes and were exploited here if they obeyed the simple criteria listed above. After cDNA synthesis, one microliter of a 1:5 dilution of the single-stranded cDNA then was amplified by PCR by using the same or a single, alternative primer. Amplification profiles were generated by using the following cycling parameters: an initial cycle of 95°C for 5 min, 37°C for 2 min, and 72°C for 2 min followed by 35 cycles of 95°C for 45 sec, 45°C for 1 min, and 72°C for 90 sec. Three microliters of each pool was checked for complexity on 8% silver-stained polyacrylamide gels. Product pools with a single, predominant product (~1%) reflecting the amplification of a highly abundant gene were not processed further. The remaining amplification pools with multiple bands then were cloned into pUC18 by using the Sureclone kit (Amersham Pharmacia). Minipreps for sequencing the inserts were prepared by alkali lysis or boiling preparations and sequenced by using the Perkin-Elmer Big-Dye reagent kit with ABI377 sequencers. In general, 50-200 sequences were determined from each amplification profile.

Genomic alignment of ORESTES using BLAST.

Specific alignment criteria were used. When aligning ORESTES to the genome, an E-value of $E = 10^{-30}$ was used.

Construction of in-house mapped transcript database: mysql relational database "map4_bh1".

Genome mapping of cDNAs – masking of genomic contigs provided by NCBI was used. MEGABLAST [144] was used to align pairs of genomic and transcribed sequences. Only pairs that aligned over at least 45% of total sequence length and with exons presenting more than 93% identity were considered. The extension bh1 (best hit 1) refers to the subtraction from the mapped data-set of 98,820 sequences [117] considered to be contamination, along with the stringent alignment criteria.

cDNA clustering – cDNA clusters were generated on the basis of the coordinates of cDNA alignments to human genomic sequences. Two sequences were clustered together if they presented at least one exon presenting a common exon/intron boundary (allowing ± 5 bp of difference). Sequences that did not present introns had to overlap at least 100 bp of another sequence in the cluster to be grouped together. This resulted in a locally constructed transcript database called map4_bh1 consisting of 3,475,517 transcripts grouped together in 318,275 distinct clusters.

RESULTS

Selection of transcripts that are possible candidates for novel genes.

BLAST alignment [3] of the three different colon transcriptomes was done following the flow-chart as depicted in figure 1. The different steps in the procedure serve two purposes: (i) selection of transcripts that are not yet represented and characterized in the public databases of expressed sequences, and (ii) test the validity of our prediction by aligning the candidates to the genome and grouping them together in clusters. The starting point of these analyses were three colon transcriptomes (Table 1). The program BLASTN was run sequentially for each ORESTES from the normal set, the mismatch repair proficient tumor set and the mismatch repair deficient tumor ORESTES transcriptome. First, all ORESTES sequences were searched against a full-length database. Then, the remaining sequences were searched against a transcript contig database. Finally, the remaining ORESTES that showed no matches against neither the full-length dataset nor the contigs dataset, served as a starting point to define candidates for either colon restricted or low abundance transcripts. In table 1, an idea is given of the size of the initial datasets.

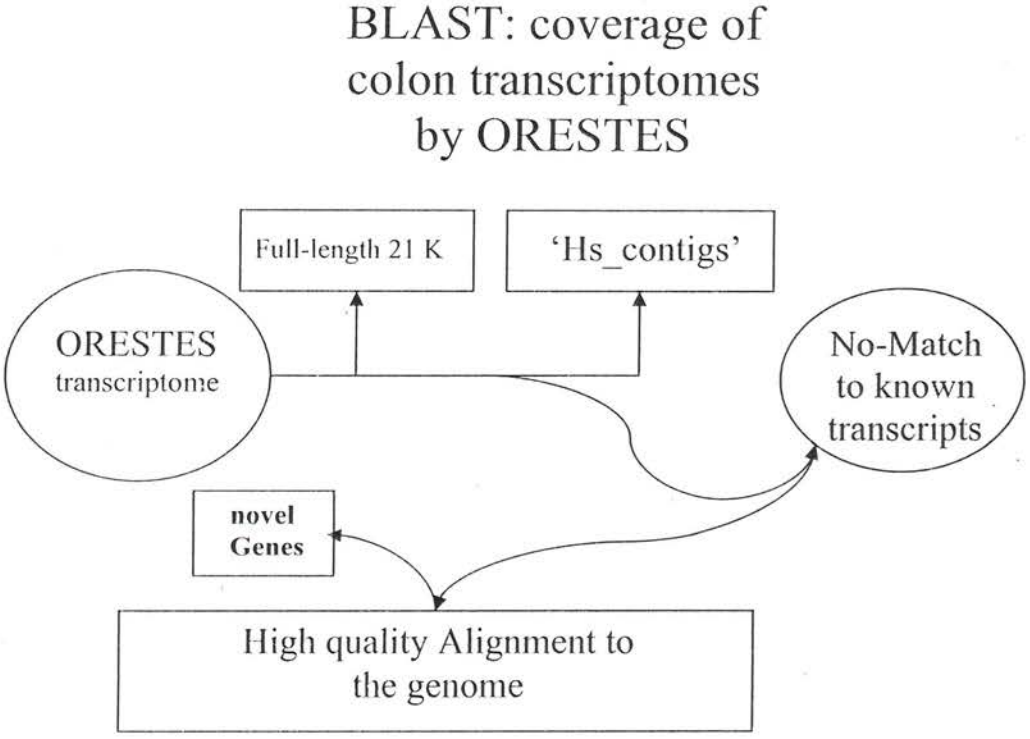

Figure 1 - Flow-chart of the procedure followed in this work.

Table 1 - Number of ORESTES sequences and libraries in the three transcriptomes.

Colon Transcriptomes		N°. of sequences	N°. of libraries
Normal		21,509	131
Tumor	Microsatellite stable	12,549	45
	Microsatellite unstable	21,509	163

Evaluation of contamination

First, an in-house constructed pipeline for cleaning contamination (mitochondrial, ribosomal, vector (PUC18) and bacterial DNA, see Table 2 below) was used. The sequences from the colon transcriptomes having a high stringent alignment with contaminating sequences were discarded from further analysis. Then, a library-level analysis to screen large sets of transcript data was used according to Sorek e Safer [117] in order to evaluate contamination with genomic DNA, pre-mRNA and non-canonical introns (see materials and methods: map4_bh1). Here, information was used from clustering and assembly of an entire EST dataset to analyze each EST library. A dataset of 52 libraries containing 98,872 sequences was used as a reference set for contamination. In order to measure the effect of different types of contamination, each type of contamination was documented before and after the alignment to the genome, while the actual genomic alignment was done without the exclusion of the contaminated portion of sequences.

Table 2 - Relative contribution of contamination by mitochondrial, ribosomal, vector (PUC18) and bacterial DNA.

