Desenvolvimento de um método de diagnóstico molecular para o câncer gástrico baseado na análise do perfil de expressão gênica através da metodologia de cDNA array

SIBELE INÁCIO MEIRELES

Tese de Doutorado apresentada à
Fundação Antônio Prudente
para obtenção de título de Doutora em Ciências

Área de concentração: Oncologia

São Paulo
2003

Desenvolvimento de um método de diagnóstico molecular para o câncer gástrico baseado na análise do perfil de expressão gênica através da metodologia de cDNA array

Sibele Inácio Meireles

Dissertação apresentada à
Fundação Antônio Prudente
para obtenção do Grau de Doutor em Ciências.

Área de concentração: Oncologia

São Paulo
2003

Desenvolvimento de um método de diagnóstico molecular para o câncer gástrico baseado na análise do perfil de expressão gênica através da metodologia de cDNA array

Sibele Inácio Meireles

Dissertação apresentada à
Fundação Antônio Prudente
para obtenção do Grau de Doutor em Ciências.

Área de concentração: Oncologia

Orientador: Luiz Fernando Lima Reis
Co-orientador: Alex Fiorini de Carvalho

São Paulo
2003

DEDICATÓRIA

A todos que fazem parte da minha vida,
amigos e familiares, e em especial,
aos meus pais e
ao Luiz.

AGRADECIMENTOS

Ao **Prof. Luiz Fernando Lima Reis**, Chefe do Laboratório de Genoma Funcional e Diretor da Pós-graduação. Meu eterno agradecimento ao meu orientador, ao respeitado pesquisador, amigo, pai, conselheiro... Você é um exemplo de sabedoria e dedicação ao trabalho e aos alunos. Obrigada Luiz, Maria Auxiliadora. Nanda, Felipe e Maria Paula, pelo carinho e preciosos momentos de uma convivência familiar.

Ao **LaBRI** (Laboratorio de Biologia da Resposta Inflamatoria), meu laboratório e segundo lar! Obrigada a TODOS! Agradeço sincera e profundamente, pela ajuda em muitos experimentos; pelos valiosos conselhos profissionais e pessoais; pelos programas computacionais que aprendi a usar; pela ajuda em consultas em bancos de dados; formatação de texto; pela ajuda e organização na coleta de amostras; pela extração dos RNAs; pelo incentivo e apoio em dias se seminário; pela ajuda na elaboração e revisão de slides e textos; pela alegria dos momentos de confraternização; pela amizade; compreensão; confiança; amor; pelas carinhosas canções improvisadas; pelas críticas e sugestões; pelos bolos de banana, baunilha, bombocado, formigueiro; pão de mel, docinhos e amendoins; pela acolhida familiar; comemorações de aniversário; anúncios de supermercado; abraços; aulas de matemática, japonês e inglês... Será impossível descrever tudo o que recebi em 5 anos de convivência, vocês me ensinaram lições para a vida inteira e são um pedaço de mim. Vai ser difícil viver longe de vocês... Adriana e Alex, Regina, Gustavo, Bia, Luciana, Abrantes, Suzana, Lara, Waleska, Bianca, Ana Helena, Vladmir, Emerson, Ana Coló, Chamberlein, Sarah, Mariana, Elaine, Lepique, Lud, Aristóbolo e Valéria.

À Patrícia e Katyana, que um dia se tornaram irmãs e nunca mais sairão do meu coração.

À Anna Cristina e Elisângela, pelo sequênciamento de DNA, amizade e carinho.

Aos pesquisadores do IME-USP (Instituto de Matematica e Estatistica da Universidade de Sao Paulo), **Eduardo Jordão Neves, Roberto Hirata** e ao aluno **Elier Broche Cristo**. Sem o conhecimento e a ajuda de vocês este trabalho estaria incompleto. Obrigada pela colaboração indispensável e valiosa.

Ao Dr **Fernando Augusto Soares**, chefe do Departamento de Anatomia Patológica. Responsável pela rigorosa análise de histologia de todas as amostras utilizadas neste trabalho e pelos experimentos maravilhosos de imunoistoquímica. Agradeço também a Dra **Maria Dirlei Begnami,** pela grande ajuda nos experimentos de imunoistoquímica e à **Alexandra Cardozo**, pela eficiência e profissionalismo.

Ao Dr. André Montagnini, chefe do Departamento de Cirurgia Abdominal, **Dra. Cláudia Zitron,** chefe do Departamento de Endoscopia Digestiva **e Dra. Adriane Pelosof,** que acreditaram neste trabalho, possibitaram a coleta de amostras de estômago e muito contribuíram com a sua realização.

Ao **Prof Dr. Ricardo Brentani**, Diretor do Hospital do Câncer e do Instituto Ludwig de Pesquisa sobre o Câncer, pela administração exemplar deste centro de excelência em tratamento e pesquisa do câncer.

À **FAPESP** (Fundação de Amparo à Pesquisa do Estado de São Paulo) pela bolsa concedida e suporte financeiro.

Aos membros da banca de qualificação, **Dr Ismael Dale C. G. Da Silva , Prof. Maria Aparecida Nagai, Prof Luiz Armando de Marco e Dr. Andrew Simpson** pelo acompanhamento durante a realização deste trabalho.

Ao **Carlos Ferreira Nascimento, Myuki Fukuda** e **Severino** que ajudaram na coleta dos tecidos e preparação de lâminas com muito zelo e habilidade.

À coordenadora **Ana Maria Kuninari** e a secretária **Márcia Hiratani,** funcionárias da pós-graduação, pela ajuda de todas as horas com muita competência e alegria.

À **Aline Pacífico,** que além de ser uma excelente profissional, é sempre atenciosa, alegre e carinhosa.

À **Suely Francisco** e aos demais funcionário da biblioteca da Fundação Antônio Prudente, pelo suporte bibliográfico e disponibilidade constante.

Aos **colegas do Instituto Ludwig**, que muitas vezes ajudaram nos experimentos, com boa vontade e amizade.

Aos **funcionários do Instituto Ludwig**, técnicos e administrativos, que trouxeram condições privilegiadas de trabalho.

Às minhas **amigas, tios e primos**, sempre presentes, oferecendo ajuda, atenção e carinho.

À **minha família**, meus pais, Sr Francisco e D. Marlene, irmãos Simone, Sérgio e Sinara e cunhados(a) Tatiana, Leo e Vinícios. Obrigada pelo amor, pelos sacrifícios, incentivo e compreensão.

Ao **Dai**, *mais* que ajudar na revisão de texto, você trouxe apoio, compreensão, amizade, carinho... Obrigada!

A todos, o meu eterno e sincero: Obrigada !!! Se eu não tivesse pessoas a quem agradecer, certamente não teria chegado até aqui.

Agradeço a **Deus** pela vida e por colocar todos vocês em meu caminho.

RESUMO

Meireles, SI. **Desenvolvimento de um método de diagnóstico molecular para o câncer gástrico baseado na análise do perfil de expressão gênica através da metodologia de *cDNA array*.** São Paulo; 2003. [Tese de Doutorado-Fundação Antônio Prudente].

O Câncer Gástrico é uma das principais causas de morte por câncer no Brasil e no mundo. Um dos fatores que contribuíram para esta alta taxa de mortalidade é o diagnóstico tardio da doença. Assim, o desenvolvimento de métodos que permitem o diagnóstico precoce do câncer gástrico podem contribuir para melhorar o prognóstico dos pacientes com esta doença. Neste trabalho, utilizamos a técnica de *cDNA array* e analisamos o perfil de expressão gênica em mucosa gástrica normal (N), gastrite (G), metaplasia intestinal (M) e tumor gástrico (T). Com base nos genes diferencialmente expressos, procuramos desenvolver uma ferramenta de diagnóstico molecular para as lesões na mucosa gástrica. Na primeira fase, utilizamos um *cDNA array,* contendo cerca de 4.500 elementos e comparamos o perfil de expressão gênica em seis amostras de mucosa gástrica normal e seis amostras de mucosa gástrica tumoral. Identificamos 80 cDNAs, diferencialmente expressos entre esses dois tecidos e ao utilizar o método de agrupamento *Self Organizing Map* (SOM), o perfil de expressão desses cDNAs permitiu separar o grupo de amostras normais do grupo de amostras tumorais. Na segunda fase foi analisado o perfil de expressão gênica de 376 genes distintos em 99 fragmentos de estômago, representando N (n=28), G (n=21), M (n=22) e T (n=28). Neste arranjo, os genes incluídos correspondem àqueles selecionados na análise anterior. A este arranjo foram acrescentados genes relacionados ao câncer, segundo dados da literatura. Através de comparações pareadas dessas amostras, identificamos 42 genes diferencialmente expressos com p< 0,0009, de acordo com o teste de Wilcoxon. Utilizando um algoritmo de agrupamento (*k-means*), através do perfil de expressão de 18 genes foi possível agrupar essas amostras em quatro grupos: um grupo para a maioria de N e G, dois grupos distintos para a maioria de T e M, respectivamente, e um quarto grupo bastante heterogêneo, contendo seis amostras representando quatro tipos de tecido. Para identificar trios de genes capazes de classificar uma amostra, utilizamos o algoritmo Discriminador Linear de Fisher. Encontramos inúmeros classificadores capazes de distinguir entre N e T, porém poucos classificadores que distinguem entre G e T ou M e T. Identificamos amostras de metaplasia intestinal cujo perfil molecular se assemelha àquele típico de amostras tumorais. Propomos, então, a utilização desses marcadores para que possamos avaliar a evolução e o prognóstico de pacientes com metaplasia intestinal.

ABSTRACT

Meireles, SI. **Development of a molecular diagnosis tool for gastric cancer based on gene expression profile using the cDNA array methodology.** São Paulo; 2003. [Tese de Doutorado-Fundação Antônio Prudente].

Gastric cancer is one of the leading causes of cancer-related death in Brazil and in the world. High frequency of gastric cancer-related death is mainly due to late-stage diagnosis. Hence, new tools aimed to early diagnostic would have a positive impact in the outcome of the disease. Using cDNA arrays, we analised the expression profile of normal gastric mucosa (N), gastritis (G), intestinal metaplasia (M) and gastric tumor (T). Based on the differentially expressed genes, we developed diagnosis tools for identification of lesions in gastric mucosa. In a first step, we used cDNA arrays having around 4,500 elements to compare the expression profile of six samples of normal gastric mucosa and six samples of tumor gastric mucosa. Eighty differentially expressed cDNAs were identified and, using Self Organizing Map (SOM), their expression profile allowed the precise separation of the normal from the tumor sample groups. In a second step, the expression profile of 376 distinct genes, derived from the first analysis and plus a set of known altered genes in human cancer according to the literature, were analised in 99 gastric fragments represented by: N (n=28), G (n=21), M (n=22) and T (n=28). Pair wise comparisons between these samples allowed the identification of 42 differentially expressed genes with p<0.0009 in a Wilcoxon test. Using the clustering algorithm k-means, the expression profile of 18 genes allowed the clustering of the majority of N and G samples in a distinct group, the majority of M and T samples in two aditional groups and fourth heterogeneous group. We then applied Fisher's linear discriminat to identify trios of genes that could be used to build classifiers for class distinction. A lager number of classifiers could distinguish between NxT whereas, for the distinction of GxT and MxT, fewer classifiers were identified. Importantly, it was possible identify samples of intestinal metaplasia whose expression pattern resembled that of an adenocarcinoma and can now be used for follow-up of patients in order to determine their potential as prognostic test for malignant transformation.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

CDH1	do inglês “cadherin 1, type 1”
cDNA	DNA complementar
cDNA arrays	Arranjos de cDNA
CLTC	do inglês “Clathrin, Heavy Polypeptide (Hc)”
CTNNB1	do inglês “Catenin beta 1”
CTSB	do inglês “Cathepsin B”
ESTs	do inglês “Expressed Sequence Tags”
G6PD	do inglês “Glucose-6-Phosphate Dehydrogenase”
HPRT	do inglês “Hypoxanthine-Guanine Phosphoribosyltransferase”
IL-1	do inglês “Interleucin 1”
IL1-β	do inglês “Interleucin1-β”
LDHA	do inglês “Lactate Dehydrogenase A”
MMP2	do inglês “Matrix Metalloproteinase 2”
NBS1	do inglês “Nijmegen Breakage Syndrome 1”
OMS	Organização Mundial da Saúde
ORESTES	do inglês “Open Reading Frame ESTs”
RPL10	do inglês “Ribosomal Protein L10”
RT-PCR	do inglês “Reverse Transcription-Polimerase Chain Reaction”
SOM	do inglês “Self Organizing Maps”
SVD	do inglês “Singular Values Discriminator”
SVM	do inglês “Support Vector Machine”
TBP	do inglês “TATA Binding Protein”
TP53	do inglês “Tumor Protein p53

ÍNDICE

1. INTRODUÇÃO

1.1. O Câncer

O câncer é um processo de múltiplas etapas e reflete o acúmulo de alterações genéticas, desencadeadas, principalmente, por defeitos herdados ou adquiridos em genes relacionados ao reparo de DNA e controle do ciclo-celular. No câncer, alguns genes sofrem mutações, que resultam em um ganho de função ou que resultam em perda de função, como no caso dos oncogenes e dos genes supressores de tumor, respectivamente. Foram propostos seis importantes defeitos relacionados à fisiologia celular que juntos desencadeiam o desenvolvimento do câncer: auto-suficiência para estímulo do crescimento celular, ausência de inibição de crescimento por sinais gerados no contato entre células, ausência de resposta ao estímulo de apoptose, potencial ilimitado de replicação celular, manutenção de angiogênese, capacidade de invasão tecidual e metástase (HANAHAN et al. 2000).

O diagnóstico precoce possui um impacto muito grande no controle da doença (ETZIONI et al. 2003), pois permite identificar o câncer ainda restrito ao sítio primário, aumentando as chances de cura e reduzindo a mortalidade e a morbidade. A detecção precoce do câncer pode se beneficiar do uso de novas tecnologias de análise molecular e de ferramentas de bioinformática. Estas tecnologias permitem identificar inúmeros marcadores moleculares que são úteis não só para o entendimento do processo de oncogênese, mas também para a detecção precoce e classificação do câncer (YEATMAN , 2003).

1.2. O Câncer Gástrico

O câncer gástrico é um dos tumores de maior taxa de mortalidade no mundo (STADTLANDER et al. 1999). No Brasil, o câncer gástrico ocupa, respectivamente, entre homens e mulheres, a quarta e quinta colocação em incidência, e a terceira e quinta colocação em mortalidade por câncer. O Instituto Nacional do Câncer estimou que em 2003 haveria 20.640 novos casos de câncer gástrico diagnosticados no Brasil, com um número de óbitos estimado em 11.145 casos/ano (MINISTÉRIO DA SAÚDE , 2003). Mundialmente, as taxas de incidência e mortalidade por câncer gástrico estão diminuindo, provavelmente por causa de mudança de hábitos alimentares da população e da facilidade na conservação de alimentos. Porém, por razões ainda não bem esclarecidas, esta diminuição tem sido limitada aos tumores da porção distal do estômago (região do corpo e antro) e o número de pacientes com câncer proximal (região do cárdia) e da junção gastro esofágica tem aumentado, significativamente, desde meados da década de 80 (DEVESA et al. 1998).

A maioria dos casos de câncer gástrico são diagnosticados tardiamente, especialmente, devido à falta de sintomas específicos que caracterizem a doença. O diagnóstico tardio dificulta o tratamento curativo, elevando os índices de morbidade e mortalidade da doença. Assim, a precocidade no diagnóstico é essencial, pois permite uma conduta mais efetiva no tratamento, aumentando a sobrevida e a taxa de cura (OLIVEIRA et al. 1998).

O principal tipo de câncer gástrico é o adenocarcinoma, responsável por 90-95% dos casos. A classificação mais utilizada para esses tumores é a classificação de Lauren, que divide o adenocarcinoma gástrico em dois principais tipos histológicos: intestinal e difuso. Alguns tumores apresentam caracteristicas dos dois tipos histológicos e são denominados do tipo misto (LAUREN , 1965; MING , 1977).

No adenocarcinoma do tipo intestinal as células cancerosas formam glândulas que variam de bem diferenciadas a moderadamente diferenciadas. É mais freqüente em relação ao tipo difuso, mais comum em homens do que em mulheres e mais freqüente em idosos. Representa o tipo histológico dominante em áreas onde o câncer gástrico é epidêmico, sugerindo um fator etiológico. A patogênese do adenocarcinoma do tipo intestinal tem sido associada à presença de lesões precursoras tais como a gastrite crônica, gastrite atrófica, metaplasia intestinal e displasia (CORREA et al. 1994). Tanto as lesões precursoras quanto o adenocarcinoma gástrico estão freqüentemente associadas à infecção por *Helicobacter pylori* (PEEK, Jr. et al. 2002).

Os fatores que levam ao maior risco de desenvolvimento do câncer em pacientes infectados por *H. pylori* são relacionados à infecção por certas cepas, fatores ambientais e fatores do hospedeiro (idade da aquisição da infecção, resposta imune, mudanças na secreção ácida). Cepas de *H. Pylori* que possuem a ilha de patogenicidade *cag*A$^+$ induzem a produção de citocinas pró-inflamatórias como IL-8 (interleucina-8), desencadeando uma resposta inflamatória na mucosa gástrica. Entre outros efeitos, a resposta inflamatória por celulas polimorfonucleares libera radicais livres, como o óxido nítrico (NO), que promovem danos no DNA e aumentam o risco de desenvolvimento do câncer (PEEK, Jr. et al. 2002). Com relação aos fatores do hospedeiro, um importante estudo demonstrou que indivíduos portadores de polimorfismos na região cromossômica do gene que codifica a IL-1 (Interleucina-1) apresentam, quando infectados por *H. pylori,* risco maior de desenvolvimento de gastrite atrófica e câncer gástrico (EL OMAR et al. 2000). As evidências discutidas acima poderiam explicar, em parte, porque apenas alguns indivíduos infectados por *H. pylori* desenvolvem câncer gástrico.

O adenocarcinoma gástrico do tipo difuso apresenta células não organizadas em estruturas glandulares, que infiltram de maneira difusa pela parede gástrica. O adenocarcinoma gástrico do tipo difuso acomete pacientes mais jovens e na mesma proporção entre mulheres e homens. Este tipo de tumor não tem sido associado a lesões precursoras como a metaplasia intestinal e possui uma alta associação à herança familial, devido a mutações no gene da e-caderina (CDH1) (GUILFORD et al. 1998).

Estudos moleculares em câncer gástrico têm demonstrado que um grande número de alterações genéticas podem estar envolvidas no processo de carcinogênese no estômago. Estas evidências indicam também que podem existir diferentes mecanismos moleculares entre o carcinoma tipo intestinal e o carcinoma do tipo difuso (EL RIFAI et al. 2002). O conhecimento das alterações moleculares envolvidas no processo de carcinogênese gástrica pode ser utilizado para identificar marcadores genéticos indicativos na evolução do processo de transformação celular e podem ser úteis na criação de ferramentas de diagnóstico molecular e prognóstico.

1.3. Utilização da Técnica de Microarranjos de cDNA no estudo do câncer gástrico

Os arranjos de cDNA (do inglês *cDNA arrays*) são arranjos de centenas a milhares de fragmentos de cDNA em superfícies de *nylon* ou vidro. Eles permitem avaliar a expressão gênica de um grande número de genes em um único experimento. Esta tecnologia tem mostrado ser uma ferramenta indispensável em estudos de expressão gênica em câncer (LIOTTA et al. 2000; PUSZTAI et al. 2003; RAMASWAMY et al. 2002).

Os estudos envolvendo a utilização da técnica de *cDNA microarray* em câncer gástrico surgiram recentemente e a maioria dos estudos analisou o perfil de expressão gênica em mucosa gástrica normal e tumoral (EL RIFAI et al. 2001; INOUE et al. 2002; JI et al. 2002; JUNG et al. 2000; LEE et al. 2002; LIU et al. 2002; MEIRELES et al. 2003; WANG et al. 2002). Poucos trabalhos compararam os genes diferencialmente expressos entre os tumores do tipo intestinal e do tipo difuso (BOUSSIOUTAS et al. 2003). Hasegawa e colaboradores estudaram especificamente o adenocarcinoma do tipo intestinal correlacionando-o com a presença de comprometimento linfonodal (HASEGAWA et al. 2002). No entanto, nenhum dos trabalhos até agora apresentados tiveram como objetivo específico a construção de classificadores moleculares que possam contribuir para o diagnóstico precoce do câncer gástrico.

2. JUSTIFICATIVA

O Câncer Gástrico é uma das principais causas de morte por câncer no Brasil e no mundo. Um dos fatores que contribuem para esta alta taxa de mortalidade é o diagnóstico tardio da doença. Na maioria dos casos, os tumores diagnosticados em fase avançada apresentam metástase em outros órgãos, comprometendo o sucesso no tratamento, levando a um pior prognóstico e alta mortalidade. Prejudicando ainda mais o prognóstico de pacientes com doença avançada, o câncer gástrico, em especial o adenocarcinoma, responde muito mal ao tratamento quimioterápico. Medidas voltadas para promover o diagnóstico precoce do câncer gástrico podem contribuir para melhorar o prognóstico destes pacientes. Entre estas medidas, está o desenvolvimento de ferramentas de diagnóstico molecular para a detecção de alterações malignas na mucosa gástrica, especialmente numa fase precoce do desenvolvimento do tumor. Realizamos este estudo, visando identificar marcadores moleculares em amostras de estômago normal e de diferentes lesões na mucosa gástrica. Esses marcadores foram utilizados para a criação de classificadores moleculares capazes de distinguir entre lesões não-malignas, pré-malignas e malignas.

O regimento do curso de Pós-Graduação da Fundação Antônio Prudente possibilita ao aluno apresentar a tese no formato tradicional ou no formato simplificado. Optamos por apresentar essa tese no formato simplificado, anexando os artigos oriundos dessa pesquisa. Incluímos, também, a discussão de alguns aspectos metodológicos que consideramos relevantes para complementar a análise final, descrita no artigo.

3. OBJETIVOS

3.1. Objetivo geral

Desenvolver um método de diagnóstico molecular para o câncer gástrico, baseado na análise do perfil de expressão gênica através da metodologia de *cDNA array*.

3.2. Objetivos específicos

- Construir *cDNA arrays,* contendo fragmentos de cDNA (ORESTES) gerados pelo Projeto Genoma de Câncer.
- Identificar genes diferencialmente expressos em mucosa gástrica normal, gastrite, mucosa gástrica com metaplasia intestinal e tumoral, obtidas no Hospital do Câncer AC Camargo através de procedimento cirúrgico e biópsia pré-cirúrgica.
- Criar classificadores moleculares capazes de distinguir entre mucosa gástrica normal, gastrite, mucosa gástrica com metaplasia intestinal e tumoral.
- Identificar genes diferencialmente expressos em adenocarcinoma do tipo intestinal e adenocarcinoma do tipo difuso.

4. DISCUSSÃO

Utilizando a técnica de *cDNA array*, estudamos o perfil de expressão gênica em mucosa normal de estômago, gastrite, metaplasia intestinal e adenocarcinoma gástrico. Os resultados mais relevantes foram discutidos nas publicações em anexo. Nesta seção, estamos abordando os aspectos importantes de nossas análises e alguns resultados, obtidos durante a fase de implementação da técnica de *cDNA array* e de análise dos dados.

Para construção de *cDNA arrays* utilizamos fragmentos de sequências de cDNA (ORESTES) gerados no Projeto Genoma de Câncer (CAMARGO et al. 2001). Neste projeto, concluído em Outubro de 2001, foram estudados tumores de relevância no Brasil, incluindo o câncer gástrico. Foram geradas ORESTES (*Open Reading Frames EST Sequences*) que são ESTs da porção central dos mRNAs, correspondente à região codificadora do gene (DIAS et al. 2000). Os clones bacterianos que contêm estas ORESTES foram congelados e estocados em nosso laboratório. Os *cDNA arrays* utilizados inicialmente em nosso estudo foram construídos com os primeiros clones ORESTES disponíveis no estoque gerados a partir de RNA extraído de estômago normal e tumoral. Posteriormente, estas sequências foram reunidas em um segundo arranjo contendo sequências adicionais geradas a partir de RNA extraído de tumor ou tecido normal de mama e de cabeça e pescoço, totalizando cerca de 4.500 ORESTES. Este arranjo maior foi utilizado no primeiro artigo. No segundo artigo foi construído um terceiro *cDNA array*, contendo os 141 clones identificados no primeiro artigo acrescido de uma coleção de genes com função relacionada a processos tumorais em diferentes órgãos, com base em dados da literatura, totalizando 376 genes distintos. Neste último array, todos os clones foram fixados em triplicata e, sempre que possível, foram selecionados dois fragmentos diferentes de cDNA para cada gene.

