(11) Número de Publicação: **PT 2468885 E**

(51) Classificação Internacional:
C12Q 1/68 (2015.01)

(12) **FASCÍCULO DE PATENTE DE INVENÇÃO**

(22) Data de pedido: **2007.07.20**

(30) Prioridade(s): **2006.07.20 EP 06291176**

(43) Data de publicação do pedido: **2012.06.27**

(45) Data e BPI da concessão: **2015.08.19**
232/2015

(73) Titular(es):
TRANSMEDI SA
15 RUE DU BOIS DE LA CHAMPELLE 54500 VANDOEUVRE LES NANCY **FR**

(72) Inventor(es):
BERNARD BIHAIN **FR**

(74) Mandatário:
ALBERTO HERMÍNIO MANIQUE CANELAS
RUA VÍCTOR CORDON, 14 1249-103 LISBOA **PT**

(54) Epígrafe: **INFIDELIDADE DA TRANSCRIÇÃO, SUA DETECÇÃO E UTILIZAÇÕES**

(57) Resumo:
PRESENTE INVENTO ESTÁ RELACIONADO COM A IDENTIFICAÇÃO DE UM NOVO MECANISMO DE INFIDELIDADE DA TRANSCRIÇÃO NAS CÉLULAS. O INVENTO PROPORCIONA COMPOSIÇÕES E MÉTODOS PARA DETECTAR O NÍVEL DE INFIDELIDADE DA TRANSCRIÇÃO NUMA AMOSTRA, ASSIM COMO A SUA UTILIZAÇÃO, E.G., PARA TERAPÊUTICA, DIAGNÓSTICO, FARMACOGENÉTICA OU DESENHO DE FÁRMACOS. COMO SERÁ DESCRITO, O INVENTO É PARTICULARMENTE ADEQUADO PARA A DETECÇÃO, MONITORIZAÇÃO OU TRATAMENTO DE DISTÚRBIOS PROLIFERATIVOS CELULARES, PARA O DESENHO E/OU RASTREIO DE FÁRMACOS, PARA ESTABELECIMENTO DO PERFIL DE DOENTES OU DOENÇAS, PREVISÃO DA GRAVIDADE DAS DOENÇAS E AVALIAÇÃO DA EFICÁCIA DE FÁRMACOS.

RESUMO

"INFIDELIDADE DA TRANSCRIÇÃO, SUA DETECÇÃO E UTILIZAÇÕES"

O presente invento está relacionado com a identificação de um novo mecanismo de infidelidade da transcrição nas células. O invento proporciona composições e métodos para detectar o nível de infidelidade da transcrição numa amostra, assim como a sua utilização, *e.g.*, para terapêutica, diagnóstico, farmacogenética ou desenho de fármacos. Como será descrito, o invento é particularmente adequado para a detecção, monitorização ou tratamento de distúrbios proliferativos celulares, para o desenho e/ou rastreio de fármacos, para estabelecimento do perfil de doentes ou doenças, previsão da gravidade das doenças e avaliação da eficácia de fármacos.

DESCRIÇÃO

"INFIDELIDADE DA TRANSCRIÇÃO, DETECÇÃO E SUAS UTILIZAÇÕES"

O presente invento está relacionado com péptidos e composições para terapêutica, diagnóstico, farmacogenética ou desenho de fármacos. Como será descrito, o invento é particularmente adequado para a detecção, monitorização ou tratamento de distúrbios de proliferação celular.

Introdução ao invento

O DNA e o RNA são macromoléculas que servem para armazenar informação genómica no núcleo da célula (DNA) e para transferir a informação genómica para o citoplasma, após um processo designado transcrição que gera RNA mensageiro de forma a produzir proteína. Ambos são polímeros lineares compostos por monómeros designados nucleótidos. O DNA e o RNA representam combinações de quatro tipos de nucleótidos. Todos os nucleótidos têm uma estrutura comum: um grupo fosfato ligado por uma ligação fosfodiéster a uma pentose que por sua vez está ligada a uma base orgânica: adenina, citosina, timina e guanina (A, C, T, G) para o DNA e adenina, citosina, uracilo e guanina (A, C, U, G) para o RNA. No RNA, a pentose é ribose e no DNA é desoxirribose. A molécula de DNA está organizada como uma cadeia dupla de nucleótidos ordenados, arranjados como

uma dupla hélice. As moléculas de RNA são polinucleótidos de cadeia simples que, essencialmente, dão 4 tipos de moléculas: 1) RNA mensageiro (mRNA), que são nucleótidos de cadeia simples transcritos a partir do DNA e traduzidos em proteínas no citoplasma, 2) RNA de transferência (tRNA), que adopta uma estrutura tridimensional bem definida e se liga a aminoácidos específicos (AA), os quais são transferidos para as cadeias polipeptídicas para formar proteínas com sequências específicas de aminoácidos num processo designado tradução, 3) RNA ribossomal (rRNA), os quais são moléculas maiores que, em conjunto com proteínas específicas, constituem o ribossoma, uma estrutura que permite a montagem dos AA em proteína seguindo a informação proporcionada pela sequência especificada por codões (um codão = 3 nucleótidos) presentes numa determinada ordem no mRNA e 4) pequenos RNAs reguladores referidos como RNA não codificador (ncRNA).

A sequência de mRNA é então transformada numa linguagem diferente, *i.e.* a linguagem dos AA, através de um processo referido como tradução.

Demonstramos aqui pela primeira vez que a fidelidade da transcrição, *i.e.*, a transferência da informação do DNA para mRNA, é dramaticamente reduzida em células patológicas, particularmente nas células de cancro. Esta ausência de fidelidade da transcrição nas células de cancro é um fenómeno geral que afecta todos os cancros e a maioria dos genes testados nos exemplos descritos, assim

como a maioria dos transcritos presentes em bases de dados disponíveis e que tem várias implicações imediatas e importantes. Em particular, a descoberta da infidelidade da transcrição permite a melhoria das abordagens usadas na proteómica e na transcriptómica para a investigação de condições patológicas. A descoberta de infidelidade da transcrição também permite uma melhor classificação das doenças relativamente à sua gravidade e melhora o prognóstico do efeito terapêutico e permite desenhar novos métodos para o rastreio da eficácia de fármacos com base na sua capacidade para corrigir uma ausência ou modificação da fidelidade da transcrição. O presente invento possui assim grandes implicações e utilidade no diagnóstico e no desenvolvimento de fármacos, assim como no tratamento, detecção e monitorização de doentes e eficácia de fármacos.

Para compreender o impacto da descoberta descrita abaixo, é importante reconhecer que, actualmente, se pensa ser necessária fidelidade absoluta na transcrição do DNA para RNA para o funcionamento normal da célula. No entanto, actualmente persiste uma questão importante por resolver. De facto, a sequenciação e anotação do genoma humano levou à noção de que são codificados 30-40 milhares de genes. Esta estimativa fica bastante abaixo do número de proteínas que podem ser identificadas pelos métodos de proteómica: até 300 mil proteínas foram identificadas. Para conciliar estas diferenças, é proposto que um único gene possa produzir vários mRNAs codificadores de proteínas diferentes através de um processo designado "splicing" alternativo. O

"splicing" alternativo remove diferentes partes do RNA pré-mensageiro (pmRNA), conduzindo assim à remoção de diferentes elementos correspondendo a intrões específicos, e produz diferentes mRNAs maduros. A análise da base de dados RefSeq indica que contém 11259 transcritos correspondendo a 6946 genes. Assim, o "splicing" alternativo pode aumentar a heterogeneidade dos transcritos em 76%. O mecanismo que é aqui descrito é diferente do "splicing" alternativo e induz uma heterogeneidade de proteínas muito maior. De facto, demonstramos a ocorrência de modificações não aleatórias da sequência de mRNA que resultam em alterações nos AA codificados, introdução de codões de paragem prematuros que dão isoformas mais curtas das proteínas, alterações dos codões de paragem naturais implicando introdução de novas sequências codificadoras que originam isoformas de proteínas de maiores dimensões. A introdução de espaços e de inserções modifica assim a grelha de leitura do mRNA e cria sequências proteicas desconhecidas. Os dados descritos neste invento mostram que este fenómeno da infidelidade da transcrição está presente em células normais mas dramaticamente aumentado em células patológicas, tais como células derivadas de cancro. É assim proposto que a infidelidade da transcrição (TI) contribua para a diversificação da informação transferida do DNA para o RNA. Os nossos resultados ainda estabelecem que a infidelidade da transcrição não ocorre ao acaso mas segue regras específicas em que é importante o contexto envolvente das bases afectadas pelo evento de TI.

Poderá ser posta ainda a hipótese de um outro mecanismo para explicar que a heterogeneidade do RNA não ocorre ao nível genómico mas antes ao nível do RNA: a edição de RNA. No entanto, a edição de RNA não consegue explicar dois tipos de acontecimentos TI, introdução de um espaço e inserção. A edição de RNA caracteriza-se por alterações de bases pós-transcrição no mRNA e no tRNA em eucariotas. A grande maioria dos acontecimentos de substituição por edição consiste em conversões de C para U ou de A para I (lido como G) (Gott, J.M. & Emeson, R.B. (2000) Annu Rev Genet 34, 499-531 - Maas, S & Rich, A. (2000) Bioessays 22, 790-802. - Niswender, C. M. (1998) Cell Mol Life Sci 54, 946-64). Ver também Klimek-Tomzak *et al.*, British J. of Cancer Vol. 94, 10 (2006) pp 586-592); Cappione *et al.*, AJHG vol. 60, 2 (1997) pp 305-312; ou Maas *et al.*, PNAS vol. 98, 25 (2001) pp 14687-14692). Diferentes trabalhos têm mostrado que estas conversões surgem através de mecanismos de desaminação, catalisados por desaminases de adenosina e de citidina. Um outro evento de substituição por edição é a conversão de "U" para "C". Apesar da teoria da reversibilidade microscópica ditar que a reacção da desaminase de citidina possa ocorrer em sentido inverso para gerar esta conversão, também foi proposto que uma actividade do tipo sintetase de CTP possa ser responsável. Foi demonstrado em vários exemplos que a edição de RNA pode ser específica das células do cancro versus células normais. Ainda, a taxa deste fenómeno pode ser afectada nas condições de cancro.

No contexto do cancro, observámos ao nível do cDNA 5,7% de alterações C→T, 9,2% de T→C e 4,7% de A→G. Assim, a edição de mRNA não pode ser responsável por mais de 20% das substituições de bases isoladas aqui descritas. De facto, as substituições de bases mais comuns são: A→C (24,6%) e T→G (16,8%) as quais representam alterações de famílias de bases que não são explicadas por processos enzimáticos de edição de RNA humanos conhecidos.

Ainda, espera-se que a infidelidade da transcrição cause deleções de bases e/ou inserções através de um mecanismo (ou mecanismos) diferentes da edição de mRNA. Ainda, um estudo recente mostra que o número de registos de dbSNPs (base da dados Single Nucleotide Polymorphism database) é de facto locais de edição (Eisenberg, E. *et al.* (2005) Nucleic Acids Res. 33 (14), 4612-7). Na nossa abordagem não consideramos todos os SNPs da dbSNP, sendo portanto excluídos SNPs conhecidos ou falsos SNPs correspondendo a locais de edição. O mecanismo para explicar a heterogeneidade de mRNA no cancro é assim a infidelidade da transcrição.

Um outro estudo (Ruiz *et al.*, 2004, Clinical Cancer research 10(8), 2560-2567) está relacionado com a identificação de determinantes de células T auxiliares do antigénio carcinoembrionário, considerados potenciais ligandos das moléculas HLA-DR. No entanto, este documento não descreve sequências nem um péptido sintético, como reivindicado pelos inventores.

Sumário do presente invento

O presente invento está relacionado com péptidos sintéticos, composições e utilizações dos mesmos para fins terapêuticos e de diagnóstico, *e.g.*, para detectar e/ou tratar doenças causadas por proliferação celular. O objecto deste invento está relacionado com o uso de um péptido sintético como definido nas reivindicações.

O invento também divulga a utilização de um composto que altera (*e.g.*, reduz ou aumenta) a taxa de infidelidade da transcrição de um gene de mamífero para a produção de uma composição farmacêutica para usar num método de tratamento de um ser humano ou animal, particularmente para o tratamento de doenças tais como distúrbios de proliferação celular incluindo sem limitações, cancros, doenças imunológicas, doenças inflamatórias ou envelhecimento.

O invento também divulga métodos e produtos (tais como sondas, sequências iniciadoras, anticorpos ou seus derivados) para a detecção ou medição (o nível de) da infidelidade da transcrição numa amostra, assim como os kits correspondentes.

O invento também divulga métodos de identificação de sequências de infidelidade da transcrição em proteínas ou ácidos nucleicos, assim como a sua utilização, *e.g.*,

como marcadores, imunogénios e/ou para gerar ligandos específicos.

Neste contexto, o invento divulga um método de identificação e/ou produção de biomarcadores, o método compreendendo a identificação, numa amostra de um indivíduo, da presença de locais de infidelidade da transcrição numa proteína ou ácido nucleico alvo e, facultativamente, determinação da sequência dos referidos locais de infidelidade da transcrição.

O invento também descreve um método de identificação e/ou produção de um ligando específico para uma característica ou condição patológica, o método compreendendo a identificação, numa amostra de um indivíduo tendo a referida característica ou condição patológica, da presença de pelo menos um local de infidelidade da transcrição numa ou mais proteínas ou ácidos nucleicos alvo, facultativamente determinação da sequência de pelo menos um desses locais de infidelidade da transcrição e produção de um ligando que se liga especificamente à referida proteína (ou domínio) ou ácido nucleico criado pela infidelidade de transcrição.

O invento é particularmente adequado para a identificação de biomarcadores de distúrbios de proliferação celular, tais como cancros, doenças imunológicas, doenças inflamatórias ou envelhecimento. É particularmente útil para a produção de ligandos que são específicos de tais

distúrbios em seres mamíferos, em particular ligandos que possam detectar a presença ou gravidade de um distúrbio de proliferação celular num indivíduo.

Este invento também divulga um ligando que se liga especificamente a uma proteína (ou domínio) ou ácido nucleico criado por infidelidade da transcrição. O ligando pode ser um anticorpo (ou qualquer fragmento (como seja Fab, Fab', CDR, etc.) ou seu derivado, como descrito mais à frente), que se liga especificamente a uma proteína (ou domínio) criado por infidelidade da transcrição. O ligando pode também ser um ácido nucleico que se liga especificamente a um ácido nucleico criado pela infidelidade da transcrição (*e.g.*, sonda, sequência iniciadora, RNAi (RNA de interferência), etc.).

Este invento também divulga um péptido compreendendo um domínio de uma proteína criada por infidelidade de transcrição, particularmente de uma proteína de mamífero, mais de preferência de uma proteína humana. O péptido é tipicamente um péptido sintético, *i.e.*, um péptido que foi preparado artificialmente, *e.g.*, através de síntese química, síntese de aminoácidos e/ou extensão, digestão de proteínas, montagem de péptidos, expressão recombinante, etc. O péptido tipicamente compreende a sequência de um fragmento C-terminal da referida proteína. O péptido de preferência compreende menos de 100, 80, 75, 70, 65, 60, 50, 45, 40, 35, 30, 25 ou mesmo 20 aminoácidos (ainda que noutras realizações, o comprimento do péptido

possa ser maior). A proteína pode ser uma proteína da superfície celular (*e.g.*, um receptor, etc.), uma proteína secretada (*e.g.*, uma proteína do plasma, etc.) ou uma proteína intracelular.

Um outro aspecto deste invento reside na utilização de um péptido criado por infidelidade da transcrição como definido atrás, como imunogénio.

Um outro objecto do invento reside no uso de um péptido sintético como definido atrás, de 100 ou menos aminoácidos de comprimento, e, em particular, compreendendo uma sequência seleccionada entre SEQ ID NOS: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 para detecção ou monitorização de distúrbios de proliferação celular.

O invento também descreve uma composição de vacina compreendendo uma parte de uma proteína criada por infidelidade de transcrição. Um outro objecto do invento reside numa composição de vacina compreendendo um péptido sintético, como definido atrás, de 100 ou menos aminoácidos de comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOs: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 e, facultativamente, um veículo, excipiente e/ou adjuvante adequado.

Este invento também descreve um métpdp de produção de um anticorpo, o método compreendendo a imunização de um mamífero não humano com um péptido criado

pela infidelidade de transcrição. Um outro objecto deste invento reside num método de produção de um anticorpo, o método compreendendo imunização de um mamífero não humano com um péptido sintético de 100 ou menos aminoácidos de comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOs: 1 a 5, 7 a 13, 15 a 18 e 20 a 32, como definido atrás, e recuperação dos anticorpos que se ligam ao referido péptido, ou correspondentes células produtoras de anticorpo. Facultativamente, podem ser produzidos derivados do anticorpo.

O invento também divulga um método de produção de um anticorpo, o método compreendendo (i) identificação de um domínio de uma proteína criado por infidelidade da transcrição e (ii) produção de um anticorpo que se liga especificamente ao referido domínio. Facultativamente, podem ser produzidos derivados do anticorpo.

O invento também divulga um anticorpo que se liga especificamente a uma parte de uma proteína criada por infidelidade da transcrição, ou um derivado de tal anticorpo tendo substancialmente a mesma especificidade de antigénio. Em particular, o invento reside num anticorpo, ou seu derivado, que especificamente se liga a um péptido sintéticos de 100 ou menos aminoácidos de comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOs: 1 a 5, 7 a 13, 15 a 18 e 20 a 32. O anticorpo pode ser policlonal ou monoclonal. O termo derivado inclui qualquer fragmento (como seja Fab, Fab', CDR, etc.) ou outros

derivados tais como anticorpos de cadeia simples, anticorpos bifuncionais, anticorpos humanizados, anticorpos humanos, anticorpos quiméricos, etc.

Um outro aspecto deste invento está relacionado com um anticorpo ou seu derivado, como definido atrás, que é conjugado com uma molécula. A molécula pode ser um fármaco, uma marca, uma molécula tóxica, um isótopo radioactivo, etc.

O invento também divulga a utilização de um anticorpo, ou seu derivado como atrás definido, para a detecção ou quantificação (*e.g.*, *in vitro*) da infidelidade da transcrição de um gene.

O invento também está relacionado com a utilização de um anticorpo conjugado, ou seu derivado como definido atrás, como medicamento ou reagente de diagnóstico.

O invento também divulga um dispositivo ou produto compreendendo, imobilizado no suporte, um reagente que se liga especificamente a uma proteína ou ácido nucleico criado pela infidelidade de transcrição. O reagente pode ser, *e.g.*, uma sonda ou um anticorpo ou seu derivado.

Este invento também divulga um método para causar ou induzir ou estimular a infidelidade da transcrição numa célula ou tecido ou organismo. Tal método pode compreender, tipicamente, a introdução de um local de infidelidade da

transcrição em sequências genómicas normais, *e.g.*, por meio de um vector de terapia génica. Tal modificação pode resultar na destruição de uma célula ou tecido. Em particular, é possível criar uma grelha de leitura aberta em qualquer sequência de um gene que resulte na produção de proteínas ou compostos tóxicos para a célula. É igualmente possível induzir a expressão, numa célula doente (ou alvo), de um biomarcador específico que pode ser atingido usando moléculas tóxicas ou terapêuticas. Usando esta abordagem, é assim possível causar morte celular ou terapia quando a infidelidade de transcrição ocorre ou excede uma determinada taxa.

Este invento também divulga a utilização de um ácido nucleico que possui locais de infidelidade da transcrição para a produção de uma composição farmacêutica para usar num método de tratamento de um ser humano ou animal. O gene pode ser qualquer gene, incluindo um gene de mamífero ou um gene de um agente patogénico, como seja um gene viral, um gene bacteriano, etc. Neste contexto, o invento está relacionado com um método de tratamento de uma doença causada por um agente patogénico, o método compreendendo a indução ou a estimulação da infidelidade de transcrição de um gene codificado pelo referido agente patogénico.

Numa realização particular, o invento está relacionado com uma molécula de ácido nucleico sintético codificadora de um péptido de 100 ou menos aminoácidos de

comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOs: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 para a detecção ou monitorização de distúrbios de proliferação celular.

O invento também divulga métodos de produção de polipéptidos recombinantes *in vitro* com infidelidade de transcrição reduzida. Tais métodos permitem uma redução na micro-heterogeneidade dos polipéptidos recombinantes. O método compreende a utilização de células hospedeiras compreendendo um ácido nucleico recombinante com utilização de codões adaptada, para reduzir a ocorrência de infidelidade de transcrição. O método pode também incluir a utilização de qualquer composto ou tratamento que reduza a ocorrência de infidelidade da transcrição. O método pode ser usado em células hospedeiras procarióticas ou eucarióticas, *e.g.*, em estirpes de *E. coli* ou CHO.

O invento pode ser usado num ser mamífero, particularmente um ser humano, para detectar, monitorizar ou tratar uma variedade de condições patológicas associadas a distúrbios de proliferação celular (*e.g.*, cancros) e/ou para produzir, desenhar ou testar fármacos terapeuticamente activos.

Legendas das Figuras

Fig. 1: Princípio da construção de bibliotecas de cDNA e sequenciação.

Fig. 2: (a-q) sequências de mRNA de referência usadas na análise. Para evitar a distorção do "blast", a cauda poli(A) das sequências de mRNA de referência foram sistematicamente removidas. Fig. 2 apresenta uma lista dos genes testados.

Fig. 3: ficheiros típicos gerados com MegaBLAST.

Fig. 4: Percentagem de desvio da sequência de nucleótidos em qualquer posição para cada um dos genes estudados.

Fig. 5: TPT1, VIM = número de ESTs e análise do teste de proporções.

Fig. 6: Variações em ESTs antes e após a aplicação sequencial de filtros electrónicos. Clip 400 e efeitos de remoção de linhas celulares.

Fig. 7: Contexto do DNA: a) Efeito da composição de bases do pmRNA na heterogeneidade de bases b0. b) Composição de bases substituintes para cada base substituída. c) Repartição de bases afectadas e substituintes dentro de testes estatisticamente significativos C>N. d) Efeito da composição de bases do pmRNA na correspondente base substituinte para repetição de b-1 ou b+1

Fig. 8: mRNA virtual variante e proteína para VIM.

Fig. 9: Impacto codificador

Fig. 10: Proteínas de interesse = sequências de aminoácidos quando o codão de paragem é estatisticamente afectado.

Fig. 11: DHPLC = princípio e limites da detecção.

Fig. 12: DHPLC = sequências iniciadoras e produtos de PCR esperados.

Fig. 13: Lista de 60 proteínas estudadas.

Fig. 14: Produção de anticorpos.

Fig. 15: Resultados de DHPLC.

Fig. 16: sensibilidade de Sanger.

Fig. 17: Detecção de APOAII e APOCII PSP em plasma de doentes cancerosos.

Fig. 18: PSP potencial nas proteínas do plasma.

Fig. 19: Avaliação das previsões de substituições da sequência de mRNA.

Descrição detalhada do invento

Definições

Infidelidade da transcrição

O termo infidelidade de transcrição designa um novo mecanismo pelo qual várias moléculas de RNA distintas são produzidas numa célula a partir de uma única sequência de gene. Este novo mecanismo identificado afecta potencialmente qualquer gene, é aleatório e segue regras particulares, como será aqui descrito. Como se mostra nos exemplos, a infidelidade da transcrição pode introduzir substituições, deleções e inserções em moléculas de RNA, criando assim uma diversidade de proteínas a partir de um único gene. A infidelidade da transcrição pode também afectar as sequências de RNA não codificadoras, modulando assim as suas funções. Medição, modulação ou eliminação da infidelidade de transcrição representa assim uma nova abordagem para a detecção ou tratamento de distúrbios, assim como para o desenvolvimento de fármacos.

Local de infidelidade da transcrição

O termo local de infidelidade da transcrição designa uma sequência e/ou posição afectada pela infidelidade da transcrição. Isto pode ser um ácido nucleico ou domínio de aminoácidos que possui pelo menos uma modificação gerada como resultado da infidelidade da transcrição. Tal modificação pode resultar *e.g.*, de uma

mudança da grelha de leitura (inserção e/ou deleção), da introdução ou supressão dos codões de paragem, da introdução de um novo codão de paragem, da substituição de um ou mais nucleótidos (implicando ou não alteração de AA), etc. Um local de infidelidade da transcrição pode compreender uma ou várias variações de sequências resultantes da infidelidade de transcrição. Um local de infidelidade de transcrição pode compreender desde, *e.g.*, 1 resíduo de nucleótido ou aminoácido modificado até, *e.g.*, 150 resíduos de nucleótidos ou aminoácidos modificados, ou mesmo mais. O local de infidelidade de transcrição tipicamente difere da sequência resultante da transcrição fiel em pelo menos uma modificação de nucleótido ou de aminoácido (*e.g.*, substituição, deleção, inserção, inversão, etc.). Uma proteína ou um domínio criado por infidelidade da transcrição (proteína TI) tipicamente compreende entre 1 e 50 aminoácidos (ou mesmo mais), com pelo menos um resíduo de aminoácido modificado. Uma sequência de ácido nucleico de infidelidade da transcrição tipicamente compreende entre 1 e 150 nucleótidos (ou mesmo mais), com pelo menos um resíduo de nucleótido modificado.

Regras e identificação de infidelidade da transcrição

Com base em técnicas específicas de detecção e/ou regras particulares de infidelidade da transcrição como definido no presente pedido de patente, é agora possível prever e identificar para qualquer gene ou proteína, Locais de Infidelidade da Transcrição. Estes locais podem também

ser obtidos através do alinhamento de sequências disponíveis para um determinado RNA e identificação das substituições de bases. Podem em última análise ser validadas por várias técnicas, *e.g.*, através da utilização de ligandos específicos para uma proteína TI teórica.

Um método particular para identificação de um local de infidelidade da transcrição compreende:

- obtenção da sequência de um determinado gene, molécula de RNA ou de cDNA, ou uma porção da mesma e
- identificação, dentro da referida sequência, da presença de uma alteração de nucleótido (*e.g.*, substituição, deleção, inserção, inversão) resultante da infidelidade da transcrição, seguindo as regras de infidelidade da transcrição, *e.g.*, como discutido em (iii) abaixo.

Um outro método particular para a identificação de um local de infidelidade da transcrição compreende:

- obtenção da sequência de uma determinado proteína ou porção da mesma e
- identificação, dentro da referida sequência, da presença de uma alteração de aminoácido (*e.g.*, substituição, deleção, inserção, inversão) resultante da infidelidade da transcrição, seguindo as regras de infidelidade da transcrição, *e.g.*, como discutido em (iii) abaixo.

A presença de uma alteração resultante da infidelidade da transcrição numa molécula pode ser identificada num processo de três passos compreendendo:

(i) identificação de locais de infidelidade da transcrição com uma máquina de aprendizagem baseada na discriminação quadrática em factores de Análise Factorial de Múltipla Correspondência (MCFA)

(ii) identificação da categoria dos locais de infidelidade da transcrição (b-1 ou b+1 ou outros) com o mesmo método e

(iii) previsão da base de substituição usando as seguintes regras de infidelidade da transcrição:

Para qualquer base presente como evento singular ("singleton"), *i.e.*, qualquer base que não é precedida ou seguida por si, então a base substituída é com grande probabilidade idêntica à base anterior àquela que é substituída (no caso da categoria b-1) ou idêntica à base seguinte àquela que é substituída (no caso da categoria b+1). Por exemplo, CAT tornar-se-á CCT (regra b-1); ATG tornar-se-á AGG (regra b+1);

Quando as substituições ocorrem dentro de três A consecutivos, a substituição do segundo A vai preferencialmente para C;

Quando as substituições ocorrem em três T consecutivos, a substituição do segundo T vai preferencialmente para C, depois para A, depois para G;

Os segmentos de C e de G são raramente substituídos na segunda posição, mas se houver qualquer substi-

tuição, o segundo C irá preferencialmente para A e substituição do segundo G vai preferencialmente para C;

Para os outros casos, a base da substituição é preferencialmente C.

Outros métodos para identificar locais de infidelidade da transcrição dentro da sequência de qualquer proteína ou ácido nucleico alvo compreendem, *e.g.*, comparação da marca da sequência expressa existente (ESTs), amplificação directa das sequências identificadas e sequenciação dos produtos de amplificação com método de sequenciação específico (Sanger, química, piro-sequenciação). Tal abordagem produz sequências variantes que envolvem alterações de AA, modificações do comprimento da proteína, etc. Podem ser desenhados oligonucleótidos que selectivamente correspondam a locais de infidelidade da transcrição existente em mRNA ou em cDNA com locais de infidelidade da transcrição a partir de um conjunto de sequências. Podem ser igualmente produzidos anticorpos contra a proteína TI.

Como alternativa, é possível identificar um local de infidelidade da transcrição por meio de análise de sequências ou de regras de bioinformática. A sequência resultante pode ser clonada em qualquer vector de transfecção. É igualmente possível desenhar uma construção compreendendo tal sequência de infidelidade da transcrição na mesma grelha de leitura de um gene repórter, que será transcrito e traduzido apenas se ocorrer infidelidade da

transcrição. O gene repórter pode ser uma enzima ou proteína fluorescente ou qualquer gene que torne a célula alvo sensível ou resistente à exposição a uma toxina.

Um método detalhado que permite a identificação de locais de infidelidade da transcrição (ou substituições) inclui o método TDG como descrito abaixo:

Uma técnica descrita por Pan e Weissman e Liu *et al.* (PNAS, 2002, 99(14), 9346-9351 e Anal Biochem., 2006 Sep 1;356(1):117-24) pode ser adaptada para detecção dos locais de infidelidade da transcrição em RNA. Esta técnica baseia-se na utilização de uma enzima, glicosilase de timidina do DNA (TDG) que é capaz de separadamente enriquecer duplas cadeias de DNA contendo desemparelhamentos a partir de misturas complexas. Estas enzimas reconhecem especificamente desemparelhamentos de nucleótidos e geram locais abásicos (a ligação entre a desoxirribose e uma das bases do DNA é clivada). Em seguida, TDG pode ligar reversivelmente estes locais abásicos e assim pode ser usada para purificação por afinidade de fragmentos de DNA contendo desemparelhamentos. Numa experiência típica, o RNA é extraído a partir de dois tipos de células (normais e cancerosas) e sujeito a transcrição reversa para preparar cDNA de cadeia dupla. Após desnaturação por aquecimento e renaturação lenta para cada amostra, é gerado cDNA com desemparelhamentos e pode ser separado a partir de duplas hélices perfeitas. Este cDNA pode então ser analisado por sequenciação directa ou comparado por DHPLC (ver mais à frente).

São descritas outras técnicas que podem ser adaptadas para identificar locais de infidelidade da transcrição ou mutações, *e.g.*, em US 6329147; US 4979330; WO 02/077286 ou US 6120992.

Tipicamente, os métodos de identificação dos locais de infidelidade da transcrição ainda compreendem um passo de validação da sequência do local de infidelidade da transcrição através da produção de uma molécula compreendendo a sequência identificada, geração de um ligando que especificamente se liga à referida molécula e verificação, numa amostra biológica, da presença de um antigénio especificamente reconhecido pelo referido ligando.

Uma sequência de infidelidade da transcrição típica de um ácido nucleico é uma sequência entre 1 e 150 nucleótidos de comprimento compreendendo a substituição, deleção ou adição de pelo menos uma base causada pela infidelidade da transcrição, de preferência seguindo uma regra como descrita no Exemplo 5. A sequência de infidelidade da transcrição pode ser maior.

Um domínio de uma proteína TI é uma sequência de aminoácidos entre 1 e 50 aminoácidos de comprimento compreendendo a substituição, deleção ou adição de pelo menos um aminoácido (ou mais alterações devido à modificação na grelha codificadora resultante da deleção ou inserção de bases) causada pela infidelidade da transcrição, de preferência seguindo uma regra como descrita no Exemplo 5.

O domínio de uma proteína TI pode ser maior. Exemplos de domínios das proteínas TI são proporcionados na secção experimental, *e.g.*, nas Figuras 13 e 14.

Detecção ou medição da infidelidade da transcrição

Este invento baseia-se na descoberta inesperada de uma taxa elevada de anormalidades da sequência natural que ocorre durante ou logo após a transcrição do DNA para RNA. O processo pode existir nas células normais mas aumenta grandemente nas células patológicas, *e.g.*, células cancerosas. A descoberta de que a transcrição DNA em RNA introduz variações da sequência seguindo regras que são diferentes das definidas por complementaridades de bases uma a uma permite o desenho racional de novos reagentes (*e.g.*, sondas, sequências iniciadoras, anticorpos, aptâmeros, etc.) que são específicos de ácido nucleico ou proteína criada por infidelidade da transcrição. Tais reagentes podem assim permitir detectar ou medir infidelidade da transcrição e discriminar entre condições normais ou de doença, assim como direccionar moléculas para células que apresentem infidelidade da transcrição.