Colon Transcriptomes	mitochondrial	Ribosomal	Vector (PUC18)	bacterial
Normal	13 %	7 %	0 %	0.1 %
Microsatellite stable tumor	14 %	10 %	0 %	0.02 %
Microsatellite unstable tumor	12 %	10 %	0 %	0.1 %

Alignment against a full-length database

A full-length database of 21,000 full-length transcripts was used in this step of the analysis. Due to the overlap between the normal colon transcriptome dataset with the full-length dataset, the BLAST alignment resulted in 8,698 ORESTES that have a defined sequence identity with 3,561 distinct full-length sequences. Similarly, for the microsatellite unstable transcriptome, the alignment resulted in 9,507 ORESTES sequences that corresponded to 3,575 distinct full-length sequences. In the same way, for the microsatellite stable transcriptome, the alignment resulted in 5,675 ORESTES sequences that corresponded to 2,147 distinct full-length sequences (Table 3).

Table 3 - Number of hits against full-length database.

Colon Transcriptomes		Nº. of ORESTES sequences	Nº. of distinct full-length sequences
Normal		8,698	3,561
Tumor	Microsatellite stable	5.675	2,147
	Microsatellite unstable	9,507	3,575

Alignment against a transcript contig database

The remaining sequences that did not match any full-length transcript were searched against a database containing the consensus sequence for all UniGene clusters [92] using the nucleotide version of the trEST database (downloaded version 12/09/02, containing 153,882 sequences) (Table 4). For the normal dataset, 3,302 distinct ORESTES matched 1,141 distinct 'UniGene cluster contigs'. Similarly, 2,335 distinct sequences from the microsatellite stable tumor data set matched 656 distinct 'UniGene cluster contigs'.In the same way, 3,769 distinct sequences from the microsatellite unstable tumor data set matched 1,119 distinct 'UniGene cluster contigs'.

Table 4 - Number of hits against a database containing the consensus sequences for all UniGene clusters.

Colon Transcriptomes		No. of ORESTES sequences	No. of distinct UniGene cluster contigs
Normal		3,302	1,141
Tumor	Microsatellite stable	2,335	656
	Microsatellite unstable	3,769	1,119

High quality alignment to the total human genome sequence.

Of the 9,509 normal ORESTES sequences, 5,283 showed a high quality (E = 10^{-30}) match to the total human genome sequence (version august 2002). Of the 4,539 mismatch repair proficient tumor ORESTES sequences, 2,593 showed a high quality match to the total human genome sequence. Finally, of the 8,233 mismatch repair deficient tumor ORESTES sequences, 4,212 showed a high quality match to the total human genome sequence. As a control of the BLAST alignment stringency, the number of transcripts still ocurring in UniGene clusters was determined by searching in a database downloaded from the ncbi collections of contributed molecular biology data (ftp://ftp.ncbi.nih.gov/repository/UniGene/), that was constructed into an in-house mysql relational database. Of the 5,283 normal ORESTES sequences, 5,092 sequences were not found in any UniGene cluster. Of the 2,593 mismatch repair proficient tumor ORESTES sequences, 2,517 sequences were not found in any UniGene cluster. Of the 4,212 mismatch repair deficient tumor ORESTES sequences, 4,132 sequences were not found in any UniGene cluster (Table 5). The

sequences not occurring in any UniGene cluster were analysed further in the following procedures.

Table 5 - Number of hits against the human genome.

Colon Transcriptomes		Nº of ORESTES-genome alignments	Nº of ORESTES not in UniGene clusters
Normal		5,283	5,092
Tumor	Microsatellite stable	2,593	2,517
	Microsatellite unstable	4,212	4,132

At this point, an evaluation of overlap in datasets was made between a colon longSAGE [104] library and the ORESTES sequences. Of the long SAGE dataset only one 20 bp tag matched to the same genomic coordinates of one of our ORESTES alignments. The low depth of transcript coverage of the longSAGE library contributes to this extremely low number of overlap between the longSAGE and ORESTES transcriptomes.

Evaluation of contamination.

Contamination of our datasets was evaluated according to Sorek and Safer [117]. The dataset of 52 libraries containing 98,872 was used as a reference set for contamination. None of the three colon ORESTES transcriptomes were made up of

libraries that corresponded to any of the 52 libraries composed by Sorek and Safer [117].

As many sequences will be singletons (clusters containing only a single transcript), an evaluation was made of the contribution of singletons to the total dataset (Table 6).

Table 6 - Number of singletons in the dataset for each colon transcriptome.

Colon Transcriptomes		Nº of singletons (%)
Normal		1,526 (29%)
Tumor	Microsatellite stable	613 (24 %)
	Microsatellite unstable	1,204 (29 %)

Next, an analysis was made of the expression classes to which each transcript could be attributed.

Classification in expression classes.

In order to obtain an initial idea of expression level, an "in house transcript mapping database" ('map4_bh1') as described in materials and methods was used in the following analyses. Camargo et al. [18] have used a full-length transcript database with expression level classification based on cluster size (UniGene) before, and the same rationale was applied in using our in-house transcript mapping database. Using our 'map4_bh1' database in this phase of the procedure serves two purposes: (i) applying more stringent criteria in selecting for candidates, as in this database only transcripts that aligned over at least 45% of total sequence length and with exons presenting more than 93% identity were considered, and (ii) assessing the relative

contribution of the remaining sequences to each of the different abundance classses. To this end, four different abundance classes were devised based on cluster size using our in-house transcript mapping database. The classes are termed rare, poor, moderate and high and were defined in the following way.

Class	Cluster size (number of transcripts in cluster)
Rare	$1 \leq X < 10$
Poor	$10 \leq X < 20$
Moderate	$20 \leq X < 100$
High	$X \geq 100$

Table 7 - Relative contribution of transcripts to each of the different abundance classses.

Class	Number of representative clusters in each class		
	Normal	Rer+	Rer-
Rare	1,617	1,388	659
Poor	30	32	13
Moderate	58	40	15
High	97	42	31
Total	1,802	1,502	718

Normal	=	Normal colon transcriptome
Rer -	=	'Replication error negative' transcriptome, originated from microsatellite stable, DNA repair proficient colorectal cancer cell-line
Rer +	=	'Replication error positive' transcriptome, originated from microsatellite unstable, DNA repair deficient colorectal cancer cell-line

Table 8 - Relative contribution of singletons to the rare expression class.

Class	Number of singletons in relation to the total number of clusters in the rare expression class		
	Normal	Rer+	Rer-
Rare	1,030 (63 %)	910 (65 %)	378 (57 %)

The clusters considered in this part of the analysis are mutually exclusive, meaning that clusters represented in one of the three transcriptomes were not found in either one of the other two transcriptomes, suggesting that they were specifically expressed in that particular phenotype. In figure 2, a venn diagram shows the partial overlap in clusters between each of the three transcriptomes, showing little overlap. All three transcriptomes only have two clusters in common. In assessing the normal and microsatellite unstable (msi or rer+) transcriptome, twelve clusters appeared common to both. Observing the normal and microsatellite stable (mss or rer-) transcriptome, thirteen clusters appeared common to both. Then, when analysing the microsatellite unstable (msi or rer+) transcriptome and microsatellite stable (mss or

rer-) transcriptome, nine clusters appeared common to both. When considering the clusters common to the different transcriptomes, they appeared to contain clusters pertaining to the moderate and high abundance classes, whereas the mutually exclusive classes contained predominantly clusters pertaining to the rare and poor classes.