No início de nossos estudos, os genes diferencialmente expressos foram identificados pela razão do valor médio de intensidade de hibridação em dois conjuntos de amostras (Normal e Tumor). Porém, as análises preliminares não foram consideradas no primeiro artigo, pois passamos a utilizar testes estatísticos para a busca por genes diferencialmente expressos, levando em consideração, além da razão, a variância de cada gene nas populações estudadas. Este importante progresso foi possível graças à colaboração dos pesquisadores do Instituto de Matemática e Estatística da Universidade de São Paulo. Estes critérios são mais adequados do que apenas utilizar a razão. Experimentalmente, não é possível garantir que um valor de razão pré-determinado seja suficiente para definir se um gene é ou não diferencialmente expresso sem considerar a variabilidade da expressão deste gene em todas as amostras. Além disso, a razão elimina toda a informação sobre o valor absoluto do nível de expressão gênica. Assim, no primeiro artigo, passamos a utilizar a medida *tnc,* que corresponde à diferença entre os valores médios de expressão entre amostras normais e tumorais dividido pelo desvio padrão nas duas populações (fórmula apresentada no primeiro artigo). Valor positivo de *tnc* demonstra que um determinado gene possui maior nível de expressão em amostra normal ao passo que o valor negativo de *tnc* demonstra que um determinado gene possui maior nível de expressão em amostra tumoral. Na figura 1, mostramos um exemplo extraído do banco de dados referente ao primeiro artigo, onde selecionamos 5 genes com base na razão de expressão sem considerar a variância e 5 genes onde, além da razão, a variância foi considerada.

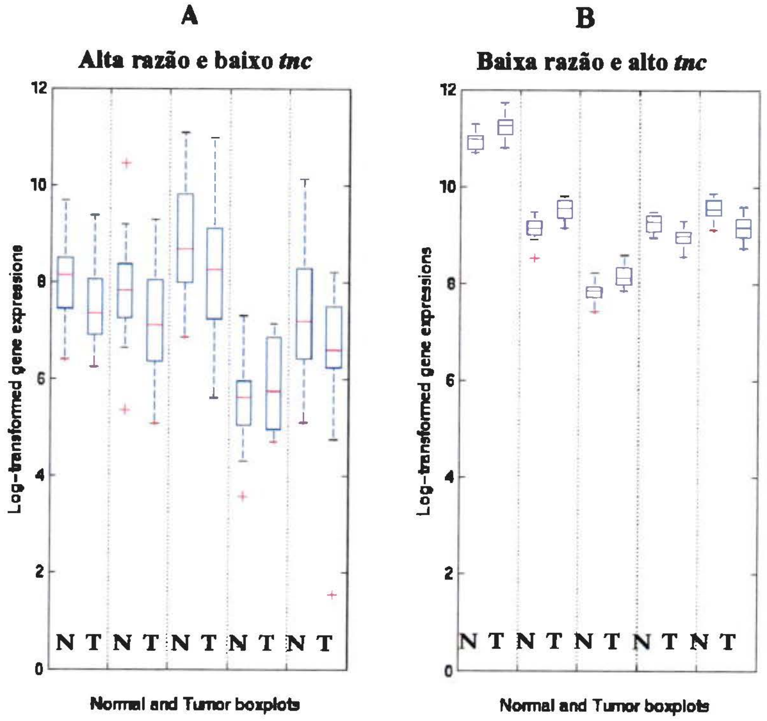


Figura 1 - Comparação entre razão e *tnc*.

Nesta figura estão apresentados *boxplot* para valores de expressão gênica obtidos por hibridização de microarranjos de cDNA. Os gráficos representam valores obtidos em amostras normais e amostras tumorais para 10 genes distintos. Em A, estão apresentados 5 genes que possuem alta razão e baixo *tnc*. Em B, estão apresentados 5 genes que possuem baixa razão e alto *tnc*.

No primeiro artigo, identificamos um painel de 80 cDNAs diferencialmente expressos entre amostras normais e tumorais cujo valor de *tnc* foi ≥ 3,5. O perfil de expressão destes genes em todas as amostras, permitiu classificar perfeitamente o grupo normal e o grupo tumoral, utilizando o método de análise computacional *Self Organizing Map* (SOM). Além disso, utilizando o método de busca supervisionada *Support Vector Machine* (SVM) identificamos, entre os 80 cDNAS, inúmeros trios de genes, cujos perfis de expressão em uma única amostra também podem ser utilizados para definir sua origem normal ou tumoral. Um dos trios, formado pela β-catenina, clatrina e receptor-α de ácido retinóico, é composto por genes que possuem um papel importante na carcinogênese gástrica, como descrito no artigo. O perfil de expressão dos genes contidos neste trio também permitiu a perfeita classificação de 42 elementos de um total de 49 amostras de tecido normal e de tecido tumoral pertencentes ao grupo de pacientes analisados no segundo artigo (Figura 2).

Através da técnica de RT-PCR, avaliamos o nível de expressão de 10 genes contidos no painel de 80 genes identificados como sendo diferencialmente expressos, utilizando 7 amostras pareadas de mucosa gástrica normal e tumoral, distintas daquelas empregadas no primeiro artigo. Sete desses genes foram confirmados, incluindo os genes NBS1 e CLTC, RPL10, EEF1A1, TARDBP, HSPCA e uma EST (AW812624). Identificamos uma variabilidade no perfil de expressão destes genes mesmo utilizando três diferentes genes normalizadores (β-actina, α-tubulina e TBP). Esta dificuldade de normalização foi discutida por Lee e cols (Lee, PD e cols. 2002) cujo estudo mostra que muitos genes considerados clássicos "*housekeeping genes*" como o GAPDH, β-actina, α2-microglobulina, α-tubulina, G6PD, LDHA e HPRT, apresentaram uma alta variabilidade de expresão tanto entre amostras de um experimento como entre diferentes experimentos.

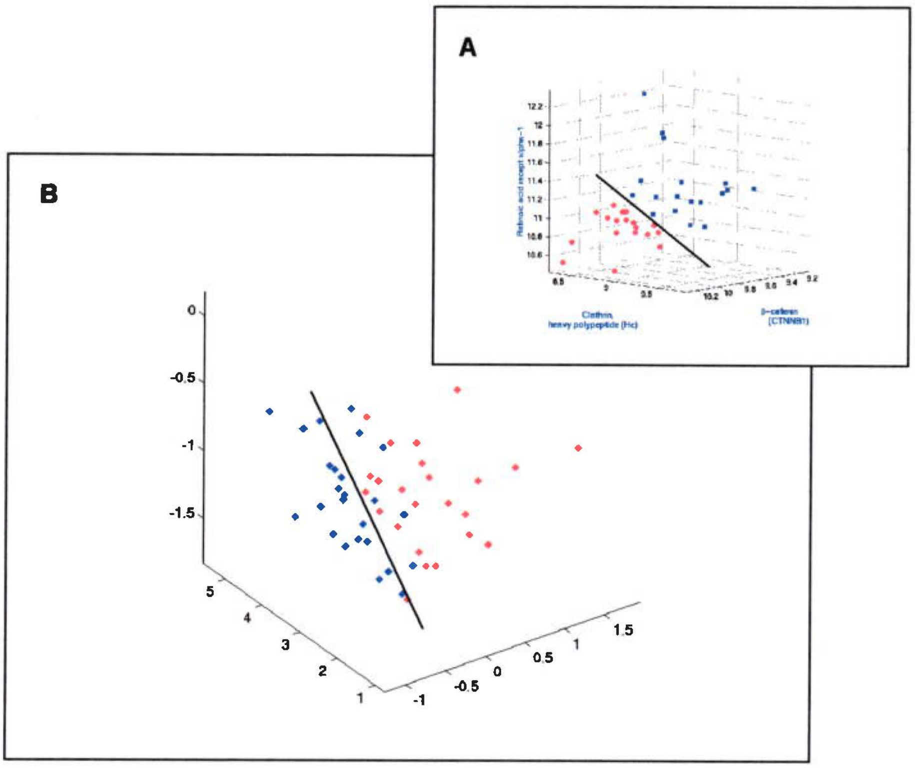


Figura 2 - Comparação da classificação molecular pelo trio de genes RAR-α1, CLTC e CTNNB1, utilizando dois grupos distintos de amostras.

Em A, mostramos a figura publicada no primeiro artigo com a perfeita classificação de 6 amostras de tecido normal (azul) e 6 amostras de tecido tumoral (vermelho), representadas em triplicata. Em B, mostramos a classificação de 25 amostras de tecido normal (azul) e 24 amostras de tecido tumoral (vermelho) utilizando o mesmo trio de genes em A.

No segundo artigo, analisamos um total de 99 amostras, representadas por mucosa normal (N), gastrite (G), mucosa com metaplasia intestinal (M) e tumor gástrico (T). Apresentamos uma lista de 42 genes expressos diferencialmente entre dois grupos quaisquer com p<0,0009, de acordo com o teste de Wilcoxon.

Comparamos, através do sinal obtido para o valor de *tnc* (positivo ou negativo), o perfil de expressão de 123 genes contidos entre os 141 genes identificados no primeiro artigo, com o perfil de expressão dos mesmos, nas amostras analisadas no segundo artigo. Os resultados são muito semelhantes entre os grupos de amostras e apenas 22 genes apresentaram um sinal oposto para o valor de *tnc*. Além da confirmação feita através do *tnc*, apresentamos na figura 3, três genes (CTNNB1, NBS1 e CLTC), cujo perfil de expressão, nas amostras do segundo artigo, está de acordo com o primeiro artigo. CTNNB1 e NBS1 são genes mais expressos em tecido tumoral e o gene CLTC foi menos expresso em tecido tumoral, segundo nossa análise.

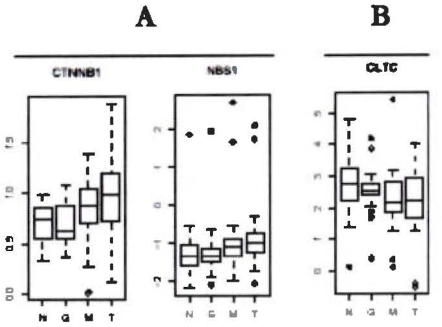


Figura 3 - Perfil de expressão dos genes CTNNB1, NBS1, CLTC, nas amostras analisadas no segundo artigo.

O perfil de expressão de CTNNB1, NBS1, CLTC está representado nos gráficos *boxplot* para as amostras N (Normais), G (Gastrite), M (Metaplasia Intestinal) e T (tumor). Em A estão agrupados os genes que apresentaram valor negativo de *tnc*, ou seja, apresentam maior expressão em tecido tumoral em relação ao tecido normal na análise do artigo 1. Em B está representado um gene que apresentou valor negativo de *tnc*, ou seja, menor expressão em tecido tumoral em relação ao tecido normal na análise do artigo 1.

Para a primeira tentativa de agrupamento das 99 amostras, selecionamos os 6 genes mais diferencialmente expressos em cada comparação, o que resultou em 18 genes distintos. Utilizando o algoritmo *K-means,* separamos as 99 amostras em quatro grupos (figura 2 do segundo artigo). As amostras correspondentes a tecidos normais e com gastrite não foram distinguíveis entre si e fizeram parte do primeiro grupo. Este resultado está de acordo com a ausência de genes diferencialmente expressos entre essas amostras como apresentados na figura 1 do segundo artigo. A maioria das amostras com metaplasia intestinal e tumorais foram separadas em dois grupos distintos, correspondendo ao segundo e ao quarto grupos. O terceiro grupo da figura contêm apenas 6 elementos que representam as quatro classes de amostras. Esta análise demonstra uma grande concordância entre a classificação molecular e a classificação histológica, mas mostra claramente a existência de amostras que, apesar de histologicamente classificadas como mucosa normal, gastrite, metaplasia intestinal ou tumor, apresentam o perfil de expressão gênica distinto de seus pares.

Para a criação de classificadores moleculares, utilizamos o método Discriminador Linear de Fisher para fazer uma busca exaustiva por trios de genes que pudessem ser usados na distinção N, G, M e T, através de comparações pareadas (NxT, GxT, MxT, NxM e GxM). Os trios encontrados para avaliar o total de 99 amostras foram ordenados de acordo com o SVD, do inglês: *Singular Values Discriminator.* O SVD avalia a razão da distância entre e dentro de grupos. Os melhores classificadores moleculares são aqueles que possuem alto SVD, pois apresentam a maior distância entre dois grupos diferentes e a distribuição mais densa entre as amostras de cada grupo. Encontramos 100 genes distintos, contidos no conjunto de trios que separam NxT, GxT e MxT. Apenas o gene CTSB (catepsina B) foi encontrado em comum nos três conjuntos de trios. Isto demonstra que os genes que distinguem tecido normal do tecido tumoral são, em sua maioria, diferentes dos genes que distinguem gastrite ou metaplasia intestinal do tecido tumoral. A lista contendo os 100 trios de genes que compõem o classificados NxT está apresentada na Tabela 1.

Nos classificadores NxT identificamos discrepância em relação à classificação histológica de 7 amostras não-tumorais que foram classificadas como tumorais GF63-N, GF59 e GH880-G BIO133, GH828, BIO124 e BIO136-M (Tabela 2). As amostras GH880 (G) e GH828 (M) também foram classificadas com tumor no classificador MxT. As amostras GF63 (N), GF59 (G), BIO133 e BIO136 (M) foram agrupadas na análise de cluster (figura 2, segundo artigo) dentro de um grupo que contêm apenas 6 elementos, representando amostras N, G, M e T. Assim, duas análises distintas (classificadores e análise de agrupamento) mostraram que algumas amostras possuem alterações moleculares, diferenciando-as das demais que compõem o mesmo grupo. Este erro classificatório pode estar refletindo alterações moleculares que ainda não influenciaram as características morfológicas do tecido e podem ser úteis no diagnóstico molecular precoce de lesões pré-malignas.

Tabela 1 - Trios de genes que classificam as amostras como normal ou tumoral

Esta tabela contêm 100 trios de genes que classificam as amostras como normal ou tumoral. Trios com maior valor SVD possuem a melhor separação entre os dois grupos de amostras.

Gene 1	Gene 2	Gene 3	SVD	Gene 1	Gene 2	Gene 3	SVD
RPL14	XBP1	COL4A1	14.94	HS.323910	XBP1	COL4A1	13.265
CTSB	XBP1	COL4A1	14.222	U1SNRNPBP	XBP1	COL4A1	13.263
KRT17	XBP1	COL4A1	14.124	IL8	XBP1	COL4A1	13.248
XBP1	COL4A1	HS.82318	14.015	NME1	XBP1	COL4A1	13.246
HS.232400	XBP1	COL4A1	13.905	MGB1	XBP1	COL4A1	13.243
XBP1	HS.21330	COL4A1	13.826	XBP1	COL4A1	CASP3	13.23
NoM.5677	HS.297095	COL4A1	13.823	XBP1	COL4A1	HS.351348	13.227
ELK1	XBP1	COL4A1	13.805	XBP1	COL4A1	DAF	13.219
XBP1	COL4A1	GCP3	13.789	TUBB2	XBP1	COL4A1	13.214
MA4	XBP1	COL4A1	13.768	FLI1	XBP1	COL4A1	13.213
XBP1	KIAA0970	COL4A1	13.753	XBP1	COL4A1	COL4A2	13.203
LCK	XBP1	COL4A1	13.674	XBP1	LOC56997	COL4A1	13.198
CTSB	TR	COL4A1	13.655	LAMB1	XBP1	COL4A1	13.196
XBP1	COL4A1	HGF	13.654	XBP1	COL4A1	HS.177781	13.195
XBP1	COL4A1	MMP10	13.644	HS.84905	XBP1	COL4A1	13.19
XBP1	COL4A1	HS.169610	13.589	XBP1	COL4A1	HS.5648	13.182
XBP1	COL4A1	TYMS	13.538	SERPINB2	XBP1	COL4A1	13.175
XBP1	COL4A1	NSEP1	13.533	XBP1	COL4A1	PCNA	13.164
VAV1	XBP1	COL4A1	13.528	XBP1	COL4A1	NRG1	13.162
LAMA2	XBP1	COL4A1	13.524	TARDBP	XBP1	COL4A1	13.158
DTR	XBP1	COL4A1	13.522	TUBB	XBP1	COL4A1	13.154
VEGFB	XBP1	COL4A1	13.507	XBP1	COL4A1	SERPINE1	13.153
NF1	XBP1	COL4A1	13.501	CTSB	COL4A1	RPL10	13.149
XBP1	KRT18	COL4A1	13.483	XBP1	COL4A1	NoM.12614	13.146
XBP1	COL4A1	HS.89603	13.483	HS.83583	XBP1	COL4A1	13.141
HS.85112	XBP1	COL4A1	13.45	FN1	XBP1	COL4A1	13.14
HS.81134	XBP1	COL4A1	13.447	CTSB	TOP2A	COL4A1	13.136
CTSB	TARDBP	HS.303023	13.437	CTSB	HS.303023	HS.80976	13.133
MMP14	XBP1	COL4A1	13.436	NME2	XBP1	COL4A1	13.132
XBP1	COL4A1	CASP9	13.436	POLR2I	CLTC	COL4A1	13.103
DPH2L1	XBP1	COL4A1	13.417	XBP1	DKFZP547I224	COL4A1	13.103
XBP1	COL4A1	KRT4	13.415	CTSB	ZFP95	COL4A1	13.089
XBP1	COL4A1	HS.4745	13.414	XBP1	COL4A1	HS.193716	13.087
NoM.4275	XBP1	COL4A1	13.413	KRT7	XBP1	COL4A1	13.082
XBP1	COL4A1	TGFB2	13.406	CTSB	TR	COL4A2	13.081
HS.326445	COL4A1	IGFBP6	13.398	GJA1	XBP1	COL4A1	13.08
SNAP25	XBP1	COL4A1	13.396	XBP1	COL4A1	HS.31210	13.08
CLTC	XBP1	COL4A1	13.38	POLR2I	CTSB	POLR2A	13.075
XBP1	COL4A1	PLAU	13.372	HS.195850	XBP1	COL4A1	13.072
HS.326445	COL4A1	TNFRSF6	13.364	XBP1	COL4A1	KPNA2	13.066
XBP1	COL4A1	JUN	13.353	XBP1	COL4A1	KRT6B	13.063
JUND	XBP1	COL4A1	13.347	XBP1	HS.348423	COL4A1	13.056
FGFR4	XBP1	COL4A1	13.338	HS.319378	XBP1	COL4A1	13.053
XBP1	COL4A1	FGFR2	13.309	XBP1	COL4A1	PABPC4	13.05
XBP1	BIRC5	COL4A1	13.295	CTSB	NoM.9304	COL4A1	13.046
XBP1	NoM.146	COL4A1	13.292	XBP1	HS.291904	COL4A1	13.044
ITGB4	XBP1	COL4A1	13.285	XBP1	COL4A1	HS.349305	13.038
XBP1	COL4A1	LGALS3BP	13.284	XBP1	COL4A1	HS.222015	13.037
IL1B	XBP1	COL4A1	13.278	POLR2A	COL4A1	HS.7243	13.034

Tabela 2 - Classificação de 99 amostras segundo 100 trios de genes que separam NxT

As amostras analisadas no segundo artigo foram classificadas de acordo com o trios que separam NxT. A escala de 0-100 refere-se ao número de trios que classificam uma amostra como tumoral.

Normal	Trios	Gastrite	Trios	Metaplasia	Trios	Tumor	Trios
BIO102	0	BIO114	0	BIO111	4	BIO108	100
BIO103	0	BIO123	0	BIO133	88	BIO93	100
BIO107	0	BIO131	0	BIO138	6	GF48	100
BIO109	0	BIO46	0	BIO45	0	GF50	100
BIO118	0	BIO50	0	BIO54	0	GF52	97
BIO48	0	BIO57	0	BIO69	4	GF54	100
BIO61	0	BIO63	1	GF46	5	GF56	100
BIO62	0	BIO72	0	GH828	99	GF58	99
BIO90	0	BIO73	0	GH834	1	GH695	100
BIO94	0	BIO98	0	GH966	0	GH831	99
GF47	0	BIO99	0	GH976	25	GH833	100
GF57	0	GF49	0	BIO195	0	GH843	100
GH882	1	GF53	1	BIO215	0	GH877	100
GH883	1	GF59	66	BIO124	44	GH881	100
GH910	1	GH880	98	BIO136	91	GH965	91
GH968	0	GH980	0	BIO140	5	GH967	100
GH972	0	BIO216	0	BIO141	5	GH969	100
GH978	0	BIO077	0	BIO145	4	GH973	100
BIO151	0	BIO113	5	BIO148	0	GH977	100
BIO213	0	BIO161	5	BIO174	1	GH979	100
BIO220	0	GH840	1	BIO226	5	GF62	100
GF63	93			GH832	3	GF68	98
GF67	0					GF70	100
BIO199	0					GF72	99
BIO218	0					BIO228	100
BIO230	0					GF066	100
BIO231	0					GH971	92
GH970	0					GH975	100

É importante analisar amostras de metaplasia intestinal que apresentam perfil molecular semelhante ao tumor. Apesar de ser uma lesão com maior risco de malignização, nem todas evoluem para o câncer, além disso, o perfil molecular desta lesão não tem sido muito estudado. Alterações moleculares que se assemelham ao câncer podem indicar o potencial maligno de algumas lesões, como discutido acima. As amostras, BIO133, BIO124, BIO136 e GH828 foram classificadas como tumor por vários classificadores moleculares NXT. Apesar da revisão do diagnóstico histológico dessas amostras não demonstrar nenhuma evidência de alteração maligna, nossas análises indicam que existem alterações moleculares semelhantes ao tumor. Baseado neste fato, sugerimos que estas amostras possuem um maior potencial de malignização e será fundamental acompanhar estes pacientes com maior rigor. A possibilidade de acompanhamento de pacientes com metaplasia intestinal com o comportamento descrito acima é a principal perspectiva do nosso trabalho.

Finalmente, identificamos genes diferencialmente expressos entre adenocarcinoma do tipo intestinal e adenocarcinoma do tipo difuso, para entender melhor os mecanismos moleculares que levam ao surgimento de cada um destes tipos histológicos. O gene VHL, responsável pela síndrome de Von Hippel Lindau, foi relacionado tanto ao tipo intestinal quanto ao tipo difuso e ainda não havia sido relacionado ao câncer gástrico. A síndrome de Von Hippel Lindau está associada ao desenvolvimentos de tumores malignos e benignos em diversos órgãos. Os tipos de tumores mais freqüentes são os hemangiblastomas de retina e cerebelo, carcinoma de células renais, feocromocitoma e tumores pancreáticos. O câncer gástrico ainda não havia sido correlacionado ao perfil de expressão do gene VHL.

5. CONCLUSÕES

Utilizamos a metodologia de *cDNA array* para analisar o perfil de expressão gênica em amostras de mucosa gástrica normal, gastrite, metaplasia intestinal e adenocarcinoma de estômago.

Através de testes estatísticos não paramétricos, comparamos a expressão gênica entre duas amostras e identificamos os genes diferencialmente expressos entre mucosa gástrica normal, gastrite, mucosa gástrica com metaplasia intestinal e tumoral.

Através de análises de agrupamento identificamos padrões de expressão gênica capazes de separar os diferentes grupos de amostras.

Desenvolvemos classificadores moleculares compostos por trios de genes capazes de distinguir entre NxT, GxT, MxT, NxM e GxM.

Identificamos amostras de metaplasia intestinal cujo perfil molecular foi distinto de outras amostras de metaplasia intestinal e semelhante ao tumor.