Ainda, o invento também permite o desenho racional de reagentes (*e.g.* sondas, sequências iniciadoras, anticorpos, aptâmeros) que são preditivos da gravidade da doença (*e.g.*, cancro). De facto, espera-se que uma maior taxa de infidelidade da transcrição esteja correlacionada

com a gravidade da doença e esta taxa aumenta progressivamente à medida que mais e mais produtos de genes são afectados. A medição directa da infidelidade da transcrição em células doentes usando métodos aqui descritos ou qualquer outra tecnologia permite detectar células em diferentes fases de progressão da doença em qualquer tecido. A medição precisa destas variações na expressão de genes (transcritos e proteínas) também melhora a capacidade de avaliar a eficácia de fármacos relativamente à gravidade da doença. Actualmente, várias técnicas de microarranjos permitem a medição de alterações na expressão de genes. No entanto, a capacidade e, mais importante, a reprodutibilidade destes resultados normalmente não são suficientes. Podemos agora postular que uma grande quantidade de variabilidade destas experiências de transcriptómica é causada em parte pela infidelidade da transcrição conforme descoberto pelos inventores. Assim, a descoberta deste fenómeno comum permite o desenho de reagentes para a expressão de genes que sejam minimamente afectados pela infidelidade da transcrição ou que directamente reflictam a infidelidade da transcrição que ocorra em sequências específicas seguindo as regras descritas no presente pedido.

É igualmente possível e mesmo provável que a infidelidade da transcrição seja modulada pela taxa de transcrição. Especulamos que uma maior expressão de um determinado gene em condições patológicas aumente a TI e assim aumente a heterogeneidade da proteína.

A infidelidade da transcrição pode ser detectada ou medida usando uma série de técnicas conhecidas *per se* na área, as quais podem ser adaptadas ao presente invento. Em particular, a infidelidade da transcrição pode ser medida usando reagentes específicos para ácido nucleico ou proteína criados pela infidelidade da transcrição, tais como sondas específicas, sequências iniciadoras ou anti-corpos, por electroforese, análise da migração em gel, espectrometria, etc., que permitem a detecção entre várias formas de uma proteína ou ácido nucleico. A taxa de infidelidade da transcrição pode ser determinada através da avaliação do número de genes numa célula que estão sujeitos a infidelidade de transcrição e/ou o número de locais de infidelidade de transcrição gerados para um determinado gene. Tal taxa pode ser determinada por comparação do nível de ácido nucleico ou proteína criado pela infidelidade de transcrição numa amostra a testar com o obtido numa amostra de referência.

Detecção da infidelidade da transcrição ao nível dos ácidos nucleicos

Como discutido atrás, a infidelidade da trans-crição introduz variações da sequência nas moléculas de RNA. A detecção ou medição da infidelidade da transcrição pode assim ser conseguida através da detecção da presença ou quantidade (absoluta ou relativa) de tais variações da sequência em RNAs codificadores de um ou vários genes ou numa célula completa ou tecido ou amostra. A detecção da

infidelidade da transcrição em ácidos nucleicos pode ser realizada por várias técnicas tais como hibridação, amplificação, formação de heteroduplexes, etc.

Virtualmente todas as tecnologias usadas para a identificação de nucleótidos presentes em diversos tipos de tecido baseiam-se num processo designado de hibridação. A hibridação é crítica para tecnologias tais como micro-arranjos usadas para identificar polimorfismos e pesquisa de variação na expressão de genes. Esta técnica é aplicada para medir o nível de expressão de genes. A hibridação de sondas oligonucleotídicas é igualmente usada para subtrair sequências de gene abundantemente expressos de forma a permitir um estudo mais adequado dos mensageiros com menor abundância - este procedimento é designado hibridação subtractiva. A hibridação é igualmente o primeiro passo da reacção de amplificação de genes normalmente usada nas reacção de PCR na forma directa ou após aplicação de transcriptase reversa para converter a sequência da mensagem de uma sequência tipo RNA para uma tipo DNA.

O mecanismo básico subjacente à hibridação é que uma sequência de nucleótidos de cadeia simples se pode ligar a uma outra desde que as respectivas sequências sejam complementares. Tanto o comprimento como a ordem específica de cada uma das 4 bases definem a eficácia da hibridação e a especificidade da sequência. Isto implica que cada base específica se liga numa forma não covalente à sua base complementar: A a T e C a G e *vice versa*. Esta ligação não

covalente numa combinação de sequência é condicionada pela ordem das bases. Assim, na prática, uma sequência definida de bases liga-se a uma sequência complementar única. Esta reacção de ligação específica proporciona a base da hibridação e permite a identificação de qualquer sequência complementar em qualquer mistura de nucleótidos. A eficácia da hibridação é determinada pelo número de bases que, na ordem apropriada, emparelham umas com as outras (é tolerado alguma grau de desemparelhamento) e determinada pelas condições da experiência tais como a restringência do tampão de ligação, o teor relativo de CG versus TA - CG e TA são ligados por 3 e 2 ligações de hidrogénio, respectivamente - e pela temperatura de fusão, *i.e.* a temperatura que é necessária para permitir a dissociação de 50% do DNA de cadeia dupla.

A detecção da infidelidade da transcrição usando hibridação tipicamente compreende colocação de uma amostra em contacto com uma sonda de ácido nucleico que é específica para uma sequência com infidelidade de transcrição e detecção da presença (ou quantidade) de híbridos formados, a referida presença ou quantidade sendo uma indicação directa da presença ou taxa de infidelidade da transcrição. Numa realização preferida, o método usa uma série de sondas de ácido nucleico que são específicas para sequências com infidelidade transcrição distintas de um ou vários genes, respectivamente. As sondas de ácidos nucleicos específicas para uma sequência com infidelidade de transcrição podem ser preparadas para qualquer gene usando a informação da

sequência e as regras de substituição, inserção ou delecção de nucleótidos descritas no presente pedido de patente (ver, *e.g.*, Exemplo 5).

Uma "sonda" de ácido nucleico refere-se a um ácido nucleico ou oligonucleótido tendo uma sequência polinucleotídica que é capaz de hibridar selectivamente com uma sequência com infidelidade de transcrição ou com o seu complemento e que é adequada para a detecção da presença (ou quantificação da mesma) numa amostra contendo a referida sequência ou complemento. As sondas são, de preferência, perfeitamente complementares de uma sequência com infidelidade de transcrição, no entanto pode ser tolerado algum desemparelhamento. As sondas tipicamente compreendem ácidos nucleicos de cadeia simples entre 8 e 1500 nucleótidos de comprimento, por exemplo entre 10 e 1000, mais de preferência entre 10 e 800, tipicamente entre 20 e 700. Deverá ser entendido que podem ser igualmente usadas sondas mais longas. Uma sonda preferida deste invento é uma molécula de ácido nucleico de cadeia simples com 8 a 400 nucleótidos de comprimento, que pode hibridar especificamente com uma sequência tendo infidelidade de transcrição.

Tipicamente, uma sonda de ácido nucleico hibrida selectivamente com uma região de uma molécula de ácido nucleico que não possui uma sequência de infidelidade da transcrição.

A selectividade, quando usada em associação com hibridação de ácido nucleico, indica que a hibridação da sonda com a sequência alvo é distinta da hibridação da referida sonda com uma outra sequência. Neste contexto, ainda que a complementaridade perfeita entre a sonda e a sequência alvo não seja necessária, deve ser suficientemente alta para permitir a hibridação.

As sondas tipicamente compreendem uma sequência que é complementar de uma sequência de infidelidade da transcrição numa molécula de RNA que codifica uma proteína da superfície celular ou uma proteína secretada. Exemplos específicos de sondas deste invento possuem uma sequência complementar de uma sequência com infidelidade de transcrição ou codificadora de uma sequência peptídica de qualquer uma de SEQ ID NOS: 1 a 32.

A sequência das sondas pode ser modificada, *e.g.*, quimicamente, de forma a, *e.g.*, aumentar a estabilidade dos híbridos (*e.g.*, grupos intercalantes ou modificados, tais como alcoxirribonucleótidos 2') ou para marcar a sonda. Exemplos típicos de marcas incluem, sem limitação, radioactividade, fluorescência, luminescência, marcação enzimática e similares. A sonda pode ser hibridada com o ácido nucleico alvo em solução, suspensão ou ligada a um suporte sólido, como seja sem limitação, uma esfera, coluna, placa, substrato (para produzir arranjos ou chips de ácidos nucleicos), etc.

A infidelidade de transcrição pode ser detectada ou medida através de amplificação selectiva usando sequências iniciadoras específicas. A detecção da infidelidade de transcrição através de amplificação selectiva tipicamente compreende colocação de uma amostra em contacto com uma sequência iniciadora de ácido nucleico que especificamente amplifica uma sequência com infidelidade de transcrição e detecção da presença (ou quantidade) dos produtos de amplificação formados, a referida presença ou quantidade sendo uma indicação directa da presença ou taxa de infidelidade de transcrição. Numa realização preferida, o método usa uma série de sequências iniciadoras de ácido nucleico que permitem a amplificação específica da sequência com infidelidade de transcrição distinta de um ou vários genes, respectivamente. A amplificação pode ser realizada de acordo com várias técnicas conhecidas *per se* na área, tais como, sem limitação, reacção em cadeia da polimerase (PCR), reacção em cadeia da ligase (LCR), amplificação mediada por transcrição (TMA), amplificação por deslocamento de cadeias (DAS) e amplificação baseada em sequências de ácidos nucleicos (NASBA).

As sequências iniciadoras de ácido nucleico específicas para uma sequência com infidelidade da transcrição podem ser preparadas para qualquer gene usando a informação de sequências e as regras de substituição, inserção ou deleção de nucleótidos descritas no presente pedido de patente (ver, *e.g.*, Exemplo 5).

O termo "sequência iniciadora" designa um ácido nucleico ou oligonucleótido tendo uma sequência poli-nucleotídica que é capaz de hibridar selectivamente com uma sequência com infidelidade de transcrição ou com um seu complemento, ou com uma região de um ácido nucleico que flanqueia um local de infidelidade da transcrição e que é adequado para a amplificação da totalidade ou de uma porção do referido local de infidelidade da transcrição numa amostra contendo a referida sequência ou o seu complemento. Sequências iniciadoras típicas deste invento são moléculas de ácido nucleico de cadeia simples de aproximadamente 5 a 60 nucleótidos de comprimento, mais de preferência cerca de 8 a cerca de 50 nucleótidos de comprimento, ainda preferencialmente cerca de 10 a 40, 35, 30 ou 25 nucleótidos de comprimento. Prefere-se a complementaridade perfeita para assegurar elevada especificidade. No entanto, pode ser tolerado algum desemparelhamento, como discutido atrás para as sondas.

O termo "flanqueia" indica que a região deverá estar localizada a uma distância do local de infidelidade da transcrição que seja compatível com as actividades de polimerase convencionais, *e.g.*, não acima de 250 pb, de preferência não excedendo 200, 150, 100 ou, mais preferencialmente, 50 pb a montante do referido local.

Este invento também divulga pelo menos um par de sequências iniciadoras de ácido nucleico, em que o referido

par de sequências iniciadoras compreende uma sequência iniciadora com sentido e uma sequência iniciadora inversa e em que as referidas sequências iniciadoras com sentido e inversas permitem a amplificação selectiva de uma sequência com infidelidade de transcrição ou da sequência exactamente complementar. (*e.g.*, na ***Fig 12***).

A infidelidade de transcrição é medida por Cromatografia Líquida de Alta Resolução Desnaturante (DHPLC). O princípio deste método está descrito *e.g.*, na ***Fig 11***. Basicamente, os produtos de amplificação são desnaturados e re-emparelhados. Durante o re-emparelhamento, são formados homoduplexes e heteroduplexes, como resultado da presença de sequências com infidelidade da transcrição. A mistura é então analisada por DHPLC. Uma vez que as estruturas de DNA dos heteroduplexes e dos homoduplexes são diferentes à temperatura da análise, são eluidos diferencialmente e a sua quantidade (relativa) pode ser avaliada.

Como verificação das sequências com infidelidade da transcrição existentes em células de cancro com maior frequência que em células normais, as infidelidades da transcrição previstas para genes seleccionados (ENO1, GAPDH, TMSB4X) foram amplificadas por RT-PCR usando os oligonucleótidos indicados (ver Exemplo 6). As variações dos homo- e heteroduplexes foram verificadas (Fig. 15). Qualquer outra técnica adequada para a detecção de ácidos nucleicos pode ser usada ou adaptada para usar no presente

invento, para detectar, quantificar ou monitorizar a infidelidade de transcrição.

Detecção da infidelidade de transcrição ao nível das proteínas

Devido à infidelidade da transcrição conduzir a alterações na sequência proteica, a presença ou nível de infidelidade de transcrição pode também ser medida através da detecção da presença ou quantidade de proteínas TI.

Várias técnicas conhecidas *per se* na área podem ser usadas e/ou adaptadas para medir a infidelidade da transcrição em proteínas. Em particular, uma vez que a infidelidade da transcrição causa modificações das proteínas conduzindo a alterações directas no seu comportamento, estas alterações podem ser detectadas por *e.g.*, electroforese em gel bidimensional e espectrometria de massa ou por desorção/ionização a laser estimulada na superfície. Como uma verificação de que a proteína TI está presente no plasma humano, mostramos o perfil de espectrometria de massa de um péptido localizado antes do codão de paragem canónico de ApoAII. Os dados de espectrometria de massa mostram que a Arginina substitui um codão de paragem canónico. Ainda, uma vez que a infidelidade da transcrição causa a ocorrência de isoformas da proteína mais longas e mais curtas (celular e plasmática) (como resultado do codão de paragem prematuro devido à substituição de bases ou a uma alteração da grelha

de leitura no caso das deleções ou inserções), tais isoformas podem ser detectadas ou quantificadas usando ligandos específicos das mesmas e/ou estratégias de sequenciação de proteínas. Neste contexto, demonstramos que as alterações no comprimento da proteína e a sequência de AA primária ocorrem predominantemente no domínio do extremo carboxilo da proteína. Assim, as estratégias de sequenciação deverão ser dirigidas para o extremo C das proteínas, uma região anteriormente insuspeita (de facto, os métodos de sequenciação directa de proteínas presentemente usados podem ser conseguidos apenas começando nos AA do extremo NH_2).

A detecção da infidelidade da transcrição usando ligandos específicos tipicamente compreende a colocação de uma amostra em contacto com um ligando que é específico de um domínio de uma proteína TI e detecção da presença (ou quantidade) do complexo formado, a referida presença ou quantidade sendo uma indicação directa da presença ou taxa de infidelidade da transcrição. Numa realização preferida, o método usa uma série de ligandos que são específicos de domínios distintos de uma ou várias proteínas TI, respectivamente. Os ligandos específicos para estes domínios podem ser preparados para qualquer proteína usando a informação de sequências e as regras de substituição de aminoácidos descritas no presente pedido de patente (ver, *e.g.*, Exemplo 5). O ligando pode ser usado na forma solúvel ou como revestimento de uma superfície ou suporte.

O invento também divulga ligando que selectivamente se ligam a um domínio de uma proteína TI. Diferentes tipos de ligandos podem ser considerados, tais como anticorpos específicos, moléculas sintéticas, aptâmeros, péptidos e similares.

O ligando pode ser um anticorpo ou um fragmento ou derivado do mesmo. Assim, um aspecto particular deste invento reside num anticorpo que se liga especificamente a um domínio de uma proteína TI.

Dentro do contexto deste invento, um anticorpo designa um anticorpo policlonal, um anticorpo monoclonal, assim como fragmentos do mesmo ou seus derivados tendo substancialmente a mesma especificidade de antigénio. Fragmentos incluem, *e.g.*, Fab, Fab'_2, regiões CDR, etc. Os derivados incluem anticorpos de cadeia simples, anticorpos humanizados, anticorpos humanos, anticorpos polifuncionais, etc.

Os anticorpos contra domínios das proteínas TI podem ser produzidos segundo procedimentos conhecidos de um modo geral na área. Por exemplo, os anticorpos policlonais podem ser produzidos através da injecção do domínio da proteína TI (*e.g.*, como proteína ou péptido), sozinho ou acoplado a uma proteína adequada, num animal não humano. Após um período adequado, o animal é sangrado, o soro recuperado e purificado por técnicas conhecidas na área (Paul, W.E. "Fundamental Immunology" Second Ed. Ravenn

Press, NY, p. 176, 1989; Harlow *et al.*, "Antibodies: A laboratory manual", CSH Press, 1988; Ward *et al.* (Nature 341 (1989) 544). Os péptidos com a mesma sequência das proteínas TI podem ser produzidos por procedimentos conhecidos de um modo geral na área. Tais péptidos podem ser acoplados a um suporte adequado e usados para a detecção de auto-anticorpos presentes em amostras biológicas (por exemplo fluidos corporais).

Os anticorpos monoclonais contra domínios das proteínas TI podem ser preparados, por exemplo, pela técnica de Kohler-Millstein (2) (Kohler-Millstein, Galfre, G., e Milstein, C, Methods Enz. 73 p. 1 (1981)) envolvendo a fusão de um linfócito B imune com células de mieloma. Por exemplo, um imunogénio como atrás descrito pode ser injectado num mamífero não humano como descrito. Subsequentemente, o baço é removido e a fusão com mieloma realizada de acordo com uma variedade de métodos. As células de hibridoma resultantes podem então ser testadas relativamente à secreção de anticorpos contra domínios de proteínas TI.

Um anticorpo "selectivo" para um domínio particular de proteínas TI designa um anticorpo cuja ligação ao referido domínio (ou um seu fragmento contendo o epitopo) pode ser discriminado de forma reprodutível de entre ligação não específica (*i.e.*, da ligação a um outro antigénio, particularmente à proteína nativa não contendo o referido domínio). Os anticorpos selectivos para os domí-

nios das proteínas TI permitem a detecção da presença de proteínas contendo tais domínios numa amostra.

Diagnóstico

O presente invento divulga a realização de ensaios de detecção ou diagnóstico que podem ser usados, entre outras coisas, para detectar a presença, ausência, predisposição, risco ou gravidade de uma doença a partir de uma amostra derivada de um indivíduo. O termo "diagnóstico" deverá ser entendido como incluindo métodos de farmaco-genética, prognóstico, etc.

O invento divulga um método de detecção *in vitro* ou *ex vivo* da presença, ausência, predisposição, risco ou gravidade de doenças numa amostra biológica, de preferência, uma amostra biológica humana, compreendendo a colocação da referida amostra em contacto com um ligando que especificamente se liga a um local de infidelidade da transcrição e determinação da formação de um complexo.

O invento também divulga um método de detecção da presença ausência, predisposição, risco ou gravidade de cancros num indivíduo, o método compreendendo a colocação *in vitro* ou *ex vivo* de uma amostra do indivíduo em contacto com um ligando que se liga especificamente a um local de infidelidade da transcrição expresso pelas células de cancro e determinação da formação de um complexo.

Este invento também divulga um método de avaliação da resposta de um indivíduo a um tratamento de cancro, o método compreendendo detecção da presença ou taxa de infidelidade da transcrição numa amostra do indivíduo em diferentes tempos antes e durante o curso do tratamento.

Este invento também divulga um método de determinação da eficácia de um tratamento de uma doença de cancro, o método compreendendo (i) obtenção de uma amostra de tecido do indivíduo durante ou após o referido tratamento, (ii) determinação da presença e/ou taxa de infidelidade da transcrição na referida amostra e (iii) comparação da referida presença e/ou taxa da quantidade de infidelidade da transcrição com uma amostra de referência do referido indivíduo retirada antes do tratamento ou numa fase inicial deste.

A presença (ou aumento) da infidelidade da transcrição numa amostra é indicativo da presença, predisposição ou fase da progressão de uma doença de cancro. Assim, o invento permite o desenho de intervenção terapêutica adequada, o qual é mais eficaz e ajustado. Igualmente, esta determinação ao nível pré-sintomático permite que seja aplicado um regime preventivo.

Os métodos de diagnóstico do presente invento podem ser realizados *in vitro*, *ex vivo* ou *in vivo*, de preferência *in vitro* ou *ex vivo*. A amostra pode ser qualquer amostra biológica derivada de um indivíduo, a qual

possui ácidos nucleicos ou polipéptidos, conforme adequado. Exemplos de tais amostras incluem fluidos corporais, tecidos, amostras de células, órgãos, biopsias, etc. As amostras mais preferidas são sangue, plasma, soro, saliva, urina, fluido seminal e similares. A amostra pode ser tratada antes da execução do método, de forma a permitir ou aumentar a disponibilidade dos ácidos nucleicos ou polipéptidos a testar. Os tratamentos podem incluir, por exemplo um ou mais do seguinte: lise celular (*e.g.*, mecânica, física, química, etc.), centrifugação, extracção, cromatografia em coluna, e similares.

Um método divulgado no presente invento consiste na determinação da presença em amostras humanas de anticorpos dirigidos contra os novos péptidos produzidos pela infidelidade da transcrição e geração de estimulação imunitária conduzindo a produção de anticorpos. Um segundo método do invento é dirigido à detecção de células portadoras de estrutura imunológica dirigida contra péptidos da infidelidade da transcrição.

Desenho de fármacos e terapia

Como discutido atrás, o invento permite o desenho (ou rastreio) de novos fármacos através da avaliação da capacidade de uma molécula candidata para modular a infidelidade da transcrição. Tais métodos incluem os ensaios de ligação e/ou ensaios funcionais (actividade) e podem ser realizados *in vitro* (*e.g.*, em sistemas celulares ou em ensaios não celulares), em animais, etc.

Este invento divulga um método de selecção, caracterização, rastreio ou optimização de um composto biologicamente activo, o referido método compreendendo a determinação *in vitro* se um composto a testar modula a infidelidade da transcrição. A modulação da infidelidade da transcrição pode ser avaliada relativamente a um gene ou proteína particular, ou relativamente a uma série pré-definida de genes ou proteínas, ou globalmente.

O presente invento também divulga um método de selecção, caracterização, rastreio ou optimização de um composto biologicamente activo, o referido método compreendendo a colocação *in vitro* de um composto a testar em contacto com um gene e determinação da capacidade do referido composto a testar para modular a produção, a partir do referido gene, de moléculas de RNA contendo locais de infidelidade da transcrição.

O presente invento também divulga um método de selecção, caracterização, rastreio e/ou optimização de um composto biologicamente activo, o referido método compreendendo o contacto de um composto a testar com uma célula e determinação, na referida célula, se o composto a testar modula a infidelidade da transcrição.

Os ensaios de rastreio atrás referidos podem ser realizados em qualquer dispositivo adequado, como sejam placas, tubos, frascos, etc. Tipicamente, o ensaio é reali-

zado em placas de microtitulação de múltiplos alvéolos. Usando o presente invento, vários compostos a testar podem ser testados em paralelo. Ainda, o composto a testar pode ser de origem, natureza e composição diversa. Pode ser qualquer substância orgânica ou inorgânica, como sejam lípidos, péptidos, polipéptidos, ácidos nucleicos, moléculas pequenas, isolados ou misturados com outras substâncias. Os compostos podem constituir a totalidade ou parte de uma biblioteca combinatória de compostos, por exemplo.

Os compostos que modulam a infidelidade da transcrição do presente invento possuem muitas utilidades, tais como utilidades terapêuticas, para reduzir ou aumentar a infidelidade da transcrição numa célula. Tais compostos podem ser usados para tratar (ou prevenir) doenças causadas ou associadas a infidelidade anormal da transcrição, tais como distúrbios proliferativos (*e.g.*, cancros), doenças imunonológicas, envelhecimento, doenças inflamatórias, etc.

Outros compostos do presente invento são compostos que têm como alvo a infidelidade da transcrição, *i.e.*, compostos que selectivamente se ligam a locais de infidelidade da transcrição. Tais compostos (*e.g.*, anticorpos, RNAi, etc.) podem ser usados como reagentes de diagnóstico ou como agentes terapêuticos, por si sós ou conjugados com uma marca.

Os compostos que modulam ou têm como alvo a infidelidade da transcrição do presente invento podem ser

administrados por qualquer via adequada, incluindo a oral, ou por entrega sistémica, injecções intravenosas, intra-arteriais, intracerebrais ou intratecais. A dosagem pode variar dentro de limites largos e terá de ser ajustada às necessidades individuais em cada caso particular, dependendo de vários factores conhecidos dos familiarizados com a área. Pode ser usada qualquer forma de dosagem aceitável em termos farmacêuticos conhecida na área, como seja qualquer solução, suspensão, pó, gel, etc., incluindo solução isotónica, soluções tamponadas e salinas, etc. Os compostos podem ser administrados sozinhos, mas são geralmente administrados com um veículo farmacêutico, de acordo com a prática farmacêutica convencional (como descrito em Remington's Pharmaceutical Sciences, Mack Publishing). Os anticorpos podem ser administrados de acordo com métodos e protocolos conhecidos *per se* na área, os quais são presentemente usados em ensaios clínicos e terapias com seres humanos.

Efeito da infidelidade da transcrição no funcionamento normal da célula

De modo geral, é aceite que as proteínas são responsáveis pela maioria das funções celulares. No entanto, é igualmente claro que o RNA, conhecido como RNA não codificador, são também moléculas funcionais por si sós. Estes transcritos, cuja importância foi anteriormente subestimada, estão presentes em muitos organismos e regulam uma série de funções celulares. Entre eles, pode-se

encontrar rRNA, tRNA, tmRNA (RNA transmensageiro), mas também outro RNA não codificador (snoRNA, snRNA, etc.) que intervêm nas modificações pós-transcrição do RNA, no "splicing" e noutras funções celulares. As funções do RNA e das proteínas podem ser afectadas pela ausência de fidelidade da transcrição descrita neste pedido. A sequência de DNA é transcrita em mRNA no núcleo pela acção de uma proteína designada polimerase II do RNA que reconhece a sequência da molécula de DNA e sintetiza um polímero de nucleótidos de cadeia simples que é complementar da matriz de DNA original. A ordem e tipo de bases do RNA em qualquer posição é determinada pela ordem e tipo de bases presentes na cadeia de DNA que serve de matriz. O extremo 5'do RNA transcrito, o qual corresponde à primeira base do mRNA, está ligada a 7-metilguanilato ligado ao nucleótido inicial constituindo assim o CAP 5' que protege o RNAS da degradação. Esta modificação ocorre antes da transcrição estar completa. O processamento no extremo 3' do transcrito primário envolve a clivagem por uma endonuclease para dar um grupo hidroxilo 3' livre a que é adicionado um segmento de resíduos de adenina por uma enzima designada polimerase de poli(A). A cauda poli(A) resultante contém 100-250 nucleótidos. O passo final do processamento é o "splicing", o qual consiste na clivagem do transcrito primário dos elementos correspondentes às sequências intrónicas seguido de ligação dos exões. Este processo é controlado por numerosas proteínas e ncRNA, algum do qual se liga ao RNA. É possível que esta ligação seja um passo que possa ser afectado pela falta de

fidelidade da transcrição. De facto, todas estas proteínas são elas próprias codificadas por RNA específico; a função destas proteínas é assim potencialmente afectada pela falta de fidelidade da transcrição. O processo altamente complexo da maturação do RNA leva à produção de mRNA que será exportado do núcleo de forma a ocorrer tradução. É importante reconhecer que, nesta fase, nem todos os tripletos de nucleótidos que estão presentes no mRNA maduro serão traduzidos em AA. O RNA mensageiro maduro possui regiões não codificadoras referidas como regiões não traduzidas 5' e 3' que são importantes em termos funcionais na determinação da estabilidade global do mRNA e noutros papéis na tradução. De forma a eficientemente copiar a informação contida no DNA na sequência correcta de AA da proteína, é necessária fidelidade absoluta da transcrição. De facto, qualquer erro que ocorra durante os processos de transcrição, ou durante a maturação, potencialmente resultarão na introdução de alterações na estabilidade do mRNA e, finalmente, numa variação na sequência primária de AA. Estas variações na sequência proteica devem pois exacerbar o fenómeno de infidelidade da transcrição através da alteração da sequência da proteína envolvida na iniciação da transcrição, na própria transcrição, na adição do cap 5' do RNA, na poliadenilação 3', no "splicing" de RNA e/ou na exportação do RNA. Assim, a demonstração neste invento que a infidelidade da transcrição afecta um grande número de genes isolados de células de cancro abre a possibilidade que o fenómeno possa exacerbar-se conduzindo ao aumento progressivo da gravidade da doença. A variação

nas sequências de RNA introduzida pela infidelidade da transcrição pode ter consequências imediatas na função celular. De facto a introdução de bases no transcrito primário de RNA que não são complementares das da matriz de DNA tem consequências potenciais imediatas na sequência primária de AA da proteína resultante. Quando a alteração afecta as primeiras 2 bases do codão, tipicamente, exerce um impacto directo na sequência de AA. Variações que afectam a terceira base do codão terão um impacto menor na sequência de AA da proteína devido ao código genético ser degenerado. Assim, as alterações das bases localizadas na terceira posição do codão podem não influenciar directamente a sequência de AA da proteína. Com base nos dados descritos abaixo, pensamos que a infidelidade da transcrição seja um fenómeno que afecta todas as bases independentemente da sua posição no codão e assim têm impacto na sequência primária de AA da proteína. As alterações na proteína podem ser neutras ou causar uma modificação da função da proteína, seja o aumento ou a perda de actividade. O fenómeno da infidelidade da transcrição é predominantemente observado após completada a codificação de pelo menos as primeiras 400-500 bases de mRNA maduro; o extremo 5' do mRNA é relativamente menos afectado. Pensamos que fragmentos mais curtos assim como fragmentos mais longos da proteína estarão presentes em células de cancro, assim como no plasma de doentes com cancro. Estas isoformas mais curtas e mais longas podem ser directamente deduzidas das regras de infidelidade da transcrição descritas abaixo. Isto permite a produção de

métodos desenhados racionalmente para a identificação de marcas específicas de cancro ou de gravidade de doença imunológica. Devido à taxa de transcrição da maioria dos genes ser controlada para se adaptar às necessidades celulares, é proposto que a infidelidade da transcrição cause alterações directamente na expressão de genes devido a função excessiva ou nula da proteína. Assim, o fenómeno da infidelidade da transcrição exerce um efeito profundo na capacidade celular para realizar a sua função. A identificação deste defeito como uma característica comum à maioria dos genes isolados a partir de todos os tipos de células de cancro proporciona assim uma procura racional de novos métodos que permitem a medição quantitativa da taxa de infidelidade da transcrição em qualquer célula determinada. Este ensaio de rastreio permite testar os novos fármacos capazes de limitar a infidelidade da transcrição, prevenindo assim a progressão da doença e restaura a função celular normal.

Infidelidade da transcrição e patologia

O presente invento mostra que a infidelidade da transcrição conduz a alterações importantes na sequência da proteína. A função de qualquer proteína é determinada pela sua estrutura tridimensional que está directamente dependente da sua sequência de AA. Proteínas com um AA variante, proteínas mais curtas ou proteínas mais longas devem resultar numa modificação profunda da actividade proteica ou não ter qualquer efeito. Foram descritos

exemplos de proteínas modificadas que inibem significativamente a função dos seus homólogos normais. A infidelidade da transcrição é um fenómeno que afecta um grande número de genes, originando assim um grande número de proteínas. Devido a várias destas proteínas participarem na manutenção da estrutura de DNA estável e participarem na reparação de DNA, uma função defectiva destas proteínas pode resultar em reparação defectiva do DNA e resultar numa capacidade significativamente diminuída das células para reparar com êxito qualquer DNA danificado. Devido à fidelidade da transcrição em células normais ser devida à actividade de vários complexos proteicos que controlam a transcrição, é muito possível que pequenas alterações iniciais destas proteínas possam progressivamente exacerbar o fenómeno. Isto resultará em último caso numa maior taxa de infidelidade de transcrição que consequentemente suprirá o estatuto de diferenciação celular e conduzirá a formas cada vez mais graves da doença. A demonstração que a infidelidade da transcrição de facto ocorre em células de cancro e conduz a uma larga diversificação de proteínas codificadas por qualquer gene abre uma nova área para o rastreio de novas sondas de diagnóstico e desenho de novos alvos terapêuticos.

O presente invento divulga um novo fenómeno que contribui para a diversificação da informação presente no DNA e que segue regras específicas. Este processo aleatório de infidelidade da transcrição é grandemente exacerbado em células de cancro mas também ocorre em células normais

seguindo as mesmas regras. Este mecanismo introduz novas bases no mRNA causando assim alterações profundas na mensagem que será traduzida ao nível dos ribossomas. Com grande probabilidade, o fenómeno é muito geral devido a estar presente na maioria dos genes testados. A consequência imediata da infidelidade da transcrição é que um único gene possa produzir muito mais proteínas do que o suspeito. Assim, a nossa descoberta tem o potencial para explicar, em parte, a discrepância relativa entre o número limitado de genes presentes no genoma e o grande número de proteínas presentes em amostras biológicas. Mostrámos que a infidelidade da transcrição é um processo não aleatório e é governado por regras específicas, algumas das quais já aqui foram descritas. Algumas serão reveladas através de uma melhor caracterização do processo usando métodos bioinformáticos e biológicos. Focámos aqui a implicação da infidelidade da transcrição no cancro. No entanto, pensamos que o processo é geral e que este pode contribuir para gerar diversidade de proteínas. Esta diversificação de proteínas deve exercer uma influência significativa na capacidade do sistema imunitário para reconhecer um padrão proteico específico. Neste contexto, pensamos que a infidelidade da transcrição desempenha um papel na patogénese de doenças imunológicas. Considerámos aqui os eventos de infidelidade da transcrição que conduzem à substituição de bases isoladas. Focámo-nos nas substituições de bases isoladas devido a este mecanismo ter poucas probabilidades de causar destabilização significativa da estrutura do mRNA. Assim, as substituições de bases isoladas não deverão interferir com a tradução, mas serão reconhecidas como uma

nova mensagem autêntica. Veremos ainda que se espera igualmente que a infidelidade da transcrição cause deleções e/ou inserções de bases. O mesmo método descrito atrás detectará mecanismos alternativos de infidelidade da transcrição. O novo processo descrito neste pedido possui várias aplicações práticas imediatas.