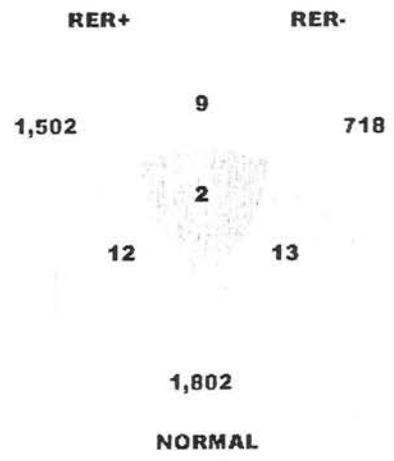

Fig. 2 Overlap between the three different colon transcriptomes.

DISCUSSION

To date, there is no consensus regarding transcriptome size yet [45,62,121], which is a problem related to the gene number problem. Since the publication of the human genome [63,137] the number of genes in the human genome has remained a controversial issue [29,32,37,57,70]. Problems relating to uncertainty in prediction of gene numbers and transcriptome size will only be resolved by integrating computational predictions with human curation and experimental validation [48,67,114]. Hence, in taking a major barrier in the issue of redundancy of gene expression, the *in silico* identification through the analysis of the expressed transcripts from the ORESTES technology can be complementary to other gene discovery strategies [25-27]. As most genes in a human cell are rarely expressed, the analysis of gene expression profiles remains incomplete. Ideally, complete gene expression profiles would be depicted by a list containing all expressed genes. However, the total number of genes expressed in human cells is still unknown, and there is still no consensus regarding the contribution of rarer transcripts to the total coverage of the human transcriptome. The determination of the biologically significant low abundance transcripts in human cells remains uncharted territory. In our analysis we used three expression data resources prepared from colorectal cell-lines. The ORESTES methodology was used to generate 21,509 normal sequences, 12,549 microsatellite stable tumor sequences and 21,509 microsatellite unstable tumor sequences. To determine the proportion of sequences that correspond to known transcripts, we used both a full-length database and a database containing the consensus sequence for all UniGene clusters. As a

result, 56% of the total number of normal ORESTES sequences were represented by 4,702 distinct known transcripts of the full-length and UniGene contig database. Similarly, 64% of the microsatellite stable tumor transcriptome corresponded to 2,803 known transcripts. In the same way, 62% of the microsatellite unstable tumor transcriptome were represented by 4,694 known transcripts. In summary, little more than half of all three transcriptomes is covered by the well annotated public transcript databases, as was expected considering conservative estimates made before by Camargo et al.[18] that indicate that ORESTES sample equally among high and low abundance genes. The ORESTES sequences that map exclusively to the genome and not to known transcripts were divided in four different abundance classes, rare, poor, moderate and high. Of these remaining sequences between 91% and 94% were classified in the rare and low classes. These clusters serve as the candidates for low abundance transcripts that are likely to be tissue and tumorigenesis specific in colorectal cancer. Their very low expression levels make a statistical evaluation of *in silico* differential expression cumbersome, since roughly half of these clusters are made up of singletons. Nevertheless, as our analysis shows, the high numbers of clusters in the mutually exclusive sets (fig 2.) suggest that the nature of the transcriptional control could be specific to the pathological cell phenotype. Indeed, benchmark validations could show that our candidates are good candidates for the different expression classes.

The high number of singletons that was observed in our analyses could well be explained by the following. Predictions based on a considerably large human transcriptome data set indicate that roughly 70% of all human genes are expressed with concentrations of less than one transcript in the cytoplasm [24,135]. In this context,

the high number of singletons in our analysis could well be explained. Here, the concept of tissue specific expression should be considered in the context of stochastic gene expression where each gene of the genome is considered to be expressed at all times, albeit in low levels. Transcription in eukaryotic cells is thought to occur with pulses of messenger RNA produced in a probabilistic manner [24,87], a phenomenon that might well explain the high number of singletons in our candidate data set. The former together with varying abundance in mismatch repair deficient colorectal cancer for comparison of different colorectal cancer phenotypes gives more insight into the problem of genes with low levels of expression and tissue-specific expression.

The candidate lists of novel tissue-specific genes can serve as a reference for the designing of primers in amplification reactions with real time PCR. Based on the aligned sequences, primers can be designed that span a maximum length of the ORESTES sequence from the candidate lists. This way, comparisons can be made between the expression levels of the candidate list and expression levels of predicted differentially expressed genes in mismatch repair deficient and proficient tumor cells, leading to a better understanding of expression levels of these novel tissue-specific genes.

Following a confirmation of the candidates of low expression by real time PCR, a selection of these candidates can be made to validate expression in other tissues in order to provide further evidence that these novel transcripts are indeed restricted to colorectal tissue. Eventually, these validated candidates can be put to use in a directed strategy for transcript finishing as done by Camargo et al.[18]. They have tested the transcript finishing approach before in full-length cDNA sequence

determination by "closing gaps" for human transcripts partially represented by ORESTES sequences.

In using our three-step approach for tissue-specific gene discovery, the advantages are many: (i) when using ORESTES, we analyse the roughly 75% part of the transcriptome that is biased towards the less abundant messages. Increasing the depth of transcript coverage generated by the ORESTES can thus greatly enhance to an increase in coverage of the human transcriptome through the rare and low abundance messages (ii) in filtering from the analysis only the unknown part of the human transcriptome and using transcriptomes of specific phenotypes, it's possible to discover tissue-specific candidate genes of great interest in oncology, especially in the field of colorectal cancer.

In summary, the strength of our approach lies in the combination of computational approaches with low cost benchmark ORESTES generation of tissue and disease phenotype specific transcript sequences. Combining this with experimental validation through real time PCR makes the procedure a cost effective and attractive alternative to highly sophisticated but more expensive technologies such as MPSS [14,15,55] for gene discovery.

4 DISCUSSÃO

Primeiramente, no transcriptoma mamário, três diferentes fontes de dados de expressão foram usados na nossa análise. Primeiramente, mais de 50,000 ESTs humanos gerados com a metodologia ORESTES foram usados. Estas seqüências são originadas de tecido mamário normal e tumoral. O banco de dados inteiro do UniGene e 4 bibliotecas de SAGE (dois de tecido mamário normal, um da linhagem celular MCF7 oriundo de tecido mamário tumoral e um de uma massa de tumor de mama) foram então usados como uma forma de identificar candidatos super-expressos em tumor de mama. Usando esta estratégia, identificamos um conjunto de 154 genes candidatos que podem estar super-expressos em tumor de mama. Destes 154 genes, 99 foram confirmados pelos dados de SAGE como sendo super-expressos em tumor de mama, ao analisarmos os dados disponíveis no banco de dados de SAGEmap. Uma caracterização funcional destes 99 genes mostrou que grande parte é relacionada à estabilidade de DNA e reparo de DNA. Estes achados são intrigantes, levando em consideração as implicações de BRCA1 e BRCA2 no processo de reparo de DNA [9]. No mínimo 2 genes que foram identificados nesta análise parecem interagir com BRCA1 [10,143]. Além destes achados, vários estudos relatam que a instabilidade de reparo de DNA está relacionada ao desenvolvimento de tumores de mama [69].