Identificamos genes que alteram o perfil de expressão de maneira progressiva e genes diferencialmente expressos entre adenocarcinoma do tipo intestinal do tipo difuso. Entre os genes identificados, destaca-se o gene VHL (Von Hippel Lindau) que foi relacionado aos dois tipos de tumor e ainda não tinha sido relacionado ao câncer gástrico.

6. PERSPECTIVAS

Neste trabalho, apresentamos um grupo de classificadores moleculares que distinguem o tecido normal, tumoral, gastrite e metaplasia intestinal entre si. Em continuidade a este estudo, é necessário testar esses classificadores em outro conjunto de amostras e avaliar se os mesmos são aplicáveis para qualquer amostra da população. Também é importante correlacionar a presença de *Helicobacter pylori* com a classificação molecular das lesões na mucosa gástrica, incluindo o adenocarcinoma, tanto do tipo intestinal quanto do tipo difuso.

Pacientes que possuem lesões com maior risco para o desenvolvimento de câncer gástrico, como a metaplasia intestinal, são acompanhados periodicamente. Os classificadores construídos neste trabalho permitem, em princípio, detectar alterações moleculares em metaplasia intestinal que indicam um potencial maligno. O acompanhamento dos pacientes, com metaplasia intestinal e com estas alterações moleculares, poderá avaliar se os mesmos irão desenvolver tumor gástrico com maior freqüência em relação a outros que não apresentaram tais alterações. Além disso, os classificadores aqui apresentados podem ser avaliados em relação aos subtipos de metaplasia intestinal (Tipo I, II e III) que estão associados ao maior ou menor risco para o desenvolvimento de tumor. Finalmente, validando-se essa ferramenta de diagnóstico molecular, será possível contribuir, significativamente, para o diagnóstico precoce do tumor.

Identificamos inúmeros genes, cujo perfil de expressão está alterado em diferentes lesões na mucosa gástrica. Paralelamente à identificação de alterações em nível de expressão gênica, é necessário avaliar a tradução gênica, através da identificação da proteína codificada por um determinado gene. Estudos imunohistoquímicos já foram iniciados em colaboração com o Departamento de Anatomia Patológica do Hospital do Câncer, utilizando a técnica de *tissue array* para analisar mucosa gástrica normal, gastrite, metaplasia intestinal, além de adenocarcinoma gástrico do tipo intestinal e do tipo difuso (Figura 4). Já confirmamos a expressão alterada de alguns dos genes descritos nesse trabalho, como a β-catenina. Além disso, estes marcadores estão sendo utilizados para estudar o tumor primário, o tumor metastático e também a via de metastatização tumoral (Tese de doutorado em andamento do aluno Alberto Siqueira Igreja). Um dos genes alvo de interesse é o VHL, que possui um papel importante no desenvolvimento da Síndrome de Von Hippel Lindau, porém ainda não foi relacionado ao câncer gástrico.

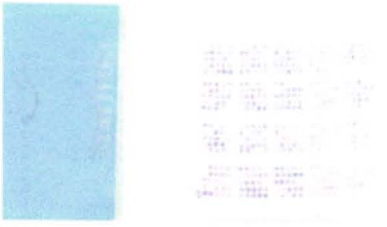


Figura 4 – *Tissue array* de amostras de estômago.

Exemplo da lämina de *tissue array* corada com Hematoxilia e Eosina. Este arranjo contëm 25 amostras de mucosa gastrica normal, 50 amostras de gastrite, 25 amostras de metaplasia intestinal, 100 amostras de adenocarcinoma intestinal do tipo intestinal e 100 amostras de adenocarcinoma do tipo difuso.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Boussioutas A, Li H, Liu J et. al. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res 2003 63: 2569-2577.

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: 12103-12108.

Correa P, Chen VW. Gastric cancer. Cancer Surv 1994 19-20: 55-76.

Devesa SS, Blot WJ, Fraumeni JF, Jr. Changing patterns in the incidence of esophageal and gastric carcinoma in the United States. Cancer 1998 83: 2049-2053.

Dias NE, 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: 3491-3496.

El Omar EM, Carrington M, Chow WH et. al. Interleukin-1 polymorphisms associated with increased risk of gastric cancer. Nature 2000 404: 398-402.

El Rifai W, Frierson HF, Jr., Harper JC, Powell SM, Knuutila S. Expression profiling of gastric adenocarcinoma using cDNA array. Int J Cancer 2001 92: 832-838.

El Rifai W, Powell SM. Molecular biology of gastric cancer. Semin Radiat Oncol 2002 12: 128-140.

Etzioni R, Urban N, Ramsey S et. al. The case for early detection. Nat Rev Cancer 2003 3: 243-252.

Guilford P, Hopkins J, Harraway J et. al. E-cadherin germline mutations in familial gastric cancer. Nature 1998 392: 402-405.

Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000 100: 57-70.

Hasegawa S, Furukawa Y, Li M et. al. Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23,040 genes. Cancer Res 2002 62: 7012-7017.

Inoue H, Matsuyama A, Mimori K, Ueo H, Mori M. Prognostic score of gastric cancer determined by cDNA microarray. Clin Cancer Res 2002 8: 3475-3479.

Ji J, Chen X, Leung SY et. al. Comprehensive analysis of the gene expression profiles in human gastric cancer cell lines. Oncogene 2002 21: 6549-6556.

Jung MH, Kim SC, Jeon GA et. al. Identification of differentially expressed genes in normal and tumor human gastric tissue. Genomics 2000 69: 281-286.

Lauren P. The two histological main types of gastric carcinoma: difuse and so-called intestinal-type carcinoma. Acta Pathol Microbiol Scand 1965 64: 31-49.

Lee S, Baek M, Yang H et. al. Identification of genes differentially expressed between gastric cancers and normal gastric mucosa with cDNA microarrays. Cancer Lett 2002 184: 197-206.

Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet 2000 1: 48-56.

Liu LX, Liu ZH, Jiang HC et. al. Profiling of differentially expressed genes in human gastric carcinoma by cDNA expression array. World J Gastroenterol 2002 8: 580-585.

Meireles SI, Carvalho AF, Hirata R et. al. Differentially expressed genes in gastric tumors identified by cDNA array. Cancer Lett 2003 190: 199-211.

Ming SC. Gastric carcinoma. A pathobiological classification. Cancer 1977 39: 2475-2485.

Ministério da Saúde. Estimativa da Incidência e Mortalidade por câncer no Brasil. Rio de Janeiro: INCA 2003.

Oliveira FJ, Ferrao H, Furtado E, Batista H, Conceicao L. Early gastric cancer: Report of 58 cases. Gastric Cancer 1998 1: 51-56.

Peek RM, Jr., Blaser MJ. Helicobacter pylori and gastrointestinal tract adenocarcinomas. Nat Rev Cancer 2002 2: 28-37.

Pusztai L, Ayers M, Stec J, Hortobagyi GN. Clinical application of cDNA microarrays in oncology. Oncologist 2003 8: 252-258.

Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol 2002 20: 1932-1941.

Stadtlander CT, Waterbor JW. Molecular epidemiology, pathogenesis and prevention of gastric cancer. Carcinogenesis 1999 20: 2195-2208.

Wang J, Chen S. Screening and identification of gastric adenocarcinoma metastasis-related genes using cDNA microarray coupled to FDD-PCR. J Cancer Res Clin Oncol 2002 128: 547-553.

Yeatman TJ. The future of clinical cancer management: one tumor, one chip. Am Surg 2003 69: 41-44.

ANEXOS

Anexo 1: Differentially expressed genes in gastric tumors identified by cDNA array.

Cancer Lett, 2003; 190(2):199-211.

Anexo 2: Molecular Classifiers for Gastric Cancer and Nonmalignant Diseases of the Gastric Mucosa.

Cancer Research 64, 1255-1265, February 15, 2004

ANEXO 1

DIFFERENTIALLY EXPRESSED GENES IN GASTRIC TUMORS IDENTIFIED BY CDNA ARRAY.

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT

Cancer Letters 190 (2003) 199–211

CANCER Letters

www.elsevier.com/locate/canlet

Differentially expressed genes in gastric tumors identified by cDNA array

Sibele I. Meireles[a,b], Alex F. Carvalho[b], Roberto Hirata Jr[c,d], André L. Montagnini[a], Waleska K. Martins[b], Franco B. Runza[b], Beatriz S. Stolf[b,e], Lara Termini[a], Chamberlein E.M. Neto[b], Ricardo L.A. Silva[b], Fernando A. Soares[a], E. Jordão Neves[c], Luiz F.L. Reis[a,b,*]

[a] *Hospital do Câncer A.C. Camargo, Rua Professor Antonio Prudente 109, 01509-010 São Paulo, SP, Brazil*
[b] *Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente 109, 01509-010 São Paulo, SP, Brazil*
[c] *Instituto de Matemática e Estatística and BIOINFO, Universidade de São Paulo, São Paulo, SP, Brazil*
[d] *SENAC College of Computer Science and Technology, São Paulo, SP, Brazil*
[e] *Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil*

Received 21 August 2002; received in revised form 8 October 2002; accepted 10 October 2002

Abstract

Using cDNA fragments from the FAPESP/LICR Cancer Genome Project, we constructed a cDNA array having 4512 elements and determined gene expression in six normal and six tumor gastric tissues. Using *t*-statistics, we identified 80 cDNAs whose expression in normal and tumor samples differed more than 3.5 sample standard deviations. Using Self-Organizing Map, the expression profile of these cDNAs allowed perfect separation of malignant and non-malignant samples. Using the supervised learning procedure Support Vector Machine, we identified trios of cDNAs that could be used to classify samples as normal or tumor, based on single-array analysis. Finally, we identified genes with altered linear correlation when their expression in normal and tumor samples were compared. Further investigation concerning the function of these genes could contribute to the understanding of gastric carcinogenesis and may prove useful in molecular diagnostics.

Keywords: Gene expression; Gastric cancer; cDNA array; Molecular marker

1. Introduction

During the last 10 years, the incidence of gastric cancer is declining worldwide but, nevertheless, it is still a tumor of high incidence [1]. Worldwide, tumors of the stomach are the fourth in incidence and second in cancer-related deaths (IARC home page: http://www.dep.iarc.fr).

At the molecular level, little is known about the mechanisms involved in gastric carcinogenesis. As established for tumors in general, it was proposed that, for gastric adenocarcinomas, accumulation of genetic alterations in a multistep fashion would correlate with disease progression and differences between diffuse and intestinal type adenocarcinomas would be linked to distinct mutation pathways [2]. These genetic

* Corresponding author. Tel.: +55-11-3207-4922; fax: +55-11-3207-7001.

E-mail address: lreis@ludwig.org.br (L.F.L. Reis).

0304-3835/02/$ - see front matter © 2002 Elsevier Science Ireland Ltd. All rights reserved.
doi:10.1016/S0304-3835(02)00587-6

alterations can be either chromosomal aberrations or confined to mutations in one or more genes.

For chromosomal aberrations, several studies applying comparative genomic hybridization identified the 20q region as the most frequent gain. Other frequent gains were observed at 6p, 7q, 8q, and 17q and losses were at 4q, 5q, 9p, and 18q [3–5]. A high level amplification of the region 17q12-21 was observed in the intestinal type of tumors [4] and fluorescent in situ hybridization analyses using probes for either gastrin or ERBB2 revealed that both genes were simultaneously amplified [6]. Using at least two highly polymorphic microsatellite markers for each nonacrocentric chromosomal arm, an exhaustive scanning for loss of heterozygosity (LOH) revealed significant LOH at several loci such as 3p, 4p, 5q, 8p, 9p, 13q, 17p, and 18q, suggesting the presence of potential tumor suppressor genes [7].

Altered expression of genes known to play a role in oncogenic transformation has also been detected in gastric cancer, either in freshly isolated tissue or in cell lines. It is well documented that mutations in the p53 gene is a frequent event in gastric cancer and detected in as much as 50% of advanced cases [8,9]. Interestingly, p53 knockout mice, carrying either one or two mutated alleles appear to be more sensitive to experimental *Helicobacter* infection [10]. Other genes with altered expression or frequently amplified in gastric cancer are cErbB2 and c-met [11], TGF-βII receptor [12], e-Cadherin [13,14], β-Catenin [15], among others.

Another tumor type of gastric cancer that accounts for 2% of the cases is designated GIST (gastrointestinal stromal tumor) and comprises the majority of gastrointestinal mesenchymal tumors (reviewed by Miettinen and colleagues [16]). At the molecular levels, GIST is commonly associated with losses in chromosomes 14 and 22 whereas gain or high-level amplification is observed in 3q, 8q, 5p, and Xp [17,18]. Mutations in the c-Kit gene have been frequently associated with GIST [19,20] and these tumors showed a remarkably homogeneous gene expression profile [21].

More recently, several groups described the utilization of high throughput methodology in order to identify genes differentially expressed in gastric cancer [3,22–24].

The FAPESP/LICR Human Genome Cancer Project finished a major effort in sequencing over 1 100 000 ORESTES (open reading frame ESTs) derived from various tumor types and a significant proportion of yet unknown sequences were generated [25]. Taking advantage of the clone collection generated by this project, we constructed a cDNA array and searched for genes differentially expressed in normal versus tumor gastric mucosa and searched for differentially expressed genes that could distinguish between normal and tumor tissues. Detailed analysis of the genes could help in understanding the molecular events related to gastric carcinogens and also, could bring some improvement towards diagnostics and prognostics of gastric cancer.

2. Materials and methods

2.1. Tissue specimens and RNA extraction

Fresh tissues from surgically resected gastric cancers were collected by the Gastric Surgery Department from Hospital do Câncer AC Camargo, São Paulo. All patients signed an informed consent and the project was approved by the in-house ethics committee. Six gastric tumors (four adenocarcinomas and two gastrointestinal stromal tumors) and six, not paired, disease-free gastric mucosa were used. Disease-free tissue from tumor margins or obtained from radical gastrectomy was considered as normal tissue. At the time of RNA extraction, histological confirmation of normal or tumor status was performed by hematoxylin–eosin staining of frozen sections. The frozen sections were also used for dissection of samples in order to enrich for tumor cells (see Fig. 1, upper panels). Only samples with at least 70% of tumor tissue and negative for infiltrating inflammatory cells were further processed. In the case of normal samples, only gastric mucosa was used. Total RNA was extracted using TRIzol Reagent (Life Technologies, Grand Island, NY) following the procedure recommended by manufacturer.

2.2. Production of cDNA arrays

A collection of 4512 ORESTES fragments derived from the FAPESP/LICR Human Cancer Genome Project [25] was immobilized in nylon membranes.

As positive control for labeling and hybridization, we spotted, in serial dilutions, a cDNA corresponding to a fragment of the lambda phage Q gene. Bacterial clones were grown in LB medium containing 7.5% glycerol and, from each clone, the cDNA insert was amplified by polymerase chain reaction (PCR), using M13 reverse and forward primers in a final volume of 100 μl. From all 4512 PCR products, 5 μl were fractionated through a 1% agarose gel in order to quality control DNA products and the remaining 95 μl were purified with QIAquick 96 PCR purification kit (Qiagen) or Sephadex G50 (Amershan Pharmacia). Purified DNA was printed onto nylon membranes by Flexys robot (Genomic Solutions, UK) using a 96 flat pinhead in 96 blocks of 7 × 7 elements.

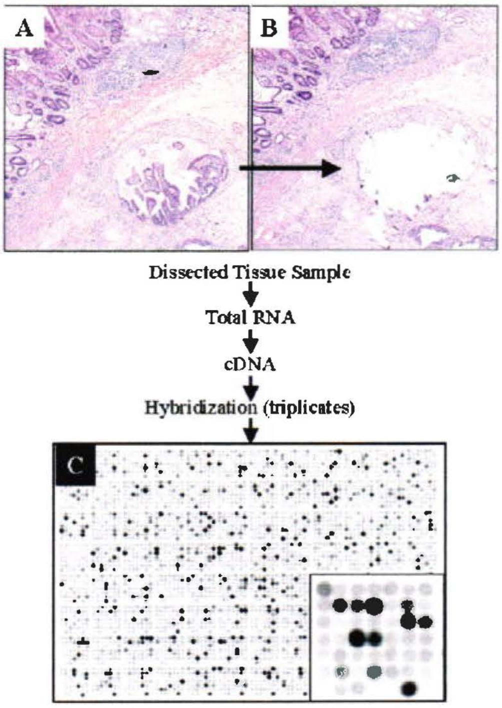


Fig. 1. Identification of genes differentially expressed in gastric tumors. Schematic representation of our experimental design. (A,B) Representation of the dissecting procedure were only the invasive portion of an adenocarcinoma was excised and processed for RNA extraction. Each RNA sample was hybridized with three identical arrays. (C) Representative image of our nylon array after hybridization with $[\alpha\text{-}^{33}P]$dCTP-labeled cDNA. Signals were captured on a phosphorimager and data was acquired by the ArrayVision software using.gel files.

2.3. Labeling, hybridization, and scanning of arrays

Thirty micrograms of total RNA were contaminated with a defined concentration of synthetic, polyadenylated RNA corresponding to the lambda phage Q gene. To this mix, we added 2.0 μg $(dT)_{15}$ in a final volume of 11 μl of water, and the mix was heated to 70 °C for 10 min and subsequently cooled to 43 °C. Reverse transcription was performed in a total volume of 50 μl using Superscript II reverse transcriptase (Life Technologies Inc.) for 2 h at 43 °C in the presence of 0.25 mM each of dATP, dGTP and dTTP, 1.66 μM-dCTP and 30 μCi of $[\alpha\text{-}^{33}P]$dCTP (3000 Ci/mmol; Amersham, UK). Subsequently, 1.5 μl 1% SDS, 1.5 μl 0.5 M EDTA and 3 μl 3 M NaOH were added and the RNA was hydrolyzed for 30 min at 65 °C and 15 min at room temperature. The solution was then neutralized with 1.5 μl 1 M Tris–HCl (pH 8) and 4.5 μl 2 M HCl. Probes were purified by gel chromatography (BioSpin 6; Bio-Rad). Prior to hybridization, the solution was boiled for 2 min, and then cooled on ice. Arrays were prehybridized for at least 1 h in 0.25 M Na_2HPO_4 (pH 7.2), SDS 7%, BSA 1%, 1 mM EDTA. Hybridization was conducted in the same buffer at 65 °C overnight [26]. For each cDNA sample, three identical membranes were hybridized simultaneously (normal samples correspond to membranes 1–18 and tumor samples correspond to membranes 19–36). The filters were then washed for 30 min in 0.5 M Na_2HPO_4 (pH 7.2), SDS 1%, 1 mM EDTA and image acquired by a phosphorimager (Molecular Dynamics Storm Imager, Molecular Dynamics, USA).

2.4. Data acquisition

Data acquisition was performed with the ArrayVision software (Amersham, UK), using.gel files. To quantify signal intensities of the hybridized spots, a template composed by equal-sized ellipses were drawn around all spots. Following the identification of the spots, the software calculated the spot-intensity value and array background intensity.

2.5. Data normalization

The background from a given array was subtracted from all 4512 spot-intensity values and we considered in our analysis only the 4388 genes with positive background-corrected values across all 36 arrays. Next we normalized the data in all arrays using total energy (4388 spots) on each given array.

2.6. Statistical analysis

Single genes with difference in expression when comparing normal and tumor samples were identified by their t-values, denoted by t_{nc}, which is the difference between Normal and Tumor sample-mean log-transformed gene expressions, standardized by the corresponding sample standard deviation. For a given gene, say gene k, and $j = 1, \ldots, 18$, let N_j^k denote its log-transformed gene expression on the j^{th} normal sample and T_j^k its log-transformed gene expression on the j^{th} tumor sample. The t_{nc} value for gene k is computed as follows:

$$t_{nc}^{k} = \frac{\overline{N^k} - \overline{T^k}}{\sqrt{\dfrac{S_{N^k}^2}{18} + \dfrac{S_{T^k}^2}{18}}}$$

where $\overline{N^k}$ (respectively $\overline{T^k}$) denotes gene k sample mean expression value in normal (tumor) arrays, S_{N^k} (respectively S_{T^k}) denotes its sample standard deviation in normal (tumor) arrays. We choose the normalization term for t_{nc} as in Ref. [27] even though our experimental setting is different because it penalises strongly replica measurement errors and therefore provides a simple and yet stringent statistics to evaluate differences in gene expression.

Elements with t_{nc} values equal or higher than 3.5 in absolute values were considered as differentially expressed. This set of cDNAs was then analyzed by Self-Organizing Map (SOM) and hierarchical clustering algorithms, both implemented in Matlab (MathWorks) neural networks and statistics toolboxes.

To find pairs and trios of genes that would allow perfect linear separation of Normal and Tumor samples we used a supervised learning technique known as Support Vector Machines, also implemented in Matlab (Cawley, G.C., Support Vector MachineToolbox v0.50, http://theoval.sys.uea.ac.uk/ ~ gcc/svm/toolbox, Support Vector Machine toolbox for Matlab Version 2.4, August, 2001, copyright Anton Schwaighofer (2001) mailto: anton.schwaighofer@gmx.net).

Write $(N^k = N_1^k, N_2^k, \ldots N_{18}^k)$ and $(T^k = T_1^k, T_2^k, \ldots T_{18}^k)$ for the vectors of expressions of gene k, respectively, among normal and tumor samples. To look for pairs of genes whose coordinated patterns of expression would change in comparing the two conditions we computed, for each pair of genes k and l, their Pearson linear correlation coefficient among normals, $\text{corr}(N^k, N^l)$, and among tumors, $\text{corr}(T^k, T^l)$.

3. Results

In order to determine the profile of gene expression in gastric tissues, we isolated total RNA from six tumor samples and from six samples of disease free gastric mucosa. For each sample, three identical nylon arrays were simultaneously hybridized, giving 18 membranes corresponding to normal tissue and 18 membranes corresponding to tumor tissues. Fig. 1 represents a scheme of our experimental design.

3.1. Identification of 80 cDNAs differentially expressed in gastric cancer

The data obtained from all 36 membranes were normalized by total energy as described in Section 2. Therefore, after normalization, all our 36 arrays have the same total expression values and one can meaningfully compare gene expressions from different arrays [27,28].

With normalized data, we computed the t-statistic, t_{nc}, for each single cDNA. In Fig. 2A, we plotted data from all 4388 cDNA clones based on their t_{nc} value. Fig. 2B represents the histogram with the t_{nc} values and in Fig. 2C we represent a quantile–quantile plot of this data versus theoretical quantiles from a normal distribution. The heavy tails of the empirical distribution of t_{nc} indicate the presence of several genes whose expression levels differ between normal and tumor samples.