Optimização de experiências de proteómica para identificar novos biomarcadores específicos de doença e desenhar novos anticorpos que serão dirigidos contra proteínas específicas de doença

Com base nos ensinamentos do presente pedido, é agora possível prever as alterações que podem ocorrer em qualquer sequência proteica como resultado da infidelidade da transcrição e assim prever alterações nas propriedades fisicoquímicas tais como massas moleculares aparentes e pontos isoeléctricos. Isto pode originar padrões proteicos modificados em géis bidimensionais e em espectrometria de massa. Devido à infidelidade da transcrição afectar a maioria dos genes, é agora possível focar numa série limitada de proteínas e caracterizar as suas alterações de forma a identificar biomarcadores específicos de doença (*e.g.*, cancro). Os algoritmos de infidelidade da transcrição podem acelerar o processo de descoberta de biomarcadores. Como alternativa, é possível desenhar anticorpos específicos dirigidos contra domínios de proteínas TI, *i.e.*, os domínios das proteínas que possuem AA modificados, assim como novos AA adicionais que prolongam a proteína

nativa gerados pela infidelidade da transcrição. Os anticorpos dirigidos contra estes domínios serão específicos da variante proteica e portanto dirigidos a proteínas específicas da doença (*e.g.*, cancro) presentes em fluidos corporais, na superfície celular ou dentro de células de cancro. Estes anticorpos serão usados para detectar proteínas características de uma condição patológica em fluidos corporais, para detectar células doentes circulantes no plasma, para identificar células doentes em amostras histológicas, para avaliar a gravidade da doença e também para dirigir medicações para células doentes específicas. Isto pode ser conseguido através de transferência genética da sequência susceptível de infidelidade da transcrição através de um vector de terapia génica, *e.g.* adenovírus, lipossomas, etc. Como alternativa, desenhámos péptidos específicos tendo a mesma sequência que um domínio resultante de TI. Eles podem ser usados para detectar autoanticorpos específicos e naturais dirigidos contra os domínios de proteínas TI presentes nos fluidos corporais.

A infidelidade da transcrição ocorre não aleatoriamente e frequentemente afecta os codões de paragem do mRNA numa maior proporção dos genes testados. Ainda, devido a poder-se estimar que até 6% de qualquer população determinada de mRNA contém estas sequências adicionais, pode-se portanto detectar numa determinada célula de cancro uma população específica de proteínas que apresenta uma nova sequência no extremo carboxilo. Atingir estas sequências não era até agora concebível devido a não haver

um processo descrito, excepto para raras doenças genéticas que afectam o DNA, capaz de introduzir em grelha sequências que imediatamente se seguem a codões de paragem canónicos. A existência destas sequências escondidas não era até agora suspeita. A descoberta da infidelidade da transcrição conducente a substituição de bases isoladas ocorrendo em codões de paragem canónicos revela a existência de tais sequências codificadoras e ainda mostra que a sua presença está aumentada nas células de cancro devido ao aumento da infidelidade da transcrição nestas células. Assim, é possível identificar e atingir estas sequências normalmente ocultas para desenhar novos fármacos específicos de cancro.

A TI relacionada com deleção ou inserção de bases isoladas que ocorre na sequência codificadora leva a uma mudança da grelha de leitura. Como convergência, as sequências na proteína correspondente são modificadas (alteração na sequência de AA e no comprimento da proteína). Pode-se assim detectar numa determinada célula de cancro uma população específica de proteínas que apresenta uma nova sequência. Assim, é possível identificar e atingir estas sequências normalmente escondidas para desenhar novos fármacos específicos de cancro.

Devido à existência de novas sequências de proteína ou de sequências modificadas, e considerando que estas alterações estão presentes na proteína presente na superfície da célula, é proposto usar estas sequências de forma a vacinar contra sequências específicas de doença. Estas vacinações serão usadas para induzir respostas

imunitárias em doentes diagnosticados com qualquer forma de doença (*e.g.*, de cancro) ou para iniciar respostas imunitárias preventivas em doentes com risco aumentado de desenvolver doenças específicas (*e.g.*, cancro) devido à predisposição genética ou devido a um aumento da exposição a uma toxina ou risco ambiental.

Assim, este invento também divulga um método de detecção da presença ou fase de uma doença num indivíduo, o método compreendendo avaliação (*in vitro* ou *ex vivo*) da presença ou taxa da infidelidade da transcrição numa amostra do referido indivíduo, a referida presença ou taxa sendo uma indicação da presença ou fase de uma doença no referido indivíduo. A amostra pode ser qualquer tecido, célula, fluido, biopsia, etc., contendo ácidos nucleicos e/ou proteínas. A amostra pode ser tratada antes da reacção, *i.e.*, por diluição, concentração, lise, etc. O método tipicamente compreende avaliação da infidelidade da transcrição numa série de genes ou proteínas distintos, tais como proteínas do plasma, proteínas da superfície celular, etc.

Este invento divulga um método de tratamento de um indivíduo necessitado do mesmo, o método compreendendo a administração ao referido indivíduo de uma quantidade eficaz de um composto que altera (*e.g.* reduz ou aumenta) a taxa de infidelidade de transcrição de um gene de mamífero.

Este invento também divulga a utilização de um composto que altera (*e.g.*, reduz ou aumenta) a taxa de infidelidade da transcrição de um gene de mamífero para a produção de uma composição farmacêutica para usar num método de tratamento de um ser humano ou animal, particularmente para o tratamento de um distúrbio de proliferação celular, como sejam cancros e doenças imunitárias.

O invento também divulga um método de avaliação da eficácia de um fármaco ou candidato a fármaco, o método compreendendo um passo de avaliação se o referido fármaco altera a taxa de infidelidade da transcrição de um gene de mamífero, tal alteração sendo uma indicação da eficácia do fármaco.

O invento ainda divulga métodos e produtos (tais como sondas, sequências iniciadoras, anticorpos ou derivados dos mesmos), para detecção ou medição (o nível de) infidelidade de transcrição numa amostra, assim como os kits correspondentes.

O invento também divulga um método de identificação e/ou produção de biomarcadores, o método compreendendo a identificação, numa amostra de um indivíduo, da presença de locais de infidelidade da transcrição numa proteína alvo, RNA ou gene e, facultativamente, determinação da sequência dos referidos locais de infidelidade da transcrição. Numa realização particular e

preferida, a proteína alvo é uma proteína da superfície celular ou uma proteína secretada, particularmente uma proteína da superfície celular ou uma proteína do plasma.

O invento divulga um método de produção ou identificação de ligandos específicos para uma característica ou condição patológica, o método envolvendo a identificação, numa amostra de um indivíduo tendo a referida característica ou condição patológica, a presença de locais de infidelidade da transcrição numa ou mais proteínas alvo, RNA ou gene, facultativamente determinação da sequência dos referidos locais de infidelidade da transcrição e produção (a) de ligandos que especificamente se ligam aos referidos locais de infidelidade de transcrição.

O invento é particularmente adequado para a identificação de biomarcadores de distúrbios de proliferação celular, tais como cancros, doenças imunitárias, inflamação ou envelhecimento. É particularmente útil para a produção de ligandos que sejam específicos de tais distúrbios em mamíferos, em particular ligandos que possam detectar a presença ou gravidade de um distúrbio de proliferação celular num indivíduo.

Em particular, este invento divulga um péptido (sintético) compreendendo um local de infidelidade da transcrição de uma proteína, particularmente de uma

proteína de mamífero, mais de preferência de uma proteína humana. O péptido tipicamente compreende um fragmento interno de proteína ou a sequência de um fragmento C-terminal da referida proteína. O péptido preferencialmente compreende menos de 100, 80, 75, 70, 65, 60, 50, 45, 40, 35, 30, 25 ou mesmo 20 aminoácidos. A proteína pode ser uma proteína da superfície celular (*e.g.*, um receptor, etc.), uma proteína secretada (*e.g.*, uma proteína do plasma) ou uma proteína intracelular. São proporcionados exemplos de tais proteínas do plasma, *e.g.*, na Fig. 13, e incluem apolipoproteínas (*e.g.*, AI, AII, CI, CII, CIII, D, E), componentes do complemento (*e.g.*, CIs, C3, C7), proteína C reactiva, inibidores de peptidases serpinas, fibrinogénio (*e.g.*, FGA1, FGA2), plasminogénio, transferrina, transtirretina, etc. Exemplos de receptores da superfície celular incluem *e.g.*, receptores de citocinas e receptores de hormonas, etc. Numa realização específica, o invento está relacionado com um péptido sintético compreendendo um local de infidelidade da transcrição de uma proteína humana abundante, do plasma ou receptor da superfície celular como listado atrás.

Exemplos específicos de tais péptidos estão descritos nas Figura 10, 14 e 18. Mais especificamente, exemplos de péptidos sintéticos do presente invento compreendem a totalidade ou um fragmento das sequências de aminoácidos que se seguem:

QMWQLFWIYHLSS (TPT1) (SEQ ID NO: 1)KLHTLSAAIYYQQB (VIM) (SEQ II NO: 2)
DFLSNK (RPS6) (SEQ ID NO: 3)MYTVEFSVHKNN (RPL7A) (SEQ ID NO: 4)NGSLGDMSDLCT (RPS4X) (SEQ ID NO: 5)ASG (FTH1) (SEQ ID NO: 6)
EPSEPSDF (FTL) (SEQ ID NO: 7)
APSIFPTLPAKPGTKQPRSPVTALSLHMLLMVSSAPSCGLIQTVSSFTVYIFTL (TPI1) (SEQ ID NO:
8)ARHGRDEEVWHRKHSHHFVQAWAWVGGLVCWPRKCHMRSTLISSLDSLLPVIPHRTEAEWV
VMFDRRH (AHSG) (SEQ ID NO: 9)
RSKAYSSVFLFRWCKANTLSKKHKFL (ALB) (SEQ ID NO: 10)
GLDSTRALENEMTV (APCS) (SEQ ID NO: 11)
GARRRPPSRCSE (APOA1) (SEQ ID NO: 12)
SVQTIVFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 13)
IVFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 14)QPDPPSVDKGRVPYSPDPPGSD (APOC2) (SEQ ID NO: 15)
DPPSVDKGRVPYSPD (APOCII) (SEQ ID NO: 16)
DLNTPSPPAYPSCELLGSCNLQGCPCRLLKRDSILSALLPHLMPGPPPGMLASQ (APOC3) (SEQ ID NO: 17)
PGSTGRLHPLHVTSASLSPTPPPPHKDKPINHDKGS (APOD) (SEQ ID NO: 18)
TPKPAAMRPHATPCLLPPRSLQRETLSPPQPSSWGGP (APOE) (SEQ ID NO: 19)
EARVGGNVGSQTQ (AZGP1) (SEQ ID NO: 20)
PSVLHTARGPRMPRPPLAPAGREPDHLPC (CHI3L1) (SEQ ID NO: 21)
DVDVAFAPTGASESSSPQDELQPPRESSARHQVTRPQPPGPQLRPASPRSGSCTLTLDSAAHG
NRIAPACN (CLU) (SEQ ID NO: 22)
NVIPLKRKMMNTLN (HRG) (SEQ ID NO: 23)
TPAARLMWSSNMPYFAQKTAKDMTSSWLQPRFIFLFVVN (IGFBP3) (SEQ ID NO: 24
GWGVFLLNPMAGGHAPTIISWEERQSWEIDGSHSSLLSLLCLWATLPTPLLSQ (INHA) (SEQ ID NO: 25)
TGPTHHSPSPSISTWCLVPVHSVNKKP (KLK11) (SEQ ID NO: 26)
WTPEPLLQPLSHPLPPAHPLGQQRL (PKM2) (SEQ ID NO: 27)
LDGRQSDALTHLEAGTWVGI (PLG) (SEQ ID NO: 28)
SLPSSSGALSKELGMQAGCLGLWAQPGPCAPSGHGMCGPVCLSLEGDSDSLCSSHMHRGPWTL
SGGSWAS (SERPINA3) (SEQ ID NO: 29)
NLRGRAATKVKMGTQMIHEFALVSLAQVVCANHVCLHSSVLPCVLNKK (TF) (SEQ ID NO: 30)
GPAPPRPAPAGPAPPRPAPAALPMGAVFKDTRAPSPPGAPLKMERGLRISVSLGACLGSPSLT
PHSHSLSLPLCLLLPVCTIPLPGIKAQGTSGEHYCS (TGFB1) (SEQ ID NO: 31)
GTSPPVDLKDEGWDFM (TTR) (SEQ ID NO: 32)

Um outro aspecto deste invento reside na utilização de um péptido de 100 ou menos aminoácidos de comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOS: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 para detecção ou monitorização de distúrbios de proliferação celular.

Um outro aspecto deste invento reside no uso de um péptido, como descrito atrás, como um imunogénio. O

invento também descreve uma composição de vacina compreendendo um péptido compreendendo um local de infidelidade da transcrição, como definido atrás, e facultativamente um veículo, excipiente e/ou adjuvante adequado.

O invento também divulga um dispositivo ou produto compreendendo, imobilizado num suporte, um reagente que especificamente se liga a um local de infidelidade da transcrição. O reagente pode ser *e.g.*, uma sonda ou um anticorpo ou seu derivado.

O invento pode ser usado em qualquer indivíduo mamífero, particularmente qualquer indivíduo humano, para detectar, monitorizar ou tratar uma variedade de condições patológicas associadas a distúrbios de proliferação celular (*e.g.*, cancros), e/ou produzir, desenhar ou testar fármacos terapeuticamente activos.

Produção e utilização de agentes que visam atingir locais de infidelidade da transcrição

Usando os ensinamentos do presente invento, é possível produzir agentes que podem atingir locais de infidelidade da transcrição, *e.g.*, que se ligam a proteínas ou ácidos nucleicos contendo um local de infidelidade da transcrição ou a células ou tecidos contendo ou expressando tais proteínas ou ácidos nucleicos. O agente com este alvo pode ser um anticorpo (ou um derivado do mesmo), uma sonda, uma sequência iniciadora, um aptâmero, etc., como descrito atrás.

Tais agentes para atingir alvos podem ser usados *e.g.*, como agentes de diagnóstico, para detectar, monitorizar, etc. a infidelidade da transcrição numa amostra, tecido, indivíduo, etc. Para esse fim, o agente pode ser acoplado a um grupo marcador, como seja uma marca radioactiva, uma enzima, uma marca fluorescente, um corante luminescente, etc.

Este invento divulga um método de produção de um agente que tem como alvo a infidelidade da transcrição, o método compreendendo (i) identificação de um local de infidelidade da transcrição de uma proteína ou ácido nucleico e (ii) produção de um agente que selectivamente se liga ao referido local. O local de infidelidade da transcrição pode ser identificado *e.g.*, através do alinhamento de sequências disponível para uma determinada molécula de RNA e identificação de variações da sequência, particularmente no extremo 3'. Os locais de infidelidade da transcrição podem também ser identificados através da aplicação das regras de infidelidade da transcrição, como descrito no presente pedido (ver *e.g.*, Exemplo 5) ou novas regras posteriormente descritas, a qualquer seguência de gene. A molécula de péptido ou de ácido nucleico compreendendo o referido local identificado pode então ser produzida usando métodos convencionais. Numa realização preferida, o passo (i) do método compreende a identificação de um local de infidelidade da transcrição de uma proteína secretada ou da superfície celular (*e.g.*, um receptor,

etc.) ou de um ácido nucleico. De facto, as proteínas secretadas e as proteínas da superfície celular podem ser facilmente atingidas usando um agente que as tem como alvo, o qual é colocado em contacto com uma célula ou administrado a um indivíduo. Exemplos de domínios de tais proteínas TI estão descritos nos exemplos.

Produção e utilização de uma vacina que provoca ou estimula ou inibe uma resposta imunitária contra um domínio de uma proteína TI

Usando os ensinamentos do presente invento, é possível produzir uma composição de vacina que induz ou inibe uma resposta imunitária contra proteínas TI ou contra células ou tecidos contendo ou expressando tais proteínas TI ou correspondentes ácidos nucleicos.

Tipicamente, a composição de vacina compreende, como imunogénio, uma molécula compreendendo a sequência de um local de infidelidade da transcrição. A composição de vacina pode compreender qualquer veículo, excipiente ou adjuvante farmaceuticamente aceitável. A composição de vacina pode ser usada para gerar uma resposta imunitária contra uma célula ou tecido doente expressando proteínas TI, resultando numa destruição da referida célula ou tecido pelo sistema imunitário. Como alternativa, a composição de vacina pode ser usada para induzir uma tolerância relativamente aos locais de infidelidade de transcrição envolvidos em doenças auto-imunes.

O invento também divulga um método de produção de um imunogénio que causa, estimula ou inibe uma resposta imunitárias contra um domínio de proteínas TI, ou contra uma célula ou tecido expressando tal proteína, o método compreendendo (i) identificação de um local de infidelidade da transcrição de uma proteína ou ácido nucleico e (ii) produção de um péptido compreendendo esse local ou variante do mesmo. O local de infidelidade da transcrição pode ser identificado *e.g.*, através do alinhamento de sequências disponíveis para uma determinada molécula de RNA e identificação de variações de sequências, particularmente no extremo 3'. O local de infidelidade da transcrição pode também ser identificado através da aplicação das regras de infidelidade da transcrição, como descrito no presente pedido de patente (ver *e.g.*, Exemplo 5) ou novas regras mais tarde descritas, a qualquer sequência de gene. A molécula de péptido compreendendo o referido local identificado pode então ser produzida usando métodos convencionais. Numa realização preferida, o invento está relacionado com um método de produção de um anticorpo, compreendendo imunização de um mamífero não humano com um péptido de acordo com o invento, e recuperação dos anticorpos que se ligam ao referido péptido ou correspondentes células produtoras dos anticorpos.

Redução da infidelidade da transcrição em sistemas produtores de proteínas recombinantes

Este invento também divulga um método de prevenção ou redução da infidelidade da transcrição que pode ocorrer em sistemas de produção recombinantes, particularmente em sistemas de produção de proteínas terapêuticas. De facto, tal infidelidade da transcrição pode baixar o rendimento da produção do sistema ou resultar na produção de misturas ou de proteínas não caracterizadas que podem apresentar vários perfis de actividade. É assim recomendado e altamente importante ser capaz de reduzir a ocorrência ou taxa de infidelidade da transcrição em sistemas de produção recombinantes, como sejam sistemas de produção bacterianos (se as regras de infidelidade da transcrição forem as mesmas para os sistemas procarióticos), como sistemas de produção eucarióticos (*e.g.*, leveduras, células de mamífero, etc). Tal redução pode ser conseguida *e.g.*, através da adaptação da sequência da molécula de ácido nucleico codificadora e/ou através da inibição de moléculas de RNA geradas através da infidelidade da transcrição. Como alternativa, as proteínas contendo locais de infidelidade da transcrição podem ser removidas do meio.

Redução da sensibilidade à infidelidade da transcrição para um gene

Este invento também divulga um método de prevenção ou redução de TI que pode ocorrer num gene. De facto, tal TI pode afectar a expressão ou a actividade de uma proteína, ou resultar na produção de misturas de

proteínas não caracterizadas que podem apresentar vários perfis de actividade. É pois recomendado e altamente importante ser capaz de reduzir a ocorrência ou taxa de TI para o referido gene. Tal redução pode ser conseguida *e.g.*, através da modificação da sequência do gene (terapia génica) e/ou através da inibição de moléculas de RNA geradas através de infidelidade da transcrição. Como alternativa, as proteínas contendo locais TI podem ser especificamente degradadas (*e.g.*, especificamente marcadas para degradação).

Medição da infidelidade da transcrição ao nível do RNA

A transcriptómica é a ciência que mede a variação nos níveis de RNA (de preferência mRNA) numa variedade de condições patológicas, de entre as quais o cancro é a mais frequente. A transcriptómica baseia-se na hibridação de cDNA com uma série específica de oligonucleótidos que identificam sub-séries pré-definidas de genes. A eficácia de hibridação está dependente de sequências específicas de qualquer RNA determinado que será sujeito a transcrição reversa para cDNA. A introdução de variação de sequências insuspeita num determinado RNA reduzirá a eficácia da hibridação e portanto causará variação no sinal emitido pelos chips transcriptómicos. A descoberta de que a infidelidade da transcrição causa alterações da sequência de bases do RNA tem duas consequências imediatas: primeiro,

permite a optimização dos chips transcriptómicos de forma a minimizar a consequência do desemparelhamento de bases devido à infidelidade da transcrição e assim melhora a precisão da experiência transcriptómica. De facto, a actual limitação das experiências de transcriptómica é a sua falta de reprodutibilidade entre estudos. Presentemente pensamos que uma parte importante da referida variabilidade seja causada pela infidelidade da transcrição. A compreensão das regras de infidelidade da transcrição como descrito atrás permite agora o desenho de chips de micro-arranjos que, especificamente, medem a taxa da infidelidade da transcrição em qualquer mistura de cDNA obtida a partir de células normais ou doentes. A monitorização da taxa de infidelidade da transcrição ao nível do RNA permite o rastreio, num elevado número de amostras, da taxa relativa de infidelidade da transcrição que ocorre numa determinada célula. Isto proporciona informação essencial para determinar a gravidade da doença, define as estratégias terapêuticas e classifica as células doentes de acordo com o seu perfil de sensibilização farmacológica. O mesmo rastreio também permite testar a eficácia de novos fármacos, *e.g.*, na terapia de cancro.

Outras aplicações da infidelidade da transcrição

A infidelidade da transcrição é um mecanismo natural agora descoberto que adiciona diversificação ao bem estabelecido código genético. Observámos que a infidelidade da transcrição ocorre, ainda que em taxas baixas, mesmo em

tecidos normais. Propomos que a infidelidade da transcrição contribua para explicar a baixa abundância relativa de genes identificados no genoma e uma muito maior diversidade de proteínas presentes nas células vivas de fluidos biológicos e tecidos.

Propomos que a infidelidade da transcrição sirva uma função específica ao nível imunológico. Propomos ainda que o sistema imunitário se baseia em parte numa determinada taxa de infidelidade da transcrição que afecta preferencialmente genes específicos de forma a identificar o que é próprio. Assim, a anormalidade na taxa de infidelidade da transcrição que ocorre por razões ainda desconhecidas contribui para a patogénese de doenças auto-imunes e alérgicas. As diferenças inter-individuais na infidelidade da transcrição deverão ser também condicionadas pelos polimorfismos de genes do sistema imunitário, determinando portanto a taxa de infidelidade da transcrição nas células circulantes do sistema imunitário, *e.g.*, linfócitos T ou B, e deverão também contribuir para a avaliação da adequabilidade do dador e do aceitador dos enxertos na determinação da gravidade das doenças de enxerto versus hospedeiro.

Também postulamos que a infidelidade da transcrição ocorra numa taxa mais elevada durante o processo de envelhecimento normal. Isto poderá criar condições de redução progressiva do desempenho que afecta todas as enzimas em geral e, mais especificamente, as proteínas que estão

envolvidas na replicação celular, reparação de DNA e manutenção da taxa normal da fidelidade da transcrição. O conceito atrás descrito deverá portanto redireccionar a pesquisa para mecanismos de envelhecimento. Métodos bioquímicos, proteómicos, transcriptómicos e outros que permitem o rastreio de sequências de mRNA e de cDNA e fidelidade relativamente ao especificado pelo DNA deverá portanto ser usado para quantificar a taxa de infidelidade da transcrição que ocorre num determinado indivíduo. A avaliação de novas moléculas que reduzem a taxa de infidelidade da transcrição será pois susceptível do rastreio biológico e conduzirá ao desenvolvimento de fármacos redutores da velocidade de envelhecimento.

Outros aspectos e vantagens do invento serão descritos na secção experimental que se segue.

Secção experimental

Exemplo 1. Princípio da construção de uma biblioteca de cDNA típica e sequenciação (ver Fig. 1)

O primeiro passo na preparação de uma biblioteca de DNA complementar (cDNA) é isolar o mRNA maduro a partir do tipo de célula ou tecido com interesse. Devido à sua cauda poli(A), é fácil obter uma mistura de todo o mRNA celular através de hibridação com oligo dT complementar ligado covalentemente a uma matriz. O mRNA ligado é então eluido com um tampão de baixo teor de sal. A cauda poli(A)

do mRNA hibrida então com oligo dT na presença de uma transcripase reversa, uma enzima que sintetiza uma cadeia de DNA complementar a partir de uma matriz de mRNA. Isto dá origem a nucleótidos de cadeia dupla contendo a matriz de mRNA original e a sua sequência de DNA complementar. O DNA de cadeia simples é em seguida obtido através da remoção da cadeia de RNA por tratamento com bases ou pela acção da RNase H. Uma série de dG é então adicionada ao extremo 3' do DNA de cadeia simples pela acção de uma enzima designada transferase terminal, uma polimerase de DNA que não requer uma matriz mas que adiciona desoxioligonucleótido ao extremo 3' livre de cada cadeia de cDNA. O oligo dG híbrida com oligo dC, o qual actua como uma sequência iniciadora para sintetizar, através da polimerase de DNA, uma cadeia de DNA complementar da cadeia de cDNA original. Estas reacções produzem uma molécula de DNA de cadeia dupla completa correspondendo às moléculas de mRNA encontradas na preparação original. Cada uma destas moléculas de DNA de cadeia dupla é normalmente referida como cDNA, cada um contendo uma cadeia dupla oligo dC-oligo dG num extremo e uma região de cadeia dupla oligo dT-oligo dA no outro extremo. Este DNA é então protegido por metilação em locais de restrição. Pequenos elementos de ligação para restrição são então ligados a ambos os extremos. Estes são segmentos de DNA sintéticos de cadeia dupla que possuem o local de reconhecimento para uma enzima de restrição particular. A ligação é efectuada pela ligase de DNA do bacteriófago T4 que pode ligar moléculas de DNA de cadeia dupla "com extremos cerses". O DNA de cadeia dupla com extremos cerses

resultante, com um local de restrição em cada extremidade, é então tratado com enzima de restrição que cria um extremo coesivo. O passo final na construção das bibliotecas de cDNA é a ligação de cadeia dupla cortada com enzimas de restrição com um plasmídeo específico que é usado para transfectar uma bactéria. As bactérias recombinantes são então crescidas para produzir uma biblioteca de plasmídeos - na presença de antibióticos correspondendo a resistência a antibióticos específicos do plasmídeo. Cada um dos clones é portador de um cDNA derivado de uma única parte do mRNA. Cada um destes clones é então isolado e sequenciado usando métodos de sequenciação clássicos. Uma corrida típica de sequenciação começa no local de inserção e dá sequências de 400 a 800 pares de bases para cada clone. Esta sequência serve como matriz para o início da segunda corrida de sequenciação. Esta progressão para a frente leva a sequenciação progressiva do inserto completo do plasmídeo. Os resultados de sequenciação de numerosas cDNAs designados ESTs foram depositados em várias bases de dados públicas.

Exemplo 2. Anotação das bases de dados

As bases de dados EST possuem informação de sequências que correspondem à sequência de cDNA obtida a partir de bibliotecas de cDNA e portanto correspondem essencialmente à sequência de mRNA individual presente em qualquer altura no tecido que foi usado para produzir estas bibliotecas. A qualidade destas sequências foi questionada por vários motivos. Primeiro, como atrás descrito, o

processo de produção de bibliotecas de cDNA inicialmente baseou-se bastante na presença de uma cauda poli(A) no extremo 3' do mRNA eucariótico. Segundo, os mRNAs são moléculas muito frágeis que são facilmente digeridas por nucleases fortemente abundantes designadas RNases. Terceiro, enquanto se construíram e sequenciaram estas bibliotecas, foi prestada pouca atenção à qualidade do material original usado e ao seu armazenamento. Devido a isto, as sequências ESTs foram usadas para anotar informação genómica, *i.e.*, para determinar se um segmentos identificado e totalmente sequenciado do DNA genómico codifica qualquer mRNA específico. Neste contexto, as sequências ESTs foram úteis de forma a identificar as sequências genómicas codificadoras. No entanto, foi dada pouca atenção à informação gerada pela própria sequência EST. De facto, a sequência de DNA genómico é considerada muito mais fiável, com fortes argumentos técnicos que apoia esta posição. Especulamos que a diversidade incluída em sequências EST deverá conter informação biologicamente, analiticamente ou clinicamente relevante. De facto, as bases de ESTs foram produzidas por uma série de investigadores que usaram todos vários métodos: isto levou-nos a especular que cada enviesamento metodológico deva contribuir para um nível de ruído de fundo com um determinado número de erros. No entanto, caso existam diferenças nos erros devido à fonte de material usado para gerar a biblioteca, então a diferença na taxa de erros estará directamente relacionada com a fonte subjacente.

Para testar isto, analisámos a base de dados ***est_human*** disponível do sítio NCBI ftp. Seleccionámos estas bases de dados devido a estas sequências não estarem anotadas ou curadas por ferramentas humanas ou bioinformáticas.

Usámos um sistema de identificação de biblioteca para determinar se uma EST foi obtida a partir de um tecido canceroso ou um normal. Cada biblioteca foi designada ***normal*** ou ***cancerosa***. Fazendo corresponder o número de acesso de cada EST ao identificador da biblioteca respectiva, classificámos 2,6 milhões de ESTs como as obtidas a partir de tecidos cancerosos e 2,8 milhões de ESTs como as obtidas a partir de tecidos normais. Para obter os alinhamentos de EST, usámos o programa megaBLAST 2.2.13 disponível ao público (*Basic Local Alignment Search Tool, Zheng Zhang. A greedy algorithm for nucleotide sequence alignement search*, J Comput Biol, 2000) e especificado pelos seguintes parâmetros:

-d est_human : base de dados ara pesquisa contra (todos os ESTs humanos de dbEST),
-i sequences.fasta : formato de ficheiro FASTA contendo a sequência de referência correspondente aos genes abundantemente expressos (Fig 2),
-D 2 : resultado tradicional do "blast"
-q -2 : pontuação de -2 para um desemparelhamento e nucleótido,
-r 1 : pontuação de 1 para um desemparelhamento isolado de nucleótido,
-b 100000 : número máximo de sequências para as quais o

alinhamento é descrito,
-p 90 : percentagem de identidade mínima entre EST e a sequência de referência,
-S 3 : tem em consideração ESTs que correspondem exactamente à sequência de referência,
-W 16 : comprimento do melhor emparelhamento perfeito para começar com a extensão do alinhamento
-e 10 : Valor expectável (valor de E), número de alinhamentos que se espera encontrar apenas por acaso para uma determinada sequência e uma determinada base de dados, de acordo com o modelo estocástico de Karlin e Altschul (1990),
-F F : não filtração da sequência de referência.
-o NonFiltre_Test_90.out : ficheiro do relatório megaBLAST EEE
-v 0 : número máximo de sequências para mostrar numa linha de descrição,
-R T : (T para Verdadeiro) : resultado da informação do log no final de cada resultado do megaBLAST,
-I T : mostra GI's em linhas de definição ("deflines") [T/F]

A linha de comando para o "blast" é:
Megablast -d est_human -iC:\SeqRef\TPTI\sequences.fasta

-o C:\blast\NonFiltre_test_90.out -R T -D 2 -I T -q -2 -r 1 -v 0 -b 100000 -p 90.00 -S 3 -W 16 -e 10.0 -F F

Figura 2(a-q) mostra as 17 sequências usadas para

a análise; de forma a evitar a distorção do "blast", a cauda poli(A) das sequências de mRNA de referência foram sistematicamente removidas.

Fig 2r proporciona uma lista de genes que foram usados para testar a ocorrência de variações de sequências.

Exemplo 3. Identificação de variações de sequências entre ESTs de origem normal e cancerosa.

Foram seleccionados 17 genes com base na sua larga representação nas bases de dados. Cada sequência EST foi então alinhada contra a sua sequência de mRNA curada (RefSeq) do NCBI usando o MegaBLAST. Isto cria uma matriz em que qualquer base determinada é definida pelo número de ESTs que têm uma base igual nesta posição. Medimos então a proporção de ESTs que se desviam de RefSeq em qualquer outra posição. A comparação das sequências EST alinhadas de acordo com o tecido de origem levou-nos a identificar variações de sequências que ocorrem em cada posição de base no grupo cancro ou no grupo normal. A Figura 4 mostra uma representação gráfica destas variações que ocorrem nas séries de tecido normais e de cancro para 17 genes relacionados. O exame visual na série de cancro revelou que a variação das sequências ocorreu mais frequentemente na série de cancro e ainda que o fenómeno surgiu mais predominantemente em locais específicos do mRNA. A maioria das observações situa-se entre 400 e 500 bases após o começo da sequência de mRNA, portanto predominantemente na

parte 3' do gene. O número muito elevado de variações não poderá ser explicado por SNPs. No gráfico estão apresentados SNPs putativos (□) e SNPs biologicamente validados (○); estes SNPs foram obtidos a partir da base de dados NCBI (dbSNP, construída em 12 de Setembro de 2006) (Sherry, S.T., *et al.* (2001) Nucleic Acids Res 29, 308-11). Tanto SNPs putativos como biologicamente validados conducentes a variações de ESTs (n = 442) foram excluídos de análise posterior (assim locais de edição de RNA (falsos SNPs) foram também excluídos).