Uma avaliação da função de cada um dos 99 genes mostrou uma coerência conceptual que pode refletir-se no conjunto de 55 genes que não foram confirmados por SAGE. Como este conjunto de 55 genes foi obtido a partir de um critério mais

rigoroso (> 80 % seqüências tumorais no "cluster" de UniGene), estes genes também são candidatos a serem diferencialmente expressos em tumor de mama. Avaliando a composição dos "clusters" do UniGene, observamos que 42 dos 55 genes candidatos são compostos somente por ESTs sem uma anotação funcional. Os transcritos destes genes foram possivelmente expressos em níveis baixos em tecido mamário, por terem os seus correspondentes 'clusters' compostos por poucos ESTs. Uma forte indicação para a baixa abundância de um transcrito em particular é o tamanho do seu "cluster" nos bancos de dados de UniGene. Ao avaliar o tamanho de cada um dos "clusters" em ambos os conjuntos dos 55 e 99 candidatos, uma clara distinção pode ser feita. O tamanho mediano dos "clusters" no conjunto de 55 genes é 4 seqüências por "cluster". Para o conjunto de 99 "clusters", o tamanho mediano é 353 seqüências por "cluster". A diferença significativa em tamanho de "cluster" foi uma forte indicação da diferença em abundância dos transcritos numa célula. Deve-se levar em consideração, porém, que a escolha das bibliotecas de SAGE pode levar a desvios do conjunto inicial de genes candidatos a serem diferencialmente expressos em tumor de mama. Esta desvantagem no uso de bancos de dados de SAGE pode levar a suposições errôneas se atribuirmos a expressão diferencial ao processo de tumorigênese e não por exemplo a diferenciação celular. Importante neste contexto é frisar que SAGE foi usado com o objetivo de confirmar através de uma técnica bem estabelecida os genes candidatos propostos, e não com o objetivo de propor genes candidatos.

No conjunto de 55 genes, poucos genes tinham uma anotação funcional. Se levarmos isto em consideração, será interessante cogitar o papel deles em carcinogênese. Por exemplo, Rhombotin 2 (um membro da família dos Rhombotin) é

um oncogene que está envolvido na interação com os 'retinoblastoma binding-proteins' [81]. Mutações no gene de reparo tipo 'mismatch' hMSH2 são associadas com uma expressão aberrante de Rhombotin 2. Este achado é interessante no contexto do nosso trabalho, pois identificamos a super-expressão de vários genes relacionados à estabilidade de DNA e mecanismos de reparo de DNA.

Estudando mais detalhadamente os bancos de dados da literatura do "National Center for Biotechnology Information NCBI" [89] sobre a lista dos 99 candidatos que foram confirmados por SAGE, foi possível demonstrar que para 18 dos 99 genes existem evidências de que eles são super-expressos em tumor de mama. Para mais 10 dos 99 genes, existem evidências de super-expressão em tumores de outros tecidos. Este alto número de genes validados em estudos de bancada mostra a coerência da abordagem de bioinformática usada aqui. O método tem o potencial de identificar marcadores de tumor e pode contribuir para a compreensão do processo de tumorigênese. A abordagem já foi adaptada e aplicada a um outro tipo de tumor para qual um número suficiente de seqüências está disponível, no caso o tecido de próstata.

No estudo do transcriptoma de mama, foi observado que uma considerável parte do transcriptoma mamário não foi coberta pelas transcriptomas de SAGE. Esta questão de cobertura levantou o interesse em estudar a parte desconhecida do transcriptoma humano através da metodologia ORESTES.

Apesar de existirem vários estudos que relataram abordagens em grande escala do transcriptoma de tecido mamário, a nossa abordagem tem várias novas características:

a) a bioinformática possibilitou o uso integrado de dados de diferentes fontes;

b) o uso de celulas mamárias luminais normais altamente purificadas e o uso de linhagens celulares oriundas de tecido epitelial de tumor de mama evita os problemas que surgem usando amostras de massas tumorais;

c) a metodologia ORESTES permite o estudo de genes raros no transcriptoma.

As vantagens em usar a metodologia ORESTES para a geração de ESTs são diversas:

a) a metodologia tende a selecionar os transcritos menos abundantes que correspondem a ~75% de todos os genes transcritos numa célula;

b) as seqüências são geradas a partir da porção central do transcrito e portanto têm uma maior probabilidade de alinhar com a região codificadora, desta maneira as seqüências são apropriadas para buscas de domínios proteícos;

c) apenas pequenas quantidades de RNA são necessárias para gerar as seqüências permitindo o uso de amostras de micro-dissecção.

Como um número grande de transcritos humanos novos ainda não foi identificado, a aplicação de estratégias novas que aumentam significativamente a taxa de identificação de novos genes se tornou uma necessidade. Apesar do enorme esforço em identificar transcritos, muitos genes expressos ainda não foram identificados e a taxa de identificação de novos genes diminuiu dramaticamente de 10.6 % de seqüências de EST em 1996 [42] a somente 2.7 % de seqüências de EST em 1998 [32].

Uma explicação para a estagnação da taxa de descoberta de genes raros é que genes expressos em baixos níveis têm uma menor probabilidade de serem identificados que genes expressos em altos níveis. Ao aplicar normalização com o objetivo de diminuir a redundância em transcriptomas, a maioria dos genes poderia ser identificado. Neste estudo, descrevemos a identifição destes novos genes de baixa abundância, comparando as coberturas dos transcriptomas de ORESTES com conjuntos da totalidade de bancos de dados de transcritos. Esta comparação foi incentivado, em parte, pela falta de informação dos dados de SAGE para genes de baixa abundância em bancos de dados públicos. Velculescu et al.[135] estimaram que aproximadamente 650,000 etiquetas deveriam ser seqüênciadas para identificar todos os transcritos presentes com uma única cópia na célula do ser humano. Como os dados de SAGE dependem fortemente de seqüências depositadas no GenBank, etiquetas desconhecidas somente poderão ser identificados por experimentos adicionais [25-27]. Ao contrário das ESTs convencionais, ORESTES são normalizadas por transcritos raros [18,34]. Uma proporção considerável destes transcritos raros que não são representados nas bibliotecas de expressão de SAGE, podem ser representados pelos de sets de ORESTES de candidatos de genes raros. Estes candidatos são candidatos em tumor de cólon com estabilidade e instabilidade de microsatélites que alinharam com alta qualidade no genoma