Before the application of more elaborated, but also more computer-intensive, exploratory methods, it is quite natural to first select a smaller subset of genes to

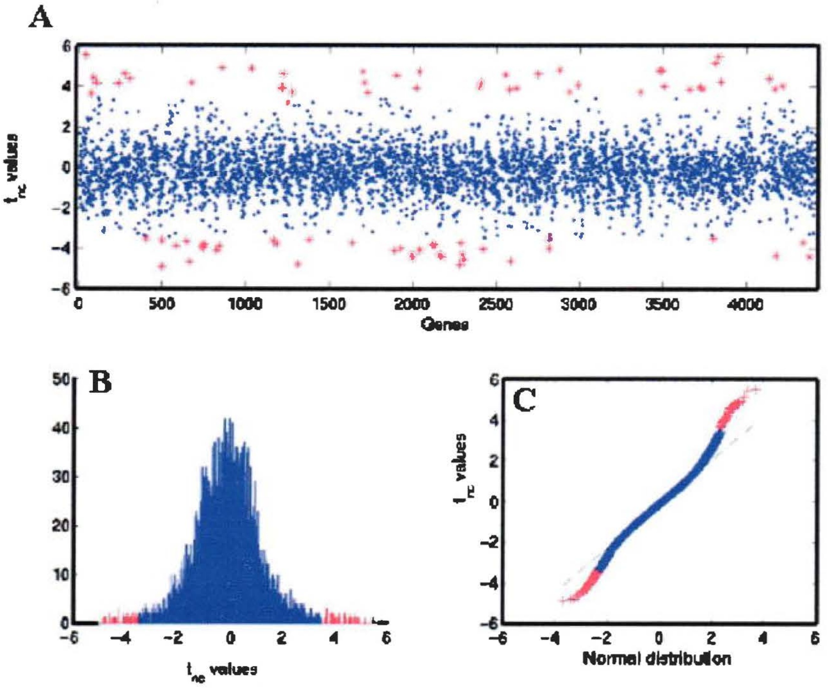


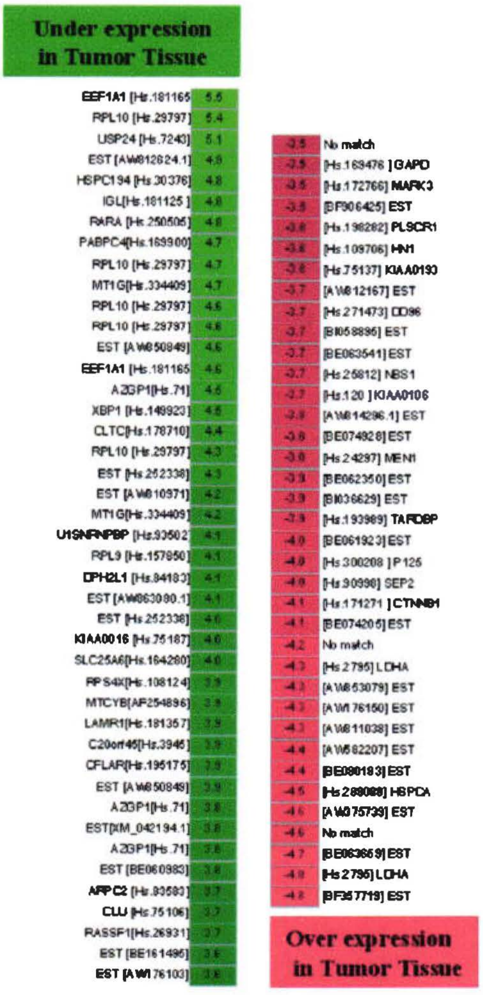
Fig. 2. Distribution of t_{nc} values for 4388 cDNAs tested in gastric tumors: Normalized data from 36 arrays hybridized with complex cDNA probes derived from normal or tumor samples were used to compute the t_{nc} value for each of the 4388 cDNAs. (A) Representation of the t_{nc} value for each single cDNA. In red, we represent 80 cDNAs with t_{nc} greater than 3.5 in absolute values. (B) Histogram representing data from (A). (C) Quantile–quantile plot of data represented in (A) against the expected value from a normal distribution.

deal with. This important step, sometimes called feature selection [29,30], was done here with the help of t_{nc} as genes with larger t_{nc}, in absolute value, are good candidates for playing a role in carcinogenesis as well as in the discrimination among normal and tumor tissues. We arbitrarily choose a threshold of 3.5 for t_{nc} we found a set of 80 cDNAs, 43 with t_{nc} larger than 3.5 (indicated in green in Fig. 3) and 37 with t_{nc} smaller than -3.5 (indicated in red in Fig. 3). All these 80 cDNAs were sequence verified.

In Fig. 3, we have a graphic representation of all 80 differentially expressed cDNAs with their respective t_{nc} value. As can be observed, five genes are represented by two or more distinct cDNA fragments. Ribosomal protein L 10 (RPL10) is represented by five cDNA clones, α2-glycoprotein 1 is represented by three clones, and metallothionein IG, Elongation Factor 1-α1, and lactate dehydrogenase A are represented by two clones. Clones representing the same gene showed very similar t_{nc} values and appeared together in the same side of Fig. 3, confirming the reproducibility of our experimental conditions and the consistence of our statistical analysis. From these 80 cDNAs, we identified 35 known genes, 31 ESTs with no functional annotation and three ORESTES sequences not yet submitted to GenBank. If a more relaxed threshold for t_{nc} is used, namely 3 instead of 3.5, 61 extra cDNAs are identified and a list with these 141 cDNAs can be visualized in our web page (http://www.array.ludwig.org.br/gastriccancer/canlettersmeireles). The sequence of all 141 cDNAs was verified experimentally.

We selected ten genes in order to experimentally confirm their differential expression in 26 new RNA samples (13 from normal tissue and 13 from tumor tissue). The levels of mRNA were estimated by RT–

PCR followed by Southern blot and phosphorimager analysis. For normalization, we used three distinct housekeeping genes (β-actin, α-tubulin, and TBP) and, for each gene, we determined its arbitrary expression unit (ratio of signal for gene/normalizing gene). A gene was considered as confirmed when the ratio of its average expression units (normal/tumor) followed its t_{nc} value. Seven of the ten genes (RPL10, CLTC, EEF1A1, TARDBP, HSPCA, NBS1, Est AW812624) could be experimentally confirmed. Nevertheless, validation of array data by RT–PCR must take in consideration the tremendous variability of housekeeping genes [31] and, more importantly, that in our case, a gene can have a high t_{nc} value even if its fold change in rather small. Similar observations were published by [32]. For instance, in our array data, β-Catenin differs only 1.3-fold between normal and tumor samples but its t_{nc} value is 4.32 due to its small SD.



Under expression in Tumor Tissue	
EEF1A1 [Hs.181165	5.5
RPL10 [Hs.29797]	5.4
USP24 [Hs.7240]	5.1
EST [AW812624.1]	4.9
HSPC194 [Hs.30376]	4.8
IGL[Hs.181125]	4.8
RARA [Hs.250505]	4.8
PABPC4[Hs.169900]	4.7
RPL10 [Hs.29797]	4.7
MT1G[Hs.334409]	4.7
RPL10 [Hs.29797]	4.6
RPL10 [Hs.29797]	4.6
EST [AW850849]	4.6
EEF1A1 [Hs.181165	4.6
A2GP1[Hs.71]	4.5
XBP1 [Hs.149923]	4.5
CLTC[Hs.178710]	4.4
RPL10 [Hs.29797]	4.3
EST [Hs.252338]	4.3
EST [AW810971]	4.2
MT1G[Hs.334409]	4.2
U1SNRNPBP [Hs.93502	4.1
RPL9 [Hs.157850]	4.1
DPH2L1 [Hs.84183]	4.1
EST [AW863090.1]	4.1
EST [Hs.252338]	4.0
KIAA0016 [Hs.75187]	4.0
SLC25A6[Hs.164280]	4.0
RPS4X[Hs.108124]	3.9
MTCYB[AF254896]	3.9
LAMR1[Hs.181357]	3.9
C20orf45[Hs.3945]	3.9
CFLAR[Hs.195175]	3.9
EST [AW850849]	3.9
A2GP1[Hs.71]	3.8
EST[XM_042194.1]	3.8
A2GP1[Hs.71]	3.8
EST [BE060983]	3.8
ARPC2 [Hs.83583]	3.7
CLU [Hs.75106]	3.7
RASSF1[Hs.26931]	3.7
EST [BE161496]	3.6
EST [AW176103]	3.6

	Over expression in Tumor Tissue
-3.5	No match
-3.5	[Hs.169476] GAPD
-3.5	[Hs.172766] MARK3
-3.5	[BF906425] EST
-3.6	[Hs.198282] PLSCR1
-3.6	[Hs.109706] HN1
-3.6	[Hs.75137] KIAA0193
-3.7	[AW812167] EST
-3.7	[Hs.271473] CD96
-3.7	[BI058895] EST
-3.7	[BE063541] EST
-3.7	[Hs.25812] NBS1
-3.7	[Hs.120] KIAA0106
-3.8	[AW814296.1] EST
-3.8	[BE074928] EST
-3.8	[Hs.24297] MEN1
-3.9	[BE062350] EST
-3.9	[BI036629] EST
-3.9	[Hs.193989] TARDBP
-4.0	[BE061923] EST
-4.0	[Hs.300208] P125
-4.0	[Hs.90998] SEP2
-4.1	[Hs.171271] CTNNB1
-4.1	[BE074205] EST
-4.2	No match
-4.3	[Hs.2795] LDHA
-4.3	[AW853079] EST
-4.3	[AW176150] EST
-4.3	[AW811038] EST
-4.4	[AW582207] EST
-4.4	[BE090183] EST
-4.5	[Hs.289088] HSPCA
-4.6	[AW075739] EST
-4.6	No match
-4.7	[BE063659] EST
-4.8	[Hs.2795] LDHA
-4.8	[BF357719] EST

Fig. 3. Genes with differential expression in gastric tumors. Based on data presented in Fig. 2A, we list genes (with respective accession numbers indicated within brackets) and indicate their t_{nc} values. In green, genes with lower expression in tumor tissue (positive t_{nc} values); in red, genes with higher expression in tumor tissue (negative t_{nc} values).

3.2. Clustering algorithms: SOM and hierarchical

After selecting the 80 cDNA clones with absolute t_{nc} value higher than 3.5, we applied a SOM algorithm [33] to identify clusters of expression profiles according to samples. Two clusters were identified and they represented a precise separation of normal and tumor samples (data not shown). When we applied a hierarchical cluster algorithm, we observed that all replicas from a given patient are grouped together, further confirming reproducibility of our data (data not shown). Next, we applied again the SOM algorithm, now to separate genes according to their expression across all 36 membranes, into six clusters. In Fig. 4, we represent these clusters and, within each cluster, we further ordered genes according to their hierarchical distance, as indicated by each dendrogram.

3.3. Genes with coordinated pattern of expression

Next we used a supervised computer learning method called Support Vector Machine (SVM) to search for trios of genes with a coordinated pattern of expression. We searched the dataset corresponding to the 80 cDNAs with t_{nc}, in absolute value, larger than 3.5 to find trios of genes whose pattern of expression in individual membranes would be such that, when plotted on three dimensional space, a plane could be found separating perfectly the 36 data points into two groups, one with 18 normal samples and another with the remaining 18 tumor samples. We found several interesting trios with this property. One trio is composed of β-Catenin, Clathrin, and Retinoic Acid

Receptor-α (Fig. 5A) and represents genes that can be mapped into a common biochemical pathway known to be implicated in gastric carcinogenesis. Another trio is composed of Ribosomal Protein L10, Humanin, and β-Catenin (Fig. 5B).

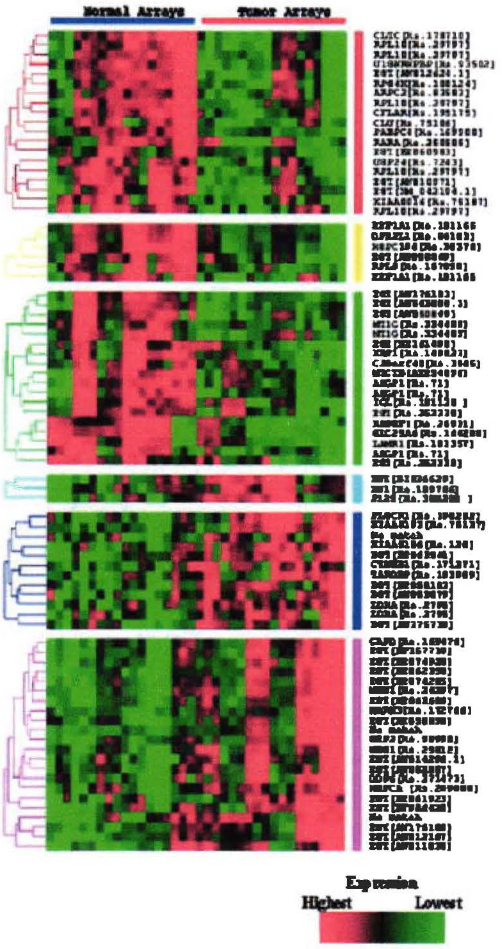


Fig. 4. Schematic representation of samples and genes clustered by Self-Organizing Map (SOM). Using the 80 cDNAs with t_{nc} higher than 3.5 in absolute values we applied SOM to cluster samples based on the expression profile of the 80 cDNAs. The resulting two clusters are represented at the top of the figure by the blue and red bars. Next, cDNAs were grouped into six clusters based on their log-transformed normalized signal intensity. For each cDNA, a maximum value is represented in bright red, minimum value in bright green and the intermediate value in black. At the left side of each cluster is a dendrogram representing hierarchical distances.

3.4. Genes with changes in their Pearson linear correlation

We also looked for pairs of genes whose pattern of expression would show changes in their Pearson linear correlation when normal and tumor samples

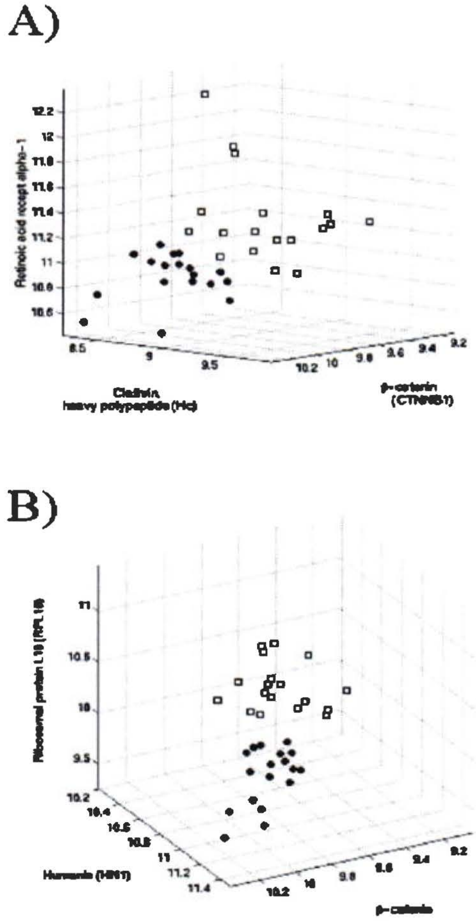


Fig. 5. Trios of genes that allow sample classification. From a gene list having cDNAs with absolute values of t_{nc} greater than 2, we applied the SVM algorithm and identified trios of genes that allowed perfect separation of all 18 normal and 18 tumor arrays. (A,B) Three-dimensional space where the log-transformed normalized signal intensity for each cDNA is plotted. Each data point in space represents one individual array; data from normal samples are represented in open squares and tumor samples are represented by dots.

were compared. For this search, we use a list of 432 cDNAs that showed t_{nc} value greater than 2.0 (in absolute value). We constructed a scatter plot where, to each pair of genes, we associate a point on the plane with coordinates given by their log-transformed normalized signal intensity on normal samples and in tumor samples. As one would expect, we found that all five clones corresponding to different cDNA fragments of Ribosomal Protein L10 had a strong positive correlation among themselves, both in normal and tumor samples. However, all five RPL10 clones showed strong positive linear correlation with MARK3 on normal samples but negative correlation on tumor samples (Fig. 6). Another gene group exhibiting this sort of correlation change is composed by Ras association domain family 1 (RASSF1), α2-glycoprotein 1 (AZGP1) and Metallothionein 1G. RASSF1 has strong positive correlation with the expressions patterns of two cDNA segments representing Metallothionein 1G in normal samples but very small correlation on tumor samples. Moreover, Metallothionein 1G has strong positive correlation with α2-glycoprotein 1 in normal samples but small correlation in tumor samples (data not shown).

4. Discussion

Gastric cancer is the second cause of cancer-related death worldwide. This observation can be explained, at least in part, by the fact that gastric cancer does not respond well to chemotherapy and/or radiotherapy, leaving surgery as the treatment of choice [34,35]. Efforts towards early diagnosis of gastric cancer are regarded as high priority since it would allow more conservative procedures, improving survival and quality of life. And as in the case of many other tumors, the molecular events related to oncogenesis of gastric cancer are not well understood. Thus, identification of genes with differential expression in gastric cancer will certainly have a positive impact in this field [36]. Such genes would be prime suspects in sharing some responsibility on the onset, development, or behavior of gastric cancer and good

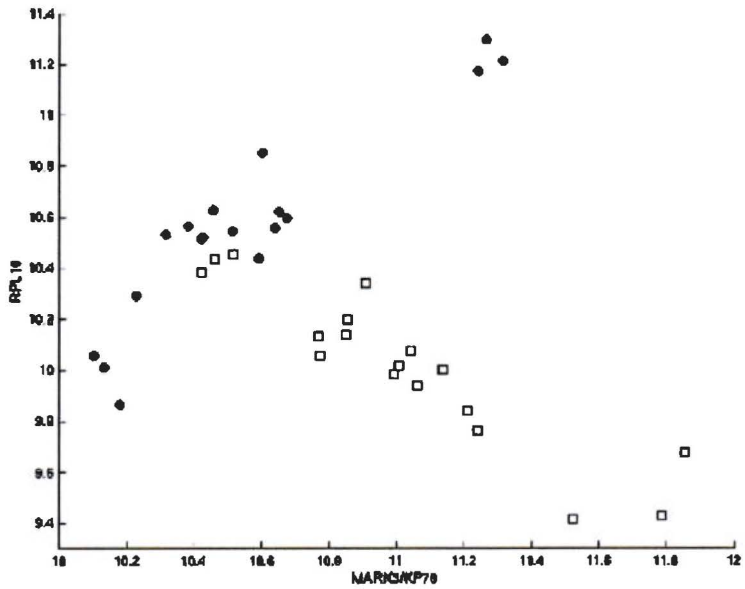


Fig. 6. Linear Pearson correlation coefficient between RPL10 and MARK3 expressions. The figure represents a scatter plot where, to each array, we associate a point on the plane representing the log-transformed normalized signal intensity for RPL10 and MARK3 on normal samples (dots) and on tumor samples (open squares).

candidates as markers for diagnosis [37]. Using cDNA arrays, we searched for genes modulated in gastric tumors and applied several statistical tools to identify correlations in their expression pattern.

We first identify single genes, whose expression would be different in normal and cancer samples. Instead of considering differences in fold expression, as is usual in the literature, we computed the *t*-statistic for each cDNA. The well-known problem in evaluating differences in expression simply by fold change is that one does not know whether a fixed value, for a given gene, is sufficiently large to characterize that gene as differently expressed without taking into account its variation of expression across all samples. By using the *t*-statistic this variation is taken into account and genes that can be considered differentially expressed would be those with larger t_{nc}, in absolute value. Among the genes with differential expression, we identified well-known tumor suppressor genes and proto-oncogenes, known to play a role in gastric cancer. We found that RASSF1, known as tumor suppressor gene [38], was underexpressed in tumor samples ($t_{nc} = 3.7$). Loss or abnormal down-regulation of RASSF1 is observed in a considerable proportion of lung, breast, ovarian, bladder, nasopharyngeal [39–44] and, more relevant, in gastric adenocarcinomas [45].

We also detected overexpression of known oncogenes. Of notice, β-Catenin was overexpressed in tumor samples ($t_{nc} = -4.1$). The role of the WNT pathway in development and oncogenesis has been widely investigated [46–48]. In the case of gastric tumors, e-Cadherins and β-Catenin are of particular importance [13,49,50]. Indeed, mutations in e-Cadherin gene has been associated with familial cases of gastric cancer [51,52].

Interestingly, two other genes involved in the WNT pathway were also found as differentially expressed in our samples. Clathrin and Retinoic Acid Receptor α were both underexpressed in tumor samples ($t_{nc} = 4.4$ and 4.8, respectively). Reduced levels of Clathrin leads to reduced recycling e-Cadherin, lowering its level at cell surface and, as consequence, more β-Catenin would be available in the cytoplasm for signaling via interaction with LEF/TCF [53,54]. Recently, it was shown that retinoic acid (RA) decreases the activity of the β-Catenin-TCF/lEF signaling pathway by inducing ubiquitin-dependent degradation of cytoplasmic β-Catenin as well as by competition with TCF for β-Catenin binding [55,56]. Thus reduced levels of Clathrin and RAR might also contribute to increased WNT signaling.

Two other genes identified as overexpressed in tumor samples might have important implications in the oncogenesis of gastric cancer, Nibrin ($t_{nc} = -3.7$) and Humanin mRNA ($t_{nc} = -3.6$). Nibrin is a member of the Mre11/Rad50/Nbs1 complex, implicated in numerous aspects of double-strand break repair, and considered as a typical tumor suppressor gene (reviewed by Wang [57]). In agreement with our data, Nibrin mRNA was also found to be augmented in GIST [21]. This is also confirmed by SAGE analysis (http://cgap.nci.nih.gov/Pathways). It is possible that, based on the findings by Paull and co-workers [58], Nibrin overexpression might favor the nucleolytic activity of the Mre11/Rad50/Nbs1 complex. Humanin was recently described as a small polypeptide that could rescue neuronal cells from specific death signals [59,60]. To the best of our knowledge, this is the first report of augmented expression of Humanin in tumor tissues and its overexpression by cancer cells could represent yet another survival signal, favoring tumor development.

Based on published observations, it is clear that molecular classification of cancer is not only feasible but also, might prove to be the method of choice to identify sub groups of a given tumor [61–65]. Having identified these 80 cDNAs, we applied other statistical tools to classify our normal and tumor samples. It has been suggested that the SOM has some important advantages for interpreting gene expression patterns, when compared to other clustering algorithms [66]. When we applied SOM to group samples, two predominant clusters of expression profile were identified and they could precisely separate normal and tumor samples (Fig. 4). Using a support vector machine algorithm, recently described as a tool to build classifiers for cancer samples [63], we performed an exhaustive search for trios of cDNAs that would allow precise separation between normal and tumor samples. We identified several trios composed by the 80 genes from Fig. 3 that, when plotted on a three-dimensional space, normal and tumor samples could be precisely separated by a plane (Fig. 5). It is possible that combination of various trios with the properties described here might have an added accuracy for molecular classification when compared to a list of differentially expressed genes as currently suggested [62,65,67]

The identification of these trios was based on their simultaneous expression levels on each given array and one could use this information to investigate whether, in such trios, the genes would fall into a common biochemical pathway or whether they belong to distinct pathways that, together, would point to some metabolic advantage for tumor cells. Indeed, the three genes from Fig. 5A, Clathrin, β-Catenin, and Retinoic Acid Receptor can all be mapped into a common pathway, as discussed earlier. In other trio (Fig. 5B), the three genes cannot be directly linked to a single pathway. RPL10, might have tumor suppressor activities and negatively regulate c-Jun activity [68]. Thus, reduced RPL10 and augmented β-Catenin in tumor samples would favor mitogenic signals, whereas elevated Humanin could provide a survival advantage for tumor cells, as mentioned above.

We also searched for genes with change in their linear Pearson correlation. This kind of analysis would allow the identification of genes whose expression occurs in a coordinated fashion in one group of samples but either are not correlated or, perhaps more interestingly, with inverse correlation in the other group. Importantly, it could be that, genes with this behavior might have low t_{nc} values in absolute numbers and thus not identified as differentially expressed. In Fig. 6, we represent the changes in linear correlation between RPL10 and MARK3. This pair of genes has a positive linear correlation in normal samples that changes to a negative linear correlation in tumor samples. As we discussed before RPL10 might function as a negative regulator c-Jun-mediated mitogenic pathway. In contrast, overexpression of β-Catenin and consequent activation of the WNT pathway activates c-Jun gene expression [69] and, possibly, MARK3 [24,70]. Hence, in tumor cells, a negative linear correlation, would favor a mitogenic signaling pathways.

Finally, it is clear that gastric adenocarcinomas and gastrointestinal stromal tumors are consequences of the transformation of different cell lineages and hence, we made no efforts neither in distinguishing nor in comparing these two tumor types. Intentionally, we simply looked for genes with conserved alterations in all tumor samples. It is not surprising that a common set of genes can be identified in two distinct tumor types. As discussed above RASSF1, Clusterin, β-Catenin, and many others genes are commonly altered in a variety of tumors. Specifically, NBS1 that we identified as overexpressed in tumor samples was also found augmented in GIST by Allander and co-workers [21]. As expected, we did identify differences in the expression profile of the two tumor types, especially in genes from cluster 1 (uppermost cluster, second and third last triplicates from the right). However, based on our findings (Figs. 5 and 6) we can suggest that, as for gastric adenocarcinomas, the WNT pathway might also be altered in GIST. We can also conclude that classifiers based on genes commonly altered in adenocarcinomas and GIST can precisely distinguish both tumor types from normal gastric mucosa (Fig. 5) and this would imply that, regardless of differences in oncogenesis, a single classifier could be applied for gastric tumors.

Taken together, the information extracted from our dataset can contribute to the better understanding of oncogenesis of gastric cancer as well as to the development of molecular-based diagnostic tools.