O passo seguinte desta análise foi testar a significância estatística das diferenças na variação de sequência que ocorrem entre ESTs de cancro e normais. Para cada posição, comparámos a proporção da base REfSeq com as das três outras bases entre grupos normais e de cancro usando um teste de proporção bilateral. Este teste foi sistematicamente aplicado desde que fossem preenchidas as seguintes condições: $n>70$ e $(n_{i.} * n_{.j})/n > 5$ $i = 1,2$; $j = 1,2$ (em que n = número de ESTs de cancro e normais para um gene, $n_{1.}$ = número de ESTs de cancro, $n_{2.}$ = número de ESTs normais, $n_{.1}$ = número de ESTs tendo a RefSeq, $n_{.2}$ = número de ESTs tendo uma substituição). Um teste estatístico é referido como sendo positivo no nível de patamar de 5% em que o valor de P correspondente é inferior a 0,05; neste caso, a hipótese nula (*i.e.* ambas as proporção são iguais) é rejeitada.

Ainda, foram considerados os dois testes de

proporção unilaterais que se seguem, de forma a precisar em qual a série a variabilidade era maior. O primeiro permitiu concluir que as variabilidades eram diferentes em ambos os grupos quando o teste estatístico era positivo, então neste caso mediu se a variabilidade era estatisticamente maior na série de cancro. Ao contrário, o segundo teste verificou a hipótese de a variabilidade ser significativamente mais elevada na série normal. Ambos os testes unilaterais foram realizados sempre que as mesmas condições dos testes bilaterais eram encontradas.

A Figura 3 mostra uma descrição de resultados típicos de "blast". Como se mostra nas Figuras 5a e 5b, para o gene representativo TPT1, o número de ESTs em qualquer posição no gene é semelhante em ambos os grupos de cancro e normal. A Figura 5c mostra que 489 de 830 testes de proporção possíveis preenchem os critérios atribuídos. Os testes de proporção que foram estatisticamente significativos ao nível de 5% estão apresentados na Figura ***5d***. Pode ser calculada uma estimativa de erro resultante de múltiplos testes, definido pelo Location Based Estimator (C. Dalmasso, P. Broet (2005) Journal de la Societé Française de Statistique, tome 146, n1-2, 2005). 26 testes de proporções positivos são devidos a substituições de bases que ocorrem no tecido normal (N>C) na Figura ***5d*** (n = 26; LBE = 33). Isto contrasta com a acumulação de testes de proporções estatisticamente significativa que ocorre devido às variações de sequências no grupo de cancro (C>N), (n = 145; LBE = 15). Uma análise semelhante foi realizada para

um segundo gene VIM (***Fig. 5e-5h***). Novamente o número de ESTs em qualquer posição determinada do gene é sobreponível nos grupos de cancro e normal. 752 de 1847 testes de proporção preenchem os critérios. Novamente, observámos uma grande diferença no número de testes de proporções positivos: n = 78 variações são devidas ao grupo normal (LBE = 50) e n = 269 variações são devidas ao grupo de cancro (LBE = 24). Repetimos a mesma análise em 17 genes que estão abundantemente presentes na base de dados ST. ***Fig. 5i*** é um resumo dos resultados estatísticos obtidos. De entre 17 genes testados, 15 apresentaram maiores variações de sequências em ESTs obtidos a partir de tecidos de cancro (***Fig. 5j***). Nesta fase, a análise estatística cobriu apenas 9 a 91% das posições de cada gene (uma média de 32%) devido a constrangimentos do teste estatístico.

A conclusão deste primeiro ciclo de análises é que as ESTs do mesmo gene eram diferentes quando comparadas de acordo com a fonte de tecido (normal versus cancro) de onde o mRNA foi extraído. Assumindo que a taxa de erros técnicos que dá a variação na sequência EST não é diferente de acordo com as origens normal ou cancerosa do tecido e considerando o facto de 15 dos 17 genes testados mostrarem variações devido à fonte de tecido, propomos que estas diferenças resultem directamente do estatuto - normal versus canceroso - das células a partir das quais as ESTs foram produzidas.

Experiências semelhantes foram realizadas para

estudar inserções e deleções. Os resultados foram obtidos após aplicação do filtro F3 (ver Exemplo 4). Uma condição mais restringente foi usada para evitar eventos não biológicos (erros de sequenciação, alinhamentos mal feitos com o Megablast …): um espaço ou uma inserção foi apenas considerado caso não existissem outras modificações numa janela de -10/+10. Para cada posição, comparámos a proporção de janelas com deleção (ou inserção) com outras janelas entre grupos normais e de cancro usando um teste de proporções bilateral como atrás descrito. ***Fig. 5k*** é um resumo dos resultados estatísticos obtidos. Dos 17 genes testados, 14 (ou 10 para inserção) apresentaram mais variações da sequência (deleções e inserções) em ESTs obtidas de tecidos de cancro (***Fig 5l***).

Exemplo 4: Verificação do excesso de variação na série de cancro C>N versus na série normal N>C

As bibliotecas derivadas de amostras de cancro ou normais foram processadas essencialmente da mesma forma. Assim, os erros aleatórios resultantes da construção das bibliotecas ou da sequenciação dos clones espera-se que ocorram com a mesma taxa em ambas as séries e não contribuam para as diferenças observadas entre os grupos de EST normal e de cancro. A análise matemática é consistente com esta interpretação (Brulliard M., e tal PNAS May 1, 2007 vol. 104 nº 18 7522-7527 - ver Fig. 7 do material suplementar).

Em seguida procurámos eliminar outras fontes de heterogeneidade de EST através da aplicação sequenciada de procedimentos de filtração com a base lógica que se segue. Os nossos requisitos iniciais foram que EST alinha com RefSeq com 100% de identidade em pelo menos 16 bases consecutivas e com ≥90% de identidade em pelo menos 50 bases. Como se mostra na Figura 6b, isto dá, quando se compara séries de cancro e normais para os 17 genes, 2281 e 725 diferenças estatisticamente significativas para C>N e N>C respectivamente (coluna F1) e distintas dos SNOs putativos ou biologicamente validados. O segundo filtro (F2) requer que esta EST alinhe com RefSeq continuamente em mais de 70% do seu comprimento. O terceiro filtro (F3) removeu ESTs com sequência mais estreitamente relacionada com paralógos e pseudogenes do que com RefSeq *bona fide*. O quarto filtro (F4) eliminou da análise as 50 primeiras e últimas bases de cada alinhamento de ESTs. Usámos este filtro para remover desemparelhamentos nas fronteiras 3' e 5' de EST que possam ser criados pelo programa MegaBLAST para optimizar mais os alinhamentos e/ou resultantes de acumulação de erros nas extremidades das ESTs alinhadas. Este último filtro (F5) normalizou os comprimentos das sequências ESTs para cada gene de forma a remover qualquer diferença de comprimento entre séries normais e de cancro. De facto observámos após aplicação dos primeiros 4 filtros, que a média do comprimento de ESTs no cancro era superior à de ESTs normais (640 ± 248 e 554 ± 229 bases, respectivamente). No entanto, observámos significativamente maior heterogeneidade de ESTs no cancro em 5 genes em que o

comprimento das ESTs não era diferentes entre as séries normal e de cancro (TPT1, VIM, HSPA8, LDHA, CALM2).

O número de testes estatisticamente significativos C>N manteve-se superior ao número de testes N>C, mas a taxa C>N decresceu de 3,15 (F1) para 2,05 (F5) (***Fig 6c***).

Para melhor confirmar que a heterogeneidade de ESTs no cancro não era devida à acumulação de erros no final das corridas de sequenciação, tivemos em consideração apenas a informação proporcionada pelas primeiras 50 bases e não mais de 450 bases de qualquer corrida de sequenciação. Após esta filtração drástica, testes C>N estatisticamente significativos foram 455 (LBE = 92) e testes N>C foram 292 (LBE = 119) para os 17 genes. Pode pois ser razoavelmente assumido que as variações das sequências causadoras de maior heterogeneidade de ESTs na série de cancro são um reflexo directo da heterogeneidade do mRNA nas células de cancro (***Fig 6d***).

Em seguida, verificámos independentemente que uma maior heterogeneidade de ESTs estatisticamente significativa persistia na série de cancro após remoção das ESTs produzidas pelas células de cancro em cultura. Após esta filtração, os testes C>N estatisticamente significativos eram 1009 (LBE = 117) e os testes N>C eram 445 (LBE = 193) para os 17 genes (***Fig. 6e***).

Exemplo 5. Decifração do código de ocorrência de

substituição de bases devido à infidelidade da transcrição em células de cancro

O nosso objectivo seguinte foi determinar se o fenómeno descrito da substituição de bases devido a infidelidade da transcrição é um fenómeno aleatório ou se segue regras específicas. Para atingir isto, focámo-nos em variações de ESTs em que C>N era estatisticamente significativo. Para evitar o enviesamento que possa ser introduzido pelos procedimentos de filtração, usámos todos os dados não filtrados. A primeira indicação que a infidelidade de transcrição não é um processo aleatório foi que as substituições de bases raramente ocorriam na região 5' dos genes testados. Para todos os genes testados, muito raramente observámos a ocorrência do fenómeno de substituição de bases nas primeiras 400-500 bases do mRNA maduro. A segunda observação que indicava que a infidelidade da transcrição não é aleatória reside na observação que existe uma diferença na composição de bases na matriz de DNA genómico observada quando da comparação de sequências a montante e a jusante dos eventos de substituição (heterogénea, H, n = 2281) com aquelas em que não foi detectado qualquer evento de substituição (não heterogéneo, NH, n = 12273). O critério para os locais NH foi variações na série de cancro inferiores a 0,5% e não estatisticamente diferentes das variações na série normal (**Fig 7a**). Nesta análise, referimo-nos à base que sofre substituição como b0, bases localizadas no extremo 5' do pmRNA são referidas como b-n e as bases situadas no extremo 3' como b+n. Por

questões de clareza, referimos nesta análise a composição de bases do pmRNA: isto corresponde ao DNA da cadeia não matriz (a cadeia não transcrita pela RNAP). Os dados mostram primeiro que nem todas as bases eram igualmente susceptíveis de variação: b0 = A (33%) ≈ T (32%) >> C (21%) >> G (14%). Ainda, as composições das 4 bases a montante e das 3 bases a jusante do local do evento eram estatisticamente diferentes (resultados da análise de Qui-quadrado) das dos locais sem variação significativa das ESTs (**Fig. 7a**, a composição básica é expressa em % e o sombreado cinzento mostra as bases enriquecidas; cinzento escuro mostra as bases raras). Especificamente, os locais em que ocorrem variações eram mais frequentemente precedidos e seguidos de A ≥ G > T ≈ C.

Assim, a ocorrência de heterogeneidade de ESTs de cancro não é aleatória, mas determinada primeiro pela natureza das bases que sofrem substituição e segundo pela natureza das bases que imediatamente precedem e se seguem ao evento.

Em seguida questionámos se a base substituinte era seleccionada ao acaso. É claro pela Figura 7b que não é o caso. Foi calculada diferença estatisticamente significativa nas proporções com a hipótese nula que a base substituinte fosse seleccionada aleatoriamente (teste de ajustamento das três bases substituintes para distribuição uniforme). A foi preferencialmente substituída por C ($p = 2,8 \times 10^{-125}$), T por G ($p = 5,7 \times 10^{-29}$) e G por A ($p = 2,2 \times$

10^{-32}). A substituição de C mostrou uma distribuição ainda mais aleatória, com uma ligeira escassez de T (p = 0,007).

Em seguida pensámos nas causas subjacentes à substituição preferencial de bases. Para se conseguir isto, distinguimos duas séries de eventos informativos e não informativos. Os eventos informativos eram situações em que a base substituída era diferente da base anterior (b-1) ou a base seguinte (b+1) (n = 1676) (***Fig 7c***). Os eventos não informativos eram situações não correspondentes a estes critérios. Quando os eventos informativos eram analisados, foram encontrados dois casos: a base substituída foi substituída por b-1 ou b+1 (79%) ou por uma outra base, diferente de b-1 e de b+1 (21%). Na primeira subsérie, a base substituinte era idêntica a b-1 (n = 799; ***Fig 7c*** painel B) ou b+1 (n = 530; ***Fig 7c*** painel C). Quando a base substituinte era b-1, então b0 = a (36%) > C (30%) >> T (21%) >> G (13%) (***Fig 7d*** painel A). A foi preferencialmente substituído por C (71% dos casos). Quando a base substituinte foi b+1, então b0 = T (47%) >> A (21%) > C (19%) >> G (13%) (**Fig 7a** painel B). T foi preferencialmente substituído por G (71%). É interessante que a importância estatística das bases envolventes foi também diferente (***Fig 7d*** painel A e ***7d*** painel B). Para as substituições b-1, o padrão de influência relativa da composição de bases foi b0 > b-1> b-2> b+1> b-3 > b+2. Para as substituições b+1, a influência relativa da base envolvente seguiu um padrão de b0> b+1 > b+2 > b-3 > b-2. 2) Na segunda subsérie de eventos informativos, a substituição de bases não correspondeu a b-1 ou a b+1 (n = 347; **Fig 7c** painel D). As

bases afectadas foram na seguinte ordem: A (47%) > T (29%) > C (14%) > G (10%). A foi mais frequentemente substituída por C (91% dos casos), T por C (50%) e A (42%), C por G (46%) e G por C (73%). Assim, quando a base substituinte não corresponde a b-1 ou a b+1, a base substituinte não é seleccionada ao acaso mas C está em grande excesso.

Em seguida considerámos a série de eventos não informativos, *i.e.*, as situações em que 1) b-1 = b+1 e em que 2) b-1 = b0 = b+1 (***Fig 7c***). Quando b-1 e b+1 eram idênticos mas diferentes de b0 (n = 339; ***Fig 7c*** painel E), as bases substituídas foram na seguinte ordem: T(34,8%) > G (23,6%) > G (23,6%) > C (21,2%) > A (20,4%) e seguiu o mesmo padrão de preferência que na Figura 7b: T→G, G→A, A→C. As substituições que ocorrem na base central da repetição de três bases idênticas (n = 266; ***Fig 7c*** painel F) foram observadas na seguinte ordem: A (46,2%) > T (36,9%) > G (10,5%) > C (6,4%). Neste caso, os eventos de substituição mais comuns foram A→C e T→C e A. As substituições raras de GGG foram mais frequentemente substituídas por GCG e CCC por CAC (***Fig 7c***).

Assim, quando as substituições ocorrem em três bases idênticas consecutivas e quando as substituições não correspondem a b-1 ou a b+1, então C é a base substituinte mais comum (***Fig 7c***). Quando a base substituinte corresponde a b-1, a substituição mais comum é A→C; quando a base substituinte corresponde a b+1, a substituição mais comum é T→G.

Pode portanto ser concluído que nem a base que sofre substituição nem a base substituinte são seleccionadas ao acaso. Ambos os fenómenos seguem padrões previsíveis definidos pela composição da base que sofre substituição e pelas bases localizadas a montante e a jusante deste evento.

Podem ser definidos algoritmos específicos para identificar com precisão a composição do motivo que determina a ocorrência da substituição de bases em qualquer sequência determinada. Em seguida separámos a série C>N em duas séries em que N é estável (sem desvio observado na mesma posição na série Normal) ou não. O desvio é considerado apenas se exceder um determinado patamar definido como a percentagem média do desvio na série Normal. A Figura 7e mostra que a assinatura -4/+3 é característica da série C>N em que N se desvia. O comprimento do contexto é menos importante na outra série. A Figura 7f mostra uma acumulação no intervalo 500-1000 que é mais importante na série C>N em que N se desvia do que na série C>N em que N é estável. Neste último caso, parece que surgem eventos mais precoces (intervalos de 0-500). As bases substituintes para os locais C>N eram semelhantes entre as duas séries (***Fig 7g***).

No caso da deleção ou da inserção, os resultados mostram que as bases omitidas (398 casos C>N) eram na seguinte ordem C (46,0%) > T (37,9%) >> A (9,3%) > G (6,8%) (***Fig 5m***) e que as bases inseridas (225 casos C>N) eram na

seguinte ordem: G (36,0%) > C (30,2%) > A (21,8%) > T (12,0%) (***Fig 5o***). Observámos que os eventos de deleção ou inserção muitas vezes ocorrem em bases idênticas conservadas. Um programa específico destinado a analisar tais eventos mostra efectivamente que 94,7% das deleções e 81% das inserções em segmentos de comprimentos variáveis. Os segmentos podem ser bipolares e simétricos (*e.g.*, AATAA) ou não (*e.g.* CCCGG) ou repetições de nucleótidos isolados (*e.g.*, AAA) (ver motivos na ***Fig 5m-5p***). No caso da deleção, as bases omitidas são idênticas a b-1 ou a b+1 (segmento) em 84% dos casos. A maioria dos motivos são dubletos ou tripletos de C>T>A≈G. Resultados semelhantes são observados no caso da inserção. De facto, os motivos de segmentos não parecem ser diferentes entre C>N e N>C. A especificidade de C>N ou de N>C viria de outra informação à volta dos segmentos. Análises adicionais em mais genes confirmarão estas observações.

Tendo em consideração os 17 genes, observámos heterogeneidade de EST na taxa de 10 por 100 bases (***Fig 5j*** e ***Fig 8a-c***). Esta taxa está em excesso relativamente a qualquer taxa descrita de mutação que ocorra no DNA genómico. Como referência, pode-se estimar que polimorfismos de nucleótidos isolados ocorrem aleatoriamente no genoma uma vez em cada 300 bases. A taxa de polimorfismos de nucleótidos isolados que afecta o DNA transcrito é muito mais inferior: é estimada que ocorra uma vez em cada 3000 bases. É claro que as mutações do DNA ocorrem mais frequentemente no cancro. No entanto, esforços profundos de

sequenciação de DNA do cancro da mama e do cólon que incluíram 14 dos 17 genes usados no nosso estudo levaram a uma taxa de mutação somática de 3,1 mutações por 106 bases (Sjoblom, T., *et al.* (2006) Science 314, 268-74). Assim, a infidelidade da transcrição que é descrita neste invento ocorre a uma taxa muito superior à das mutações que afectam o DNA. Mais importante, a maioria das substituições de bases no DNA têm consequências limitadas ao nível das proteínas devido a menos de 10% do DNA genómico ser transcrito para mRNA. Pelo contrário, a substituição de bases devido à infidelidade da transcrição muitas vezes tem consequências directas na função da proteína. De facto, 1179 de 2281 substituições aqui descritas (1548 CDS - 369 substituições silenciosas) conduziram a substituições de bases com impacto imediato na sequência primária de AA da proteína (***Fig 8d***). Mais importante, as substituições de bases significativas, que afectam o codão de paragem, foram observadas em 9 dos 17 genes testados. Antes do conceito de infidelidade da transcrição, não tinha sido no entanto proposto que as proteínas humanas possuíssem sequências codificadoras adicionais codificadas pelas sequências de RNA consideradas até então "regiões não traduzidas". Mostramos agora que a substituição de bases que ocorre em codões de paragem naturais devido à infidelidade da transcrição revela novas regiões codificadoras que codificam AA específicos. Esta nova região codificadora está em fase com a grelha de leitura aberta nativa. O codão de paragem natural é transformado numa região codificadora, O tripleto de bases seguinte é então lido como um AA e a tradução

continua com uma nova região codificadora até um novo codão de paragem ser atingido. Verificámos que é de facto o caso em 8 dos 17 genes que demonstraram infidelidade da transcrição. Os 8 continham codões de paragem alternativos em grelha com a correspondente RefSeq (GAPDH não possui um codão de paragem alternativo). Isto tem consequências imediatas porque em cada caso são criadas novas sequências codificadoras de 14, 7, 13, 15, 13, 4, 9, 55, AA em TPT1, RPS6, RPL7A, VIM, RPS4X, FTHI, FTL e TPOO respectivamente (***Fig 10***). Para além destes AA terem o potencial de criar motivos que serão grandemente aumentados em cancro, estes motivos resultarão ou não em novas funções das proteínas. Prever esta ocorrência conduz ao possível desenvolvimento de ferramentas úteis que poderão ser usadas no diagnóstico, terapêutica ou noutros objectivos. A previsão desta ocorrência conduz igualmente ao possível desenvolvimento de anticorpos específicos que reconhecerão sequências específicas do cancro no extremo terminal carboxilo da proteína. Nenhum método analítico é actualmente capaz de sequenciar directamente a proteína no extremo terminal carboxilo. É, no entanto, possível clivar as proteínas enzimaticamente e sequenciar os produtos de clivagem a partir do seu extremo NH_2. É igualmente possível analisar o teor de AA dos péptidos gerados por proteólise usando espectrometria de massa. Mostramos ainda que estes codões de paragem alternativos são igualmente afectados pela infidelidade da transcrição (7/9 genes possuem o segundo codão de paragem alternativo afectado). O mesmo fenómeno atrás descrito pode expandir mais a leitura para uma nova série de sequências.

A anotação de todas as sequências proteicas usando o nosso método revelará várias sequências de mRNA codificadoras insuspeitas que serão mais ou menos eficazmente transcritas, dependendo da utilização de codões assim como da capacidade da maquinaria da tradução para traduzir ou não correctamente a substituição da base. De facto, as substituições de bases podem levar a alterações na estrutura do mRNA, o que pode modificar a velocidade de leitura do ribossoma. No entanto, assumimos que as substituições de bases não envolvem alterações da estrutura do RNA.Com base na ocorrência de alterações do codão de paragem, estimamos nos genes afectados que até 4% do mRNA em tecidos de cancro possuam estas regiões codificadoras adicionais.

Um programa específico baseado em vários filtros pode ser usado para anotar todas as sequências proteicas relativamente à presença de um putativo péptido pós-paragem (Post Stop Peptide, PSP). Após aquisição das sequências de ácidos nucleicos correspondentes às proteínas estudadas, o programa pesquisa a presença ou não de uma sequência nucleotídica em fase após o codão de paragem canónico, com um outro codão de paragem em fase (a possibilidade de ultrapassar um ou mais codões de paragem no caso de infidelidade da transcrição afectando estes codões de paragem alternativos pode ser levado em consideração). Pode ser fixado um comprimento mínimo (*e.g.*, apenas sequências codificadoras de mais de 12 aminoácidos, mas eventualmente, o comprimento do péptido poderá ser menor) de acordo com o

critério mínimo para um padrão antigénico potencial. PSPs antigénicos são então guardados (***Fig 18***). Um passo adicional é usado para validar os candidatos com a anotação por uma máquina de aprendizagem por treino do codão de paragem canónico: de facto pode ser determinada a probabilidade de uma ou mais bases do codão de paragem poder ser substituída por infidelidade da transcrição. É igualmente possível analisar o teor de AA de péptidos gerados por proteólise usando espectrometria de massa.

Também demonstrámos que para além de criar condições que permitam a tradução de novas sequências proteicas, a infidelidade da transcrição pode introduzir codões de paragem prematuros no mRNA. Vinte e quatro codões de paragem novos que ocorrem dentro da grelha de leitura aberta canónica foram identificados dentro de 13 de 17 genes. Isto indicou que a infidelidade da transcrição pode originar a produção de proteínas mais curtas que não possuem domínios específicos. Estas proteínas truncadas devem resultar num aumento ou perda de função. A estrutura tridimensional da proteína é provavelmente afectada e criará novas entidades que poderão ser reconhecidas pelo sistema imunitário.

Finalmente, a fidelidade da transcrição nos 17 genes testados revelou que 50% de todas as substituições C>N identificadas conduzem a alterações de AA. 17% correspondem à substituição de um AA por outro da mesma família e 33% corresponde a substituições de AA de uma classe de AA diferente. Assim, a infidelidade da transcrição é capaz de

gerar proteínas com novas sequências de AA com funções ou actividade potencialmente modificadas. A previsão da infidelidade da transcrição usando as regras atrás descritas permitirá a previsão lógica de alterações no comportamento de proteínas e o resultado de experiências de proteómica.

Exemplo 6. Validações biológicas

As validações biológicas são realizadas em dois níveis: mRNA e proteínas.

Primeiro, as substituições de mRNA dos 17 genes de interesse serão detectadas em tecidos cancerosos humanos. Usámos DHPLC (Cromatografia Líquida de Alta Resolução Desnaturante), a qual é um método cromatográfico em larga escala para detectar mutações de sequências. O princípio da experiência está descrito na Figura 11a-c. Primeiro, desenvolvemos um método para testar o patamar de DHPLC Transgenómico de forma a estimar a percentagem de DNA mutado na amostra que é suficiente para permitir a formação e detecção de heteroduplexes. De facto, usámos fragmentos de PCR de 300 pb com 1 e 3 substituições de bases. Diferentes proporções destes fragmentos foram preparadas: 0%, 2,5%, 5%, 7,5%, 10%, 20% ou 50%. A DHPLC permitiu-nos distinguir DNA [normal] e [normal mais mutado] desde que a amostra possuísse 2,5 a 5% de DNA mutado (***Fig 11d***). Estes resultados indicam que fomos capazes de distinguir mRNA de tecidos normais e cancerosos para os genes de interesse, o mRNA extraído de tecidos normais e cancerosos adjacentes

(Biochain Inc) foi usado para testar três genes: GAPDH, ENO1 e TMSB4X. Como DHPLC funciona com DNA, as amostras de RNA são convertidas em cDNA usando transcriptase reversa. Escolhemos regiões que apresentam mais substituições significantes em ESTs provenientes de tecidos de cancro do que em ESTs provenientes de tecidos normais. As sequências iniciadoras usadas para amplificação estão apresentadas na Figura 12.

Numa primeira série de experiências, vários cDNA de tecido canceroso e de tecido normal adjacente (fígado, rim, mama e cólon) foram amplificados por PCR com as referidas sequências iniciadoras e injectados no sistema de DHPLC.

A temperatura da estufa foi seleccionada com o programa Navigator (Transgenomic). Os perfis foram obtidos para os genes atrás descritos e uma experiência representativa está apresentada na figura 15 para os genes ENO1, o GAPDH e TMSB4X. Como se mostra na figura 15a a 15c, os perfis de cancro são claramente diferentes dos perfis normais para os genes GAPDH e ENO1. Não foram observadas diferenças para o gene TMSB4X conforme esperávamos (muito menos locais de infidelidades da transcrição). As injecções do mesmo produto de PCR e de 2 outros produtos de PCR foram feitas em triplicado (figura 15b e 15c) e os perfis são muito reprodutíveis na mesma experiência. No entanto, a natureza da diferença não pode ser deduzida desta experiência.

Consequentemente, a continuação desta validação biológica baseada no mRNA foi efectuada. De facto, a sequenciação dos produtos de PCR obtidos a partir de tecidos cancerosos pode permitir detectar com precisão as mutações mais abundantes.

O método de sequenciação clássica de Sanger do mRNA amplificado por PCR após transcrição reversa não detecta variantes da sequência que ocorram em taxas inferiores a 15-30% numa posição específica (Fig 16). De facto, bases mutadas obtidas através da sequenciação de misturas de produtos de amplificação de sequências conhecidas mutadas e não mutadas são conseguidas para misturas de 50-50% e picos mais pequenos são detectados a partir de 15%. A piro-sequenciação e o PCR em emulsão são métodos mais sensíveis que permitem a detecção de heterogeneidade de cDNA, sendo portanto possível a análise de produtos de RT-PCR obtidos a partir de células de cancro e normais (Thomas, R.K. *et al.*, (2006) Nat Med 12, 852-5).

Um gene afectado pela infidelidade da transcrição é clonado na sua totalidade ou como um fragmento compreendendo ou não o seu local de infidelidade da transcrição. A construção com e sem codões de paragem canónicos é ligada em grelha com o gene codificador da resistência à neomicina. As células de cancro e normais são transfectadas com esta construção quimérica, primeiro transitoriamente e depois estavelmente. Prevemos que a infidelidade da

transcrição conduza a uma alteração no codão de paragem canónico, permitindo assim a tradução do gene de neomicina e criando a célula de resistência à neomicina. Previmos que as células de cancro que sejam resistentes à neomicina crescerão mais rapidamente e que a forma destas células diferirá significativamente da das células normais. Ainda, prevemos que estas células são mais invasivas e podem ser comparadas com uma fase mais tardia da progressão do cancro. Assim, esta técnica pode ser usada para determinar a fase cancerosa das diferentes células de um doente individual. A prova final que a resistência à neomicina ocorre como resultado da infidelidade da transcrição será obtida por sequenciação da construção inserida amplificada a partir de DNA genómico e mostrando que a informação genómica permanece inalterada. Também verificamos que o mRNA da célula resistente à neomicina possui um codão de paragem mutado devido à infidelidade da transcrição. Um técnica que permite a detecção de locais de infidelidade da transcrição é a construção de bibliotecas de cDNA que consiste na clonagem de cDNA obtido por transcrição reversa a partir de RNA extraído de tecidos de doentes cancerosos e normais. Cada cDNA amplificado ou não a partir de um gene específico é clonado em plasmídeos separados que são usados para transformar *Escherichia coli*. Diferentes colónias de *E. coli* são picadas e os plasmídeos são sequenciados. O número de clones que é necessário sequenciar depende da percentagem de substituições no fragmento de cDNA clonado. Uma análise estatística dar-nos-á a variação da sequência nos locais de infidelidade da transcrição. Esta técnica

pode ser melhorada. De facto, após transcrição reversa e amplificação com sequências iniciadoras específicas, o cDNA pode ser clonado num plasmídeo que é portador de um gene repórter, *e.g.*, o gene *lacZ*. O cDNA e o gene repórter são clonados em fase. Quando uma substituição surge na sequência do codão de paragem do mRNA, o gene repórter é expresso. Quando as células de *E. coli* foram transformadas com esta construção, as colónias portadoras do plasmídeo em que o cDNA está mutado no codão de paragem são seleccionadas com a expressão do gene repórter. Após o que o plasmídeo pode ser sequenciado de forma a verificar se a substituição do codão de paragem está presente.

Uma outra técnica baseada em PCR em tempo real com sequências iniciadoras específicas é usada para detectar os locais de infidelidade da transcrição. Estas sequências iniciadoras são desenhadas para corresponderem ao cDNA com variações na sequência baseadas em análise estatística. Uma segunda sequência iniciadora foi desenhada para corresponder ao cDNA sem a variação de sequência (referência). O número de variações na sequência iniciadora é muito importante e determinámos numa experiência com variação de sequências conhecida que 2 mutações fora da posição que é estudada são necessárias para se obter um sinal de fluorescência específico. Quando a sequência iniciadora é complementar do cDNA sem uma substituição da sequência (referência), o sinal de fluorescência é detectado num número específico de ciclos de amplificação. O mesmo cDNA amplificado com a sequência iniciadora que é

portador da substituição leva a um sinal fluorescente que surge com um número superior de ciclos de amplificação. A diferença entre os números de ciclos de amplificação com as 2 sequências iniciadoras é uma estimativa directa da presença de um cDNA com uma substituição da sequência. Ainda, verificámos que o método é suficientemente sensível para detectar 1% de sequências mutadas numa mistura de sequências conhecidas.

Numa experiência típica, o RNA extraído de tecidos cancerosos e normais é sujeito a transcrição reversa e amplificado com cada uma das sequências iniciadoras específicas. A diferença entre os ciclos de amplificação com ambas as sequências iniciadoras é então medida.

Finalmente, pudemos focar-nos em novas proteínas induzidas quando o codão de paragem natural é afectado. Mostrámos que o codão de paragem é significativamente afectado para 9 dos nossos 17 genes de interesse. Isto conduz a populações específicas distintas de proteínas com uma nova sequência no extremo terminal carboxilo. Estimámos que os tecidos cancerosos para os genes afectados possuíssem 4% de proteínas que são maiores que os genes normais.

Face a esta hipótese, analisámos 60 proteínas do plasma abundantes (***Fig 13***) e pesquisámos possíveis PSPs. Encontrámos 22 genes para os quais um PSP era suficientemente longo e antigénico (***Fig 18***). Testámos então

sequências proteicas putativas mais longas induzidas (***Fig 18 b-c***). Pesquisámos novos péptidos putativos no plasma obtido de indivíduos normais e cancerosos (***Fig 18***). Seleccionámos 3 das 22 proteínas com interesse: APOAI, APOAII e APOCII. Com base na análise atrás referida, esperávamos encontrar proteínas mais longas no plasma do doente com cancro (13, 16 e 17 AA mais longas). Prevê-se que estas sequências peptídicas resultantes da infidelidade da transcrição sejam imunogénicas e os anticorpos dirigidos contra estas novas sequências representem ligandos específicos para medir a infidelidade da transcrição (**Fig 14**). Uma vez que a análise de Kyte-Doolittle indica que estas sequências não são hidrofóbicas, esperamos que estas três novas proteínas sejam secretadas para a circulação (ver abaixo).