Como a maioria dos genes humanos são expressos em baixa abundância, a análise dos perfis de expressão gênica continua incompleto em todos os tecidos. Portanto, nos transcriptomas de cólon, o foco do estudo está nos genes raros do transcriptoma humano. Idealmente, os perfis de expressão gênica deveriam ser listadas conforme a sua classe de abundância, para todos os genes humanos. No

entanto não se sabe ao certo o número total de genes expressos em células humanas, e ainda não existe um consenso em relação à contribuição do número de transcritos raros à cobertura do transcriptoma humano nos diferentes tecidos. A determinação dos transcritos de baixa abundância que têm um papel biológico significativo em tecidos específicos continua sendo uma área pouco explorada. Na nossa análise nós usamos três fontes de dados relativos à expressão gênica, preparados dos transcritos de linhagens celulares de origem de tecido colorretal. A metodologia 'ORESTES' foi usada para gerar 21,509 seqüências normais, 12,549 seqüências tumorais do fenótipo com estabilidade em microsatélites, e 21,509 seqüências tumorais do fenótipo com instabilidade em microsatélites. Para determinar a parte destes transcriptomas que corresponde à transcritos bem caracterizadas, nós usamos um banco de dados de transcritos de comprimento completo, e um banco de dados contendo a seqüência consenso ('contig') de todos os 'clusters' de UniGene. Do número total de 'ORESTES' do transcriptoma normal colorretal, 56% das seqüências estavam representadas por 4,702 transcritos distintos de 'full-lengths' ou 'contigs'. De modo similar, 64% do transcriptoma tumoral com estabilidade em microsatélites corresponde às 2,803 transcritos conhecidos, e 62% do transcriptoma tumoral com instabilidade em microsatélites corresponde às 4,694 transcritos conhecidos. Portanto, um pouco mais da metade de todos os três transcriptomas é coberto pelos bancos de dados públicos bem anotados. Isto era esperado, considerando que 'ORESTES' representam igualmente os genes de alta e baixa abundância.

As seqüências que não alinharam nos transcritos conhecidos, mas que alinharam exclusivamente no genoma foram divididas em quatro diferentes classes de abundância: raro, baixo, moderado e alto. Dos 'clusters' distribuidos em classes

de abundância diferentes, entre 91% e 94% foram classificados como 'raro' e 'baixo', conforme sua classificação. Estes 'clusters' servem como candidatos para transcritos de baixo abundância que são especificamente expressos nos diferentes fenótipos do câncer colorretal em uma determinada fase da tumorigênese.

Porém, como aproximadamente metade destes 'clusters raros' é feito de 'clusters' de uma seqüência só ('singletons'), isso dificulta uma rígida avaliação estatística da expressão diferencial. No entanto, como as nossas análises mostram, o alto número de 'clusters' nos conjuntos mutuamente exclusivos, sugere-se que a natureza do controle transcripcional pode ser específica para o fenótipo patológico da célula. Portanto, a presença de um 'singleton' que sofre 'splicing' num fenótipo de câncer colorretal, junto com a ausência em ambos os outros dois transcriptomas (normal e o outro fenótipo de tumor, instabilidade ou estabilidade de microsatélites), poderá ser uma forte indicação para a sua especifidade para tecido e / ou doença.

Validações experimentais utilizando, por exemplo, 'real time PCR' podem demonstrar se a baixa quantidade relativa de transcritos, corresponde à realidade clínica.

Ao usar a nossa abordagem computacional, as vantagens para descoberta de genes especificamente expressos em tecido de cólon são as seguintes: (i) usando 'ORESTES', uma considerável parte do transcriptoma humano, correspondente as mensagens de baixa abundância é analisado, (ii) se filtrarmos somente a parte desconhecida do transcriptoma humano e usarmos transcriptomas de fenótipos específicos, será possivel descobrir candidatos a genes de grande interesse na área oncológica, especialmente na área de tumores do cólon.

5 CONCLUSÕES

No primeiro objetivo, demonstramos através de métodos computacionais a importância do estudo de expressão gênica diferencial. Também demonstramos como as aberrações em expressão afetam os fenótipos, em particular em câncer de mama.

Para atingir completamente o segundo objetivo formulado, ainda deverão ser feito validações experimentais por 'real-time PCR' de expressão de genes raros em diferentes tipos de tumores de cólon. Comparações foram feitas do transcriptoma de tumor de cólon com instabilidade em microsatélites com a sua contraparte normal, e do transcriptoma de tumor de cólon com estabilidade em microsatélites com a sua contraparte normal, e os resultados dessas análises têm um grande interesse pelo seu valor no estadiamento dos tumores coloretais na clínica. A biologia computacional permite abordar vários outros problemas neste contexto como, por exemplo, a freqüência e distribuição no transcriptoma de microsatélites em tumores de cólon. Os resultados dos estudos e achados deste projeto podem ser utilizados por outros grupos para estudos com cólon. Também podem ser feitos estudos de 'proteômica' a partir do transcriptoma de cólon. Uma outra área promissora do estudo do câncer é o desenvolvimento de anticorpos monoclonais, onde os candidatos podem servir como ponto de partida. Para o transcriptoma de cólon podem ainda ser feitos experimentos voltados a ocorrência de microsatélites em diferentes regiões codificadores dos transcritos, tanto quanto inserção de seqüências repetitivas no genoma. Um aspecto interessante a ser estudado mais profundamente é alteração de promoteres e outras

regiões reguladoras por microsatélites ou por outras seqüências repetitivas, possivelmente atavés da aplicação de "Hidden Markov Models" [5] em programas computacionais.

6 REFERÊNCIAS BIBLIOGRÁFICAS

1. Aaltonen LA, Peltomaki P, Leach FS, et al. Clues to the pathogenesis of familial colorectal cancer. **Science** 1993; 260:812-6.

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

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. **J Mol Biol** 1990; 215:403-10.

4. Altucci L, Addeo R, Cicatiello L, et al. 17beta-Estradiol induces cyclin D1 gene transcription, p36D1-p34cdk4 complex activation and p105Rb phosphorylation during mitogenic stimulation of G(1)-arrested human breast cancer cells. **Oncogene** 1996; 12: 2315-24.

5. Baldi PF, Brunak S. **Bioinformatics, the machine learning approach.** 2nd ed. Cambridge Massachusetts: The MIT press; 1998. Hidden Markov Models: The Theory; p.143-62.

6. Benatti P, Roncucci L, Ganazzi D, et al. Clinical and biologic heterogeneity of hereditary nonpolyposis colorectal cancer. **Int J Cancer** 2001; 95:323-8.

7. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. **Nucleic Acids Res** 2000; 28:15-8.

8. van den Berg A, van der Leij J, Poppema S. Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. **Nucleic Acids Res** 1999; 27:e17.

9. Bhattacharyya A, Ear US, Koller BH, Weichselbaum RR, Bishop DK. The breast cancer susceptibility gene BRCA1 is required for subnuclear assembly of Rad51 and survival following treatment with the DNA cross-linking agent cisplatin. **J Biol Chem** 2000; 275:23899-903.