Acknowledgements

We would like to thank all members of the Department of Abdominal Surgery, Hospital do Câncer for tissue collection, Anna Christina de Mattos Salim and Elizangela Monteiro, Laboratory of Genetics from LICR, for helping with sequence verification, and Carlos Ferreira and Miyuki F. da Silva for technical assistance in tissue sections. We also thank Dr. Junior Barrera for discussions, Marcel Brun for developing some of the software used in this work and Dr. Ricardo Brentani for critically reading the manuscript. This work was supported, in part, by the CEPID/FAPESP (98/1435-2). The Clone Collection is funded by the FAPESP/LICR consortium and. S.I.M., L.T., and B.S.S are supported by a predoctoral fellowship from FAPESP. R.L.A.S is supported by a post-doctoral fellowship from FAPESP.

References

[1] A.O. Chan, B.C. Wong, S.K. Lam, Gastric cancer: past, present and future, Can. J. Gastroenterol. 15 (2001) 469–474.

[2] W. Yasui, H. Yokozaki, J. Fujimoto, K. Naka, H. Kuniyasu, E. Tahara, Genetic and epigenetic alterations in multistep carcinogenesis of the stomach, J. Gastroenterol. 35 (Suppl. 12) (2000) 111–115.

[3] W. El Rifai, H.F. Frierson Jr., J.C. Harper, S.M. Powell, S. Knuutila, Expression profiling of gastric adenocarcinoma using cDNA array, Int. J. Cancer 92 (2001) 832–838.

[4] A. Kokkola, O. Monni, P. Puolakkainen, M.L. Larramendy, M. Victorzon, S. Nordling, R. Haapiainen, E. Kivilaakso, S. Knuutila, 17q12–21 amplicon, a novel recurrent genetic change in intestinal type of gastric carcinoma: a comparative genomic hybridization study, Genes Chromosomes Cancer 20 (1997) 38–43.

[5] M. Nessling, S. Solinas-Toldo, K.K. Wilgenbus, F. Borchard, P. Lichter, Mapping of chromosomal imbalances in gastric adenocarcinoma revealed amplified protooncogenes MYCN, MET, WNT2, and ERBB2, Genes Chromosomes Cancer 23 (1998) 307–316.

[6] V. Vidgren, A. Varis, A. Kokkola, O. Monni, P. Puolakkainen, S. Nordling, F. Forozan, A. Kallioniemi, M.L. Vakkari, E. Kivilaakso, S. Knuutila, Concomitant gastrin and ERBB2 gene amplifications at 17q12-q21 in the intestinal type of gastric cancer, Genes Chromosomes Cancer 24 (1999) 24–29.

[7] A.S. Yustein, J.C. Harper, G.R. Petroni, O.W. Cummings, C.A. Moskaluk, S.M. Powell, Allelotype of gastric adenocarcinoma, Cancer Res. 59 (1999) 1437–1441.

[8] G.N. Ranzani, O. Luinetti, L.S. Padovan, D. Calistri, B. Renault, M. Burrel, D. Amadori, R. Fiocca, E. Solcia, p53 gene mutations and protein nuclear accumulation are early events in intestinal type gastric cancer but late events in diffuse type, Cancer Epidemiol. Biomarkers Prev. 4 (1995) 223–231.

[9] T. Shepherd, D. Tolbert, J. Benedetti, J. Macdonald, G. Stemmermann, J. Wiest, G. De Voe, M.A. Miller, J. Wang, A. Noffsinger, C. Fenoglio-Preiser, Alterations in exon 4 of the p53 gene in gastric carcinoma, Gastroenterology 118 (2000) 1039–1044.

[10] J.G. Fox, X. Li, R.J. Cahill, K. Andrutis, A.K. Rustgi, R. Odze, T.C. Wang, Hypertrophic gastropathy in *Helicobacter felis*-infected wild-type C57BL/6 mice and p53 hemizygous transgenic mice, Gastroenterology 110 (1996) 155–166.

[11] K. Tsugawa, Y. Yonemura, Y. Hirono, S. Fushida, M. Kaji, K. Miwa, I. Miyazaki, H. Yamamoto, Amplification of the c-met, c-erbB-2 and epidermal growth factor receptor gene in human gastric cancers: correlation to clinical features, Oncology 55 (1998) 475–481.

[12] K. Park, S.J. Kim, Y.J. Bang, J.G. Park, N.K. Kim, A.B. Roberts, M.B. Sporn, Genetic changes in the transforming growth factor beta (TGF-beta) type II receptor gene in human gastric cancer cells: correlation with sensitivity to growth inhibition by TGF-beta, Proc. Natl. Acad. Sci. USA 91 (1994) 8772–8776.

[13] Y. Shimoyama, S. Hirohashi, Expression of E- and P-cadherin in gastric carcinomas, Cancer Res. 51 (1991) 2185–2192.

[14] K. Matsuura, J. Kawanishi, S. Fujii, M. Imamura, S. Hirano, M. Takeichi, Y. Niitsu, Altered expression of E-cadherin in gastric cancer tissues and carcinomatous fluid, Br. J. Cancer 66 (1992) 1122–1130.

[15] D.K. Woo, H.S. Kim, H.S. Lee, Y.H. Kang, H.K. Yang, W.H. Kim, Altered expression and mutation of beta-catenin gene in gastric carcinomas and cell lines, Int. J. Cancer 95 (2001) 108–113.

[16] M. Miettinen, M. Sarlomo-Rikala, J. Lasota, Gastrointestinal stromal tumors: recent advances in understanding of their biology, Hum. Pathol. 30 (1999) 1213–1220.

[17] W. El Rifai, M. Sarlomo-Rikala, L.C. Andersson, M. Miettinen, S. Knuutila, DNA copy number changes in gastrointestinal stromal tumors – a distinct genetic entity, Ann. Chir. Gynaecol. 87 (1998) 287–290.

[18] W. El Rifai, M. Sarlomo-Rikala, M. Miettinen, S. Knuutila, L.C. Andersson, DNA copy number losses in chromosome 14: an early change in gastrointestinal stromal tumors, Cancer Res. 56 (1996) 3230–3233.

[19] S. Sakurai, T. Fukasawa, J.M. Chong, A. Tanaka, M. Fukayama, C-kit gene abnormalities in gastrointestinal stromal tumors (tumors of interstitial cells of Cajal), Jpn. J. Cancer Res. 90 (1999) 1321–1328.

[20] R. Fukuda, N. Hamamoto, Y. Uchida, K. Furuta, T. Katsube, H. Kazumori, S. Ishihara, K. Amano, K. Adachi, M. Watanabe, Y. Kinoshita, Gastrointestinal stromal tumor with a novel mutation of KIT proto- oncogene, Intern. Med. 40 (2001) 301–303.

[21] S.V. Allander, N.N. Nupponen, M. Ringner, G. Hostetter, G.W. Maher, N. Goldberger, Y. Chen, J. Carpten, A.G. Elkahloun, P.S. Meltzer, Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile, Cancer Res. 61 (2001) 8624–8628.

[22] M.H. Jung, S.C. Kim, G.A. Jeon, S.H. Kim, Y. Kim, K.S. Choi, S.I. Park, M.K. Joe, K. Kimm, Identification of differentially expressed genes in normal and tumor human gastric tissue, Genomics 69 (2000) 281–286.

[23] Y. Hippo, M. Yashiro, M. Ishii, H. Taniguchi, S. Tsutsumi, K. Hirakawa, T. Kodama, H. Aburatani, Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes, Cancer Res. 61 (2001) 889–895.

[24] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.M. Chong, M. Fukayama, T. Kodama, H. Aburatani, Global gene expression analysis of gastric cancer by oligonucleotide microarrays, Cancer Res. 62 (2002) 233–240.

[25] A.A. Camargo, H.P. Samaia, E. Dias-Neto, D.F. Simao, I.A. Migotto, M.R. Briones, F.F. Costa, M.A. Nagai, S. Verjovski-Almeida, M.A. Zago, L.E. Andrade, H. Carrer, H.F. El Dorry, E.M. Espreafico, A. Habr-Gama, D. Giannella-Neto, G.H. Goldman, A. Gruber, C. Hackel, E.T. Kimura, R.M. Maciel, S.K. Marie, E.A. Martins, M.P. Nobrega, M.L. Paco-Larson, M.I. Pardini, G.G. Pereira, J.B. Pesquero, V. Rodrigues, S.R. Rogatto, I.D. da Silva, M.C. Sogayar, M.F. Sonati, E.H. Tajara, S.R. Valentini, F.L. Alberto, M.E. Amaral, I. Aneas, L.A. Arnaldi, A.M. de Assis, M.H. Bengtson, N.A. Bergamo, V. Bombonato, M.E. de Camargo, R.A. Canevari, D.M.

Carraro, J.M. Cerutti, M.L. Correa, R.F. Correa, M.C. Costa, C. Curcio, P.O. Hokama, A.J. Ferreira, G.K. Furuzawa, T. Gushiken, P.L. Ho, E. Kimura, J.E. Krieger, L.C. Leite, P. Majumder, M. Marins, E.R. Marques, A.S. Melo, M. Melo, C.A. Mestriner, E.C. Miracca, D.C. Miranda, A.L. Nascimento, F.G. Nobrega, E.P. Ojopi, J.R. Pandolfi, L.G. Pessoa, A.C. Prevedel, P. Rahal, C.A. Rainho, E.M. Reis, M.L. Ribeiro, N. da Ros, R.G. deSa, M.M. Sales, S.C. Sant'anna, M.L. dos Santos, A.M. da Silva, N.P. da Silva, W.A. Silva Jr., R.A. da Silveira, J.F. Sousa, D. Stecconi, F. Tsukumo, V. Valente, F. Soares, E.S. Moreira, D.N. Nunes, R.G. Correa, H. Zalcberg, A.F. Carvalho, L.F. Reis, R.R. Brentani, A.J. Simpson, S.J. deSouza, The contribution of 700 000 ORF sequence tags to the definition of the human transcriptome, Proc. Natl. Acad. Sci. USA 98 (2001) 12103–12108.

[26] G.M. Church, W. Gilbert, Genomic sequencing, Proc. Natl. Acad. Sci. USA 81 (1984) 1991–1995.

[27] M.J. Callow, S. Dudoit, E.L. Gong, T.P. Speed, E.M. Rubin, Microarray expression profiling identifies genes with altered expression in HDL-deficient mice, Genome Res. 10 (2000) 2022–2029.

[28] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, H. Herzel, Normalization strategies for cDNA microarrays, Nucleic Acids Res. 28 (2000) E47.

[29] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[30] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press, New York, 1999.

[31] P.D. Lee, R. Sladek, C.M. Greenwood, T.J. Hudson, Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies, Genome Res. 12 (2002) 292–297.

[32] W. Jin, R.M. Riley, R.D. Wolfinger, K.P. White, G. Passador-Gurgel, G. Gibson, The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*, Nat. Genet. 29 (2001) 389–395.

[33] T. Kohonen, Self-Organizing Maps, Springer, Berlin, 1997.

[34] R. De Vivo, S. Pignata, R. Palaia, V. Parisi, B. Daniele, The role of chemotherapy in the management of gastric cancer, J. Clin. Gastroenterol. 30 (2000) 364–371.

[35] H.H. Hartgrink, H.J. Bonenkamp, C.J. van de Velde, Influence of surgery on outcomes in gastric cancer, Surg. Oncol. Clin. North Am. 9 (2000) 97-viii.

[36] A. Boussioutas, D. Taupin, Towards a molecular approach to gastric cancer management, Intern. Med. J. 31 (2001) 296–303.

[37] K.F. Becker, G. Keller, H. Hoefler, The use of molecular biology in diagnosis and prognosis of gastric cancer, Surg. Oncol. 9 (2000) 5–11.

[38] K. Dreijerink, E. Braga, I. Kuzmin, L. Geil, F.M. Duh, D. Angeloni, B. Zbar, M.I. Lerman, E.J. Stanbridge, J.D. Minna, A. Protopopov, J. Li, V. Kashuba, G. Klein, E.R. Zabarovsky, The candidate tumor suppressor gene, RASSF1A, from human chromosome 3p21.3 is involved in kidney tumorigenesis, Proc. Natl. Acad. Sci. USA 98 (2001) 7504–7509.

[39] R. Dammann, C. Li, J.H. Yoon, P.L. Chin, S. Bates, G.P. Pfeifer, Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3, Nat. Genet. 25 (2000) 315–319.

[40] R. Dammann, G. Yang, G.P. Pfeifer, Hypermethylation of the cpG island of Ras association domain family 1A (RASSF1A), a putative tumor suppressor gene from the 3p21.3 locus, occurs in a large percentage of human breast cancers, Cancer Res. 61 (2001) 3105–3109.

[41] D.G. Burbee, E. Forgacs, S. Zochbauer-Muller, L. Shivakumar, K. Fong, B. Gao, D. Randle, M. Kondo, A. Virmani, S. Bader, Y. Sekido, F. Latif, S. Milchgrub, S. Toyooka, A.F. Gazdar, M.I. Lerman, E. Zabarovsky, M. White, J.D. Minna, Epigenetic inactivation of RASSF1A in lung and breast cancers and malignant phenotype suppression, J. Natl. Cancer Inst. 93 (2001) 691–699.

[42] A. Agathanggelou, S. Honorio, D.P. Macartney, A. Martinez, A. Dallol, J. Rader, P. Fullwood, A. Chauhan, R. Walker, J.A. Shaw, S. Hosoe, M.I. Lerman, J.D. Minna, E.R. Maher, F. Latif, Methylation associated inactivation of RASSF1A from region 3p21.3 in lung, breast and ovarian tumours, Oncogene 20 (2001) 1509–1518.

[43] K.W. Lo, J. Kwong, A.B. Hui, S.Y. Chan, K.F. To, A.S. Chan, L.S. Chow, P.M. Teo, P.J. Johnson, D.P. Huang, High frequency of promoter hypermethylation of RASSF1A in nasopharyngeal carcinoma, Cancer Res. 61 (2001) 3877–3881.

[44] M.G. Lee, H.Y. Kim, D.S. Byun, S.J. Lee, C.H. Lee, J.I. Kim, S.G. Chang, S.G. Chi, Frequent epigenetic inactivation of RASSF1A in human bladder carcinoma, Cancer Res. 61 (2001) 6688–6692.

[45] D.S. Byun, M.G. Lee, K.S. Chae, B.G. Ryu, S.G. Chi, Frequent epigenetic inactivation of RASSF1A by aberrant promoter hypermethylation in human gastric adenocarcinoma, Cancer Res. 61 (2001) 7034–7038.

[46] R. Cavallo, D. Rubenstein, M. Peifer, Armadillo and dTCF: a marriage made in the nucleus, Curr. Opin. Genet. Dev. 7 (1997) 459–466.

[47] M.J. Smalley, T.C. Dale, Wnt signalling in mammalian development and cancer, Cancer Metastasis Rev. 18 (1999) 215–230.

[48] M. Peifer, P. Polakis, Wnt signaling in oncogenesis and embryogenesis–a look outside the nucleus, Science 287 (2000) 1606–1609.

[49] T. Oda, Y. Kanai, T. Oyama, K. Yoshiura, Y. Shimoyama, W. Birchmeier, T. Sugimura, S. Hirohashi, E-cadherin gene mutations in human gastric carcinoma cell lines, Proc. Natl. Acad. Sci. USA 91 (1994) 1858–1862.

[50] K.F. Becker, H. Hofler, Frequent somatic allelic inactivation of the E-cadherin gene in gastric carcinomas, J. Natl. Cancer Inst. 87 (1995) 1082–1084.

[51] G. Keller, H. Vogelsang, I. Becker, J. Hutter, K. Ott, S. Candidus, T. Grundei, K.F. Becker, J. Mueller, J.R. Siewert, H. Hofler, Diffuse type gastric and lobular breast carcinoma in a familial gastric cancer patient with an E-cadherin germline mutation, Am. J. Pathol. 155 (1999) 337–342.

[52] S.A. Gayther, K.L. Gorringe, S.J. Ramus, D. Huntsman, F. Roviello, N. Grehan, J.C. Machado, E. Pinto, R. Seruca, K. Halling, P. MacLeod, S.M. Powell, C.E. Jackson, B.A. Ponder,

C. Caldas, Identification of germ-line E-cadherin mutations in gastric cancer families of European origin, Cancer Res. 58 (1998) 4086–4089.

[53] J. Behrens, J.P. von Kries, M. Kuhl, L. Bruhn, D. Wedlich, R. Grosschedl, W. Birchmeier, Functional interaction of beta-catenin with the transcription factor LEF-1, Nature 382 (1996) 638–642.

[54] T.L. Le, A.S. Yap, J.L. Stow, Recycling of E-cadherin: a potential mechanism for regulating cadherin dynamics, J. Cell Biol. 146 (1999) 219–232.

[55] S. Byers, M. Pishvaian, C. Crockett, C. Peer, A. Tozeren, M. Sporn, M. Anzano, R. Lechleider, Retinoids increase cell-cell adhesion strength, beta-catenin protein stability, and localization to the cell membrane in a breast cancer cell line: a role for serine kinase activity, Endocrinology 137 (1996) 3265–3273.

[56] V. Easwaran, M. Pishvaian, Salimuddin, S. Byers, Cross-regulation of beta-catenin-LEF/TCF and retinoid signaling pathways, Curr. Biol. 9 (1999) 1415–1418.

[57] J.Y. Wang, Cancer. New link in a web of human genes, Nature 405 (2000) 404–405.

[58] T.T. Paull, D. Cortez, B. Bowers, S.J. Elledge, M. Gellert, From the Cover: Direct DNA binding by Brca1, Proc. Natl. Acad. Sci. USA 98 (2001) 6086–6091.

[59] Y. Hashimoto, T. Niikura, H. Tajima, T. Yasukawa, H. Sudo, Y. Ito, Y. Kita, M. Kawasumi, K. Kouyama, M. Doyu, G. Sobue, T. Koide, S. Tsuji, J. Lang, K. Kurokawa, I. Nishimoto, A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta, Proc. Natl. Acad. Sci. USA 98 (2001) 6336–6341.

[60] Y. Hashimoto, T. Niikura, Y. Ito, H. Sudo, M. Hata, E. Arakawa, Y. Abe, Y. Kita, I. Nishimoto, Detailed characterization of neuroprotection by a rescue factor humanin against various Alzheimer's disease-relevant insults, J. Neurosci. 21 (2001) 9235–9245.

[61] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98 (2001) 13790–13795.

[62] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[63] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, T.R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Natl. Acad. Sci. USA 98 (2001) 15149–15154.

[64] C.M. Perou, T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Borresen-Dale, P.O. Brown, D. Botstein, Molecular portraits of human breast tumours, Nature 406 (2000) 747–752.

[65] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[66] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. USA 96 (1999) 2907–2912.

[67] L.J. 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooij, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.

[68] F.S. Monteclaro, P.K. Vogt, A Jun-binding protein related to a putative tumor suppressor, Proc. Natl. Acad. Sci. USA 90 (1993) 6726–6730.

[69] B. Mann, M. Gelos, A. Siedow, M.L. Hanski, A. Gratchev, M. Ilyas, W.F. Bodmer, M.P. Moyer, E.O. Riecken, H.J. Buhr, C. Hanski, Target genes of beta-catenin-T cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas, Proc. Natl. Acad. Sci. USA 96 (1999) 1603–1608.

[70] T. Kato, S. Satoh, H. Okabe, O. Kitahara, K. Ono, C. Kihara, T. Tanaka, T. Tsunoda, Y. Yamaoka, Y. Nakamura, Y. Furukawa, Isolation of a novel human gene, MARKL1, homologous to MARK3 and its involvement in hepatocellular carcinogenesis, Neoplasia. 3 (2001) 4–9.

ANEXO 2

MOLECULAR CLASSIFIERS FOR GASTRIC CANCER AND NONMALIGNANT DISEASES OF THE GASTRIC MUCOSA.

[CANCER RESEARCH 64, 1255–1265, February 15, 2004]

Molecular Classifiers for Gastric Cancer and Nonmalignant Diseases of the Gastric Mucosa

Sibele I. Meireles,[1,2] Elier B. Cristo,[3] Alex F. Carvalho,[1] Roberto Hirata, Jr.,[4] Adriane Pelosof,[2] Luciana I. Gomes,[1,2] Waleska K. Martins,[1,2] Maria D. Begnami,[2] Cláudia Zitron,[2] André L. Montagnini,[2] Fernando A. Soares,[2] E. Jordão Neves,[3] and Luiz F. L. Reis[1,2]

[1]Ludwig Institute for Cancer Research and [2]Hospital do Câncer A.C. Camargo; [3]BioInfo and Instituto de Matemática e Estatística da Universidade de São Paulo; and [4]SENAC College of Computer Science and Technology, São Paulo, Brazil

ABSTRACT

High incidence of gastric cancer-related death is mainly due to diagnosis at an advanced stage in addition to the lack of adequate neoadjuvant therapy. Hence, new tools aimed at early diagnosis would have a positive impact in the outcome of the disease. Using cDNA arrays having 376 genes either identified previously as altered in gastric tumors or known to be altered in human cancer, we determined expression signature of 99 tissue fragments representing normal gastric mucosa, gastritis, intestinal metaplasia, and adenocarcinomas. We first validated the array by identifying molecular markers that are associated with intestinal metaplasia, considered as a transition stage of gastric adenocarcinomas of the intestinal type as well as markers that are associated with diffuse type of gastric adenocarcinomas. Next, we applied Fisher's linear discriminant analysis in an exhaustive search of trios of genes that could be used to build classifiers for class distinction. Many classifiers could distinguish between normal and tumor samples, whereas, for the distinction of gastritis from tumor and for metaplasia from tumor, fewer classifiers were identified. Statistical validations showed that trios that discriminate between normal and tumor samples are powerful classifiers to distinguish between tumor and nontumor samples. More relevant, it was possible to identify samples of intestinal metaplasia that have expression signature resembling that of an adenocarcinoma and can now be used for follow-up of patients to determine their potential as a prognostic test for malignant transformation.

INTRODUCTION

Gastric cancer is still one of the major causes of cancer-related death worldwide, although its incidence is declining during the last decades, as reviewed previously (1). This high mortality is, at least in part, a consequence of late-stage diagnosis, due to the lack of specific symptoms at early stages of the disease. Moreover, no effective therapeutic strategy is available at advanced stages, and patients often undergo radical gastrectomy leading to high morbidity. Despite the aggressiveness of this treatment, 5-year survival rate in advanced stages is extremely poor, ranging from 5% to 15%. On the contrary, when early diagnosis is successful, there is a higher degree of resectability and better survival rates (2).

Gastric adenocarcinoma represents >95% of all gastric tumors and, following Lauren's classification (3), it can be divided into intestinal and diffuse type, according to tumor histology. These two histological types have a distinct pathology, epidemiology, and etiology. The intestinal type is more frequent and represents the dominant histological type in areas where stomach cancer is epidemic, suggesting an environmental etiology. The pathogenesis of intestinal type adenocarcinoma has been connected to precursor changes such as chronic active gastritis, multifocal atrophic gastritis, intestinal metaplasia, and dysplasia as proposed by Correa and Chen (4), and with the presence of *Helicobacter pylori* infection (5). In contrast, the diffuse type has not been related to precursor lesions and has a higher association with familial occurrence. It is widely accepted that genetic alterations in the *CDH1* (e-cadherin) gene play an important role in the oncogenesis of the diffuse-type gastric cancer (6).

The relationship between intestinal metaplasia and intestinal-type gastric adenocarcinomas has not been fully established. Some individuals with intestinal metaplasia will never develop gastric cancer. The molecular events related to progression from metaplasia to adenocarcinomas remains unknown. Certain molecular alterations have been associated with intestinal metaplasia. For example, mutation in p53 was detected in intestinal metaplasia adjacent to gastric tumors (7). Overexpression of cyclooxygenase-2 was detected in intestinal metaplasia-associated gastritis (8), and also detected in gastric cancer and in other tumors like colon cancer (9). Kang *et al.* (10) detected promoter hypermethylation in the DNA mismatch repair gene (*hMLH1*) in intestinal metaplasia, and this epigenetic alteration is related to microsatellite instability. This author also describes hypermethylation of other genes, including *p16, DAP-kinase, THBS1,* and *TIMP-3.* Recently, Boussioutas *et al.* (11) described the expression profile of 124 gastric mucosa representing gastritis, intestinal metaplasia, and adenocarcinomas, and identified a series of genes that are typically expressed in the intestinal type and could be related to tumor progression. Together, these findings demonstrate that some of the molecular events associated with intestinal metaplasia can also be detected in cancer sample and, hence, comparing the molecular alterations between nonmalignant and malignant lesions could lead to the identification of genes involved in gastric carcinogenesis. Moreover, the identification of expression signatures that correlate with adenocarcinomas could be used for follow-up of patients with intestinal metaplasia and assessment of the risk of the premalignant stages becoming malignant.