Identificação do péptido pós-paragem (PSP) de ApoAII e ApoCII devido à substituição no codão de paragem canónico

PSPs que resultam das substituições de bases em codões de paragem canónicos podem ser identificados e caracterizados da seguinte forma. São preparados anticorpos policlonais de coelho que reconhecem uma porção imunogénica do PSP em questão. Estes anticorpos anti-péptido podem ser testados por transferência de pontos usando o péptido purificado para verificar que de facto reconhecem o PSP. As transferências Western podem ser então realizadas em amostras de plasma obtidas de doentes com cancro usando os

anticorpos dirigidos contra o PSP. A Figura 17a (painel direito) mostra que o anticorpo anti-PSP ApoAII reconhece uma banda nas transferências Western nas amostras de plasma obtidas a partir de doentes com cancro da próstata, a qual não é observada quando se usa soro pré-imune de coelho como controlo negativo (Figura 17a, painel esquerdo). A banda PSP APOAII foi igualmente observada em doentes com cancro na fase metástica (Figura 17b, painel direito). Esta banda possui uma massa molecular ligeiramente superior (11,4 kDa) comparativamente com a da forma de monómero nativo de APOAII (Figura 17a, painel central, 17b, painel esquerdo). Esta massa molecular corresponde à prevista com base na sequência peptídica adicional. Podem ser igualmente efectuados géis bidimensionais de forma a caracterizar melhor esta banda.

As experiências de cromatografia de afinidade podem ser realizadas para isolar a forma PSP de ApoAII usando o anticorpo anti-PSP (***Fig 17e***). O anticorpo anti-PSP é imobilizado nas esferas da matriz e a coluna é incubada na presença de plasma ou de HDL sem lípidos, depois lavada sequencialmente para remover proteínas não especificamente ligadas e finalmente eluída com detergente ou reagentes caotrópicos. A fracção eluida é analisada por transferência Western usando os anticorpos anti-PSP e anti-ApoAII comerciais. Duas bandas de 9 kDa e 11,4 kDa são reconhecidas pelo anticorpo comercial anti-apoAII enquanto apenas a banda de 11,4 kDa é reconhecida pelo anticorpo anti-PSP. Assim, a banda de 9 kDa corresponde à forma de

ApoAII nativa e a de 11,4 kDa corresponde à forma PSP de ApoAII. A presença de ApoAII nativa na fracção eluida sugere que a proteína ApoAII possa formar dímeros com a proteína ApoAII nativa em condições não redutoras. De facto, ApoAII existe normalmente no plasma como um dímero de 2 monómeros ligados por uma ponte dissulfeto.

ApoAII está situada principalmente na fracção de HDL do plasma, a qual pode ser isolada por ultracentrifugação sequencial. A transferência Western mostra que a ApoAII PSP não é detectada na fracção d<1,07 g/ml contendo VLDL e LDL, mas antes na fracção d>1,07 g/ml contendo HDL e proteínas do plasma (***Fig 17c***). Após outros passos de ultracentrifugação para purificar e lavar HDL (d 1,07-1,21 g/ml), a transferência Western revela que a ApoAII PSP permanece associada à fracção HDL em vez da fracção d>1,21 g/ml, também referido como soro deficiente em lipoproteína (LPDS). Uma quantidade correspondente de plasma está apresentada na transferência Western para demonstrar que isto não é devido a uma diluição da fracção d>1,21 g/ml durante os passos de purificação. A separação das lipoproteínas do plasma por filtração em gel usando Superose 6B (Amersham, gE Healthcare) também demonstra que a ApoAII PSP eluiu de forma semelhante a ApoAII associada a HDL.

A associação de PSP ApoAII com HDL permitiu a purificação da forma PSP de ApoAII de forma semelhante à de ApoAII. HDL é primeiro delipidada para remover todos os lípidos. Os sedimentos de proteínas sem lípidos resultantes

foram então ressuspenso num tampão Tris 10 mM contendo ureia ou guanidina na presença ou ausência de agente redutor e aplicados numa coluna de filtração em gel (por exemplo, Superdex 200, Amersham-GE Healthcare). A transferência Western (Figura 17d) mostra que a forma PSP de ApoAII ainda está presente em HDL delipidado. Mais purificação foi conseguida por electroforese preparativa (exemplo: DE52). A forma PSP de ApoAII foi seguida por transferência Western. A forma PSP purificada de ApoAII foi então clivada enzimaticamente (tripsina) e os péptidos resultantes foram analisados em MS-MS para sequenciação completa dos AA. Os resultados mostram que o codão de paragem canónico é substituído preferencialmente por Arginina com substituição (U para C) ou (U para A) seguido de Serina, Valina, Ácido Glutâmico, Treonina, Isoleucina, Valina, Fenilalanina, Ácido Glutâmico, Prolina, ácido Glutâmico, Leucina, Alanina, Serina, Arginina. Esta é a sequência exacta de aminoácidos prevista para ocorrer após ultrapassagem do codão de paragem canónico de ApoAII. Assim, o codão de paragem canónico UGA de ApoAII foi substituído para AGA ou CGA conduzindo a Arginina. Uma hipótese ainda não explorada é que UGA seja convertido em GGA originando Glicina. Mas a detecção desta variante por espectrometria de massa está presentemente para além dos limites tecnológicos.

Um outro exemplo está ilustrado pelo PSP de ApoCII. A Figura 17f mostra uma experiência semelhante à da Figura 17a, com excepção dos Westerns serem realizados com anticorpos anti-ApoCII comerciais ou anti-PSP ApoCII. Um

procedimento semelhante ao do PSP de ApoAII pode ser seguido para ApoCII. No entanto, ApoCII é menos abundante no plasma. Uma quantidade muito maior de plasma que contém PSP ApoCII é portanto necessária de forma a obter-se quantidades suficientes para análise em MS-MS.

Conclusões

Descrevemos aqui um novo mecanismo conduzindo a substituições de bases que ocorrem principalmente no extremo 3' da sequência codificadora e na região não traduzida do mRNA. Estas substituições de bases conduzem a alterações nas sequências de AA da proteína devido a alterações da identidade de AA, introdução prematura de codões de paragem, assim como a modificação de codões de paragem naturais resultando na introdução de novas regiões codificadoras. Este fenómeno de infidelidade da transcrição poderá também afectar o mundo do ncRNA e perturbar a regulação mediada por estes RNAs. Ocorre na maioria dos genes a uma taxa que excede qualquer fenómeno descrito condutor à mutação de DNA. A infidelidade da transcrição é grandemente aumentada nas células de cancro. A infidelidade da transcrição proporciona um novo paradigma para compreender a patologia do cancro, gravidade da doença e progressão da doença. Tem implicações importantes não só para o desenho de novas experiências de transcriptómica e de proteómica, mas também para o desenvolvimento de ferramentas específicas de diagnóstico e de terapêutica.

LISTA DE SEQUÊNCIAS

<110> TRANSMEDI SA

<120> Infidelidade da transcrição, detecção e suas utilizações

<130> B0478WO

<160> 32

<170> PatentIn versão 3.3

<210> 1
<211> 13
<212> PRT
<213> Homo sapiens

<400> 1

Gln Met Trp Gln Leu Phe Trp Ile Tyr His Leu Ser Ser
1 5 10

<210> 2
<211> 14
<212> PRT
<213> Homo sapiens

<400> 2

Lys Leu His Thr Leu Ser Ala Ala Ile Tyr Tyr Gln Gln Glu
1 5 10

<210> 3
<211> 6
<212> PRT
<213> Homo sapiens

<400> 3

Asp Phe Leu Ser Asn Lys
1 5

<210> 4
<211> 12
<212> PRT
<213> Homo sapiens

<400> 4

Met Tyr Thr Val Glu Phe Ser Val His Lys Asn Asn
1 5 10

<210> 5
<211> 12
<212> PRT

<213> Homo sapiens

<400> 5
Asn Gly Ser Leu Gly Asp Met Ser Asp Leu Cys Thr
1 5 10

<210> 6
<211> 3
<212> PRT
<213> Homo sapiens

<400> 6
Ala Ser Gly
1

<210> 7
<211> 8
<212> PRT
<213> Homo sapiens

<400> 7
Glu Pro Ser Glu Pro Ser Asp Phe
1 5

<210> 8
<211> 54
<212> PRT
<213> Homo sapiens

<400> 8
Ala Pro Ser Ile Phe Pro Thr Leu Pro Ala Lys Pro Gly Thr Lys Gln
1 5 10 15

Pro Arg Ser Pro Val Thr Ala Leu Ser Leu His Met Leu Leu Met Val
20 25 30

Ser Ser Ala Pro Ser Cys Gly Leu Ile Gln Thr Val Ser Ser Phe Thr
35 40 45

Val Tyr Ile Phe Thr Leu
50

<210> 9
<211> 69
<212> PRT
<213> Homo sapiens

<400> 9
Ala Arg His Gly Arg Asp Glu Glu Val Trp His Arg Lys His Ser His
1 5 10 15

His Phe Val Gln Ala Trp Ala Trp Val Gly Gly Leu Val Cys Trp Pro
20 25 30

Arg Lys Cys His Met Arg Ser Thr Leu Ile Ser Ser Leu Asp Ser Leu
35 40 45

Leu Pro Val Ile Pro His Arg Thr Glu Ala Glu Trp Val Val Val Met
50 55 60

Phe Asp Arg Arg His
65

<210> 10
<211> 26
<212> PRT
<213> Homo sapiens

<400> 10
Arg Ser Lys Ala Tyr Ser Ser Val Phe Leu Phe Arg Trp Cys Lys Ala
1 5 10 15

Asn Thr Leu Ser Lys Lys His Lys Phe Leu
20 25

<210> 11
<211> 14
<212> PRT
<213> Homo sapiens

<400> 11

Gly Leu Asp Ser Thr Arg Ala Leu Glu Asn Glu Met Thr Val
1 5 10

<210> 12
<211> 12
<212> PRT
<213> Homo sapiens

<400> 12
Gly Ala Arg Arg Arg Pro Pro Ser Arg Cys Ser Glu
1 5 10

<210> 13
<211> 20
<212> PRT
<213> Homo sapiens

<400> 13

Ser Val Gln Thr Ile Val Phe Gln Pro Gln Leu Ala Ser Arg Thr Pro
1 5 10 15

Thr Gly Gln Ser
20

<210> 14
<211> 16
<212> PRT
<213> Homo sapiens

<400> 14
Ile Val Phe Gln Pro Gln Leu Ala Ser Arg Thr Pro Thr Gly Gln Ser
1 5 10 15

<210> 15
<211> 22
<212> PRT
<213> Homo sapiens

<400> 15
Gln Pro Asp Pro Pro Ser Val Asp Lys Gly Arg Val Pro Tyr Ser Pro
1 5 10 15

Asp Pro Pro Gly Ser Asp
20

<210> 16
<211> 15
<212> PRT
<213> Homo sapiens

<400> 16
Asp Pro Pro Ser Val Asp Lys Gly Arg Val Pro Tyr Ser Pro Asp
1 5 10 15

<210> 17
<211> 54
<212> PRT
<213> Homo sapiens

<400> 17
Asp Leu Asn Thr Pro Ser Pro Pro Ala Tyr Pro Ser Cys Glu Leu Leu
1 5 10 15

Gly Ser Cys Asn Leu Gln Gly Cys Pro Cys Arg Leu Leu Lys Arg Asp
20 25 30

Ser Ile Leu Ser Ala Leu Leu Pro His Leu Met Pro Gly Pro Pro Pro

35 40 45

Gly Met Leu Ala Ser Gln
50

<210> 18
<211> 36
<212> PRT
<213> Homo sapiens

<400> 18
Pro Gly Ser Thr Gly Arg Leu His Pro Leu His Val Thr Ser Ala Ser
1 5 10 15

Leu Ser Pro Thr Pro Pro Pro Pro His Lys Asp Lys Pro Ile Asn His
20 25 30

Asp Lys Gly Ser
35

<210> 19
<211> 37
<212> PRT
<213> Homo sapiens

<400> 19
Thr Pro Lys Pro Ala Ala Met Arg Pro His Ala Thr Pro Cys Leu Leu
1 5 10 15

Pro Pro Arg Ser Leu Gln Arg Glu Thr Leu Ser Pro Pro Gln Pro Ser
20 25 30

Ser Trp Gly Gly Pro
35

<210> 20
<211> 13
<212> PRT
<213> Homo sapiens

<400> 20
Glu Ala Arg Val Gly Gly Asn Val Gly Ser Gln Thr Gln
1 5 10

<210> 21
<211> 29
<212> PRT
<213> Homo sapiens

<400> 21

Pro Ser Val Leu His Thr Ala Arg Gly Pro Arg Met Pro Arg Pro Pro
1 5 10 15

Leu Ala Pro Ala Gly Arg Glu Pro Asp His Leu Pro Cys
20 25

<210> 22
<211> 72
<212> PRT
<213> Homo sapiens

<400> 22
Asp Val Asp Val Ala Phe Ala Pro Thr Gly Ala Ser Glu Ser Ser Ser
1 5 10 15

Pro Gln Asp Glu Leu Gln Pro Pro Arg Glu Ser Ser Ala Arg His Gln
20 25 30

Val Thr Arg Pro Gln Pro Pro Gly Pro Gln Leu Arg Pro Ala Ser Pro
35 40 45

Arg Ser Gly Ser Cys Thr Leu Thr Leu Asp Ser Ala Ala His Gly Lys
50 55 60

Asn Arg Ile Ala Pro Ala Cys Asn
65 70

<210> 23
<211> 14
<212> PRT
<213> Homo sapiens

<400> 23
Asn Val Ile Pro Leu Lys Arg Lys Met Asn Asn Thr Leu Asn
1 5 10

<210> 24
<211> 39
<212> PRT
<213> Homo sapiens

<400> 24
Thr Pro Ala Ala Arg Leu Met Trp Ser Ser Asn Met Pro Tyr Phe Ala
1 5 10 15

Gln Lys Thr Ala Lys Asp Met Thr Ser Ser Trp Leu Gln Pro Arg Phe
20 25 30

Ile Phe Leu Phe Val Val Asn
35

<210> 25
<211> 53
<212> PRT
<213> Homo sapiens

<400> 25

Gly Trp Gly Val Phe Leu Leu Asn Pro Met Ala Gly Gly His Ala Pro
1 5 10 15

Thr Ile Ile Ser Trp Glu Glu Arg Gln Ser Trp Glu Ile Asp Gly Ser
20 25 30

His Ser Ser Leu Leu Ser Leu Leu Cys Leu Trp Ala Thr Leu Pro Thr
35 40 45

Pro Leu Leu Ser Gln
50

<210> 26
<211> 27
<212> PRT
<213> Homo sapiens

<400> 26
Thr Gly Pro Thr His His Ser Pro Ser Pro Ser Ile Ser Thr Trp Cys
1 5 10 15

Leu Val Pro Val His Ser Val Asn Lys Lys Pro
20 25

<210> 27
<211> 25
<212> PRT
<213> Homo sapiens

<400> 27
Trp Thr Pro Glu Pro Leu Leu Gln Pro Leu Ser His Pro Leu Pro Pro
1 5 10 15

Ala His Pro Leu Gly Gln Gln Arg Leu
20 25

<210> 28
<211> 20
<212> PRT

<213> Homo sapiens

<400> 28
Leu Asp Gly Arg Gln Ser Asp Ala Leu Thr His Leu Glu Ala Gly Thr
1 5 10 15

Trp Val Gly Ile
20

<210> 29
<211> 71
<212> PRT
<213> Homo sapiens

<400> 29
Ser Leu Pro Ser Ser Ser Gly Ala Leu Ser Lys Glu Leu Gly Met Gln
1 5 10 15

Ala Gly Cys Leu Gly Leu Trp Ala Gln Pro Gly Pro Cys Ala Pro Ser
20 25 30

Gly His Gly Met Cys Gly Pro Val Cys Leu Ser Leu Glu Gly Asp Ser
35 40 45

Asp Ser Leu Cys Ser Ser His Met His Arg Gly Pro Trp Thr Leu Gln
50 55 60

Ser Gly Gly Ser Trp Ala Ser
65 70

<210> 30
<211> 48
<212> PRT
<213> Homo sapiens

<400> 30
Asn Leu Arg Gly Arg Ala Ala Thr Lys Val Lys Met Gly Thr Gln Met
1 5 10 15

Ile His Glu Phe Ala Leu Val Ser Leu Ala Gln Val Val Cys Ala Asn
20 25 30

His Val Cys Leu His Ser Ser Val Leu Pro Cys Val Leu Asn Lys Lys
35 40 45

<210> 31
<211> 100
<212> PRT

<213> Homo sapiens

<400> 31

Gly	Pro	Ala	Pro	Pro	Arg	Pro	Ala	Pro	Ala	Gly	Pro	Ala	Pro	Pro	Arg
1				5				10					15		

Pro	Ala	Pro	Ala	Ala	Leu	Pro	Met	Gly	Ala	Val	Phe	Lys	Asp	Thr	Arg
			20					25					30		

Ala	Pro	Ser	Pro	Pro	Gly	Ala	Pro	Leu	Lys	Met	Glu	Arg	Gly	Leu	Arg
		35					40					45			

Ile	Ser	Val	Ser	Leu	Gly	Ala	Cys	Leu	Gly	Ser	Pro	Ser	Leu	Thr	Phe
	50					55					60				

Pro	His	Ser	His	Ser	Leu	Ser	Leu	Pro	Leu	Cys	Leu	Leu	Leu	Pro	Val
65					70					75					80

Cys	Thr	Ile	Pro	Leu	Pro	Gly	Ile	Lys	Ala	Gln	Gly	Thr	Ser	Gly	Glu
				85					90					95	

His	Tyr	Cys	Ser
			100

<210> 32
<211> 16
<212> PRT
<213> Homo sapiens

<400> 32

Gly	Thr	Ser	Pro	Pro	Val	Asp	Leu	Lys	Asp	Glu	Gly	Trp	Asp	Phe	Met
1				5					10					15	

Lisboa, 12 de novembro de 2015

REIVINDICAÇÕES

1. Uso de um péptido sintético de 100 ou menos aminoácidos de comprimento, compreendendo uma sequência seleccionada entre SEQ ID NOS: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 para detecção ou monitorização de distúrbios de proliferação celular.

2. O uso de um péptido da reivindicação 1 para detecção ou monitorização de distúrbios da proliferação celular, em que o péptido é imunogénico.

3. Uma composição de vacina compreendendo um péptido da reivindicação 1 e facultativamente um veículo, excipiente e/ou adjuvante adequado.

4. Uma molécula de ácido nucleico sintética codificadora de um péptido de 100 aminoácidos ou menos compreendendo uma sequência seleccionada entre SEQ ID NOS: 1 a 5, 7 a 13, 15 a 18 e 20 a 32 para detecção ou monitorização de distúrbios de proliferação celular.

5. Um método de produção de um anticorpo, compreendendo o método a imunização de um mamífero não humano com um péptido da reivindicação 1 e recuperação dos anticorpos que se ligam ao referido péptido ou células produtoras de anticorpos correspondentes.

6. Um anticorpo, ou seu derivado, que se liga especificamente a um péptido da reivindicação 1.

7. Um anticorpo, ou seu derivado, de acordo com a reivindicação 6, em que o referido anticorpo é conjugado com uma molécula, tal como um fármaco, uma marca, uma molécula tóxica ou um isótopo radioactivo.

8. O anticorpo conjugado, ou seu derivado, da reivindicação 7, para usar como medicamento ou reagente de diagnóstico.

Lisboa, 12 de novembro de 2015

Célula ou tecido
Passo 1 Extracção de mRNA
5' AAAAAAA 3'
Passo 2 Hibridar com sequência iniciadora oligo-dT
5' AAAAAAA 3'
TTTTTTTT 5'
Passo 3 Conversão de mRNA em cDNA
AAAAAAA 3'
TTTTTTTT 5'
Passo 4 Remoção de RNA/adição de cauda poli-dG
3' GGGG TTTTTTTT 5'
Passo 5 Hibridar com sequência iniciadora oligo-dC
5' CCCC
3' GGGG TTTTTTTT 5'
Passo 6 Síntese da cadeia complementar
5' CCCC AAAAAAA 3'
3' GGGG TTTTTTTT 5'

Figura 1a

Passo 6
Síntese da cadeia complementar
Passo 7
Ligação de cDNA a adaptadores
Passo 8
Digestão com enzimas de restrição
Passo 9
Ligação ao vector de clonagem
Passo 10
Sequenciação
5' CCCC ——— AAAAAAA 3'
3' GGGG ——— TTTTTTT 5'
AGCT — CCCC ——— AAAAAAA — AGCT
TCGA — GGGG ——— TTTTTTT — TCGA
CT — CCCC ——— AAAAAAA — AG
GA — GGGG ——— TTTTTTT — TC
5' ——— 3'

Figura 1b

>gi|16507985|ref|NM_001428.2| Homo sapiens enolase 1, (alpha) (ENO1), mRNA

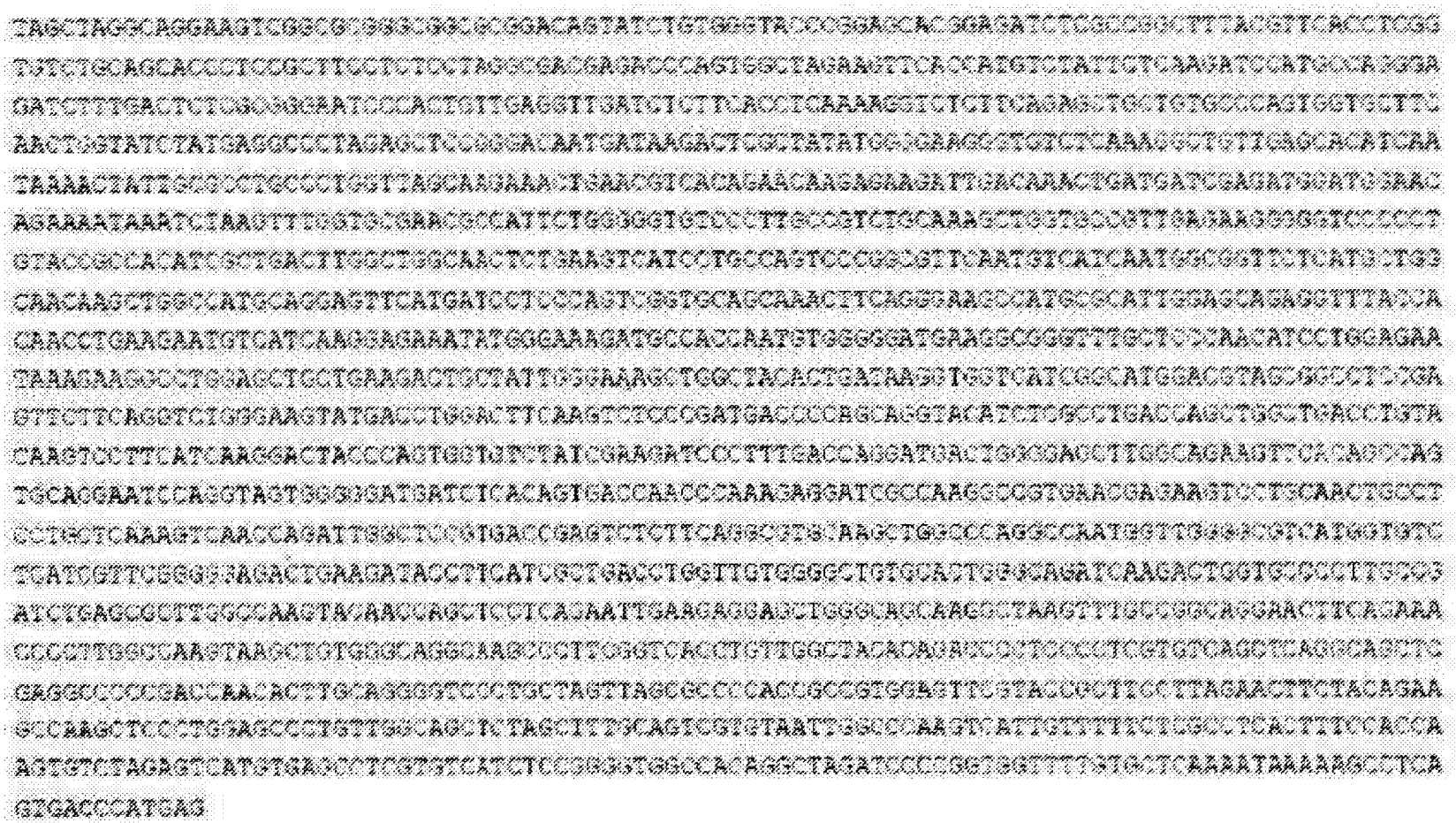


Figura 2a

>gi|56682958|ref|NM_002032.2| Homo sapiens ferritin, heavy polypeptide 1 (FTH1), mRNA

ATAAGAGACCACAAGCGACCCGCAGGGCCAGACGTTCTTCGCCGAGAGTCGTCGGGGTTTCCTGCTTCAACAGTGCTTGGACGGAAC
CCGGCGCTCGTTCCCCACCCCGGCCGGCCGCCCATAGCCAGCCCTCCGTCACCTCTTCACCGCACCCTCGGACTGCCCCAAGGCCCC
CGCCGCCGCTCCAGCGCCGCGCAGCCACCGCCGCCGCCGCCGCCTCTCCTTAGTCGCCGCCATGACGACCGCGTCCACCTCGCAGGT
GCGCCAGAACTACCACCAGGACTCAGAGGCCGCCATCAACCGCCAGATCAACCTGGAGCTCTACGCCTCCTACGTTTACCTGTCCAT
GTCTTACTACTTTGACCGCGATGATGTGGCTTTGAAGAACTTTGCCAAATACTTTCTTCACCAATCTCATGAGGAGAGGGAACATGC
TGAGAAACTGATGAAGCTGCAGAACCAACGAGGTGGCCGAATCTTCCTTCAGGATATCAAGAAACCAGACTGTGATGACTGGGAGAG
CGGGCTGAATGCAATGGAGTGTGCATTACATTTGGAAAAAAATGTGAATCAGTCACTACTGGAACTGCACAAACTGGCCACTGACAA
AAATGACCCCCATTTGTGTGACTTCATTGAGACACATTACCTGAATGAGCAGGTGAAAGCCATCAAAGAATTGGGTGACCACGTGAC
CAACTTGCGCAAGATGGGAGCGCCCGAATCTGGCTTGGCGGAATATCTCTTTGACAAGCACACCCTGGGAGACAGTGATAATGAAAG
CTAAGCCTCGGGCTAATTTCCCCATAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTACCT
TTTCTATAAGTTGTACCAAAACATCCACTTAAGTTCTTTGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACCCAGGTGTTGT
CTTTGAGGTCTTGGGATGAATCAGAAATCTATCCAGGCTATCTTCCAGATTCCTTAAGTGCCGTTGTTCAGTTCTAATCACACTAAT
CAAAAAGAAACGAGTATTTGTATTTATTAAACTCATTAGTTTGGGCAGTATACTAAGGTGTGGCTGTCTTGGATTCAGATAGAACTA
AGGGTTCCCGACTCTGAATCCAGAGTCTGAGTTAAATGTTTCCAATGGTTCAGTCTAGCTTTCACAGTTTTTATGAATAAAAGGCAT
TAAAGGCTGA

Figura 2b

>gi|56682960|ref|NM_000146.3| Homo sapiens ferritin, light polypeptide (FTL), mRNA

GCAGTTCGGCGGTCCCGCGGGTCTGTCTCTTGCTTCAACAGTGTTTGGACGGAACAGATCCGGGGACTCTCTTCCAGCCTCCGACCGC
CCTCCGATTTCCTCTCCGCTTGCAACCTCCGGGACCATCTTCTCGGCCATCTCCTGCTTCTGGGACCTGCCAGCACCGTTTTTGTGGT
TAGCTCCTTCTTGCCAACCAACCATGAGCTCCCAGATTCGTCAGAATTATTCCACCGACGTGGAGGCAGCCGTCAACAGCCTGGTCAA
TTTGTACCTGCAGGCCTCCTACACCTACCTCTCTCTGGGCTTCTATTTCGACCGCGATGATGTGGCTCTGGAAGGCGTGAGCCACTTC
TTCCGCGAATTGGCCGAGGAGAAGCGCGAGGGCTACGAGCGTCTCCTGAAGATGCAAAACCAGCGTGGCGGCCGCGCTCTCTTCCAGG
ACATCAAGAAGCCAGCTGAAGATGAGTGGGGTAAAACCCCAGACGCCATGAAAGCTGCCATGGCCCTGGAGAAAAAGCTGAACCAGGC
CCTTTTGGATCTTCATGCCCTGGGTTCTGCCCGCACGGACCCCCATCTCTGTGACTTCCTGGAGACTCACTTCCTAGATGAGGAAGTG
AAGCTTATCAAGAAGATGGGTGACCACCTGACCAACCTCCACAGGCTGGGTGGCCCGGAGGCTGGGCTGGGCGAGTATCTCTTCGAAA
GGCTCACTCTCAAGCACGACTAAGAGCCTTCTGAGCCCAGCGACTTCTGAAGGGCCCCTTGCAAAGTAATAGGGCTTCTGCCTAAGCC
TCTCCCTCCAGCCAATAGGCAGCTTTCTTAACTATCCTAACAAGCCTTGGACCAAATGGAAATAAAGCTTTTTGATGC

Figura 2c

>gi|83641890|ref|NM_002046.3| Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA

AAATTGAGCCCGCAGCCTCCCGCTTCGCTCTCTGCTCCTCCTGTTCGACAGTCAGCCGCATCTTCTTTTGCGTCGCCAGCCGAGCCACA
TCGCCTCAGACACCATGGGGAAGGTGAAGGTCGGAGTCAACGGATTTGGTCGTATTGGGCGCCTGGTCACCAGGGCTGCTTTTAACTCT
GGTAAAGTGGATATTGTTGCCATCAATGACCCCTTCATTGACCTCAACTACATGGTTTACATGTTCCAATATGATTCCACCCATGGCAAA
TTCCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTCATCAATGGAAATCCCATCACCATCTTCCAGGAGCGAGATCCCTCCAAAAT
CAAGTGGGGCGATGCTGGCGCTGAGTACGTCGTGGAGTCCACTGGCGTCTTCACCACCATGGAGAAGGCTGGGGCTCATTTGCAGGGGG
GAGCCAAAAGGGTCATCATCTCTGCCCCTCTGCTGATGCCCCCATGTTCGTCATGGGTGTGAACCATGAGAAGTATGACAACAGCCTC
ATCATCAGCAATGCCTCCTGCACCACCAACTGCTTAGCACCCCTGGCCAAGGTCATCCATGACAACTTTGGTATCGTGGAAGGACT
CATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCCTCCGGGAAACCTGTGGCGTGGGCGTGGCCGCTCTCT
AGAACATCATCCCTGCCTCTACTGGCGCTGCCAAGGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCCTTC
CGTGTCCCCACTGCCAACGTGTCAGTGGTGGACCTGACCTGCCGTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAGGTGGTGAA
GCAGGCGTCGGAGGGCCCCCTCAAGGGCATCCTGGGCTACACTGAGCACCAGGTGGTCTCCTCTGACTTCAACAGCGACACCCACTCC
TCCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGTCAAGCTCATTTCCTGGTATGACAACGAATTTGGCTACAGCAAC
AGGGTGGTGGACCTCATGGCCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGAGAG
ACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACTGAATCTCATCCCTCACACTGCCATGTAGACCCCTTGAAG
AGGGGATGCCTAGTGGAGCCCACTGTCCCCACTGATTAAAAATCACCTGTGCTCAAAAA

Figura 2d

>gi|42734684|ref|NM_006597.3| Homo sapiens heat shock 70kDa protein 8 (HSPA8), transcript variant 1, mRNA

Figura 2e

>gi|18390348|ref|NM_000972.2| Homo sapiens ribosomal protein L7a (RPL7A), mRNA

TTTCCTTTCTCCTCCTCCGGCCCAAGATGGCGAAAAGGAAGAAGGCAAGGGAAAGAAGGTGGCTCCGGCCCCAGCTTTCCTGAA
GAAGCAGGAAGCCTAAGAAAGTGGATGAATCGCCTTTTTGAGAAAAGTTCTGGCATTTTGGGACATCCAGCCCAAAAGGG
ACCTCACCGCTTTTGTAAATGGTCCCGCTATATCAGGTTGCAGCGAGAGCCATCCTCTATAAGTGGCTGAAAGTGCTTCCTGCG
ATTAACCAGTTCACCCAGGGCCCTGGACCGCCAAACAGCTACTCAGCTGCTTAAGCTGGCCACAAGTACAGCAGGACAAAGCAAGA
CAAGAAGCAGACATGTTGGCCGGTGCCGAGAACAAGGCTGGCAAAAGGGAAGGTCCCAAGAATGACCAGCCTGTCCTTCCGAGCAG
GATTAACACCGTCACCACCTTGGTGAGAACAAGAAAGCTCAATCTGGTGGTGATTGCACAGCGACGTGGATCCCATCGAGCTGGTTGTC
TTCTTGCCTGCCCTGTGTCGTAAAAATGGCGGTCCCTTACTGCATTAICAAGGGAAAGGCAAGACTGGGACGCTCTAGTCCACAGGAAGAC
CTGCCACCACTGTCGCCTTCACACAGGTGAACTCGGAAGACAAAGGCCCTTTGCCTAAGCTGGTGGAAGCTATCAAGCTACAATTACAATG
ACAGATACGATGAGATCCGCCGTCACTGGGGTGGCAATGTCCTGGGTCCCTAAGTCTGTGGTCTGTATCGTCCAAGCTCGAAAGGCTAAG
GCTAAGAACTTGCCACTAAACTGGCTTAAATGTACACTGTTGAGTTTCCTGTACATAAAATAATTGAATGAATACAATTTCCTTC