10. Bochar DA, Wang L, Beniya H, et al. BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. **Cell** 2000; 102:257-65.

11. Boguski MS, Schuler GD. ESTablishing a human transcript map. **Nat Genet** 1995; 10:369-71.

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

13. Boon K, Osorio EC, Greenhut SF, et al. An anatomy of normal and malignant gene expression. **Proc Natl Acad Sci U S A** 2002; 99:11287-92.

14. Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. **Nat Biotechnol** 2000; 18:630-4.

15. Brenner S, Williams SR, Vermaas EH et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. **Proc Natl Acad Sci U S A,** 2000; 97:1665-70.

16. Bundred NJ. Prognostic and predictive factors in breast cancer. **Cancer Treat Rev** 2001; 27:137-42.

17. Burge CB, Karlin S. Finding the genes in genomic DNA. **Curr Opin Struct Biol** 1998; 8:346-54.

18. Camargo AA, Samaia HP, Dias-Neto E, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci U S A** 2001; 98:12103-8.

19. Camargo AA, de Souza SJ, Brentani RR, Simpson AJ. Human gene discovery through experimental definition of transcribed regions of the human genome. **Curr Opin Chem Biol** 2002; 6:13-6.

20. Cancer Research Campaign. **Breast Cancer UK**. Factsheet 6.1 London Cancer Research Campaign 1996.

21. Carmeci C, Thompson DA, Ring HZ, Francke U, Weigel RJ. Identification of a gene (GPR30) with homology to the G-protein-coupled receptor superfamily associated with estrogen receptor expression in breast cancer. **Genomics** 1997; 45:607-17.

22. Carninci P, Shibata Y, Hayatsu N, et al. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. **Genome Res** 2000; 10:1617-30.

23. Charpentier AH, Bednarek AK, Daniel RL, et al. Effects of estrogen on global gene expression: identification of novel targets of estrogen action. **Cancer Res** 2000; 60:5977-83.

24. Chelly J, Concordet JP, Kaplan JC, Kahn A. Illegitimate transcription: transcription of any gene in any cell type. **Proc Natl Acad Sci U S A** 1989; 86:2617-21.

25. Chen JJ, Rowley JD, Wang SM. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. **Proc Natl Acad Sci U S A** 2000; 97:349-53.

26. Chen J, Lee S, Zhou G, Wang SM. High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. **Genes Chromosomes Cancer** 2002; 33:252-61.

27. Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. **Proc Natl Acad Sci U S A** 2002; 99:12257-62.

28. Clarke C, Titley J, Davies S, O'Hare MJ. An immunomagnetic separation method using superparamagnetic (MACS) beads for large-scale purification of human mammary luminal and myoepithelial cells. **Epithelial Cell Biol** 1994; 3:38-46.

29. Claverie JM. Gene number. What if there are only 30,000 human genes? **Science** 2001; 291:1255-7.

30. Collechi P, Santoni T, Gnesi E, et al. Cyclins of phases G1, S and G2/M are overexpressed in aneuploid mammary carcinomas. **Cytometry** 2000; 42:254-60.

31. Correa RG, de Carvalho AF, Pinheiro NA, Simpson AF, de Souza SJ. NABC1 (BCAS1): alternative splicing and downregulation in colorectal tumors. **Genomics** 2000; 65:299-302.

32. Das M, Burge CB, Park E, Colinas J, Pelletier J. Assessment of the total number of human transcription units. **Genomics** 2001; 77:71-8.

33. Dias Neto E, Harrop R, Correa-Oliveira R, Wilson RA, Pena SD, Simpson AJ. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of ESTs from nancgram quantities of mRNA. **Gene** 1997; 186:135-42.

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

35. Dihlmann S, Amler LC, Schwab M, Wenzel A. Variations in the expression of the adenomatous polyposis coli (APC) tumor suppressor gene in human cancer cell lines of different tissue origin. **Oncol Res** 1997; 9:119-27.

36. Ettenberg SA, Rubinstein YR, Banerjee P, Nau MM, Keane MM, Lipkowitz S. Cbl-b inhibits EGF-receptor-induced apoptosis by enhancing ubiquitination and degradation of activated receptors. **Mol Cell Biol Res Commun** 1999; 2:111-8.

37. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. **Nat Genet** 2000; 25:232-4.

38. Falany JL, Falany CN. Expression of cytosolic sulfotransferases in normal mammary epithelial cells and breast cancer cell lines. **Cancer Res** 1996; 56:1551-5.

39. Fisher LD, van Belle G. **Biostatistics: a methodology for the health sciences**. New York: John Wiley & Sons; 1993. Fisher's Exact Test; p. 185-87.

40. Fogel M, Friederichs J, Zeller Y, et al. CD24 is a marker for human breast carcinoma. **Cancer Lett** 1999; 143:87-94.

41. Foster KW, Frost AR, McKie-Bell P, et al. Increase of GKLF messenger RNA and protein expression during progression of breast cancer. **Cancer Res** 2000; 60:6488-95.

42. Gerhold D, Caskey CT. It's the genes! EST access to human genome content. **Bioessays** 1996; 18:973-81.

43. Gomm JJ, Browne PJ, Coope RC, Liu QY, Buluwela L, Coombes RC. Isolation of pure populations of epithelial and myoepithelial cells from the normal human mammary gland using immunomagnetic separation with Dynabeads. **Anal Biochem** 1995; 226:91-9.

44. O'Hare MJ, Ormerod MG, Monaghan P, Lane EB, Gusterson BA. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. **Differentiation** 1991; 46:209-21.

45. Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M. A question of size: the eukaryotic proteome and the problems in defining it. **Nucleic Acids Res** 2002; 30:1083-90.

46. He Q, Liang CH, Lippard SJ. Steroid hormones induce HMG1 overexpression and sensitize breast cancer cells to cisplatin and carboplatin. **Proc Natl Acad Sci U S A** 2000; 97:5768-72.

47. Hide W, Burke J, Christoffels A, Miller R. **Genome informatics.** Tokyo: University Academy Press; 1997. A novel approach towards a comprehensive consensus representation of the expressed human genome; p.187-96.

48. Hogenesch JB, Ching KA, Batalov S, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. **Cell** 2001; 106:413-5.

49. Hubank M, Schatz DG. Identifying differences in mRNA expression by representational difference analysis of cDNA. **Nucleic Acids Res** 1994; 22:5640-8.

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

51. Imoto M, Taniguchi Y, Fujiwara H, Umezawa K. Enhancement of CDP-DG:inositol transferase activity in src-and erbB2-transformed cells. **Exp Cell Res** 1994; 212:151-4.

52. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. **Nature** 1993; 363:558-61.

53. Jamieson S, Going JJ, D'Arcy R, George WD. Expression of gap junction proteins connexin 26 and connexin 43 in normal human breast and in breast tumours. **J Pathol** 1998; 184:37-43.