In this work, we used a cDNA array with 376 genes, including those 141 genes described previously by our group (12) plus other genes known to be generally altered in human cancers. We determined the expression profile in 99 tissue samples representing normal gastric mucosa, as well as gastritis, intestinal metaplasia, and adenocarcinoma of the stomach. Using Fisher's linear discriminant analysis, we identified a series of molecular classifiers that could distinguish between cancer and noncancer samples. Importantly, we also identify a series of intestinal metaplasias of which the gene expression profile resembles that of adenocarcinoma.

MATERIALS AND METHODS

Tissue Samples and RNA Preparations. Fresh tissue samples were obtained by surgery or by endoscopy at the Gastric Surgery Department and Gastric/Esophagic Endoscopy Department from Hospital do Cancer AC Camargo (São Paulo, Brazil). All of the patients signed an informed consent, and the project was approved by the in-house ethics committee. Tissue samples

Received 6/23/03; revised 12/8/03; accepted 12/10/03.

Grant support: CEPID-Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Grant 98/14335-2. E. J. Neves was partially supported by FAPESP, Grants 99/11962-9 and 99/07390-0, and by Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, Grants 41.96.0923.00 and 521097/2001-0. S. I. Meireles, E. B. Cristo, and L. I. Gomes were supported by fellowships from FAPESP.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Luiz F. L. Reis, Rua Prof. Antonio Prudente 109, Sao Paulo, SP 01509-010, Brazil. Phone: 55-11-33883228; Fax: 55-11-32077001; E-mail: lreis@ludwig.org.br.

Table 1 *Description of patients and samples*

A. Nonmalignant Samples

	Inflammation				Activity		IM		Procedure		
Group	Absent/Minimal	Discrete	Moderate	Intense	Absent	Present	Absent	Present	Biopsy	Surgery	Total
Normal	28	–	–	–	28	–	28	–	17	11	28
Gastritis/atrophy	–	8	7	6	6	15	21	–	15	06	21
Intestinal metaplasia	–	10	10	2	13	9	–	22	16	06	22

B. Tumor Samples

ID	WHO	Laurén	Grade	Size (cm)	Borrmann	Infiltration	LN[a]	TNM
BIO228	Signet-ring cell carcinoma	Diffuse	PD	NA	IV	NA	NA	NA
GF48	Signet-ring cell carcinoma	Diffuse	PD	7.5	IV	Serosa	0/7	T2N0M1
GF52	Signet-ring cell carcinoma	Diffuse	PD	5.3	II	Muscularis propria	3/40	T2N1M0
GF62	Signet-ring cell carcinoma	Diffuse	PD	3.5	III	Serosa	1/40	T3N1M0
GF66	Signet-ring cell carcinoma	Diffuse	PD	6.5	II	Serosa	28/42	T3N3M0
GF68	Signet-ring cell carcinoma	Diffuse	PD	6.0	II	Serosa	9/40	T4N2M0
GF72	Signet-ring cell carcinoma	Diffuse	PD	6.0	II	Fat tissue	4/7	T4N1M1
GH877	Signet-ring cell carcinoma	Diffuse	PD	7.5	II	Serosa	7/44	T3N2M0
GH967	Signet-ring cell carcinoma	Diffuse	PD	2.9	III	Fat tissue	0/16	T2N0M0
GH977	Signet-ring cell carcinoma	Diffuse	PD	6.5	IV	Muscularis propria	13/23	T2N2M0
BIO093	Tubular adenocarcinoma	Intestinal	PD	NA	III	NA	NA	NA
BIO108	Tubular adenocarcinoma	Intestinal	PD	NA	II	NA	NA	NA
GF50	Tubular adenocarcinoma	Intestinal	MD	10	III	Fat tissue	3/42	T4N1M0
GF54	Tubular adenocarcinoma	Intestinal	PD	4.5	III	Fat tissue	2/22	T3N1M0
GF56	Tubular adenocarcinoma	Intestinal	MD	5.5	I	Fat tissue	47/55	T4N3M1
GF58	Tubular adenocarcinoma	Intestinal	MD	7.5	II	Fat tissue	1/29	T4N1M1
GF70	Tubular adenocarcinoma	Intestinal	MD	11.8	III	Serosa	0/42	T3N0M0
GH695	Tubular adenocarcinoma	Intestinal	PD	11	III	Fat tissue	17/18	T4N3M1
GH831	Mucinous adenocarcinoma	Intestinal	MD	11	II	Fat tissue	06/18	T4N1M1
GH833	Tubular adenocarcinoma	Intestinal	MD	6	III	Fat tissue	1/4	T2N1M0
GH843	Mucinous adenocarcinoma	Intestinal	WD	8	II	Fat tissue	10/37	T3N2M0
GH881	Tubular adenocarcinoma	Intestinal	MD	7	III	Fat tissue	9/30	T4N2M0
GH965	Tubular adenocarcinoma	Intestinal	WD	5	I	Mucosa	0/7	TisN0M0
GH969	Papillary adenocarcinoma	Intestinal	WD	7	II	Serosa	11/18	T4N2M1
GH971	Tubular adenocarcinoma	Intestinal	WD	5.0	III	Mucosa	0/39	T1N0M0
GH973	Tubular adenocarcinoma	Intestinal	MD	14	II	Serosa	20/41	T3N3M0
GH975	Tubular adenocarcinoma	Intestinal	PD	9.0	III	Fat tissue	5/26	T3N1M0
GH979	Tubular adenocarcinoma	Intestinal	PD	6.5	III	Fat tissue	103/125	T4N3M0

[a] LN, lymph node; TNM, Tumor-Node-Metastasis; PD, poorly differentiated; NA, not applicable; MD, moderately differentiated; WD, well differentiated.

were either snap frozen in liquid nitrogen or collected in RNAlater, and are represented by 28 gastric adenocarcinomas (18 intestinal type and 10 diffuse type, according to Lauren classification; Ref. 3) and 71 nontumor gastric samples (28 normal gastric mucosa, 21 gastritis mucosa, and 22 intestinal metaplasia of the gastric mucosa). Detailed description of samples is presented in Table 1. Samples labeled as "GF" or "GH" were obtained from surgery, and the majority came from patients with gastric adenocarcinomas. Samples labeled as "BIO" were obtained by endoscopic biopsy, and the majority is from patients with only the indicated pathology.

At the time of RNA extraction, histological confirmation of tumor or nontumor status was performed by H&E staining of each individual sample. The frozen sections were also used for tissue dissection to enrich for tumor cells. For tumor specimens, only samples with at least 70% of tumor tissue and free of inflammatory infiltrate were additionally processed. In the case of nontumor samples, only gastric mucosa was used. Total RNA was extracted using TRIzol Reagent (Life Technologies, Inc., Grand Island, NY) following the procedure recommended by the manufacturer.

Production of cDNA Arrays. We constructed a cDNA array with 376 genes including 141 clones of genes described previously as being altered in gastric cancer (12) and 235 genes known to be altered in human cancers on the basis of available literature (complete gene list is available).[5] All of the clones correspond to ORESTES fragments derived from the Fundação de Amparo à Pesquisa do Estado de São Paulo/Ludwig Institute for Cancer Research Human Cancer Genome Project and were sequence verified. Whenever possible, each gene is represented by two clones corresponding to different regions of the complete cDNA and each cDNA clone was printed in triplicates onto nylon membranes (Flexys robot; Genomic Solutions, Ann Arbor, MI) making a total of 2400 spots. Production of the cDNA array, labeling, hybridization, and detection of signals were carried out as described previously (12). For each tissue sample, 25 μg of total RNA were radioactively labeled with [α-^{33}P]dCTP (3000 Ci/mmol; Amersham, Piscataway, NJ) and hybridized against a nylon-based cDNA array. Data acquisition was performed with the ArrayVision software (Amersham), using gel files.

Statistical Analysis. Data analysis was performed using R,[6] an open source interpreted computer language for statistical computation and graphics, and tools from the Bioconductor project,[7] adapted to our needs. Principal component analysis was performed using TMEV (13). After image acquisition and quantification (see above), spots with signal lower or equal to background were identified and excluded from the analysis. Next, background-subtracted spot intensities were normalized by global mean normalization procedure (14, 15). Replica spots representing the same gene were identified, and average signal intensity was determined.

Next, we searched our data for differentially expressed genes in the four clinical conditions. We used a nonparametric test (Mann-Whitney) to determine the *P* for each individual gene in each pair-wise comparison. For display purpose, we highlighted genes with $P \leq 0.0009$ in Figs. 1 and 3. For all of the pair-wise comparisons (Figs. 1 and 3;Table 2), genes that are overexpressed in the second entity have the $[-\log_2(P)]$ preceded by a minus signal. For clustering, we selected the nonredundant set of 6 genes with lowest *P* for each comparison and used the resulting 18 genes for clustering samples into four groups using the nonsupervised algorithm *k*-means. Once clusters were obtained, samples were organized hierarchically, based on their correlation distances (16). For classifiers, we use Fisher's linear discriminant analysis and made exhausted search of the entire dataset for trios of genes such that data points representing signal intensity for all 3 of the genes for each sample would be separated by a plane in a three-dimensional space. More precisely, for a given group of genes, this linear classification method searches for linear combinations of their expressions with large ratios of between-groups to within-groups sum of squares (16). This maximal ratio of sum of squares, or

[5] Internet address: http://array.ludwig.org.br/publications/meirelescancerresearch2003

[6] Internet address: http://cran.r-project.org/.

[7] Internet address: http://www.bioconductor.org.

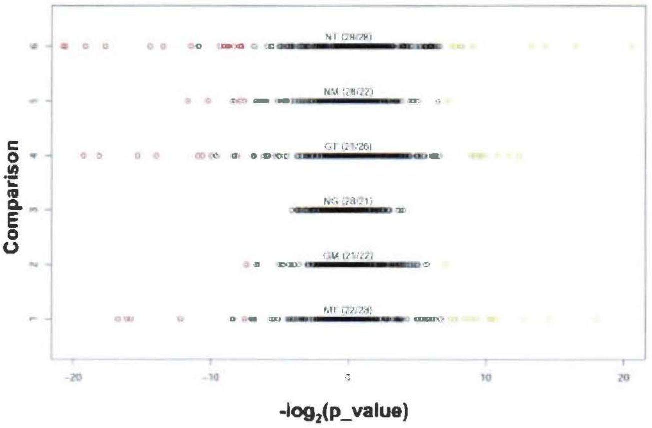


Fig. 1. Differentially expressed genes in pair-wise comparison in tissue samples of normal gastric mucosa, gastritis, gastric intestinal metaplasia, and gastric adenocarcinomas. Expression profile from samples representing normal gastric mucosa (*N*), gastritis (*G*), intestinal metaplasia (*M*), and tumor (*T*) containing 28, 21, 22, and 28 samples, respectively, were analyzed in six pair-wise comparisons, NT, NM, GT, NG, GM, and MT. Genes were distributed according to their [$-\log_2(P)$] in each pair-wise comparison. The *green* and *red dots* represent genes that showed $P < 0.0009$ (Mann-Whitney). *Green color* was used for genes with higher expression in the first entity of the comparison, and *red color* was used for genes with higher expression in the second entity of the comparison (values preceded by a minus sign).

its square root, which is denoted here by SVD (singular value decomposition), measures how well separated the two groups are. For the search of trios, the 99 samples dataset was split into two groups, one with 65 samples, used as learning set and a second, independent group, with 34 samples used for validation. Using the learning set of samples (65 samples) we performed an exhaustive search for the best classification trios for each one of the six comparisons of interest among normal, gastritis, metaplasia, and tumor. Trios were ranked according their SVD. In the case of normal *versus* tumor, where many trios were available, we only considered trios with perfect classification (1044 trios). Next, the best classifiers were tested with the remaining 34 samples with the results presented here.

Tissue Array. For the preparation of the stomach tissue microarray, all of the gastrectomy specimens were retrieved from the hospital archives. All of the tissues were fixed in formalin and, from each specimen, H&E-stained slides underwent pathological review to reconfirm diagnosis and selection of blocks. Samples were divided into five groups, histologically normal mucosa (25 cases), chronic gastritis (50 cases), intestinal metaplasia (25 cases), intestinal type adenocarcinoma (75 cases), and diffuse type adenocarcinoma (75 cases). The Lauren type of gastric carcinoma was determined by the following criteria: the intestinal type of gastric adenocarcinoma is histologically characterized by the presence of cohesive cells forming glandular and papillary structure. The diffuse type of gastric carcinoma is characterized histologically by noncohesive cells and the common presence of signet ring cells. Importantly, samples used for RNA extraction represented only a small portion of the samples comprising the tissue microarray.

For the construction of the stomach tissue microarray, new sections were obtained from the representative paraffin blocks, and all of the H&E-stained slides for these cases were reviewed. A slide with representative condition was selected from each case, and an area of one of the studied groups was circled on the slide. The corresponding formalin-fixed, paraffin-embedded blocks were retrieved, and the area corresponding to the selected area on the slide was circled on the block with a felt marker for tissue microarray construction. Using a tissue microarrayer (Beecher Instruments, Silver Spring, MD), the area of interest in the donor paraffin block was cored twice with a 0.6-mm diameter needle and transferred to a recipient paraffin block. Sections of 4 μm were cut from stomach tissue microarray block, deparaffined, dehydrated, and submitted to immunohistochemistry with a polyclonal antibody against metalloproteinase 2 (Oncogene; clone 75–7f7; dilution 1:40). For determining arbitrary units, we gave a score (1–4) for intensity of staining plus a score (1–4) for the percentage of positive cells in each tissue spot. Each tissue sample was spotted in duplicate in the slide, and four slides were measured, making a total of eight spots for each tissue sample. All four of the slides were analyzed by two independent pathologists and, for each sample, we sum the two scores (intensity and percentage) for all eight spots gave by each pathologist. Arbitrary units correspond with the average score for each tissue sample (average of 16 determinations for the corresponding eight spots).

RESULTS

Identification of Genes Differentially Expressed in Normal, Nonmalignant, and Malignant Diseases. For the identification of differentially expressed genes in 99 samples representing normal gastric mucosa ($n = 28$), gastritis ($n = 21$), intestinal metaplasia ($n = 22$), and adenocarcinomas of the intestinal type ($n = 18$) or diffuse type ($n = 10$), data from all of the hybridizations were background-corrected and normalized as described in "Materials and Methods" leaving a set of 370 genes for analysis. For each gene, we considered the average signal intensity from all of the replica spots. We first compared samples in a pair-wise manner using a nonparametric test (Mann-Whitney) to access significance. In Fig. 1 we represent all of the genes and their respective *P*s for each pair-wise comparison, indicating in color those with $Ps \leq 0.0009$. Green color was used for genes with higher expression in the first entity of the comparison, and red color was used for genes with higher expression in the second entity of the comparison. As expected, there were more differentially expressed genes when tumor samples were compared with normal, gastritis, or intestinal metaplasia and, conversely, fewer

Table 2 *Genes differentially expressed in pair-wise comparison of gastric tissue samples*

Symbol	NxT	GxT	MxT	NxM	GxM
PRPF8	20.6[a]	12.4	7.74	–[b]	–
HS.327751	16.6	11.7	–	–	–
XBP1	14.3	8.93	14.6	–	–
POLR21	13.3	10.8	10.3	–	–
LCK	9.03	–	–	–	–
BAD	8.21	–	–	–	–
IGF1R	8.17	–	–	–	–
MUC6	7.8	9.67	8.91	–	–
IGL	7.46	9.97	12.7	–	–
TYMS	–	9.44	–	–	7.07
NF1	–	9.14	–	–	–
HS.164280	–	–	9.34	–	–
KRT20	–	–	18	−11.7	−7.41
DPH2L1	–	–	8.47	–	–
CASP7	–	–	10.7	–	–
TUBB	–	–	10.4	−8.44	–
S100A14	–	–	7.41	–	–
KRT19	–	–	7.71	−8.27	–
PTSGS2	–	–	−8.47	7.25	–
CDH1	–	–	–	−7.56	–
KRT17	–	–	–	−8.43	–
TIMP1	–	−8.4	–	–	–
HDGF	−7.77	–	–	–	–
TIMP3	−7.83	−8.26	−8.43	–	–
MYC	−7.94	–	–	–	–
LAMC2	−8.31	–	–	–	–
SPP1	−8.68	−9.96	−8.34	–	–
IGFBP4	−8.82	−8.12	–	–	–
KRT7	−8.94	–	–	–	–
COL4A2	−9.07	−9.67	−7.54	–	–
FOS	−9.35	–	–	–	–
MMP2	−10.8	−9.96	–	–	–
PCNA	−10.9	–	–	–	–
LAMA4	−10.9	−8.26	–	–	–
PLS3	−11.4	−10.6	−12.2	–	–
DAF	−13.4	−9.57	–	−7.87	–
HS.177781	−14.4	−10.9	–	–	–
VIM	−17.7	−14	–	–	–
FN1	−19.1	−18.1	−15.8	–	–
CTSB	−20.6	−15.3	−7.09	−10.2	–
COL1A2	−20.7	−19.3	−16.1	–	–
COL4A1	−33.1	−27.4	−16.7	–	–

[a] For each comparison, we represent the [$-\log_2(P)$].

[b] Values preceded by a minus signal indicate overexpression in the second entity of the comparison.

genes with statistically significant differences could be identified when we compare intestinal metaplasia with normal or gastritis. Finally, no statistically significant differences ($P \leq 0.0009$) could be observed in the expression profile when we compared normal *versus* gastritis. If we consider genes with $P < 0.05$, a larger number of genes differentially expressed (153 genes) could be identified. In Table 2, we describe the identity of the 42 most differentially expressed genes for each comparison. For convenience, instead of showing the *P*s, we present minus the logarithm of the *P*, which is more directly associated with the significance as it is large for large significance (small *P*s), with a minus signal indicating whether the mean expression is higher on the second condition being compared.

We next used a nonsupervised method of clustering to determine whether the 6 genes with lowest *P*s for each comparison would be capable of grouping samples based on their expression profiles. Using the κ-means algorithm (16), samples were grouped in four clusters on the basis of the expression profile of 18 genes that were nonredundant among the 6 genes with lowest *P*s for all of the comparisons except normal × gastritis (Fig. 2). In the first group, the majority of samples representing normal gastric mucosa (green labels) or gastritis (blue labels) were clustered together with only one tumor sample (red labels) and four samples representing metaplasia (brown labels). The second cluster is composed by the majority of samples corresponding to metaplasia plus two normal and three gastritis. The third cluster is composed of a mix of all four of the tissue classes in which a higher expression of the *c-Myc* oncogene and a lower expression of the *CDKN1A* gene could be detected. The fourth cluster corresponds with the vast majority of tumor samples, either of the intestinal or the diffuse type plus one gastritis and one metaplasia. Interestingly, when we did principal component analysis along all 99 samples using the expression profile of all 370 genes, the group of 6 samples comprising the third cluster, representing all four of the tissue classes, is detached from the remaining 93 samples, additionally corroborating their unique gene expression profile (figure available elsewhere).[5]

Differentially Expressed Genes between Intestinal Type and Diffuse Type Gastric Adenocarcinoma. As proposed by Correa and Chen (4), pathogenesis of the intestinal type of gastric adenocarcinoma has been connected to precursor changes in a progressive fashion going from chronic active gastritis, multifocal atrophic gastritis, intestinal metaplasia, and dysplasia. As can be observed in Fig. 2, the fourth cluster has the majority of tumors (25 of 28), with the majority of intestinal type tumors at the bottom branch (14 of 18). Because we could observe this dichotomy in the clustering of two types of gastric adenocarcinomas, we searched for genes differentially among them as compared with normal gastric mucosa (Fig. 3). Genes with $Ps \leq 0.0009$ (Mann-Whitney Test) are denoted in green (for only one comparison) or in red (for both comparisons). Among the genes with augmented expression in the diffuse type adenocarcinomas we identified matrix metalloproteinase (MMP2), and its overexpression could contribute to the phenotypic characteristics of this tumor. Of notice, expression of the *VHL* gene was diminished on intestinal type adenocarcinomas.

Tissue Expression of MMP2. We found *MMP2* to be expressed in higher levels in both intestinal and diffuse types of adenocarcinomas (Figs. 4*A*). To confirm this observation, we generated a tissue array having 75 samples of diffuse and 75 samples of intestinal types of gastric adenocarcinomas, spotted in duplicates, plus 25 samples of normal gastric mucosa, 50 samples of chronic gastritis, and 25 samples of intestinal metaplasia (total of 500 tissue fragments). In agreement with mRNA levels, both diffuse- and intestinal-type adenocarcinomas showed stronger staining for MMP2 as compared with other entities (Fig. 4, *B* and *C*) with a statistically significant difference when we compared normal and all tumor samples (*P* of 0.036 and 0.034 for Mann-Whitney and *t* test, respectively). The fact that the majority of the samples used for the stomach tissue microarray were different from those used as source of RNA for gene expression increases the significance of these findings. In the case of diffuse-type

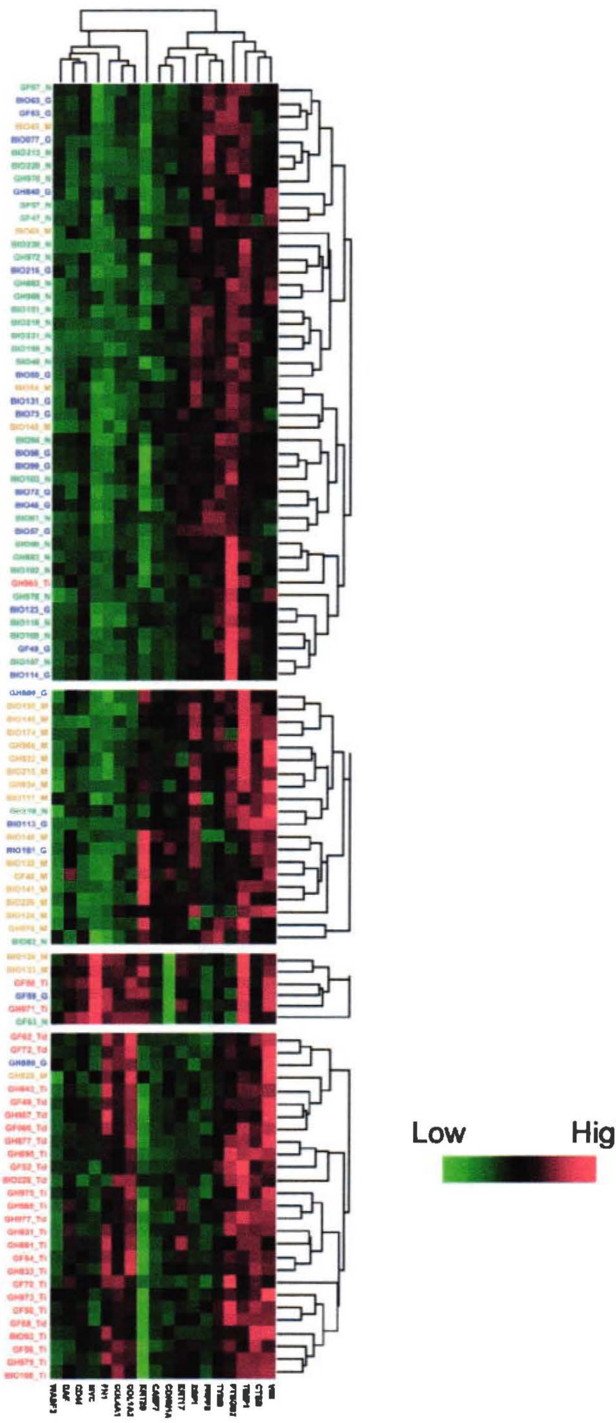


Fig. 2. Clustering of the 99 tissue samples according to the expression profile of 18 genes. Using the κ-means algorithm, 99 tissue samples representing normal gastric mucosa (*green*), gastritis (*blue*), metaplasia (*brown*), and adenocarcinomas (*red*) were grouped into four clusters on the basis of the expression profile of the nonredundant set of 18 genes representing the 6 genes with lowest *P*s for each pair-wise comparison. The *columns* represent genes ordered according to their hierarchical distances. The *red color* denotes high expression, and the *green color* denotes low expression as compared with average expression among all 99 of the samples. Within each cluster, samples were ordered on the basis of their correlation distances.