Figura 2f

>gi|39812410|ref|NM_001007.3| Homo sapiens ribosomal protein S4, X-linked (RPS4X), mRNA

GGGCGGAGCAGCTGAAAATCCGGCGCGCAGTCTCCAGCCCCAATTTCTACGCGCACCGGAAGACGGAGGTCCTCTTTCCTTGCCT
AACGCAGCCATGGCTCGTGGTCCCAAGAAGCATCTGAAGCGGGTGGCAGCTCCAAAGCATTGGATGCTGGATAAATTGACCGGTGTG
TTTGCTCCTCGTCCATCCACCGGTCCCCACAAGTTGAGAGAGTGTCTCCCCCTCATCATTTTCCTGAGGAACAGACTTAAGTATGCC
CTGACAGGAGATGAAGTAAAGAAGATTTGCATGCAGCGGTTCATTAAAATCGATGGCAAGGTCCGAACTGATATAACCTACCCTGCT
GGATTCATGGATGTCATCAGCATTGACAAGACGGGAGAGAATTTCCGTCTGATCTATGACACCAAGGGTCGCTTTGCTGTACATCGT
ATTACACCTGAGGAGGCCAAGTACAAGTTGTGCAAAGTCAGAAAGATCTTTGTGGGCACAAAAGGAATCCCTCATCTGGTGACTCAT
GATGCCCGCACCATCCGCTACCCCGATCCCCTCATCAAGGTGAATGATACCATTCAGATTGATTTGGAGACTGGCAAGATTACTGAT
TTCATCAAGTTCGACACTGGTAACCTGTGTATGGTGACTGGAGGTGCTAACCTAGGAAGAATTGGTGTGATCACCAACAGAGAGAGG
CACCCTGGATCTTTTGACGTGGTTCACGTGAAAGATGCCAATGGCAACAGCTTTGCCACTCGACTTTCCAACATTTTTGTTATTGGC
AAGGGCAACAAACCATGGATTTCTCTTCCCCGAGGAAAGGGTATCCGCCTCACCATTGCTGAAGAGAGAGACAAAAGACTGGCGGCC
AAACAGAGCAGTGGGTGAAATGGGTCCCTGGGTGACATGTCAGATCTTTGTACGTAATTAAAAATATTGTGGCAGGATTAATAGC

Figura 2g

>gi|17158043|ref|NM_001010.2| Homo sapiens ribosomal protein S6 (RPS6), mRNA

CCTCTTTTCCGTGGCGCCTCGGAGGCGTTCAGCTGCTTCAAGATGAAGCTGAACATCTCCTTCCCAGCCACTGGCTGCCAGAAACTCA
TTGAAGTGGACGATGAACGCAAACTTCGTACTTTCTATGAGAAGCGTATGGCCACAGAAGTTGCTGCTGACGCTCTGGGTGAAGAATG
GAAGGGTTATGTGGTCCGAATCAGTGGTGGGAACGACAAACAAGGTTTCCCCATGAAGCAGGGTGTCTTGACCCATGGCCGTGTCCGC
CTGCTACTGAGTAAGGGGCATTCCTGTTACAGACCAAGGAGAACTGGAGAAAGAAAGAGAAAATCAGTTCGTGGTTGCATTGTGGATG
CAAATCTGAGCGTTCTCAACTTGGTTATTGTAAAAAAAGGAGAGAAGGATATTCCTGGACTGACTGATACTACAGTGCCTCGCCGCCT
GGGCCCCAAAAGAGCTAGCAGAATCCGCAAACTTTTCAATCTCTCTAAAGAAGATGATGTCCGCCAGTATGTTGTAAGAAAGCCCTTA
AATAAAGAAGGTAAGAAACCTAGGACCAAAGCACCCAAGATTCAGCGTCTTGTTACTCCACGTGTCCTGCAGCACAAACGGCGGCGTA
TTGCTCTGAAGAAGCAGCGTACCAAGAAAAATAAAGAAGAGGCTGCAGAATATGCTAAACTTTTGGCCAAGAGAATGAAGGAGGCTAA
GGAGAAGCGCCAGGAACAAATTGCGAAGAGACGCAGACTTTCCTCTCTGCGAGCTTCTACTTCTAAGTCTGAATCCAGTCAGAAATAA
GATTTTTGAGTAACAAATAAATAAGATCAGACTCTG

Figura 2h

>gi|34328943|ref|NM_021109.2| Homo sapiens thymosin, beta 4, X-linked (TMSB4X), mRNA

ACTAACTCGGTGGTGGCCACTGCGCAGACCAGACTTCGCTCGTACTCGTGCGCCTCGCTTCGCTTTTCCTCCGCAACCATGTCTGACAAACCC
GATATGGCTGAGATCGAGAAATTCGATAAGTCGAAACTGAAGAAGACAGAGACGCAAGAGAAAAATCCACTGCCTTCCAAAGAAACGATTGA
ACAGGAGAAGCAAGCAGGCGAATCGTAATGAGGCGTGCGCCGCCAATATGCACTGTACATTCCACAAGCATTGCCTTCTTATTTTACTTCTT
TTAGCTGTTTAACTTTGTAAGATGCAAAGAGGTTGGATCAAGTTTAAATGACTGTGCTGCCCCTTTCACATCAAAGAACTACTGACAACGAA
GGCCGCGCCTGCCTTTCCCATCTGTCTATCTATCTGGCTGGCAGGGAAGGAAAGAACTTGCATGTTGGTGAAGGAAGAAGTGGGGTGGAAGA
AGTGGGGTGGGACGACAGTGAAATCTAGAGTAAAACCAAGCTGGCCCAAGGTGTCCTGCAGGCTGTAATGCAGTTTAATCAGAGTGCCATTT
TTTTTTTTGTTCAAATGATTTTAATTATTGGAATGCACAATTTTTTTAATATGCAAATAAAAAGTTTAAAAACTT

Figura 2i

>gi|4507666|ref|NM_003295.1| Homo sapiens tumor protein, translationally-controlled 1 (TPT1), mRNA

GGGCCCCGAGGGGGCTCCGGCTGCAGCGGCTCGTTTCAGGCTCGTAGCTAGGCCGCTCGTCCTTCCTCAGT
CGCCATCATGATATCTACCGGGACTCATCAGCCACATGAGATGTTCTGCGACAICTACAAGATCCGGGAGATCGGCGACGGCAGTT
GTGCTTGGAGGTGGAGGGGAAGATGGTCAGTAGAGACATTGATGACTTCCTCATTGGAAATGCCTCCGCCTGAGG
CGCCGAGAGGCGAAGCTACCGAAAGCCACATGGTCGATATTGTCATGAACCATGCACCTGCAGGAAACAAGTTTCACAAA
AGAAGCCTACAAGAAGTACATCAAAGATTACATCAAAATCAATGAAAGGAAACTTGAAGAACAGAGAACAGAAGAGTAAAACCTTT
TATGACAAGGGGCCTGCAGAACAAATGCAACACATCCTTGCTTTATTGCTGAAAACATGAATCCAGA
TGGCATGGTTGCTCTATTGGACTTACCGATGGATGGTGACCCCATATATGATTTTCTTTAAGGATGGTTAGAAATGGAAAAATG
TTAACAAATGTGCCAATTATTTGGATCTATCATAACTGGCTTCTGCTTGTCATCCACAACACCAGGACTTAAGA
CAAATGGCACTGATGTCATCTTGACTCTATTTGACTGATTTATTTGGAGTGGAGCCATTGTTTTTAAGAAAAACATG
TCATGTAGGTTGTCTAAAATAAAATGCATTTAAACTCATTTGAGAS

Figura 2j

>gi|62414288|ref|NM_003380.2| Homo sapiens vimentin (VIM), mRNA

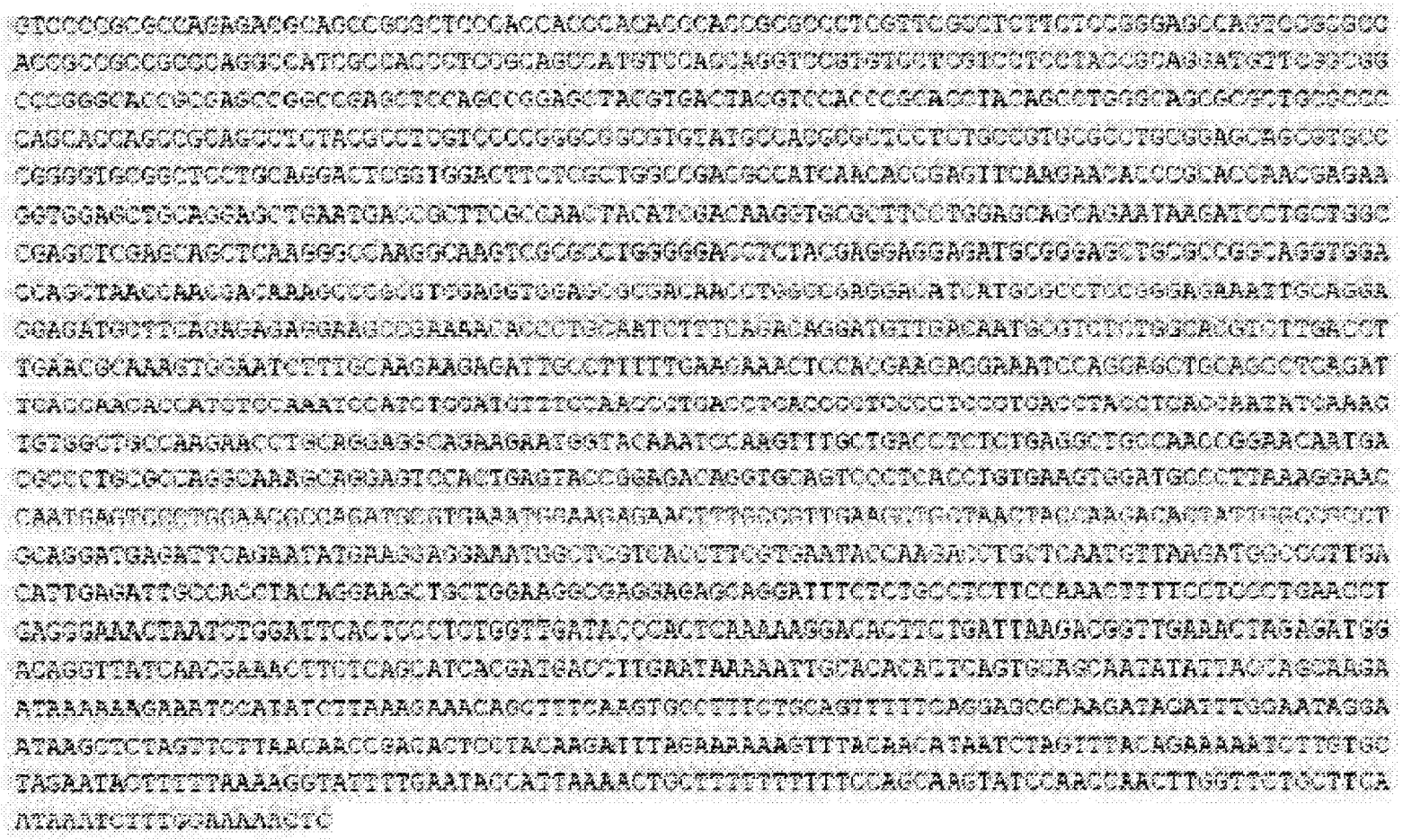


Figura 2k

>gi|6392890|ref|NM_000477.3| Homo sapiens albumin (ALB), mRNA

Figura 21

>gi|34577108|ref|NM_000034.2| Homo sapiens aldolase A, fructose-bisphosphate (ALDOA), transcript variant 1, mRNA

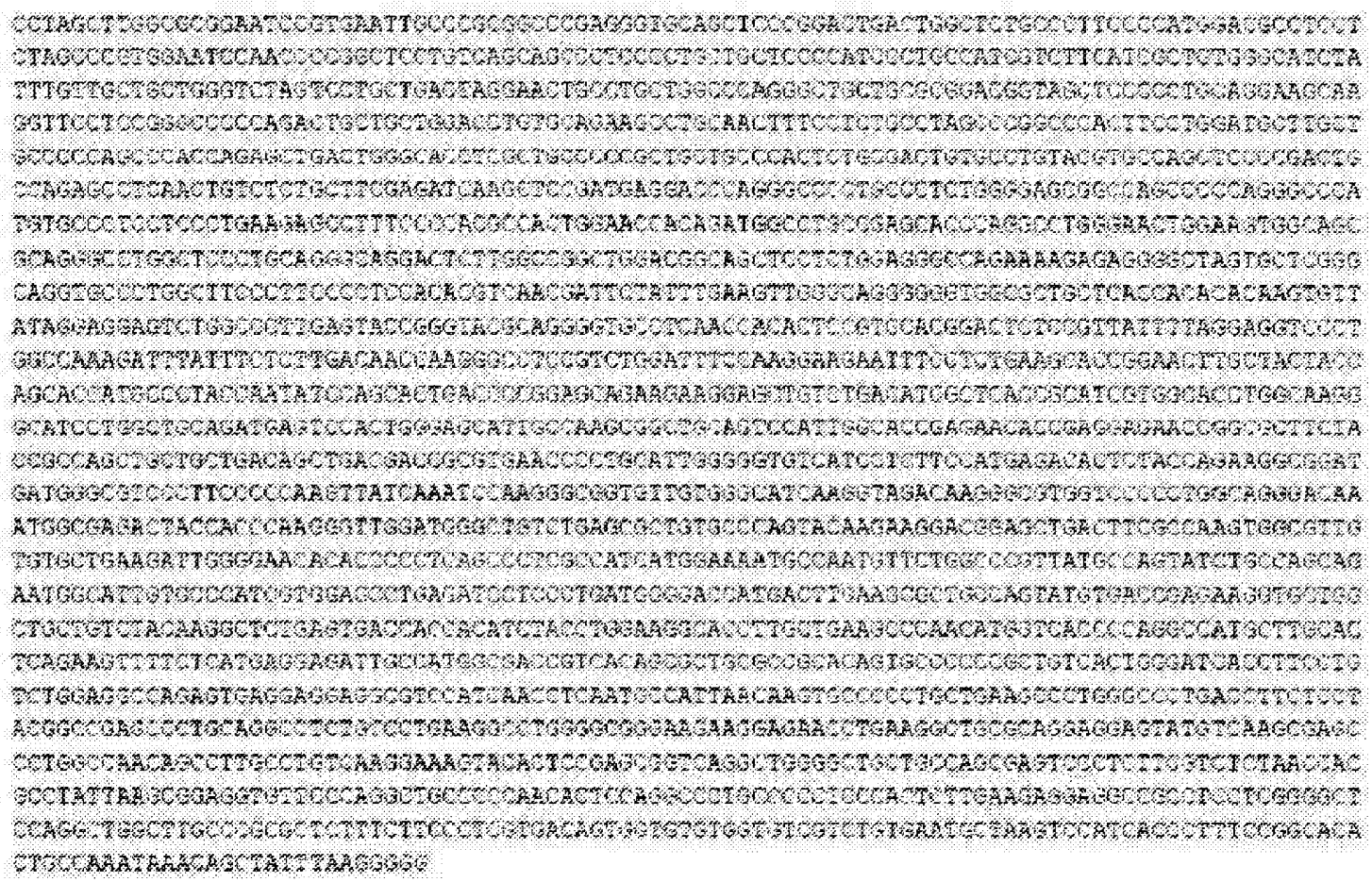


Figura 2m

>gi|50345983|ref|NM_001001937.1| Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle (ATP5A1), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA

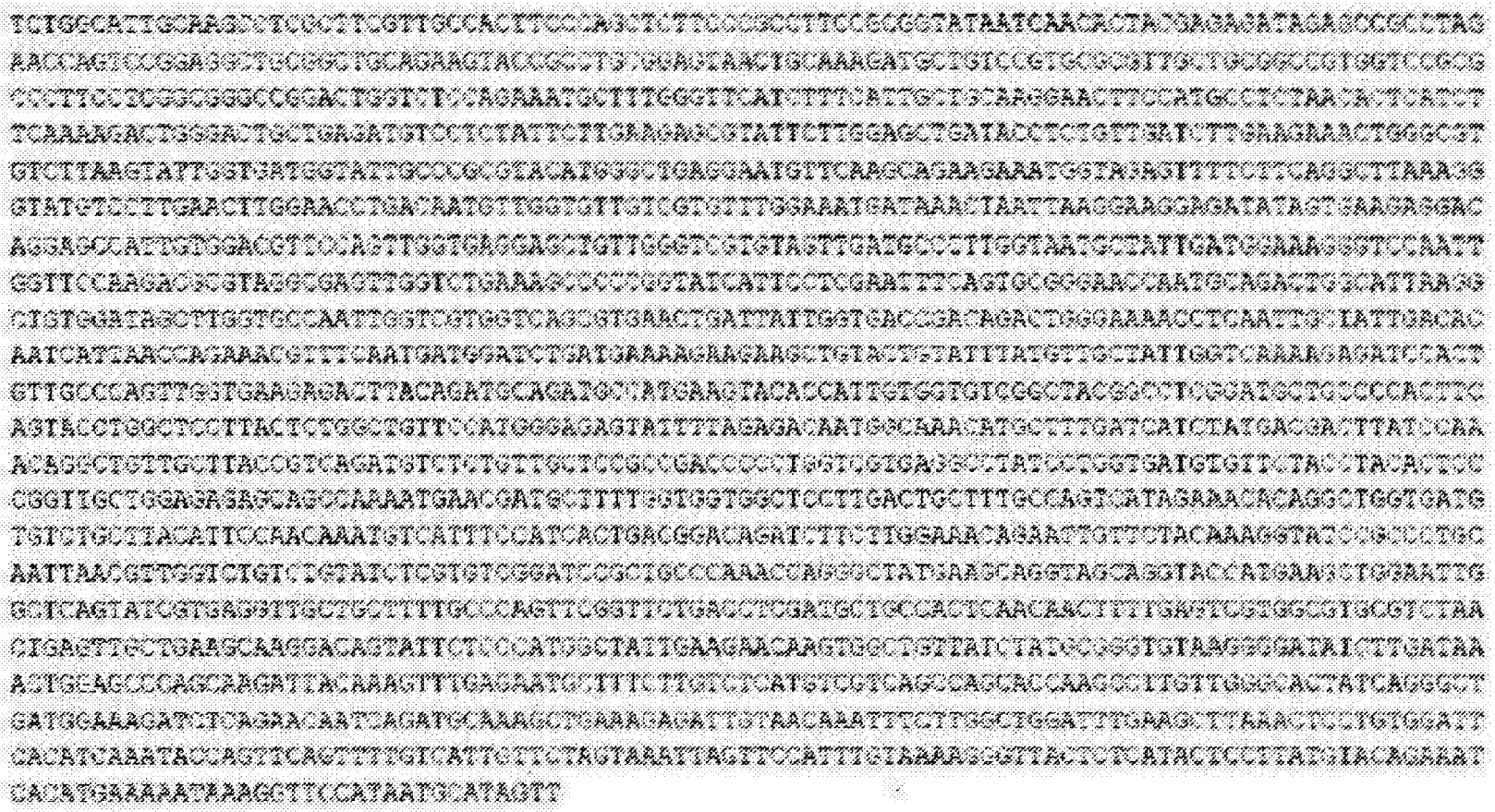


Figura 2n

>gi|20428653|ref|NM_001743.3| Homo sapiens calmodulin 2 (phosphorylase kinase, delta) (CALM2), mRNA

AGTCCGAGTGGAGAGAGCGAGCTGAGTGGTTGTGTGGTCGCGTCTCGGAAACCGGTAGCGCTTGCAGCATGGCTGACCAACTGACTGAAGA
GCAGATTGCAGAATTCAAAGAAGCTTTTTCACTATTTGACAAAGATGGTGATGGAACTATAACAACAAAGGAATTGGGAACTGTAATGAGA
TCTCTTGGGCAGAATCCCACAGAAGCAGAGTTACAGGACATGATTAATGAAGTAGATGCTGATGGTAATGGCACAATTGACTTCCCTGAAT
TTCTGACAATGATGGCAAGAAAAATGAAAGACACAGACAGTGAAGAAGAAATTAGAGAAGCATTCCGTGTGTTTGATAAGGATGGCAATGG
CTATATTAGTGCTGCAGAACTTCGCCATGTGATGACAAACCTTGGAGAGAAGTTAACAGATGAAGAAGTTGATGAAATGATCAGGGAAGCA
GATATTGATGGTGATGGTCAAGTAAACTATGAAGAGTTTGTACAAATGATGACAGCAAAGTGAAGACCTTGTACAGAATGTGTTAAATTTC
TTGTACAAAATTGTTTATTTGCCTTTTCTTTGTTTGTAACTTATCTGTAAAAGGTTTCTCCCTACTGTCAAAAAAATATGCATGTATAGTA
ATTAGGACTTCATTCCTCCATGTTTTCTTCCCTTATCTTACTGTCATTGTCCTAAAACCTTATTTTAGAAAATTGATCAAGTAACATGTTG
CATGTGGCTTACTCTGGATATATCTAAGCCCTTCTGCACATCTAAACTTAGATGGAGTTGGTCAAATGAGGGAACATCTGGGTTATGCCTT
TTTTAAAGTAGTTTTCTTTAGGAACTGTCAGCATGTTGTTGTTGAAGTGTGGAGTTGTAACTCTGCGTGGACTATGGACAGTCAACAATAT
GTACTTAAAAGTTGCACTATTGCAAAACGGGTGTATTATCCAGGTACTCGTACACTATTTTTTTGTACTGCTGGTCCTGTACCAGAAACAT
TTTCTTTTATTGTTACTTGCTTTTTAAACTTTGTTTAGCCACTTAAAATCTGCTTATGGCACAATTTGCCTCAAAATCCATTCCAAGTTGT
ATATTTGTTTTCCAATAAAAAAATTACAATTTACGC

Figura 2o

>gi|5031856|ref|NM_005566.1| Homo sapiens lactate dehydrogenase A (LDHA), mRNA

TGCTGCAGCCGCTGCCGCCGATTCCGGATCTCATTGCCACGCCCCCCGAGGACCGCCCGGACGTGCATTCCCGATTCCTTTTGGTTCCAAG
TCCAATATGCCAACTCTAAACGATCAGCTGATTTATAATCTTCTAAAGGAAGAACAGACCCCCCAGAATAAGATTACAGTTGTTGGGGTTG
GTGCTGTTGGCATGGCCTGTGCCATCAGTATCTTAATGAAGGACTTGGCAGATGAACTTGCTCTTGTTGATGTCATCGAAGACAAATTGAA
GGGAGAGATGATGGATCTCCAACATGGCAGCCTTTCCTTAGAACACCAAAGATTGTCTCTGGCAAAGACTATAATGTAACTGCAAACTCC
AAGCTGTGTCATTATCACAGCCTGGGGCACGTCAGCAAGAGGGAGAAAGCCGTCTTAATTTGGTCCAGCGTAACGTGAACATATTTAAATTCA
TCATTCCTAATGTTGTAAAATACAGCCCGAACTGCAAGTTGCTTATTGTTTCAAATCCAGTGGATATCTTGACCTACGTGGCTTGGAAGAT
AAGTGGTTTTCCCAAAAACCGTGTTATTGGAAGTGGTTGCAATCTGGATTCAGCCTGATTCCGTTACCTGATGGGGGAAAGGCTGGGAGTT
CACCCATTAAGCTGTCATGGGTGGGTCCTTGGGGAACATGGAGATTCCAGTGTGCCTGTATGGAGTGGAATGAATGTTGCTGGTGTCTCTC
TGAAGACTCTGCAGCCAGATTTAGGGACTGATAAAGATAAGGAACAGTGGAAAGAGGTTCACAAGCAGGTGGTTGAGAGTGCTTATGAGGT
GATCAAACTCAAAGGCTACACATCCTGGGCTATTGGACTCTCTGTAGCAGATTTGGCAGAGATATAATGAAGAATCTTAGGCGGGTGCAC
CCAGTTTCCACCATGATTAAGGGTCTTTACGGAATAAAGGATGATGTCTTCCTTAGTGTTCCTTGCATTTTGGGACAGAATGGAATCTCAG
ACCTTGTGAAGGTGACTCTGACTTCTGAGGAAGAGGCCCGTTTGAAGAAGAGTGCAGATACACTTTGGGGGATCCAAAAGGAGCTGCAATT
TTAAAGTCTTCTGATGTCATATCATTTCACTGTCTAGGCTACAACAGGATTCTAGGTGGAGGTTGTGCATGTTGTCCTTTTTTATCTGATCT
GTGATTAAAGCAGTAATATTTTAAGATGGACTGGCAAAAACATCAACTCCTGAAGTTAGAAATAAGAATGGTTTGTAAAATCCACAGCTAT
ATCCTGATGCTGGATGGTATTAATCTTGTGTAGTCTTCAACTGGTTAGTGTGAAATAGTTCTGCCACCTCTGACGCACCACTGCCAATGCT
GTACGTACTGCATTTGCCCCTTGAGCCAGGTGGATGTTTACCGTGTGTTATATAACTTCCTGGCTCCTTCACTGAACATGCCTAGTCCAAC
ATTTTTTCCCCAGTGAGTCACATCCTGGGATCCAGTGTATAAATCCAATATCATGTCTTGTGCATAATTCTTCCAAAGGATCTTATTTTGTG
AACTATATCAGTAGTGTACATTACCATATAATGTAAAAAGATCTACATACAAACAATGCAACCAACTATCCAAGTGTTACACCAACTAAAA
GCCCCAATAAACCTTGAACAGTG

Figura 2p

>gi|52851446|ref|NM_000365.4| Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA

CCTTCAGCGCCTCGGCTCCAGCGCCATGGCGCCCTCCAGGAAGTTCTTCGTGGGGGGAAACTGGAAGATGAACGGCGGAAGCAGAGTCTG
GGGGAGCTCATCGGCACTCTGAACGCGGCCAAGGTGCCGGCCGACACCGAGGTGGTTTGTGCTCCCCCTACTGCCTATATCGACTTCGCCC
GGCAGAAGCTAGATCCCAAGATTGCTGTGGCTGCGCAGAACTGCTACAAAGTGACTAATGGGGCTTTTACTGGGGAGATCAGCCCTGGCAT
GATCAAACACTGCGGAGCCACGTGGGTGGTCCTGGGGCACTCAGAGAGAAGGCATGTCTTTGGGGAGTCAGATGAGCTGATTGGGCAGAAA
GTGGCCCATGCTCTGGCAGAGGGACTCGGAGTAATCGCCTGCATTGGGGAGAAGCTAGATGAAAGGGAAGCTGGCATCACTCAGAAGGTTG
TTTTCGAGCAGACAAAGGTCATCGCAGATAACGTGAAGGACTGGAGCAAGGTCGTCCTGGCCTATGAGCCTGTGTGGGCCATTGGTACTGG
CAAGACTGCAACACCCCAACAGGCCCAGGAAGTACACGAGAAGCTCCGAGGATGGCTGAAGTCCAACGTCTCTGATGCGGTGCTCAGAGC
ACCCGTATCATTTATGGAGGCTCTGTGACTGGGGCAACCTGCAAGGAGCTGGCCAGCCAGCCTGATGTGGATGGCTTCCTTGTGGGTGGTG
CTTCCCTCAAGCCCGAATTCGTGGACATCATCAATGCCAAACAATGAGCCCCATCCATCTTCCCTACCCTTCCTGCCAAGCCAGGGACTAA
GCAGCCCAGAAGCCCAGTAACTGCCCTTTCCCTGCATATGCTTCTGATGGTGTCATCTGCTCCTTCCTGTGGTCTCATCCAAACTGTATCT
TCCTTTACTGTTTATATCTTCACCCTGTAATGGTTGGGACCAGGCCAATGGCTTCTCCACTTACTATAATGGTTGGAACTAAACGTCACCA
AGGTGGCTTCTCCTTGGCTGAGAGATGGAAGGCGTGGTGGCATTTGCTCCTGGGTTCCCTAGGCCCTAGTGAGGGCAGAAGAGAAACCATC
CTCTCCCTTCTTACACCGTGAGGCCAAGATCCCCTCAGAAGGCAGGAGTGCTGCCCTCTCCCATGGTGCCCGTGCCTCTGTGCTGTGTATG
TGAACCACCCATGTGAGGGAATAAAGCCTGGCACTAGG

Figura 2q

TPT1 = Proteína de tumor, controlada a nível da tradução, 1880 pb, cr 13 q12-q24
VIM = Vimentina, 1847 pb, cr 10 p13.
FTH1 = Ferritina, polipéptido pesado 1, 1228pb, cr 11 q13
RPS4X = Proteína ribossomal S, ligada a X, 955 pb, cr X q13.1
RPL7A = Proteína ribossomal L7a, 890 pb, cr 9 q34.
ENO1 = Enolase 1, 1812 pb, cr 1 p36-p36.2.
HSPA8 = Proteína de choque térmico 70 kDa 8, 2260 pb, cr 11 q24.1.
GAPDH = Desidrogenase de gliceraldeído-3-fosfato, 1310 pb, cr 12 p13.
TMSB4X = Timosina, beta 4, ligada a X, 627 pb, cr X q21.3-q22.
RPS6 = Proteína ribossomal S6, 829 pb, cr 9 p21.
FTL = Ferritina, polipéptido leve, 870 pb, cr 19 q13.3-q22.
ALB = Albumina, 2215 pb, cr 4 q11-q13.
ALDOA = Aldolase A, frutose-bifosfato, 2303 pb, cr 16q22-q24.
ATP5A1 = Sintetase de ATP, transporte de H+, complexo f1 mitocondrial, subunidade alfa 1, músculo cardíaco, 1945 pb, cr 18 q12-q21.
CALM2 = Calmodulina 2 (fosforilase cinase, delta), 1128 pb, cr 2p21.
LDHA = Desidrogenase de lactato A, 1661 pb, cr 11p15.4.
TPI1 = Isomerase de triosefosfato 1, 1220 pb, cr 12p13.