54. Jensen RA, Page DL, Holt JT. Identification of genes expressed in premalignant breast disease by microscopy-directed cloning. **Proc Natl Acad Sci U S A** 1994; 91:9257-61.

55. Jongeneel CV, Iseli C, Stevenson BJ, et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. **Proc Natl Acad Sci U S A,** 2003; 100:4702-5.

56. Kadkol SS, Brody JR, Epstein JI, Kuhajda FP, Pasternack GR. Novel nuclear phosphoprotein pp32 is highly expressed in intermediate- and high-grade prostate cancer. **Prostate** 1998; 34:231-7.

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

58. Kikuchi K, Kitamura K, Kakinoki Y, et al. Gene expressions and activities of protein phosphatases 1 alpha, 2A and 2C in hepatocarcinogenesis and regeneration after partial hepatectomy. **Cancer Detect Prev** 1997; 21:36-43.

59. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. **Cell** 1996; 87:159-70.

60. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. **Nature** 1997; 386:761,763.

61. Kinzler KW, Vogelstein B. Landscaping the cancer terrain. **Science** 1998; 280:1036-7.

62. Kuznetsov VA, Knott GD, Bonner RF. General statistics of stochastic process of gene expression in eukaryotic cells. **Genetics** 2002;161:1321-32.

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

64. Landis SH, Murray T, Bolden S, Wingo PA. Cancer Statistics, 1998. **CA Cancer J Clin** 1998; 48:6-29.

65. Lash AE, Tolstoshev CM, Wagner L, et al. SAGEMap: a public gene expression resource. **Genome Res** 2000; 10:1051-60.

66. Lau LF, Nathans D. Identification of a set of genes expressed during the G0/G1 transition of cultured mouse cells. **EMBO J** 1985; 4:3145-51.

67. Lee S, Clark T, Chen J, et al. Correct identification of genes from serial analysis of gene expression tag sequences. **Genomics** 2002; 79:598-602.

68. Leerkes MR, Caballero OL, Mackay A, et al. *In silico* comparison of the transcriptome derived from purified normal breast cells and breast tumor cell lines reveals candidate upregulated genes in breast tumor cells. **Genomics** 2002; 79:257-65.

69. Levy-Lahad E, Lahad A, Eisenberg S, et al. A single nucleotide polymorphism in the RAD51 gene modifies cancer risk in BRCA2 but not BRCA1 carriers. **Proc Natl Acad Sci U S A** 2001; 98:3232-6.

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

71. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. **Science** 1992; 257:967-71.

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

73. Liang P, Pardee AB. Recent advances in differential display. **Curr Opin Immunol** 1995; 7:274-80.

74. Lockhart DJ, Dong H, Byrne MC et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nat Biotechnol** 1996; 14:1675-80.

75. Loging WT, Lal A, Siu IM, et al. Identifying potential tumor markers and antigens by database mining and rapid expression screening. **Genome Res** 2000; 10:1393-402.

76. Loidi L, Garcia-Caballero T, Vidal A, et al. Complex regulation of prothymosin alpha in mammary tumors arising in transgenic mice. **Life Sci** 1999; 64:2125-33.

77. Lu R, Serrero G. Mediation of estrogen mitogenic effect in human breast cancer MCF-7 cells by PC-cell derived growth factor (PCDGF/granulin precursor). **Proc Natl Acad Sci U S A** 2001; 98:142-7.

78. Lynch HT, Smyrk TC. Hereditary colorectal cancer. **Semin Oncol** 1999; 26:478-84.

79. Magro G, Lanzafame S, Colombatti A. Immunohistochemical staining patterns of type VI collagen in the normal, hyperplastic, and neoplastic adult male breast. **Pathologica** 1994; 86:142-5.

80. Man MZ, Wang X, Wang Y. POWER_SAGE: comparing statistical tests for SAGE experiments. **Bioinformatics** 2000; 16:953-9

81. Mao S, Neale GA, Goorha RM. T-cell oncogene rhombotin-2 interacts with retinoblastoma-binding protein 2. **Oncogene** 1997; 14:1531-9.

82. Marin A, Lopez de Cerain A, Hamilton E, et al. DT-diaphorase and cytochrome B5 reductase in human lung and breast tumors. **Br J Cancer** 1997; 76:923-9.

83. Matsui H, Kubochi K, Okazaki I, et al. Collagen biosynthesis in gastric cancer: immunohistochemical analysis of prolyl-4-hydrolyse. **J Surg Oncol** 1999; 70:239-46.

84. Miller RT, Christoffels AJ, Burke J, et al. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. **Genome Res** 1999; 9:1143-55.

85. Nam DK, Lee S, Zhou G, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. **Proc Natl Acad Sci U S A** 2002; 99:6152-6.

86. National Cancer Institute Cancer Genome Anatomy Project (CGAP). Avaliable from <URL:http://www.ncbi.nlm.nih.gov/ncicgap/> [2003 Dec 23]

87. Ohlsson R, Paldi A, Graves JA. Did genomic imprinting and X chromosome inactivation arise from stochastic expression? **Trends Genet** 2001; 17:136-41.

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

89. OMIM™ ("Online Mendelian Inheritance in Man™") Johns Hopkins University. Avaliable from <URL:http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM> [2003 Dec 23]

90. Oskouian, B. Overexpression of fatty acid synthase in SKBR3 breast cancer cell line is mediated via a transcriptional mechanism. **Cancer Lett** 2000; 149:43-51.

91. Page MJ, Amess B, Townsend RR, et al. Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. **Proc Natl Acad Sci U S A** 1999; 96:12589-94.

92. Pagni M, Iseli C, Junier T, Falquet L, Jongeneel V, Bucher P. trEST, trGEN and Hits: access to databases of predicted protein sequences. **Nucleic Acids Res** 2001; 29:148-51.

93. Park CC, Bissell MJ, Barcellos-Hoff MH. The influence of the microenvironment on the malignant phenotype. **Mol Med Today** 2000; 6:324-9.

94. Patanjali SR, Parimoo S, Weissman SM. Construction of a uniform-abundance (normalized) cDNA library. **Proc Natl Acad Sci U S A** 1991; 88:1943-7.

95. Pedroni M, Sala E, Scarselli A, et al. Microsatellite instability and mismatch-repair protein expression in hereditary and sporadic colorectal carcinogenesis. **Cancer Res** 2001; 61:896-9.

96. Perucho M. Cancer of the microsatellite mutator phenotype. **Biol Chem** 1996; 377:675-84.

97. Pisani P., Parkin, D. M., Bray, F. and Ferlay, J. Estimates of the worldwide mortality from 25 cancers in 1990. **Int J Cancer** 1999; 83:18-29.

98. Prasad SC, Soldatenkov VA, Kuettel MR, Thraves PJ, Zou X, Dritschilo A. et al. Protein changes associated with ionizing radiation-induced apoptosis in human prostate epithelial tumor cells. **Electrophoresis** 1999; 20:1065-74.