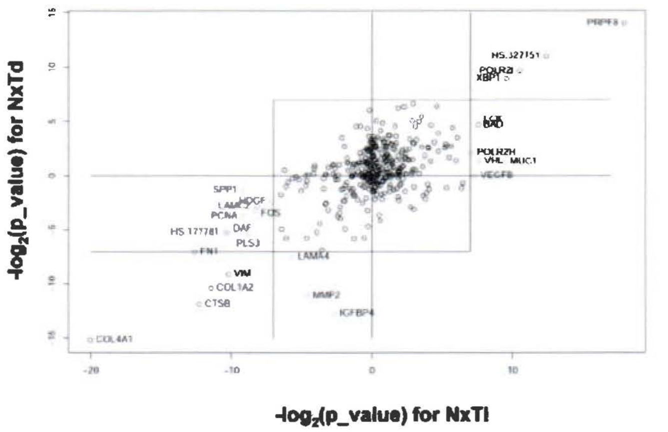


Fig. 3. Scatter plot representing genes differentially expressed between intestinal and diffuse types of gastric adenocarcinomas. The minus $\log_2$ of the *P*s for each gene in the NxTi (*X* axis) and NxTd (*Y* axis) comparisons were plotted with minus signal indicating overexpression in tumor samples. Genes with $P \leq 0.0009$ (Mann-Whitney) are denoted in *green* (for only one comparison) or in *red* (for both comparisons).

adenocarcinomas, we observed that signet ring cells failed to express MMP2 (see panel SRC in Fig. 4*B*). A pictorial case stained with a polyclonal antibody against CTNNB1 is available at the provided website.[5] This figure clearly demonstrates not only the augmented expression but also changes in cellular localization going from the cell surface in the normal sample to the nucleus in both diffuse and intestinal types of adenocarcinomas (17, 18).

Differentially Expressed Genes during the Evolution to Intestinal Type Gastric Adenocarcinoma. We next searched for genes that had their expression progressively altered according to the proposed cascade of evolution from normal gastric mucosa, gastritis, and intestinal metaplasia to intestinal type adenocarcinoma, and with *P*s ≤0.0009 in the extreme comparison *i.e.*, normal *versus* tumor of the intestinal type (tumors of the diffuse type were excluded from this comparison). The top 7 genes with increased and the top 7 genes with decreased expression are represented in Fig. 5. The 7 selected genes with increased expression pattern were *COL1A1, FN1, CTSB, COL1A2, Hs.177781, DAF,* and *VIM.* The 7 selected genes with decreased expression pattern were *PRPF8, Hs.327751, VHL, LCK, BAD, VEGFB,* and *POLR2H.*

Construction of Molecular Classifiers. As the main goal of the present work was the building of molecular classifiers, we next performed an exhaustive search for pairs and trios of genes that could be used for class distinction on the basis of the expression signature of each individual sample. Using the signal intensity of all 370 genes, we applied Fisher Linear Discriminant Analysis (16) and identified all of the possible pairs and trios of genes that could correctly separate tissue samples in each of the six possible comparisons (NxT, NxG, NxM, GxT, GxM, and MxT), allowing none, one, two, or three misclassifications. For the identification of trios we used were a set of 65 samples (learning set). We analyzed 38 samples in NxT comparison, 34 samples in NxG comparison, 29 samples in NxM comparison, 36 samples in GxT comparison, 27 samples in GxM comparison, and 31 samples in MxT comparison. The number of trios found for each comparison is shown in Table 3. We identified 1,044, 17, and 1 trios of genes that could precisely separate all of the normal, gastritis, and intestinal metaplasia from tumor samples, respectively, with perfect distinction of all of the samples. When we accepted one sample misclassified, we found a larger number of trios, 12,278, 512, and 19, for the same comparisons described above, plus 52 and 4 trios that could classify normal and gastritis from intestinal metaplasia, respectively. We could not identify a single trio that could distinguish normal from gastritis even accepting one sample misclassified.

The high number of trios that can distinguish between normal and tumor samples is, in part, because four pair of genes (*KIAA0106-COL4A1, CLTC-COL4A1, XBP1-COL4A1,* and *COL4A1- MCL1*) could precisely separate the set of normal and tumor samples used for training.

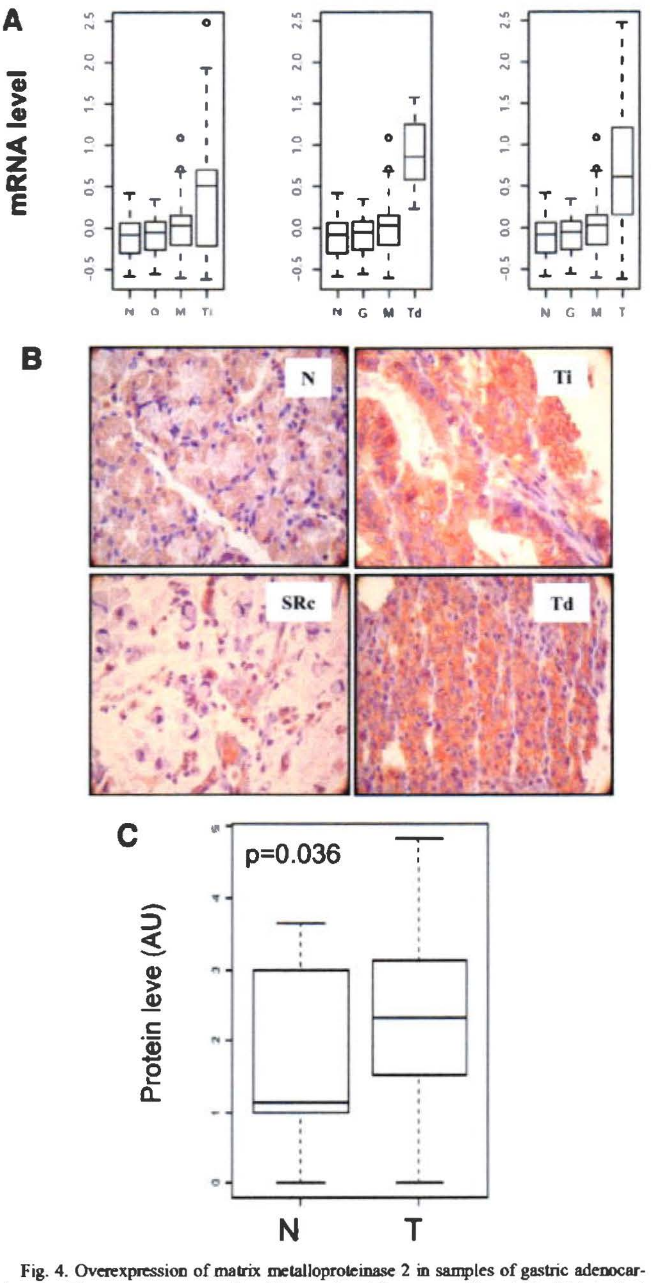


Fig. 4. Overexpression of matrix metalloproteinase 2 in samples of gastric adenocarcinomas. Tissue arrays representing 25 normal gastric mucosa, 50 gastritis, 25 intestinal metaplasia, and 75 samples of each intestinal and diffuse types of adenocarcinomas were stained with anti-matrix metalloproteinase 2 (Oncogene Science). In *A* we have box plots indicating mRNA expression levels in all four groups of samples (*N* = normal gastric mucosa; *G* = gastritis; *M* = intestinal metaplasia; *Ti* = intestinal-type adenocarcinoma; and *Td* = diffuse-type adenocarcinoma) with *P*s of 0.0097 for NxTi, 1.5×10^{-5} for NxTd, and 1.9×10^{-5} for NxT. In *B* we have representative fields of the tissue array stained for matrix metalloproteinase 2 (*n* = normal gastric mucosa; *Ti* = intestinal-type adenocarcinoma; *Td* = diffuse-type adenocarcinoma; *SRc* = signet ring cells; magnification = ×400). In *C* we represent protein levels in normal and tumor samples (diffuse plus intestinal types) in arbitrary units as defined in "Materials and Methods," as well as the corresponding *P* (Mann-Whitney).

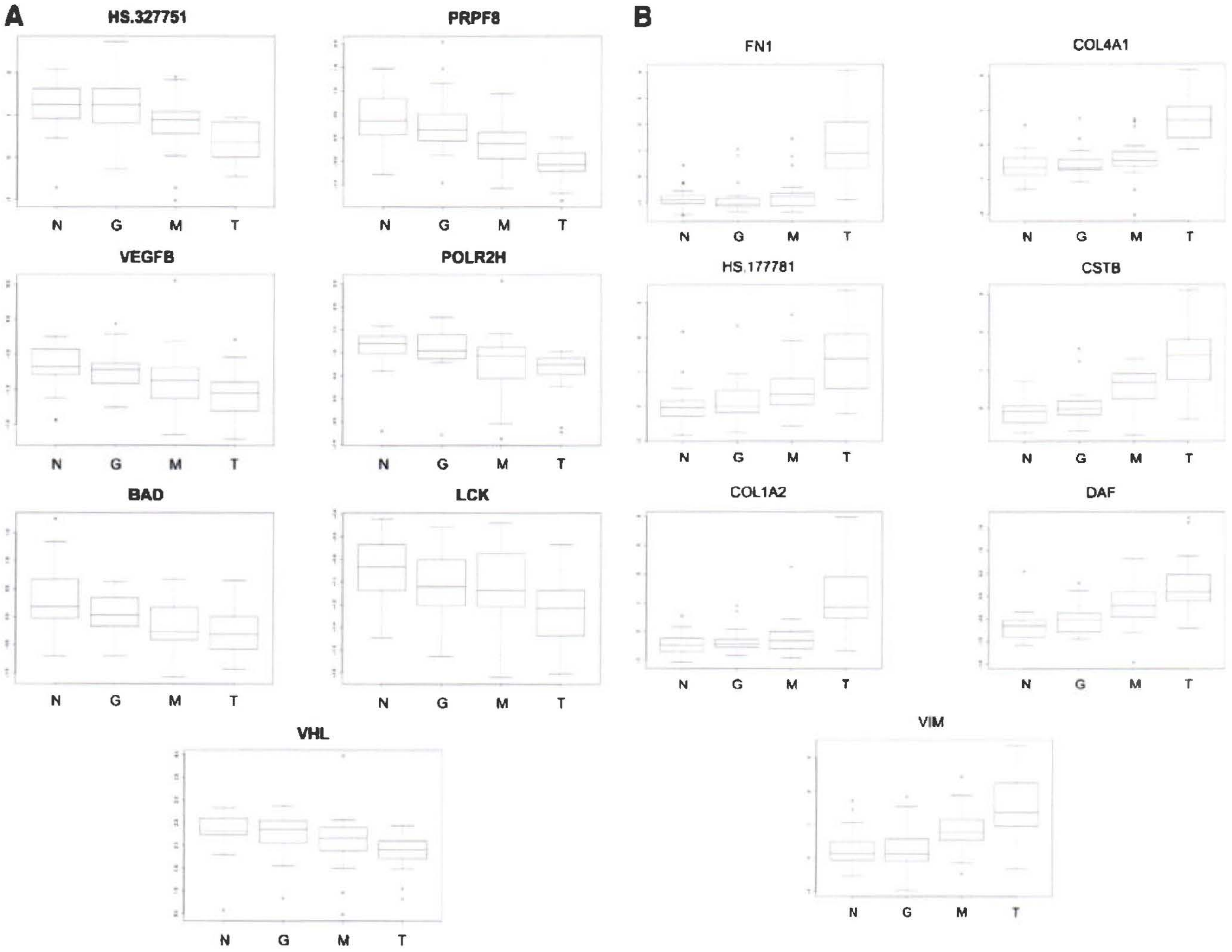


Fig. 5. Genes with progressively altered expression according to the cascade of evolution of intestinal-type adenocarcinoma. The top 7 genes with increased (*A*) or decreased (*B*) expression pattern according to the cascade of evolution from normal gastric mucosa, gastritis, intestinal metaplasia, and intestinal-type adenocarcinoma are presented. All of the genes have *Ps* ≤ 0.0009 (Mann-Whitney) for the NxTi comparison.

Next, all of the identified trios represented in Table 3 were tested again, in an independent set of 34 samples (validation set), corresponding with an increase of 18 samples in NxT comparison, 15 samples in NxG comparison, 21 samples in NxM comparison, 13 samples in GxT comparison, 16 samples in GxM comparison, and 19 samples in MxT comparison. In Fig. 6 we represent the trio with highest SVD (*COL4A1*, *XBP1*, and *RPL14*) and its performance using the training set of samples (Fig. 6*A*) and the separation of the validation set of samples (Fig. 6*B*, samples represented by stars) using the same rule defined during training. On the basis of their performance on the learning set of samples, trios were ranked by their SVD score (square root of the ratio of between groups and within groups sum of squares, as explained above). Trios with highest SVD show the least dispersed distribution of samples of a given class and the greater distance between the two classes. We selected the 100 trios with highest SVD score and again determined their performance on the validation set of samples using the rule defined during training (Fig. 6*C*) or allowed definition of new rules (Fig. 6*D*). By comparing Fig. 6, *C* and *D*, it can be notice that, when new rules are allowed, classification of samples GF63 improves, but samples BIO124 and GH971 are misclassified by a larger number of trios. The actual numbers for this heatmap is available elsewhere.[5]

Table 3 *Identification of classifiers*

Comparison	Number of misclassified samples 0	1	2	3
N X T	1,044	12,278	45,636	66,428
G X T	17	512	4,741	27,548
M X T	1	19	808	24,648
N X M	0	52	354	10,810
G X M	0	4	110	1,223
N X G	0	0	0	2

To estimate the likelihood that good classifiers as the ones we describe were the result of chance, we performed resampling experiments based on a sequential search method for classifier, restricted to the normal × tumor comparison. If N represents the total number of genes ($n = 370$ in the present situation), the method starts by selecting the best k discriminating genes among the original N genes. Denoting this set of genes by G1, we next search the k best distinct discriminating pairs of genes such that at least one of them belongs to G1. Denoting this set of pairs of genes by G2, we then proceed to search for the k best distinct discriminating trios (denoted by G3) such that at least one of its pairs belong to G2. We finally applied a similar process to get the set G4, the k best distinct discriminating groups of

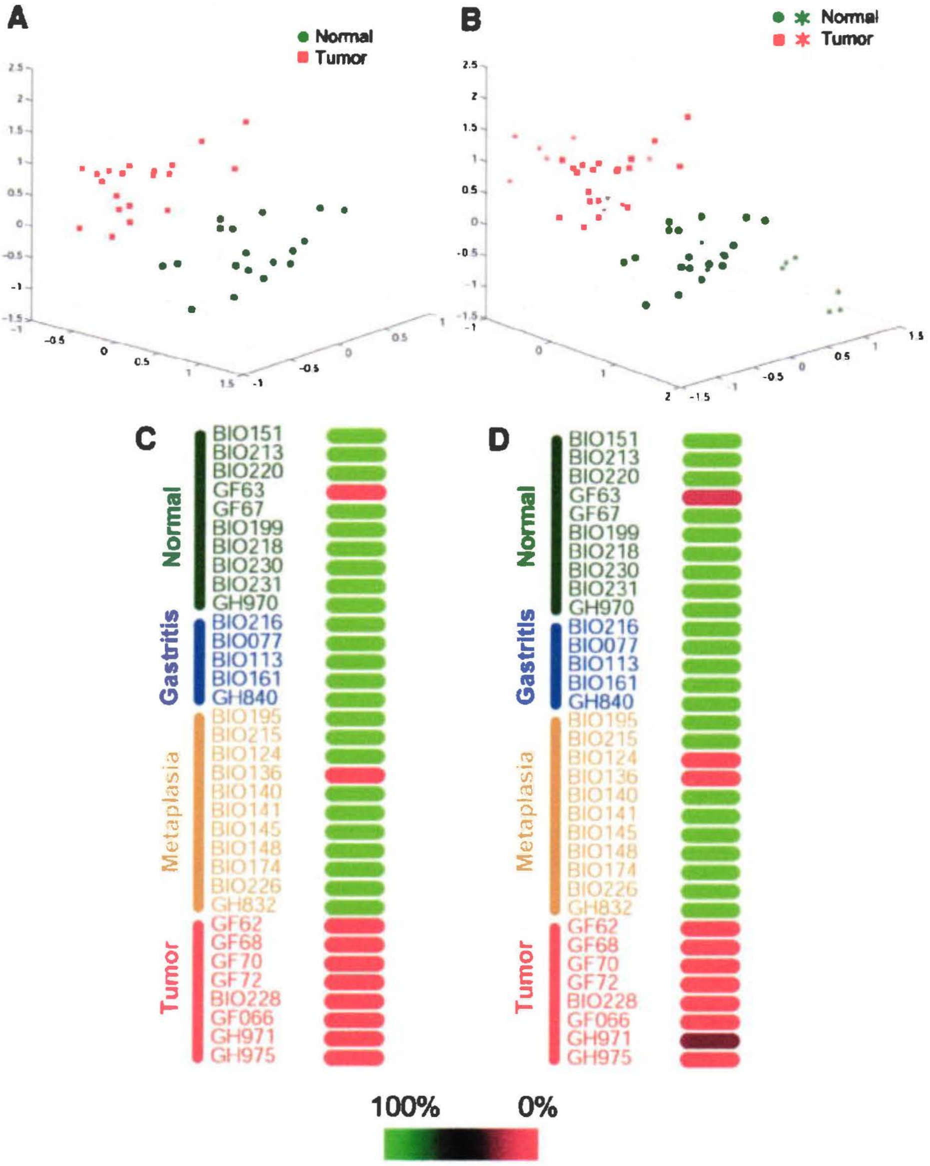


Fig. 6. Training and validation of classifiers normal and tumor samples. In *A* we represent the trio (*COL4A1*, *XBP1*, and *RPL14*) with highest SVD and its performance on the training set of normal (*green circles*) and tumor (*red squares*) samples. In *B* we represent the performance of the same trio, for class distinction of the validation set of samples (represented by *stars*) using the same rule defined during training. In *C* and *D*, we represent the performance of the top 100 trios for normal x tumors, based on their singular value decomposition, using either, the same rule defined during training (*C*) or allowing definition of new rules for each trio during validation (*D*). The exact numbers represented by the color scale can be obtained at the provided website.[5] The color scale represents frequency that a given sample was classified as normal by all classifiers.

four genes. The motivation of introducing this method is that it can be implemented much faster than the original exhaustive search, and, therefore, is amenable to bootstrap procedures. We applied the method with k = 100 for 1000 bootstrap samples and never found, for any bootstrap search, better results than those corresponding with the real dataset, which strongly suggests that our findings are not due to chance.

Molecular Classification of the 99 Gastric Tissue Samples. Having identified trios and their respective SVD, we determined the efficiency of these trios to correctly classify all 99 samples. As the number of trios for NxT classification was quite large and very redundant, we selected the top 100 trios for the NxT separation, according to their SVD. For the remaining classification, we used 17, 20, 52, and 4 trios for the GxT, MxT, NxM, and GxM, respectively. The number of genes involved in these trios is 103, 24, 38, 57, and 11, for NxT, GxT, MxT, NxM, and GxM, respectively (the identity of the genes, the structure of the trios, and the identity of misclassified samples can be obtained elsewhere).[5] Each of the 99 samples was then classified by all trios of the five comparisons. In Fig. 7, we represent the performance of four trios that can classify NxT, GxT, and MxT. For the NxT classification, we represent the trio with highest SVD with all 54 normal and tumor samples (Fig. 7*A*) and with all 99 samples (Fig. 7*B*). We also represent the best trio that precisely separates all normal and tumor samples (Fig. 7*C*), and the same trio with all 99 samples (Fig. 7*D*). In Fig. 7, *E* and *F*, we represent the trio with highest SVD for GxT and MxT separation with the corresponding samples. The performance of all of the trios against all of the samples is represented in Fig. 8. The spectrum of colors from green to red was used to denote 100% to 0% classification as the first entity of each comparison (for example, in the first line, representing NxT, 100% means all trios classifying the corresponding sample as N and 0% means all trios classifying the corresponding sample as tumor). Thus, a black bar denotes that a sample was classified by 50% of the trios as one of the two entities.

It was expected that samples pertaining to one of the two classes of the classifiers would be colored as green and red because the classifiers were built specifically for them. Indeed, in the first three columns, representing the classifiers having tumor as the second entity

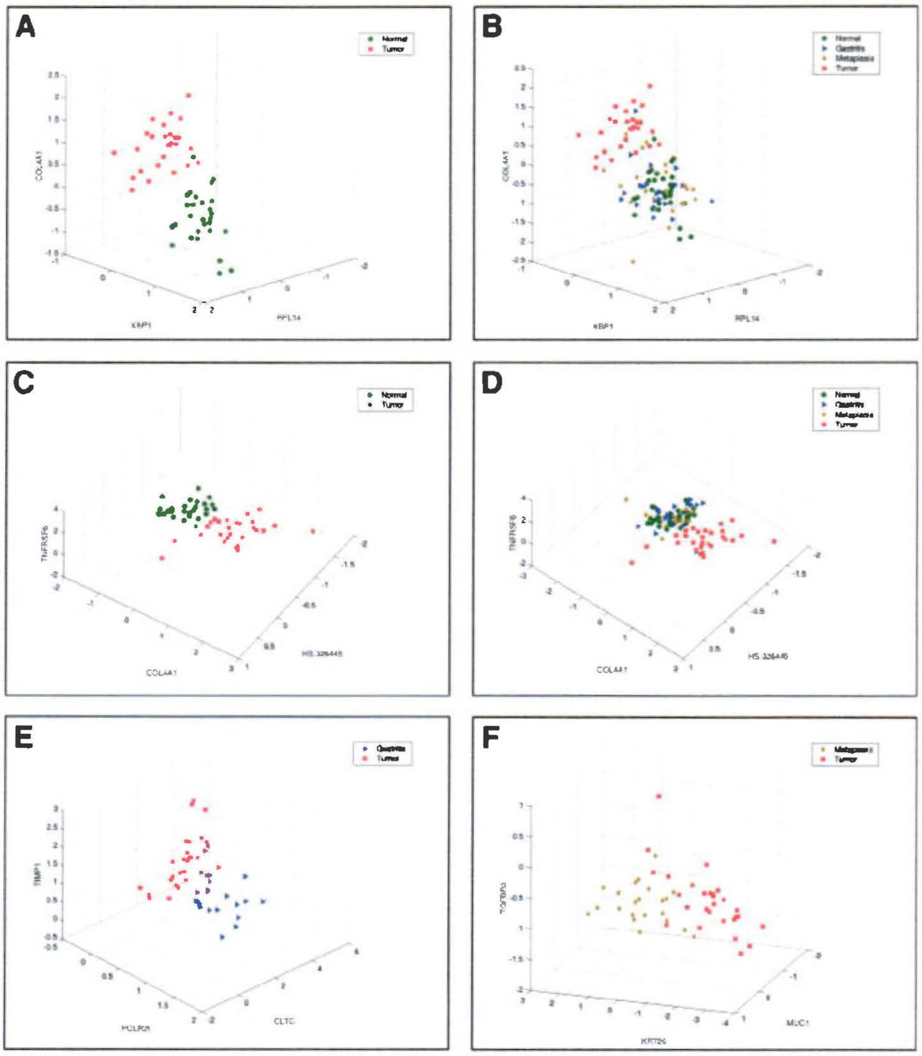


Fig. 7. Class distinction by trios identified by Fisher's linear discriminant analysis. The performance of four trios that can classify NxT, GxT, and MxT are represented. For the NxT classification, we represent the trio with highest singular value decomposition with the 54 normal and tumor samples (*A*) and with all 99 samples (*B*). We also represent one trio with perfect NxT classification with the 54 normal and tumor samples (*C*), and with all 99 samples (*D*). In *E* and *F* we represent the trios with best singular value decomposition for GxT and MxT classification with the corresponding samples. Genes are plotted according to their normalized background-corrected log intensities.

produced the vast majority of red bars at the bottom of the figure. Interestingly, there was one normal sample (GF63), two gastritis samples (GF59 and GH880), and three samples of intestinal metaplasia (BIO133, GH828, and BIO136) that were classified as a tumor by >50% of the classifiers for the NxT comparison. Samples GF63, GF59, GH880, BIO133, GH828, and BIO136 were considered as tumor by 93, 66, 98, 88, 99, and 91 trios, respectively. These samples are those that, in Fig. 2, are placed in the third and fourth clusters and, at the principal component analysis, are detached from the remaining samples (figure available).[5] Also compatible with the number of genes with significant expression differences observed in Fig. 1, there was a less clear separation of samples in the two columns representing NxM and GxM comparisons but nevertheless, a larger number of red bars can be observed in samples representing intestinal metaplasia.