Figura 2r

```
>gi|47009767|emb|BX437546.2| BX437546 Homo sapiens THYMUS Homo sapiens cDNA clone CS0CAP007YK06
          5-PRIME, mRNA sequence.
          Length = 868

 Score = 1531 bits (829), Expect = 0.0
 Identities = 829/829 (100%)
 Strand = Plus / Plus

Query: 2    ccccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 61
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12   ccccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 71

Query: 62   tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 121
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 72   tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 131

Query: 122  cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 181
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 132  cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 191

Query: 182  gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 241
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 192  gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 251

Query: 242  ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 301
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 252  ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 311

Query: 302  gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 361
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 312  gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 371
```

Figura 3a

Figura 3b

Figura 3c

```
Query: 482  aaaaactaccagttctttattggtgaaaacatgaatccagatggcatggttgctctattg 541
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 498  aaaaactaccagttctttattggtgaaaacatgaatccagatggcatggttgctctattg 557

Query: 542  gactaccgtgaggatggtgtgaccccatatatgattttctttaaggatggtttagaaatg 601
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 558  gactaccgtgaggatggtgtgaccccatatatgattttctttaaggatggtttagaaatg 617

Query: 602  gaaaaatgttaacaaatgtggcaattatttggatctatcacctgtcatcataactggct 661
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 618  gaaaaatgttaacaaatgtggcaattatttggatctatcacctgtcatcataactggct 677

Query: 662  tctgcttgtcatccacacaacaccaggacttaagacaaatgggactgatgtcatcttgag 721
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 678  tctgcttgtcatccacacaacaccaggacttaagacaaatgggactgatgtcatcttgag 737

Query: 722  ctcttcatttatttgactgtgatttatttggagtggaggcattgtttttaagaaaaaca 781
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 738  ctcttcatttatttgactgtgatttatttggagtggaggcattgtttttaagaaaaaca 797

Query: 782  tgtcatgtaggttgtctaaaaataaaatgcatttaaactcatttgagag 830
            |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 798  tgtcatgtaggttgtctaaaaataaaatgcatttaaactcatttgagag 846
```

Figura 3d

```
>gi|45696213|emb|AL520698.3| AL520698 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens
          cDNA clone CS0DB002YF24 5-PRIME, mRNA sequence.
        Length = 909

 Score = 1530 bits (828), Expect = 0.0
 Identities = 826/829 (99%)
 Strand = Plus / Plus

Query: 2   ccccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggcccgtgcttaagc 61
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 18  ccccttgagcgctgctccggctgcaccgcgctcgctccgagtttcaggcttgtgcttaagc 77

Query: 62  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 121
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 78  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 137

Query: 122 cacgatgagatgttcttcgacatctacaagatccgggagatcgcggacgggttgtgcctg 181
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 138 cacgatgagatgttcttcgacatctacaagatccgggagatcgcggacgggttgtgcctg 197

Query: 182 gaggtggagggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 241
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 198 gaggtggagggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 257

Query: 242 cgaaatgcctccgctgaaggcccgagggcgaaggtaccgaaagcacagtaatcactggt 301
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 258 cgaaatgcctccgctgaaggcccgagggcgaaggtaccgaaagcacagtaatcactggt 317

Query: 302 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 361
           |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct: 318 gtcgatattgtccatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 377

Query: 362 aagtacatcaaagattacatgaaaatccaatcaaagggaaacttgaagaacagagaccagaa 921
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 378 aagtacatcaaagattacatgaaaatccaatcaaagggaaacttgaagaacagagaccagaa 437
```

Figura 3e

Query: 422 agagcaaaaaccctttatgacagggctgcagaacaaatcaagcacacctgctgaattcc 481
Sbjct: 438 agagcaaaaaccctttatgacagggctgcagaacaaatc4agcacacctgctgaattcc 497
Query: 482 aaaaactacagttctttcatgtgaaaacatgaatcccagatggcatggtgctccattg 541
Sbjct: 498 aaaaactacagttctttcatgtgaaaacatgaatcccagatggcatggtgctccattg 557
Query: 542 gactaccgtgaggatggtgtgaccccacataatgattctttcaaggatggttcagaaatg 601
Sbjct: 558 gactaccgtgaggacggtgtgacccccacatatgattctttcaaggatggttcagaaatg 617
Query: 602 gaaaaatgttcaacaaaatggtggcaaattatttgggatctatcacctgtcatcataactggct 661
Sbjct: 618 gaaaaatgttcaacaaaatggtggcaaattatttgggatccatcacctgtcatcataactggct 677
Query: 662 tctgcttgtcatccacacaaccaccaggacttaagacaaaatgggactgatgtcatcttgag 721
Sbjct: 678 tctgcttgtcatcccacacaaccaccaggacttcaagacaaaatgggactgatgtcatcttgag 737
Query: 722 ctcttccattctatttgactgtgattcacttgggaggcatgtctttcaagaaaaaca 781
Sbjct: 738 ctcttccattctatttgactgtgaatggcactgtctttcaagaaaaaca 797
Query: 782 tgtcatgtaggttgccttaaaaatcaaaatgcattcaaactccattgaggag 830
Sbjct: 798 tgtcatgtaggttgccttaaaaatcaaaatgcattcaaactccattgaggag 846

Figura 3f

```
>gi|10105790|gb|BE717525.1| RC4-HT0778-190700-014-e04 HT0778 Homo sapiens cDNA, mRNA sequence
          Length = 341

 Score =  564 bits (305), Expect = e-158
 Identities = 314/318 (98%), Gaps = 1/318 (0%)
 Strand = Plus / Plus

Query: 347  aaagaagcctacaagaagtacatcaaagattacatgaaatcaatcaaagggaaacttgaa 406
             |||||||| |||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct: 12   aaagaagcatacaagaagtacatcaaaga-tacatgaaatcaatcaaagggaaacttgaa 70

Query: 407  gaacagagaccagaaagagtaaaacctttatgacaggggctgcagaacaaatcaagcac 466
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 71   gaacagagaccagaaagagtaaaacctttatgacaggggctgcagaacaaatcaagcac 130

Query: 467  atccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagatggc 526
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 131  atccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagatggc 190

Query: 527  atggttgctctattggactaccgtgaggatggtgtgaccccatatatgatttctcttaag 586
             |||||||||||||||| |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 191  atggttgctctattgggctaccgtgaggatggtgtgaccccatatatgatttctcttaag 250

Query: 587  gatggtttagaaatggaaaaatgttaacaaatgtggcaattattttggatctatcacctg 646
             ||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct: 251  gatggtttagaaatggaaaaatgttaacaaatgtggcaattatttgggatctatcacctg 310

Query: 647  tcatcataactggcttct 664
             ||||||||||||||||||
Sbjct: 311  tcatcataactggcttct 328
```

Figura 3g

```
>gi|10224966|gb|BE771308.1| QV1-FT0062-210600-249-a12 FT0062 Homo sapiens cDNA, mRNA sequence
          Length = 338

 Score =  562 bits (304), Expect = e-157
 Identities = 316/321 (98%), Gaps = 3/321 (0%)
 Strand = Plus / Plus

Query: 346 aaaagaagcctacaagaa-gtacatca-aagattacatgaaaatcaatcaaagggaaaactt 403
           |||||||||||||||||| |||||||   |||||||| ||||||||||||||||||||||
Sbjct: 6   aaaagaagcctacaagaacgtacatcagttgattacatg-aatcaatcaaagggaaaactt 64

Query: 404 gaagaacagagaccagaaagagtaaaacctttatgacagggctgcagaacaaaatcaag 463
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 65  gaagaacagagaccagaaagagtaaaacctttatgacagggctgcagaacaaaatcaag 124

Query: 464 cacatccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagat 523
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 125 cacatccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagat 184

Query: 524 ggcatggttgctctattggactaccgtgaggatggtgtgacccatatatgatttcttt 583
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 185 ggcatggttgctctattggactaccgtgaggatggtgtgacccatatatgatttcttt 244

Query: 584 aaggatggtttagaaatggaaaaatgttaacaaatgtggcaattatttggatctatcac 643
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 245 aaggatggtttagaaatggaaaaatgttaacaaatgtggcaattatttggatctatcac 304

Query: 644 ctgtcatcataactggcttcc 664
           ||||||||||||||||||||
Sbjct: 305 ctgtcatcataactggcttcc 325
```

Figura 3h

```
>gi|10246833|gb|BE814599.1| MR0-BN0070-070800-033-b04_1 BN0070 Homo sapiens cDNA, mRNA
          sequence.
         Length = 140

 Score =  259 bits (140), Expect = 3e-066
 Identities = 140/140 (100%)
 Strand = Plus / Plus

Query: 451 agaacaaaatcaagcacatccttgctaatttcaaaaactaccagttctttattggtgaaaa 510
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1   agaacaaaatcaagcacatccttgctaatttcaaaaactaccagttctttattggtgaaaa 60

Query: 511 catgaatccagatggcatggttgctctattggactaccgtgaggatggtgtgacccccata 570
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61  catgaatccagatggcatggttgctctattggactaccgtgaggatggtgtgacccccata 120

Query: 571 tatgattttctttaaggatg 590
           ||||||||||||||||||||
Sbjct: 121 tatgattttctttaaggatg 140
```

Figura 3i

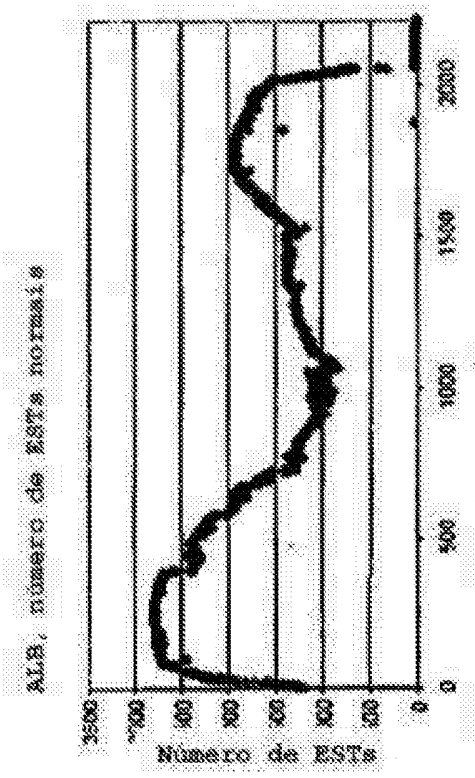


Figura 4a-2

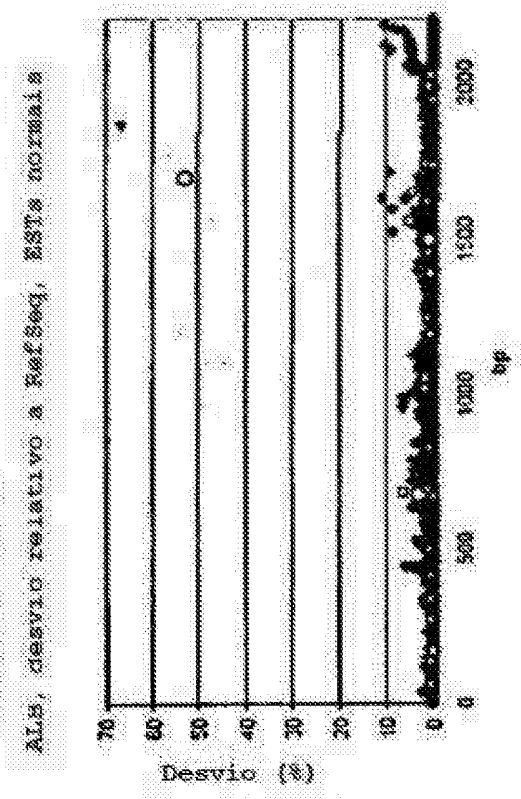


Figura 4a-4

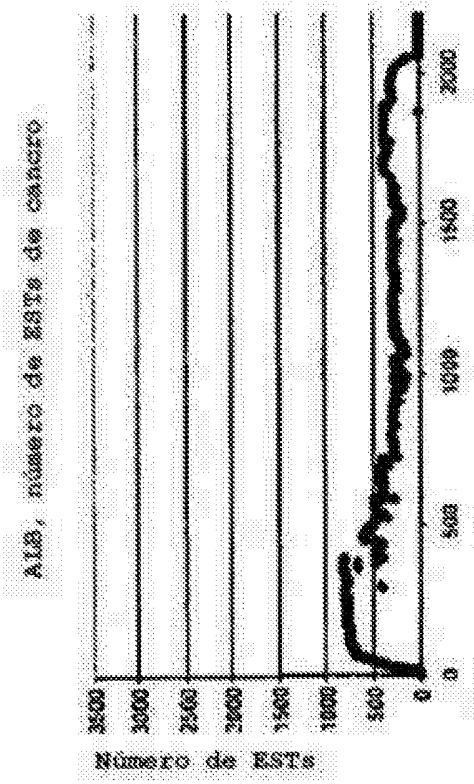


Figura 4a-1

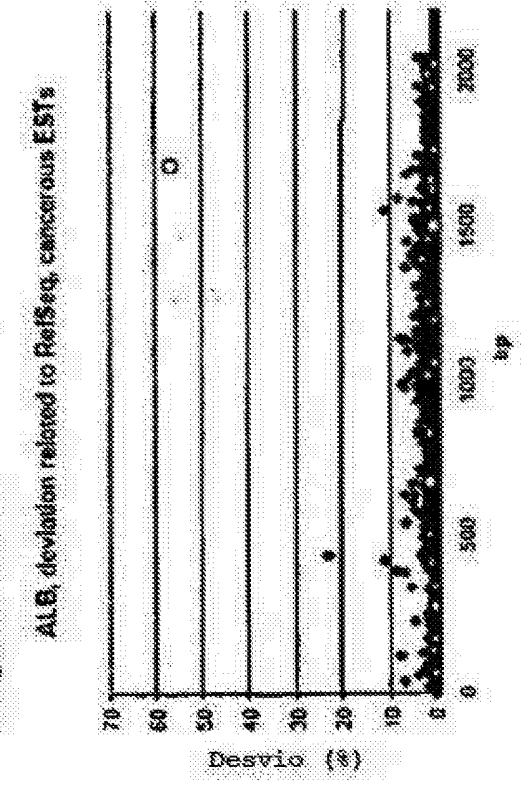


Figura 4a-3

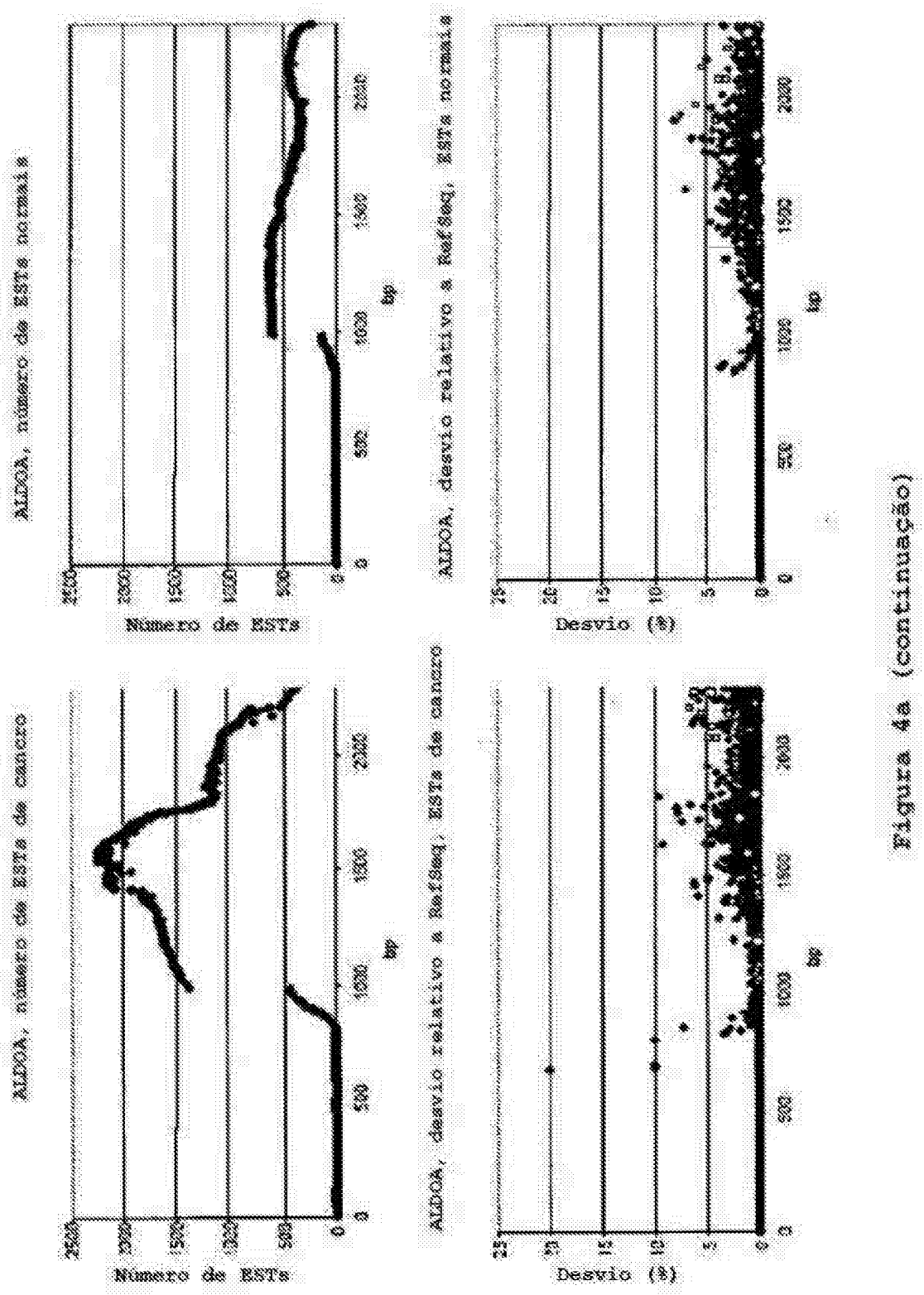


Figura 4a (continuação)

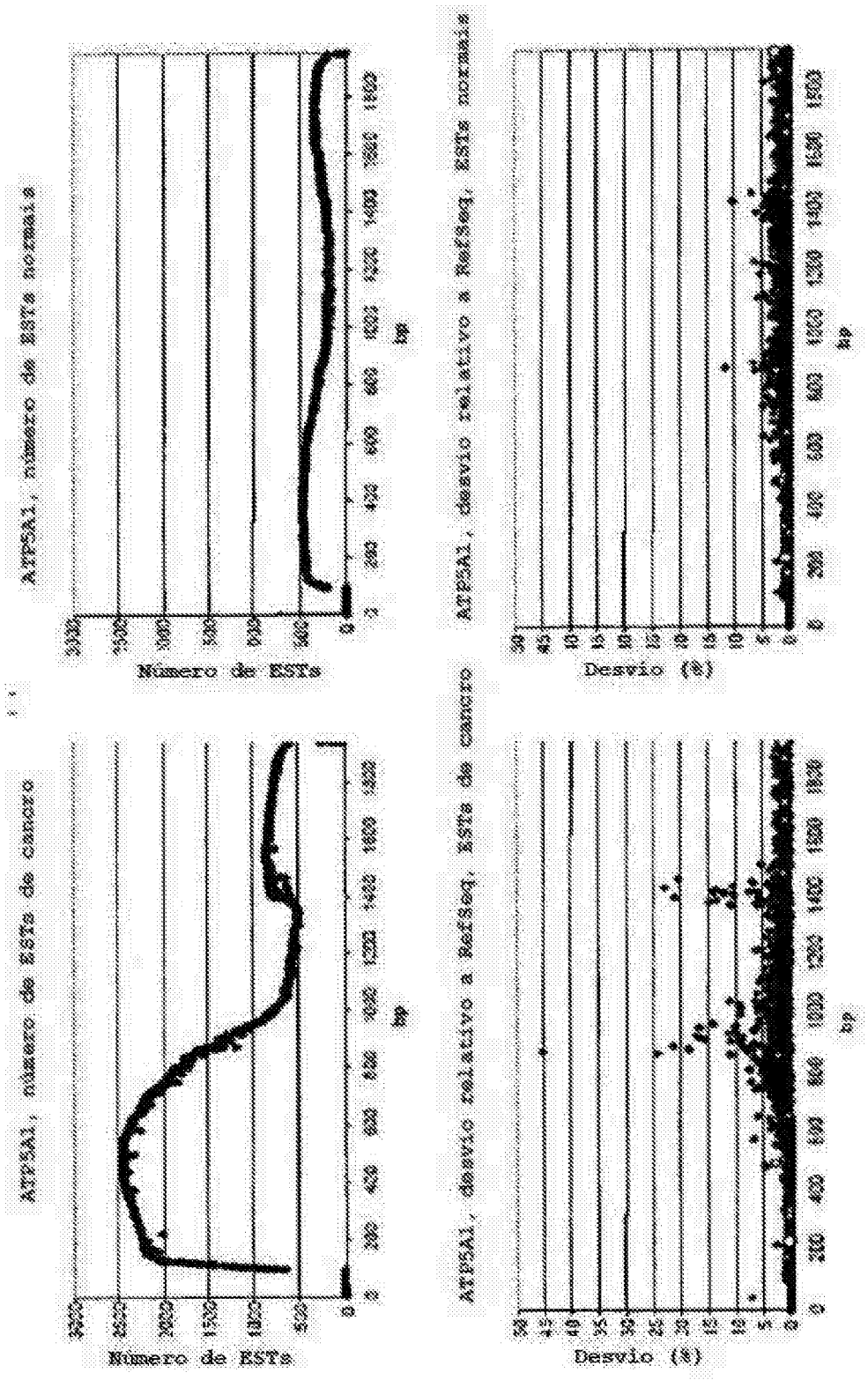


Figura 4a (continuação)

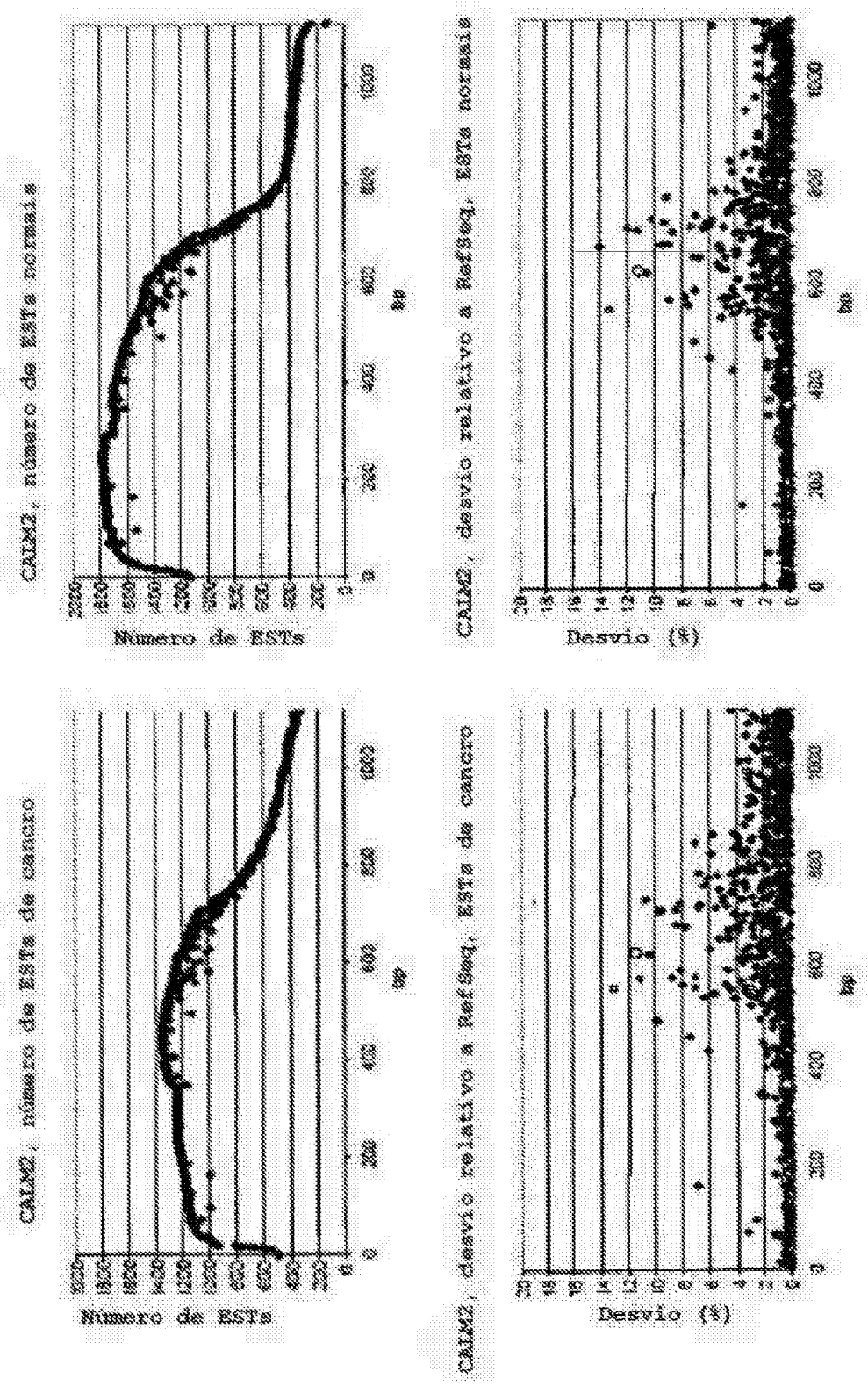


Figura 4a (continuação)

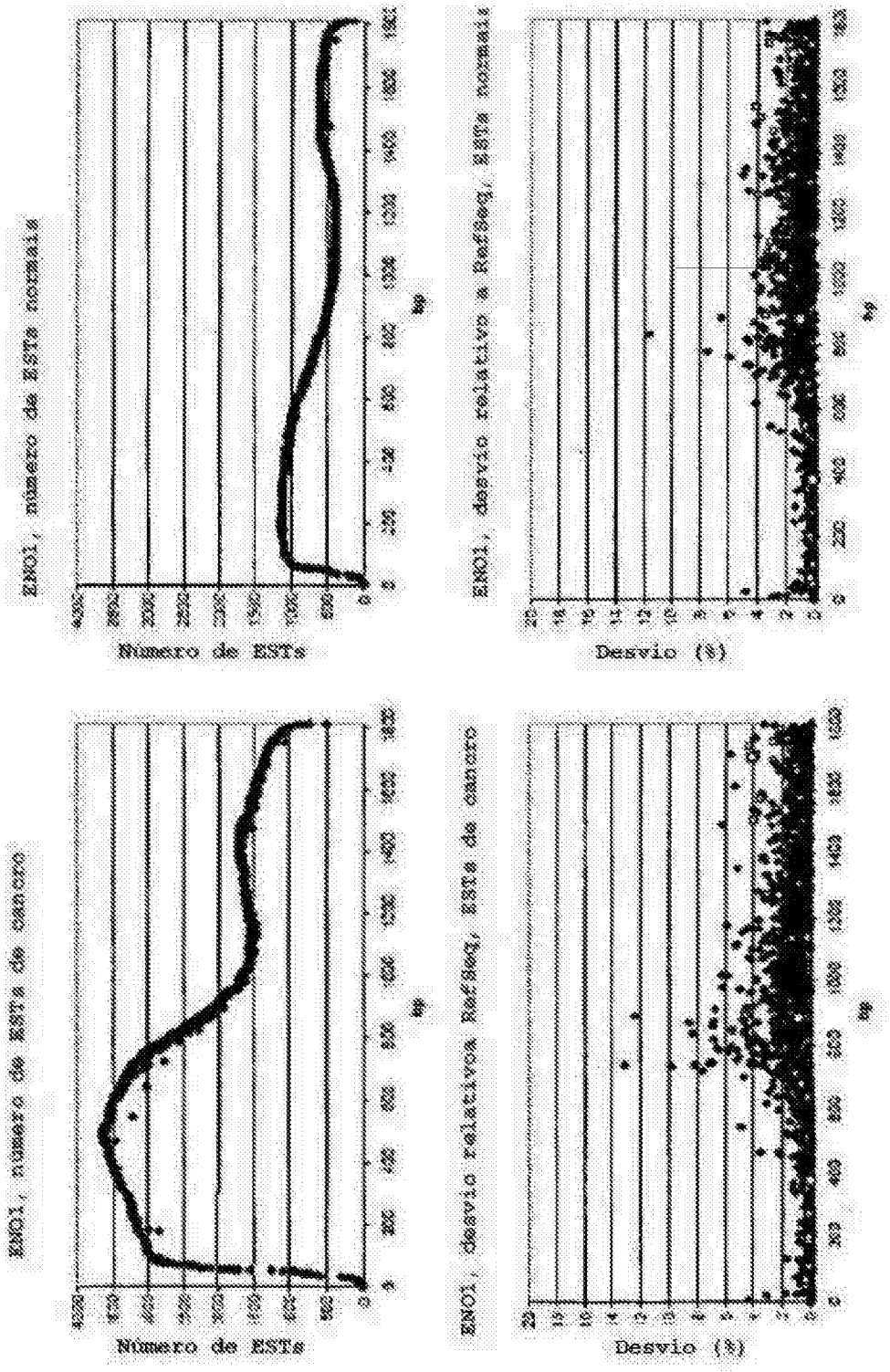

ENO1, número de ESTs normais
Número de ESTs
kp
ENO1, desvio relativo a RefSeq, ESTs normais
Desvio (%)
kp
ENO1, número de ESTs de câncer
Número de ESTs
kp
ENO1, desvio relativo RefSeq, ESTs de câncer
Desvio (%)
kp

Figura 4b

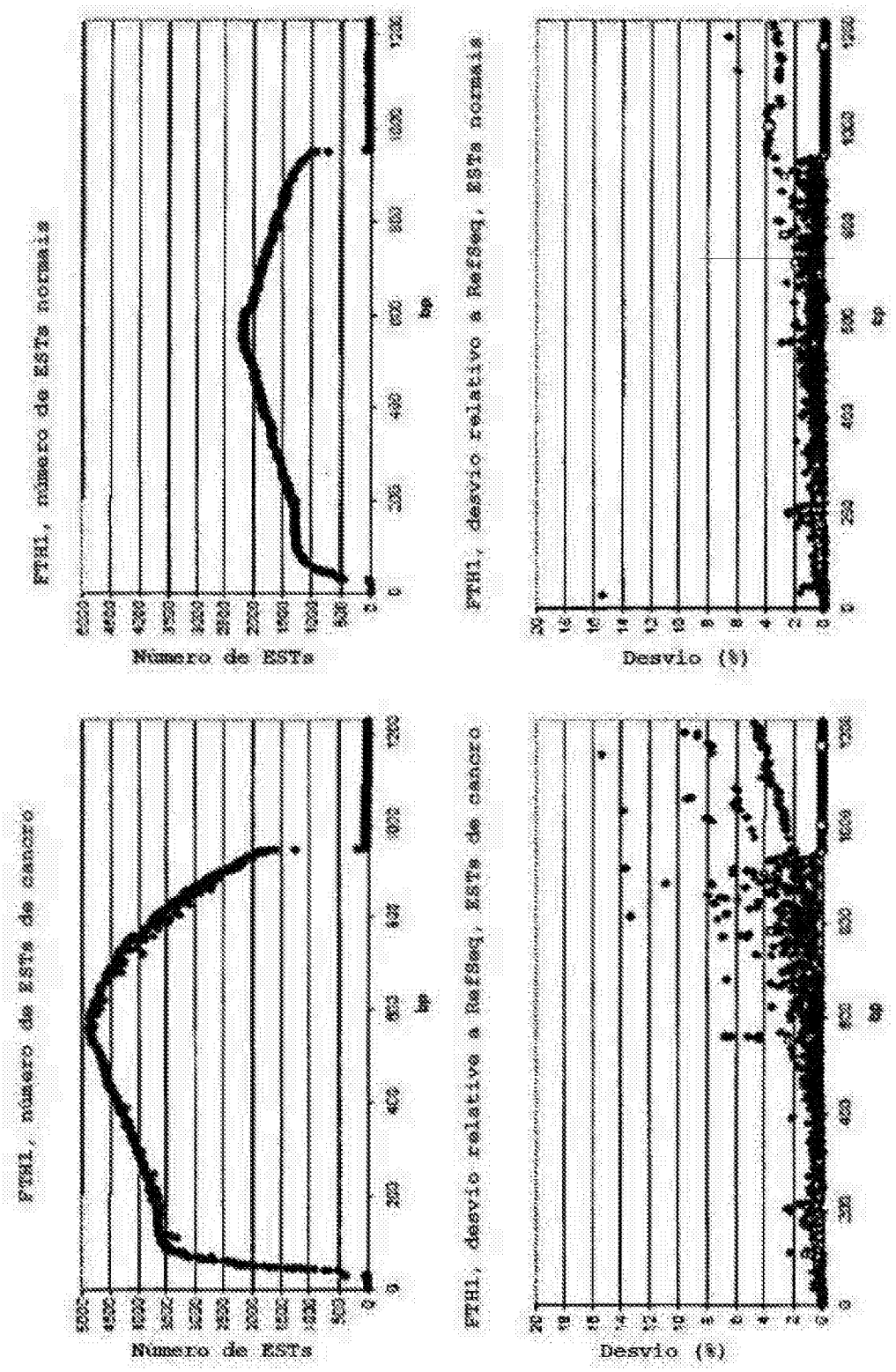


Figura 4b (continuação)