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

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

101. Riggins GJ, Strausberg RL. Genome and genetic resources from the Cancer Genome Anatomy Project. **Hum Mol Genet** 2001; 10:663-7

102. Sager R. Tumor suppressor genes: the puzzle and the promise. **Science** 1989; 246:1406-12.

103. Sager R. Expression genetics in cancer: shifting the focus from DNA to RNA. **Proc Natl Acad Sci U S A** 1997; 94:952-5.

104. Saha S, Sparks AB, Rago C, et al. Using the transcriptome to annotate the genome. **Nat Biotechnol** 2002; 20:508-12.

105. Schaefer C, Grouse L, Buetow K, Strausberg RL. A new cancer genome anatomy project web resource for the community. **Cancer J** 2001; 7:52-60.

106.Schaller G, Fuchs I, Pritze W, et al. Elevated keratin 18 protein expression indicates a favorable prognosis in patients with breast cancer. **Clin Cancer Res** 2000; 2:1879-85.

107.Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science** 1995; 270:467-70.

108.Scheurle D, DeYoung MP, Binninger DM, Page H, Jahanzeb M, Narayanan R Cancer gene discovery using digital differential display. **Cancer Res** 2000; 60:4037-43.

109.Schlosshauer PW, Brown SA, Eisinger K APC truncation and increased beta-catenin levels in a human breast cancer line. **Carcinogenesis** 2000; 21:1453-6.

110.Schmitt AO, Specht T, Beckmann G, et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. **Nucleic Acids Res** 1999; 27:4251-60.

111.Schon M, Klein CE, Hogenkamp V, Kaufmann R, Wienrich BG, Schon MP. Basal-cell adhesion molecule (B-CAM) is induced in epithelial skin tumors and inflammatory epidermis, and is expressed at cell-cell and cell-substrate contact sites. **J Invest Dermatol** 2000; 115:1047-53.

112.Scott MR, Westphal KH, Rigby PW. Activation of mouse genes in transformed cells. **Cell** 1983; 34:557-67.

113.Silva AP, Salim AC, Bulgarelli A, et al. Identification of 9 novel transcripts and two RGSL genes within the hereditary prostate cancer region (HPC1) at 1q25. **Gene** 2003; 310:49-57.

114.Simpson AJG, de Souza SJ, Camargo AA, Brentani RR. Definition of the gene content of the human genome: the need for deep experimental verification. **Comp Funct Genom** 2001; 2:169-75

115.Smith KJ, Johnson KA, Bryan TM, et al. The APC gene product in normal and tumor cells. **Proc Natl Acad Sci U S A** 1993; 90:2846-50.

116.Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. Construction and characterization of a normalized cDNA library. **Proc Natl Acad Sci U S A** 1994; 91:9228-32.

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

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

119.Stekel DJ, Git Y, Falciani F. The comparison of gene expression from multiple cDNA libraries. **Genome Res** 2000; 10:2055-61

120.Stemmer-Rachamimov AO, Wiederhold T, Nielsen GP, et al. NHE-RF, a merlin-interacting protein, is primarily expressed in luminal epithelia, proliferative endometrium, and estrogen receptor-positive breast carcinomas. **Am J Pathol** 2001; 158:57-62.

121.Stern MD, Anisimov SV, Boheler KR. Can transcriptome size be estimated from SAGE catalogs? **Bioinformatics** 2003; 19:443-8.

122.Stremmel C, Wein A, Hohenberger W, Reingruber B. DNA microarrays: a new diagnostic tool and its implications in colorectal cancer. **Int J Colorectal Dis** 2002; 17:131-6.

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

124.Strausberg RL. The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. **J Pathol** 2001; 195:31-40.

125.Strausberg RL, Greenhut SF, Grouse LH, Schaefer CF, Buetow KH. *In silico* analysis of cancer through the Cancer Genome Anatomy Project. **Trends Cell Biol** 2001; 11:S66-71.

126.Strausberg RL, Camargo AA, Riggins GJ et al. An international database and integrated analysis tools for the study of cancer gene expression. **Pharmacogenomics J** 2002; 2:156-64.

127.Taipale J, Beachy PA. The Hedgehog and Wnt signalling pathways in cancer. **Nature,** 2001; 411:349-54.

128.Thibodeau SN, French AJ, Cunningham JM, et al. Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1. **Cancer Res** 1998; 58:1713-8.

129.Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. **Proc Natl Acad Sci U S A** 1996; 93:14800-3.

130.Tomlinson I, Bodmer W. Selection, the mutation rate and cancer: ensuring that the tail does not wag the dog. **Nat Med** 1999; 5:11-2.

131. Tomlinson I, Sasieni P, Bodmer W. How many mutations in a cancer? **Am J Pathol** 2002; 160:755-8.

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

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

134. Velculescu VE. Essay: Amersham Pharmacia Biotech & Science prize. Tantalizing transcriptomes--SAGE and its use in global gene expression analysis. **Science** 1999; 286:1491-2.

135. Velculescu VE, Madden SL, Zhang L, et al. Analysis of human transcriptomes. **Nat Genet** 1999; 23:387-8.

136. Velculescu VE, Vogelstein B, Kinzler KW. Analysing uncharted transcriptomes with SAGE. **Trends Genet** 2000; 16:423-5.

137. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. **Science** 2001; 291:1304-51.

138. Wagener C, Ergun S. Angiogenic properties of the carcinoembryonic antigen-related cell adhesion molecule 1. **Exp Cell Res** 2000; 261:19-24.

139. Wang SM, Fears SC, Zhang L, Chen JJ, Rowley JD. Screening poly(dA/dT)-cDNAs for gene identification. **Proc Natl Acad Sci U S A** 2000; 97:4162-7.

140. Wen XY, Stewart AK, Sooknanan RR, et al. Identification of c-myc promotor-binding protein and X-box binding protein 1 as interleukin-6 target genes in human multiple myeloma cells. **Int J Oncol** 1999; 15:173-8.

141. Yan H, Dobbie Z, Gruber SB, et al. Small changes in expression affect predisposition to tumorigenesis. **Nat Genet** 2002; 30:25-6.

142. Zhang M, Martin KJ, Sheng S, Sager R. Expression genetics: a different approach to cancer diagnosis and prognosis. **Trends Biotechnol** 1998; 16:66-71.

143. Zhang H, Wang Q, Kajino K, Greene MI. VCP, a weak ATPase involved in multiple cellular events, interacts physically with BRCA1 in the nucleus of living cells. **DNA Cell Biol** 2000; 19:253-63.

144. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. **J Comput Biol** 2000; 7:203-14.

145. Zhang L, Zhou W, Velculescu VE, et al. Gene expression profiles in normal and cancer cells. **Science** 1997; 276:1268-72.

146. Zhao HJ, Hosoi Y, Miyachi H, et al. DNA-dependent protein kinase activity correlates with Ku70 expression and radiation sensitivity in esophageal cancer cell lines. **Clin Cancer Res** 2000; 6:1073-8.

147. Zhuo D, Zhao WD, Wright FA, et al. Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. **Genome Res** 2001; 11:904-18.