To determine whether the biological phenomena that were supporting class distinction between tumor and nontumor samples were somehow related, we identified the genes composing the top 45 trios for the NxT distinction plus the genes composing the 17 trios for GxT and the 20 trios for MxT. The number of genes was 50, 24, and 38, respectively, and only 1 gene, *CTSB*, was at the intersection of the three groups, whereas the sum of the three groups has 100 genes indicating very little redundancy among the genes.

DISCUSSION

During the last 2 years, several groups, including ourselves (12), published studies focusing on the expression profile of gastric cancer and the identification of differentially expressed genes (19–25). Whereas the majority of the published data are based on the comparison between normal and tumor tissues, less information concerning the molecular events that could establish a link between gastric cancer and intestinal metaplasia is available (11). Likewise, no systematic effort for the construction of molecular classifiers for gastric cancer that could impact on early diagnosis has been reported.

In the present work, we describe the identification of a large set of molecular classifiers that could be used for distinction between normal gastric mucosa and those representing gastritis, intestinal metaplasia, and gastric adenocarcinomas. We have also identified genes of which the gene expression profile can be correlated with the transition stages between normal and intestinal type of adenocarcinomas and genes that are differentially expressed between diffuse and intestinal type of gastric adenocarcinomas.

Before searching for classifiers, we first validated our cDNA array by searching for differentially expressed genes in the various disease states. As expected there were fewer differentially expressed genes when we compared samples representing the nontumor states than

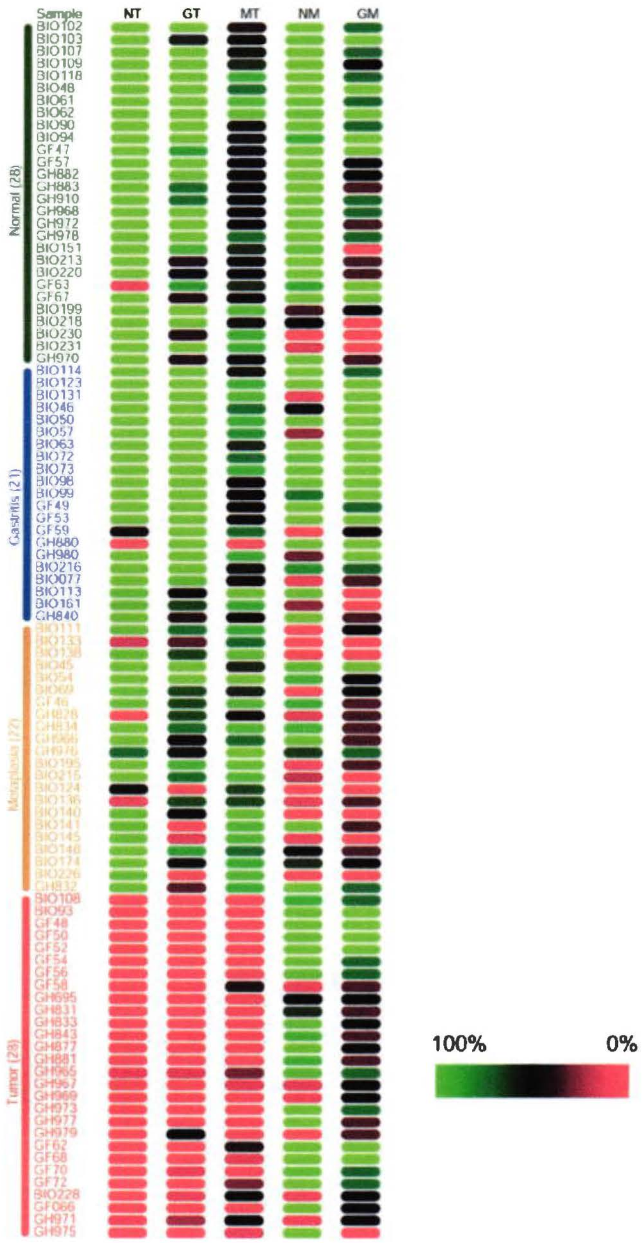


Fig. 8. Molecular classification of 99 gastric samples. Each of the 99 samples was classified by the trios according to the five comparisons. The numbers of trios were 100 for NT, 17 for GT, 20 for MT, 52 for NM, and 4 for GM. Color scale represents frequency that a given sample was classified as the first entity of the comparison by each of the trios for that comparison (100% = *green*; 0% = *red*).

comparisons having tumor samples (Fig. 1 and Table 2). As a way to confirm the correlation between our findings and the biology of the process we are dealing with, we determined the pattern of expression of genes that, based on current knowledge, should be altered in our data set. Alterations in the Wnt pathway have been found in a large set of gastric adenocarcinomas, and overexpression of β-catenin (*CTNNB1*) in gastric tumors has been described (12, 17, 18, 23, 25). We observed over expression of CTNNB1 in this collection of samples in both intestinal and diffuse types of gastric adenocarcinomas (figure available).[5] It was reported recently that MMP2 mRNA is elevated in both diffuse and intestinal types of gastric adenocarcinomas (11), and we have confirmed these observations both at the RNA and at the protein levels (Fig. 4). Also, *KRT20* is typically augmented in intestinal metaplasia and, in agreement with this notion; we found its expression augmented in relation to all other three of the tissue classes (see Table 2).

The cDNA microarray technology has been largely applied to cancer research (26), and expression profile is being increasingly used for distinction between physiological and disease states, as well as to distinguish between groups of disease samples of which the expression profile can discriminate between clinically or biologically similar entities (27–29). Using a set of 18 nonredundant genes representing the 6 genes with lowest *P*s for all six of the pair-wise comparisons, we applied the nonsupervised algorithm *k*-means to see whether our samples could be grouped accordingly. In Fig. 3, we can observe that the vast majority of samples representing normal and gastritis samples are grouped together, and most of the samples representing intestinal metaplasia and tumors groups in two other clusters. Interestingly, a small cluster with only 6 samples of all four of the entities could be observed and, by the pattern of their expression profile, it is clear that overexpression of *MYC* and lower expression of *CDKN1A* is the hallmark of this group of samples. When we did principal component analysis along all 99 samples, this group of 6 samples is detached from the remaining 93 samples (figure available).[5] Also, *KRT20* was expressed at higher levels in virtually half of the samples from the cluster where the majority of intestinal metaplasias were grouped (see upper branch). There is a controversy on the literature concerning the pattern of expression of *KRT20* in intestinal metaplasias of the gastric esophageal junction and its pattern of expression in gastric mucosa (30–35). Apparently, *KTR20* expression in intestinal metaplasias of the gastric mucosa is also variable. Importantly, we repeated the clustering of samples using self-organizing maps, also a nonsupervised algorithm, and again, samples were grouped with comparable results (data not shown).

The phenotypic and biological differences observed between intestinal- and diffuse-type gastric adenocarcinoma should be a consequence of differences in gene expression. The hallmark of diffuse-type gastric adenocarcinoma is the lack of glandular organization with spreading of tumor cells throughout the parenchyma implying the need for extracellular matrix destruction. In agreement with this notion and in agreement with data presented by Boussioutas *et al.* (11), we found *MMP2* to be expressed at higher levels in both types of tumor, with diffuse type having even higher mRNA levels than the intestinal type (Fig. 4*A*). Again, these data were confirmed by immunohistochemistry in a collection of 150 tumor samples (Fig. 4, *B* and *C*). Interestingly, signet-ring cells, frequently observed in diffuse type of adenocarcinomas, did not express MMP2 (Fig. 4*B*, signet ring cells or SRc).

To additionally contribute to the understanding of the cascade of events that takes place during the oncogenesis of intestinal type gastric adenocarcinoma, we identified genes of which the expression showed a constant increase or decrease along the cascade from normal to tumor samples with a $P \leq 0.0009$ between the normal and the tumor samples (Fig. 5). Five of the 7 genes with increased expression toward malignancy have functions related to the extracellular matrix (*COL4A1, FN1, CTSB, COL1A2,* and *VIM*). Another gene, *DAF*, also showed augmented expression from normal to gastritis, metaplasia, and tumor. *DAF* could be involved in escaping from complement, and its increase expression in gastric cancer was also observed by serial analysis of gene expression analysis (SAGE anatomical viewer[8]). Among the genes of which the expression decreased from normal to tumor samples, we identified the tumor suppressor gene *VHL*. There are few reports in the literature where the status of *VHL* gene in gastric

[8] Internet address: http://cgap.nci.nih.gov/SAGE/viewer.

cancer was investigated. Leung *et al.* (36) failed to demonstrate hypermethylation of the promoter region of *VHL* in 5 gastric cancer cell lines and 26 gastric carcinomas. Diminished expression of the *VHL* gene, assessed by immune histochemistry was found in 27 of 318 samples of gastric carcinomas (37). Of notice, there was a diminished expression of *BAD*, a proapoptotic gene, and such reduced expression could contribute to cell survival. It would be important to see whether there is a correlation between reduced *BAD* expression and expression of trefoil factor 1, known to antagonizes BAD-induced apoptosis in gastric mucosa (38). Finally it is likely that reduced expression of *LCK* just reflects the reduction of infiltrating lymphocytes, because our samples were dissected to exclude inflammatory cells from tumor samples.

Having compared tumor types and identified genes of which the pattern of expression correlates with disease progression we next searched for genes that could be used for the construction of molecular classifiers. As mentioned before, different clustering approaches were successfully used to distinguish between tumor and nontumor samples (24), morphologically similar samples (29, 11), and determine disease outcome (27, 39). Whereas a large group of statistical methods is available for such a task, (reviewed in Ref. 40), they are often based on the expression of a large set of genes and, necessarily, require data from the two sample groups to identify the set of genes where similarities and differences can be used to define the clusters. To be routinely applied, these requirements could represent potential pitfalls. An alternative would be the implementation of supervised learning procedures where classification could be based on expression signatures rather then comparative expression profile, as proposed earlier (40–42). The major advantage would be the possibility of creating a database against which the test sample would be classified, as demonstrated by Ramaswamy *et al.* (41). A known group of samples could be used for training and the resulting classifier used for prediction of an unknown sample (class prediction), as demonstrated by Golub *et al.* (29). Support vector machine, an example of supervised learning algorithm, was successfully used for class distinction by Shipp *et al.* (43) and by us (12). Several other mathematical methods such as Nearest Neighbors Classifiers and Classification Trees (44) could also be applied to search for groups of genes with different expression patterns or sample signature. It appears to us that Fisher's linear discriminant analysis (16) attains a good compromise between simplicity and performance, making it a good choice for this investigation. Moreover, this approach to identify expression signatures corresponds to the usual approach to identify differentially expressed (single) genes, based on the t-statistics.

Hence, we decided to apply Fisher's linear discriminant and, by exhausted search among all 370 genes, we identified trios of genes that could be used for class distinction of all four entities in a pair-wise manner (Table 3). First, it was interesting to observe that the closer along the cascade of disease progression two samples are, the lowest the number of trios. Also, many trios that can distinguish between normal and tumor samples can be explained by the fact that four pairs of genes (*KIAA0106-COL4A1, CLTC-COL4A1, XBP1-COL4A1,* and *COL4A1- MCL1*) could precisely separate all 38 of the normal and tumor samples used in the training set of samples. Hence, the list of trios for this comparison is highly redundant for trios composed of these four pairs.

It is noteworthy that, according to our strategy to select cDNA fragments for immobilization in the array, many genes were represented by two or even three cDNA fragments representing distinct regions of the same gene. Thus, it was interesting to determine whether classifiers based on the average signal intensity for all spots of different cDNA fragments corresponding to a given gene would be reproduced when signal intensity of replicas for a single cDNA fragment of that given gene was considered. Indeed, there was a strong correlation between classifiers identified by the two strategies.

All of the trios identified as potential classifiers using the training set of samples were first validated against an independent group of samples, the validation set, comprising 34 new samples using the same classification rule defined during training (Fig. 6*C*) or by new rules defined during validation (Fig. 6*D*). The performances of all of the trios were ranked by their SVD score. Interestingly, the trio with highest SVD, composed of *COL4A1*, *XBP1*, and *RPL14*, misclassified one normal sample (GF63), whereas a trio with lower SVD (*TNFRS6*, *COL4A1*, and *Hs.325445*) correctly classified all 54 of the normal plus tumor samples (Fig. 7, *A* and *C*, respectively). With the exception of NxT classification, for which we selected only the top 100 trios with highest SVD, the performance of all of the trios for all five of the comparisons using all 99 samples is represented in Fig. 8. It is noteworthy that, from the 150 genes present in top 45 trios for NxT plus the 17, 20, 52, and 4 trios for the GT, MT, NM, and GM, respectively, a single gene, *CTSB*, is present in the five groups of classifiers. These data strongly suggest that the biological phenomena that these genes are involved in differ for each comparison.

There is a strong correlation between data from classifiers and from cluster analysis. Samples that are misclassified by the trios are those that were not grouped in the expected cluster in Fig. 2. For the NxT comparison (Fig. 8, first column) there are 6 nontumor samples that were classified as tumor by >50% of the trios; GF 63 (N), GF59 (G), BIO133 (M), and BIO136 (M) were all grouped in the third cluster, whereas GH828 (M) and GH880 (G) fall into the fourth cluster, together with the majority of tumor samples. Together, these observations suggest that the expression profile of the 18 genes used for clustering reflects the structure detected by the signatures identified by the trios. It is important to remember that the 100 trios used for NxT classification are composed of 103 genes. Also, it is clear that the closer the samples are in the cascade of disease progression, the higher the number of misclassified samples, probably reflecting a gradual change in the pattern of gene expression. Nonetheless, there is always a predominant distribution of green and red bars for the samples that represents a given comparison.

The classifiers presented here, based on trios of genes defined by Fisher's linear discriminant analysis, were able to discriminate tissue samples representing the cascade of events related to gastric adenocarcinomas and should now be validated for class prediction. Importantly, it is now imperative to apply these classifiers to a large set of samples representing intestinal metaplasia to determine the correlation between misclassification of these samples as adenocarcinomas and the frequency they transform into malignant disease. This study will require a large collection of samples and a period of follow-up that cannot be precisely anticipated but is now under way, in a multicentric effort.

ACKNOWLEDGMENTS

We thank Carlos F. Nascimento and Sueli Nonogaki for helping with tissue array, and Dr. Ricardo Brentani for critically reading this manuscript.

REFERENCES

1. Stadtlander, C. T., and Waterbor, J. W. Molecular epidemiology, pathogenesis and prevention of gastric cancer. Carcinogenesis (Lond.), *20:* 2195–2208, 1999.
2. Oliveira, F. J., Ferrao, H., Furtado, E., Batista, H., and Conceicao, L. Early gastric cancer: report of 58 cases. Gastric Cancer, *1:* 51–56, 1998.
3. Lauren, P. The two histological main types of gastric carcinoma: difuse and so-called intestinal-type carcinoma. Acta Pathol. Microbiol. Scand., *64:* 31–49, 1965.
4. Correa, P., and Chen, V. W. Gastric cancer. Cancer Surv., *19–20:* 55–76, 1994.
5. Peek, R. M., Jr., and Blaser, M. J. Helicobacter pylori and gastrointestinal tract adenocarcinomas. Nat. Rev. Cancer, *2:* 28–37, 2002.

6. Guilford, P., Hopkins, J., Harraway, J., McLeod, M., McLeod, N., Harawira, P., Taite, H., Scoular, R., Miller, A., and Reeve, A. E. E-cadherin germline mutations in familial gastric cancer. Nature (Lond.), *392:* 402–405, 1998.
7. Shiao, Y. H., Rugge, M., Correa, P., Lehmann, H. P., and Scheer, W. D. p53 alteration in gastric precancerous lesions. Am. J. Pathol., *144:* 511–517, 1994.
8. Sung, J. J., Leung, W. K., Go, M. Y., To, K. F., Cheng, A. S., Ng, E. K., and Chan, F. K. Cyclooxygenase-2 expression in Helicobacter pylori-associated premalignant and malignant gastric lesions. Am. J. Pathol., *157:* 729–735, 2000.
9. Dubois, R. N. Review article: cyclooxygenase–a target for colon cancer prevention. Aliment. Pharmacol. Ther., *14*(Suppl. 1): 64–67, 2000.
10. Kang, G. H., Shim, Y. H., Jung, H. Y., Kim, W. H., Ro, J. Y., and Rhyu, M. G. CpG island methylation in premalignant stages of gastric carcinoma. Cancer Res., *61:* 2847–2851, 2001.
11. Boussioutas, A., Li, H., Liu, J., Waring, P., Lade, S., Holloway, A. J., Taupin, D., Gorringe, K., Haviv, I., Desmond, P. V., and Bowtell, D. D. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res., *63:* 2569–2577, 2003.
12. Meireles, S. I., Carvalho, A. F., Hirata, R., Montagnini, A. L., Martins, W. K., Runza, F. B., Stolf, B. S., Termini, L., Neto, C. E., Silva, R. L., Soares, F. A., Neves, E. J., and Reis, L. F. Differentially expressed genes in gastric tumors identified by cDNA array. Cancer Lett., *190:* 199–211, 2003.
13. Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques, *34:* 374–378, 2003.
14. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., *30:* e15, 2002.
15. Edwards, D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics, *19:* 825–833, 2003.
16. Hastie, T., Tibshirani, R., and Friedman, J. The elements of statistical learning. New York: Springer-Verlag, 2001.
17. Grabsch, H., Takeno, S., Noguchi, T., Hommel, G., Gabbert, H. E., and Mueller, W. Different patterns of β-catenin expression in gastric carcinomas: relationship with clinicopathological parameters and prognostic outcome. Histopathology, *39:* 141–149, 2001.
18. Woo, D. K., Kim, H. S., Lee, H. S., Kang, Y. H., Yang, H. K., and Kim, W. H. Altered expression and mutation of β-catenin gene in gastric carcinomas and cell lines. Int. J. Cancer, *95:* 108–113, 2001.
19. Hasegawa, S., Furukawa, Y., Li, M., Satoh, S., Kato, T., Watanabe, T., Katagiri, T., Tsunoda, T., Yamaoka, Y., and Nakamura, Y. Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23, 040 genes. Cancer Res., *62:* 7012–7017, 2002.
20. Inoue, H., Matsuyama, A., Mimori, K., Ueo, H., and Mori, M. Prognostic score of gastric cancer determined by cDNA microarray. Clin. Cancer Res., *8:* 3475–3479, 2002.
21. Liu, L. X., Liu, Z. H., Jiang, H. C., Qu, X., Zhang, W. H., Wu, L. F., Zhu, A. L., Wang, X. Q., and Wu, M. Profiling of differentially expressed genes in human gastric carcinoma by cDNA expression array. World J. Gastroenterol., *8:* 580–585, 2002.
22. Lee, S., Baek, M., Yang, H., Bang, Y. J., Kim, W. H., Ha, J. H., Kim, D. K., and Jeoung, D. I. Identification of genes differentially expressed between gastric cancers and normal gastric mucosa with cDNA microarrays. Cancer Lett., *184:* 197–206, 2002.
23. El Rifai, W., Frierson, H. F., Jr., Harper, J. C., Powell, S. M., and Knuutila, S. Expression profiling of gastric adenocarcinoma using cDNA array. Int. J. Cancer, *92:* 832–838, 2001.
24. Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J. M., Fukayama, M., Kodama, T., and Aburatani, H. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res., *62:* 233–240, 2002.
25. Mori, M., Mimori, K., Yoshikawa, Y., Shibuta, K., Utsunomiya, T., Sadanaga, N., Tanaka, F., Matsuyama, A., Inoue, H., and Sugimachi, K. Analysis of the gene-expression profile regarding the progression of human gastric carcinoma. Surgery (St. Louis), *131:* S39–S47, 2002.
26. Yeatman, T. J. The future of clinical cancer management: one tumor, one chip. Am. Surg., *69:* 41–44, 2003.
27. van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van, d., V, Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med., *347:* 1999–2009, 2002.
28. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature (Lond.), *403:* 503–511, 2000.
29. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash. DC), *286:* 531–537, 1999.
30. Couvelard, A., Cauvin, J. M., Goldfain, D., Rotenberg, A., Robaszkiewicz, M., and Flejou, J. F. Cytokeratin immunoreactivity of intestinal metaplasia at normal oesophagogastric junction indicates its aetiology. Gut, *49:* 761–766, 2001.
31. Jovanovic, I., Tzardi, M., Mouzas, I. A., Micev, M., Pesko, P., Milosavljevic, T., Zois, M., Sganzos, M., Delides, G., and Kanavaros, P. Changing pattern of cytokeratin 7 and 20 expression from normal epithelium to intestinal metaplasia of the gastric mucosa and gastroesophageal junction. Histol. Histopathol., *17:* 445–454, 2002.
32. Mohammed, I. A., Streutker, C. J., and Riddell, R. H. Utilization of cytokeratins 7 and 20 does not differentiate between Barrett's esophagus and gastric cardiac intestinal metaplasia. Mod. Pathol., *15:* 611–616, 2002.
33. Mouzas, I. A., Jovanovic, I., Milosavljevic, T., Tzardi, M., and Kanavaros, P. Cytokeratin immunoreactivity of intestinal metaplasia. Gut., *51:* 894–895, 2002.
34. Odze, R. Cytokeratin 7/20 immunostaining: Barrett's oesophagus or gastric intestinal metaplasia? Lancet, *359:* 1711–1713, 2002.
35. Ormsby, A. H., Goldblum, J. R., Rice, T. W., Richter, J. E., Falk, G. W., Vaezi, M. F., and Gramlich, T. L. Cytokeratin subsets can reliably distinguish Barrett's esophagus from intestinal metaplasia of the stomach. Hum. Pathol., *30:* 288–294, 1999.
36. Leung, W. K., Yu, J., Ng, E. K., To, K. F., Ma, P. K., Lee, T. L., Go, M. Y., Chung, S. C., and Sung, J. J. Concurrent hypermethylation of multiple tumor-related genes in gastric carcinoma and adjacent normal tissues. Cancer (Phila.), *91:* 2294–2301, 2001.
37. Lee, H. S., Lee, H. K., Kim, H. S., Yang, H. K., and Kim, W. H. Tumour suppressor gene expression correlates with gastric cancer prognosis. J. Pathol., *200:* 39–46, 2003.
38. Bossenmeyer-Pourie, C., Kannan, R., Ribieras, S., Wendling, C., Stoll, I., Thim, L., Tomasetto, C., and Rio, M. C. The trefoil factor 1 participates in gastrointestinal cell differentiation by delaying G1-S phase transition and reducing apoptosis. J. Cell Biol., *157:* 761–770, 2002.
39. Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med., *8:* 816–824, 2002.
40. Quackenbush, J. Computational analysis of microarray data. Nat. Rev. Genet., *2:* 418–427, 2001.
41. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, *98:* 15149–15154, 2001.
42. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, *16:* 906–914, 2000.
43. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med., *8:* 68–74, 2002.
44. Dudoit, S. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., *97:* 77–87, 2002.