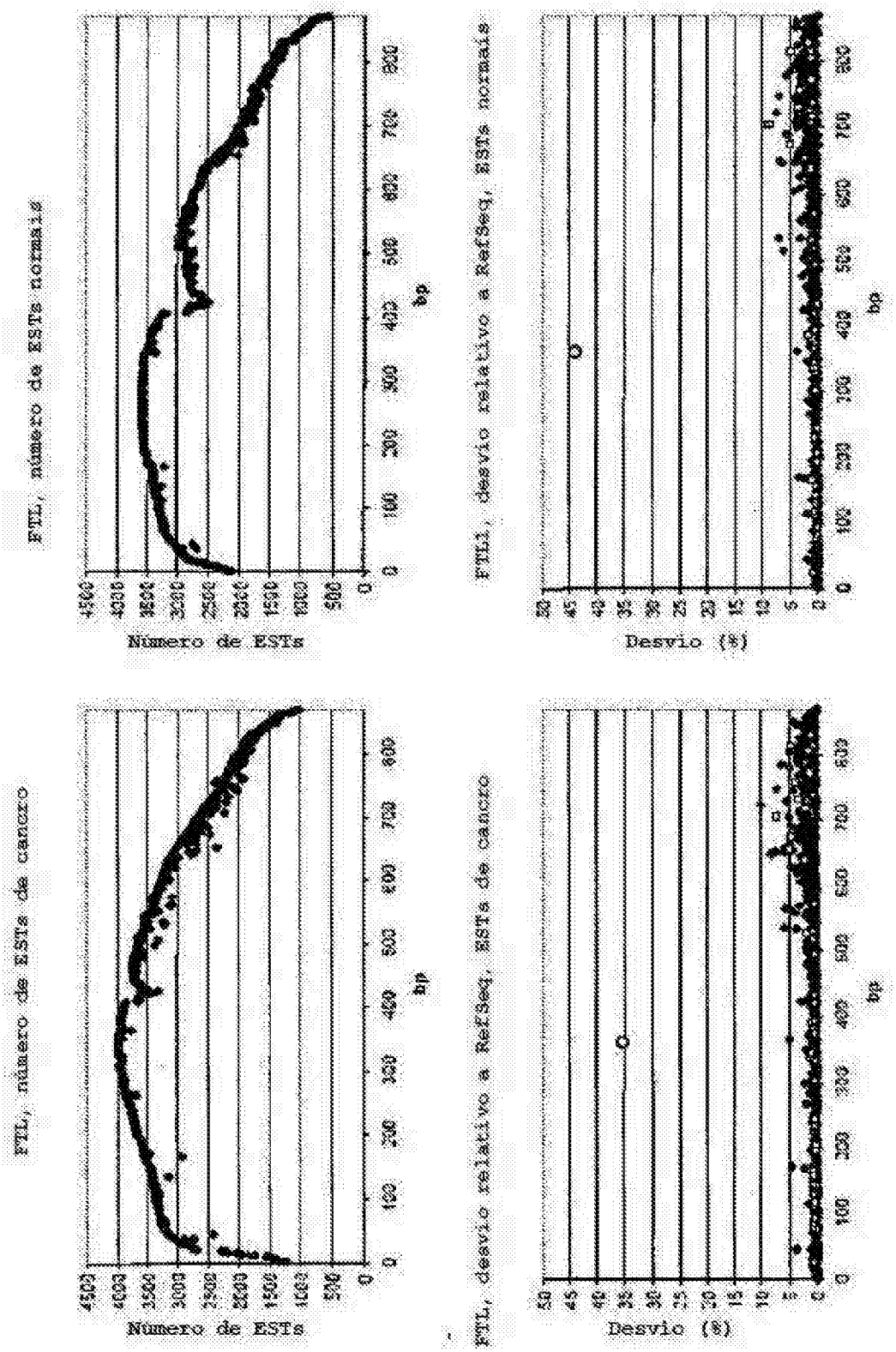


Figura 4b(continuação)

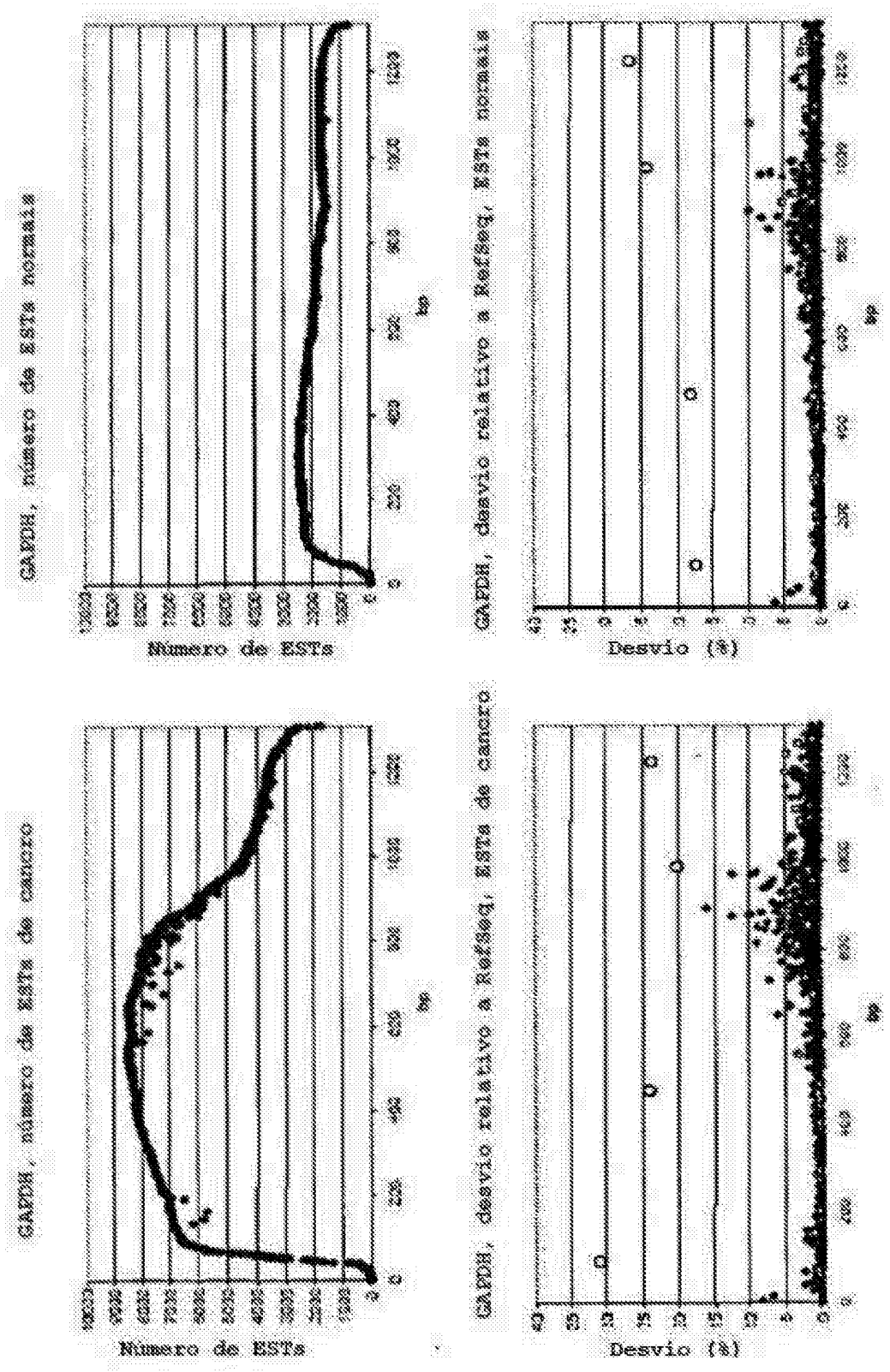


Figura 4b (continuação)

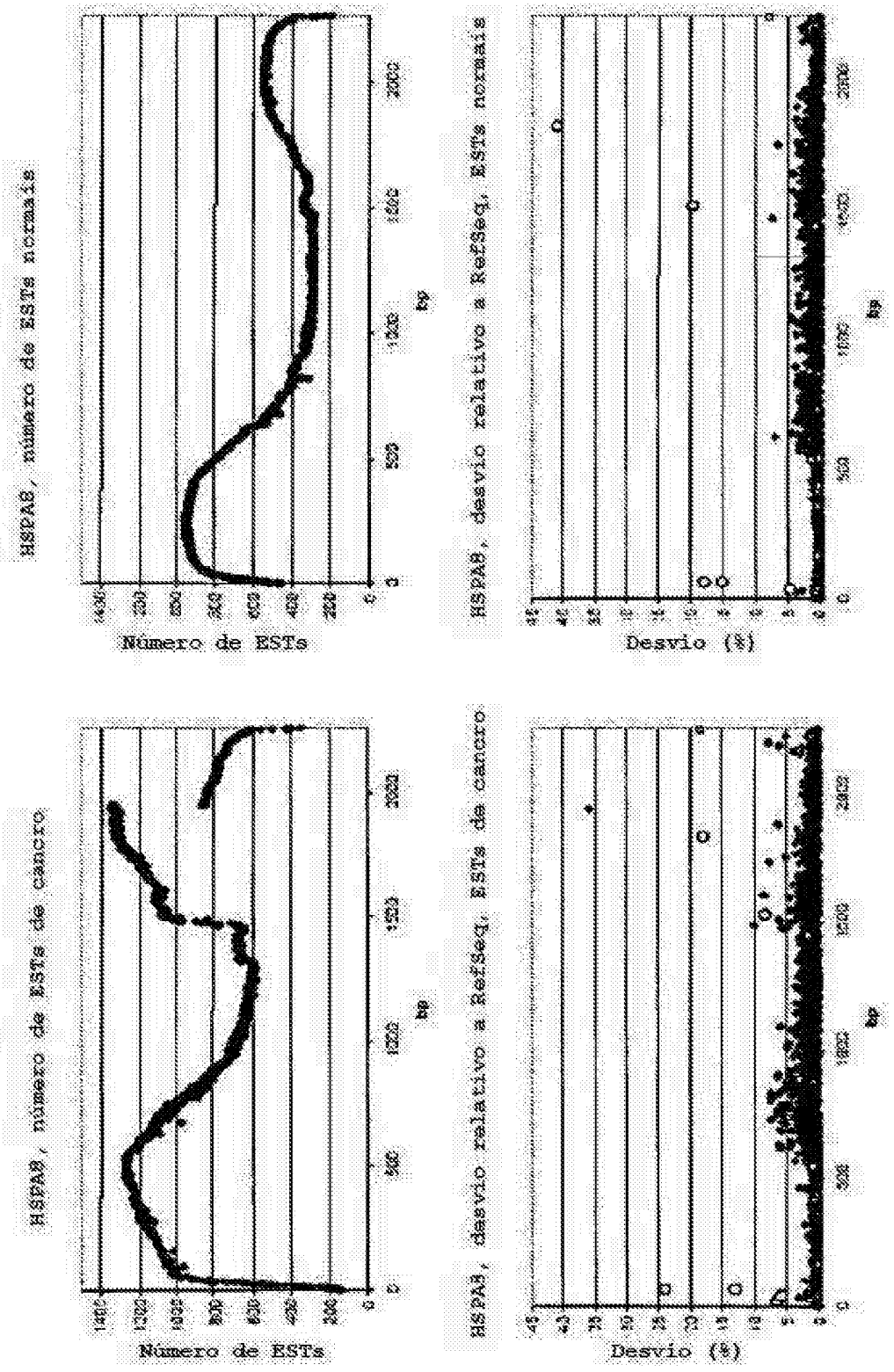

HSPA8, número de ESTs normais
Número de ESTs
HSPA8, desvio relativo a RefSeq, ESTs normais
Desvio (%)
HSPA8, número de ESTs de cancro
Número de ESTs
HSPA8, desvio relativo a RefSeq, ESTs de cancro
Desvio (%)
kb

Figura 4c

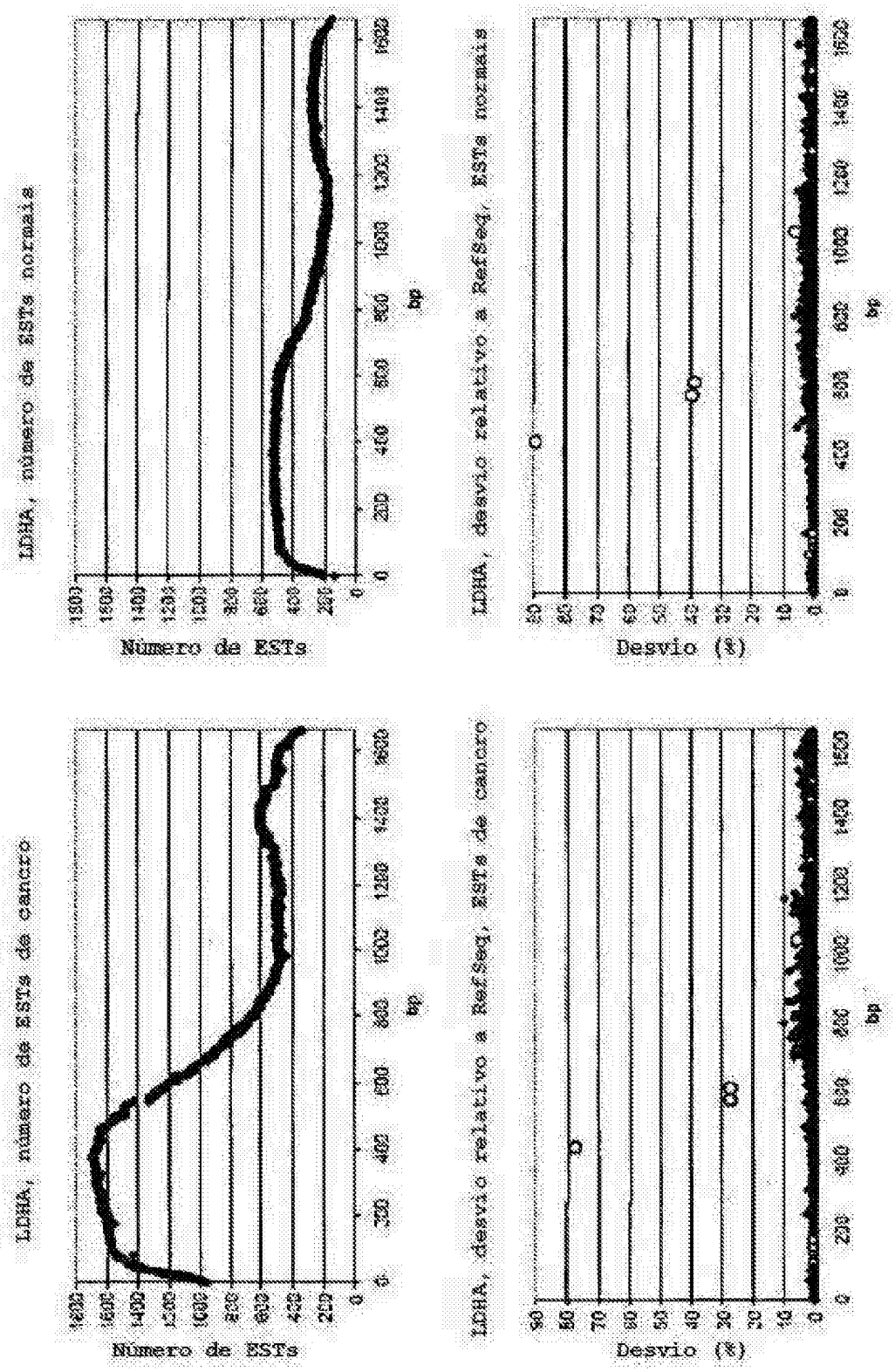


Figura 4c (continuação)

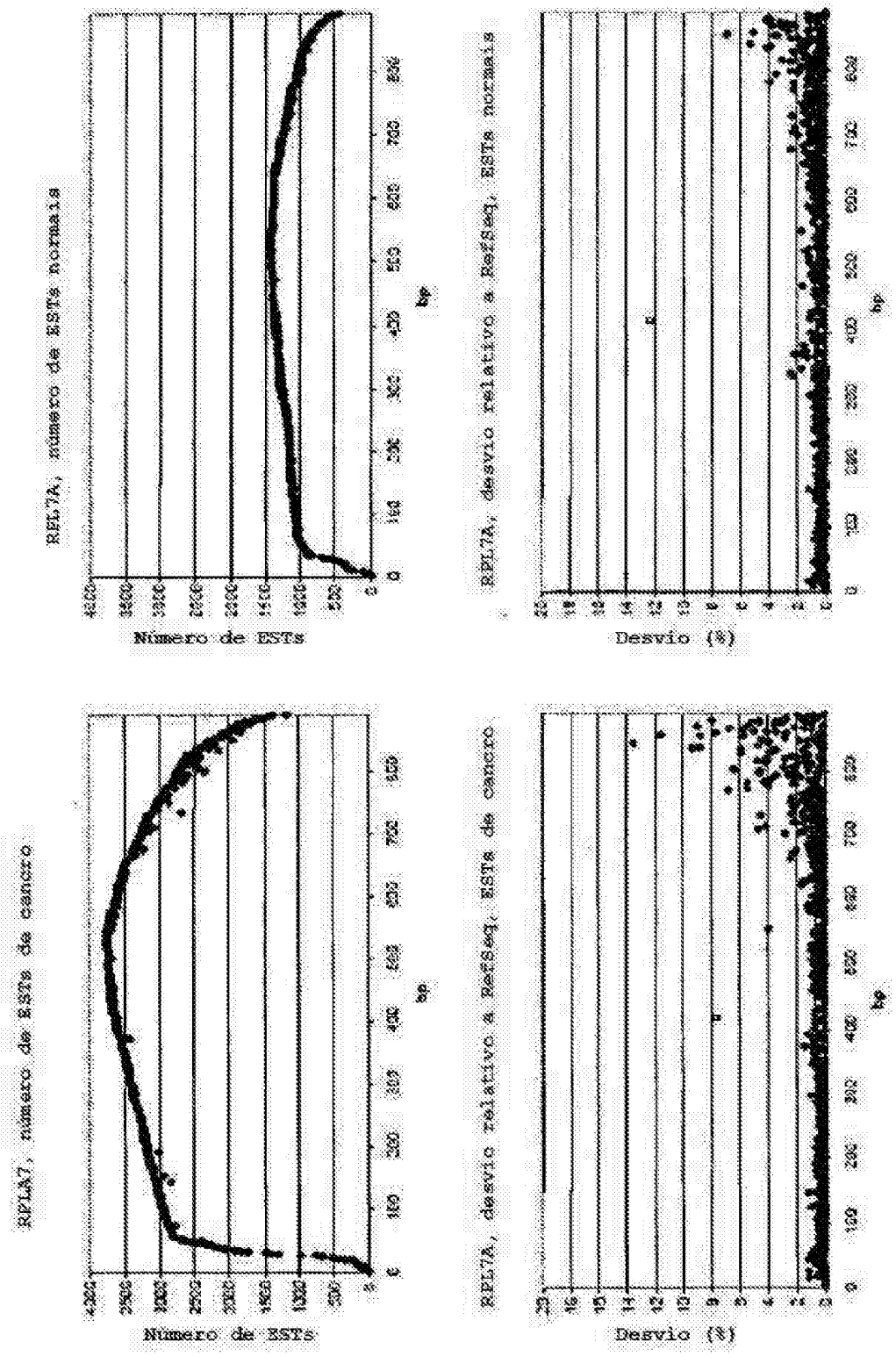


Figura 4c (continuação)

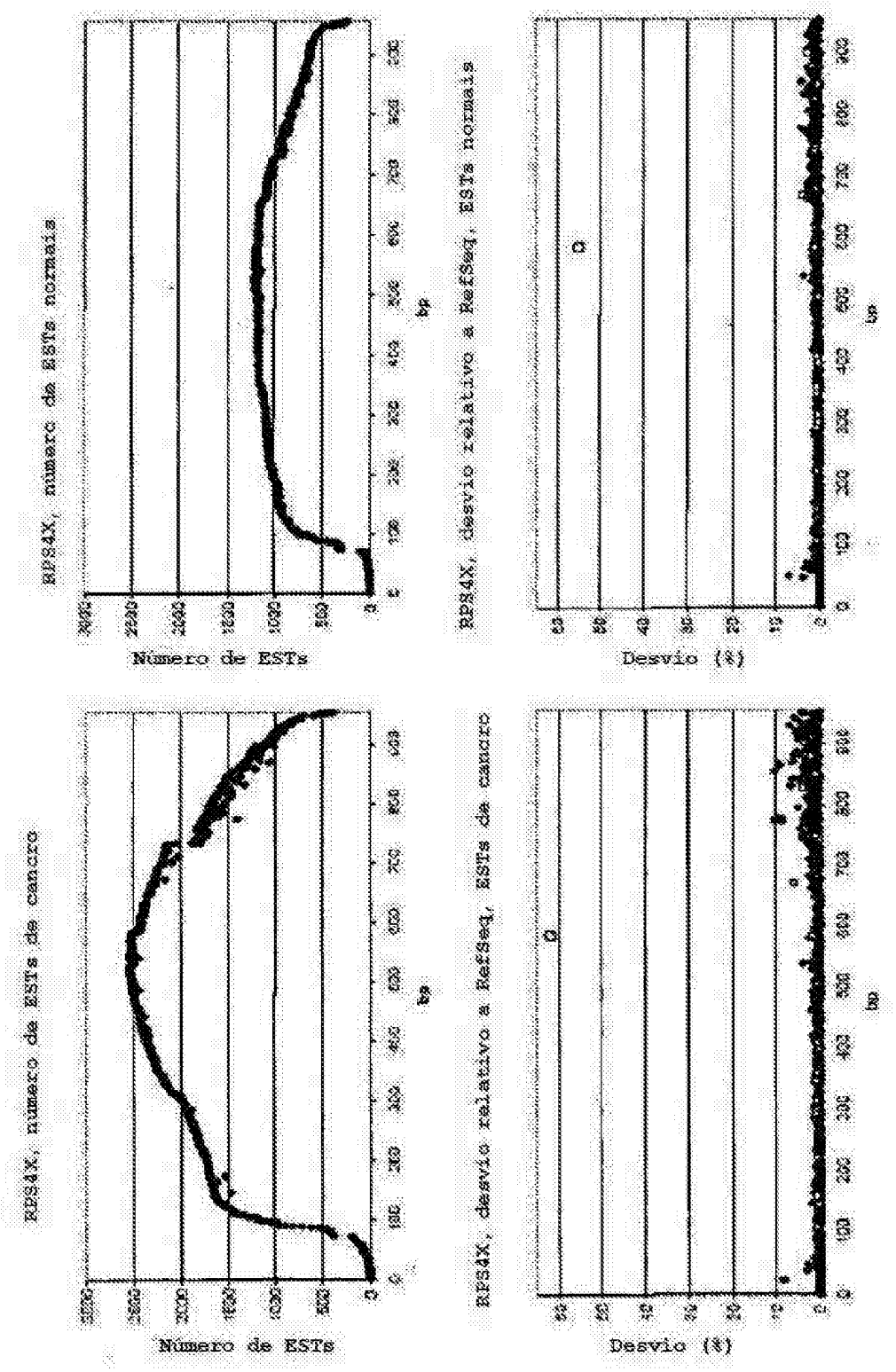


Figura 4c (continuação)

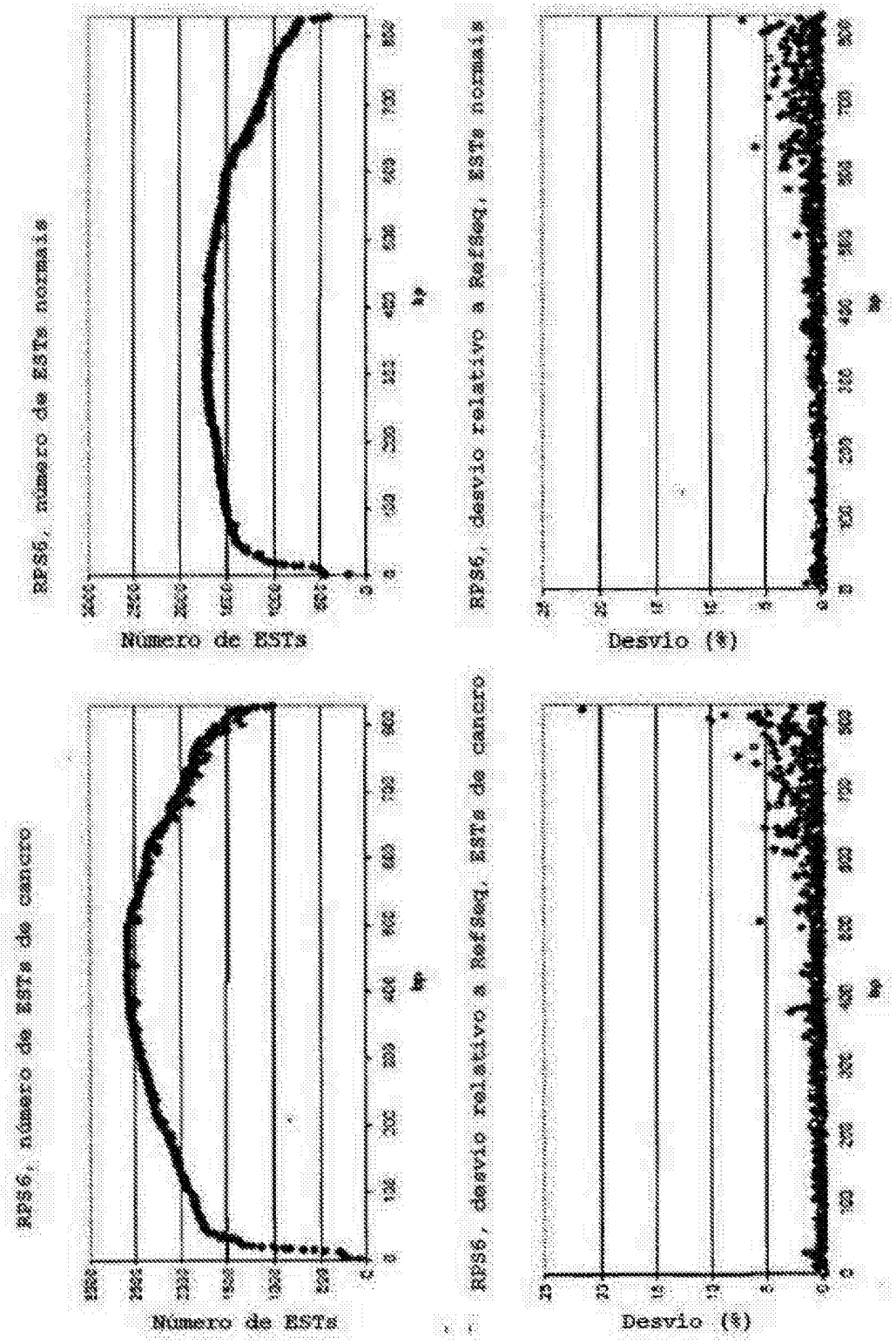

RPS6, número de ESTs normais
Número de ESTs
RPS6, desvio relativo a RefSeq, ESTs normais
Desvio (%)
RPS6, número de ESTs de cancro
Número de ESTs
RPS6, desvio relativo a RefSeq, ESTs de cancro
Desvio (%)

Figura 4d

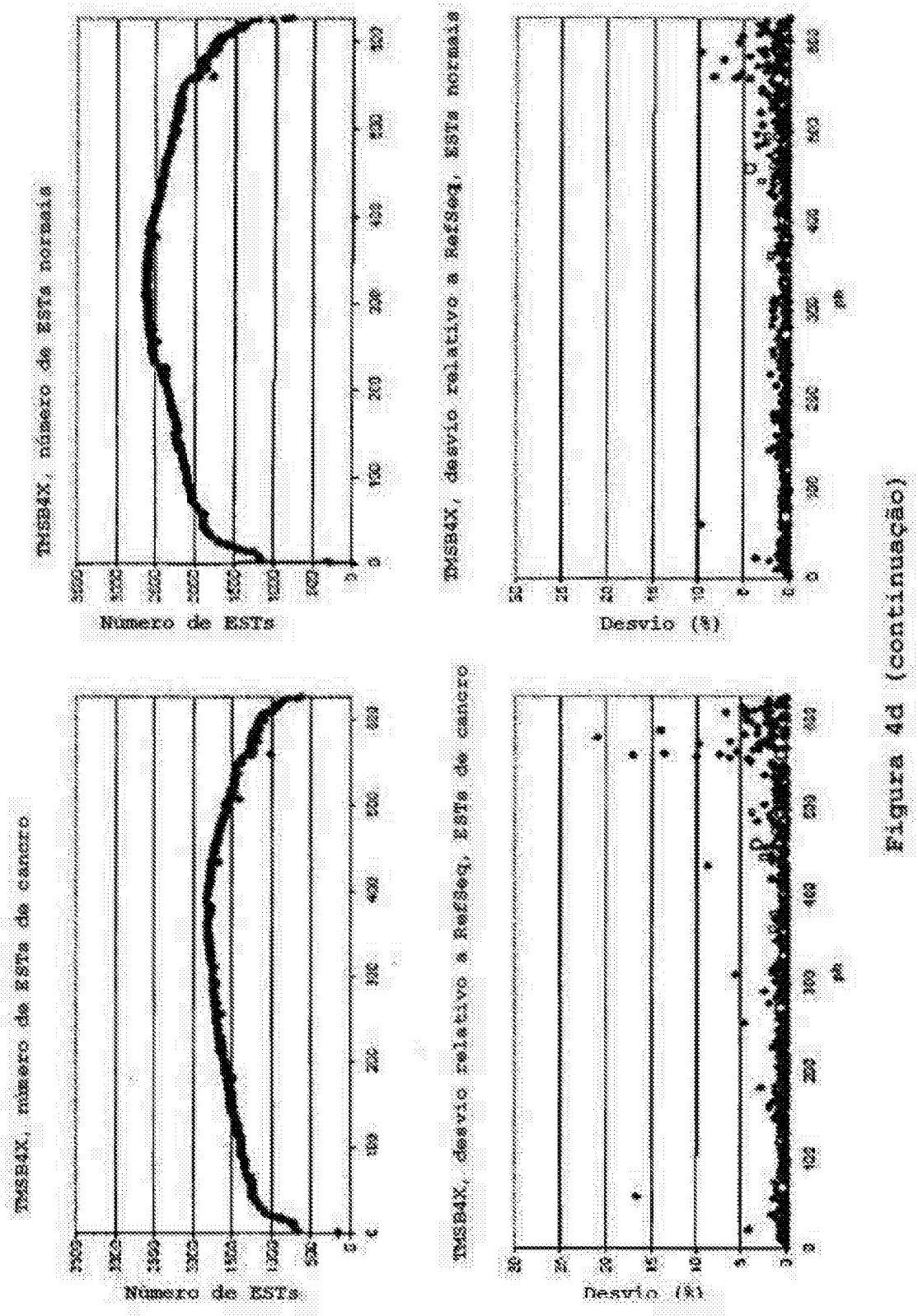


Figura 4d (continuação)

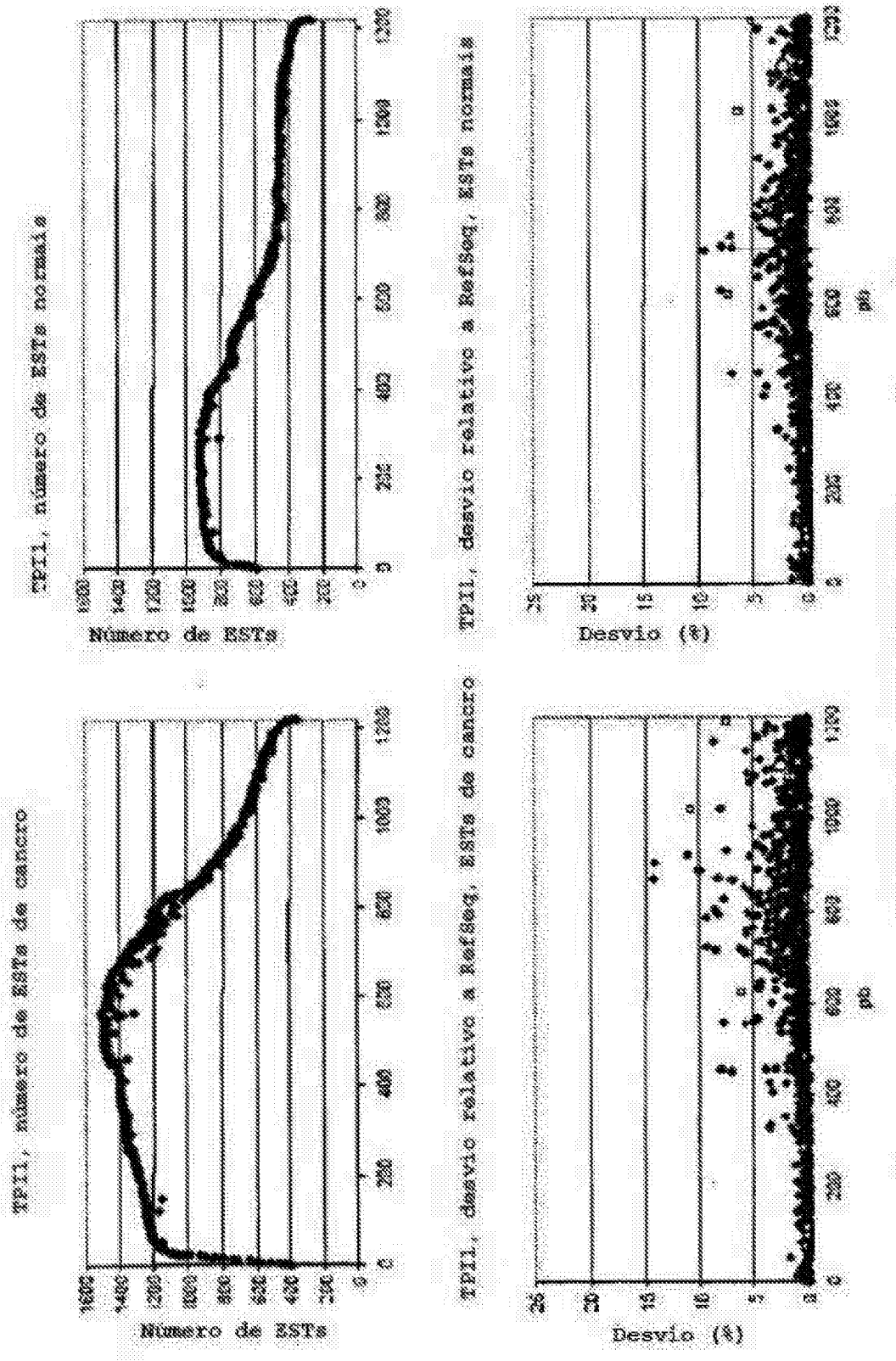


Figura 4d (continuação)

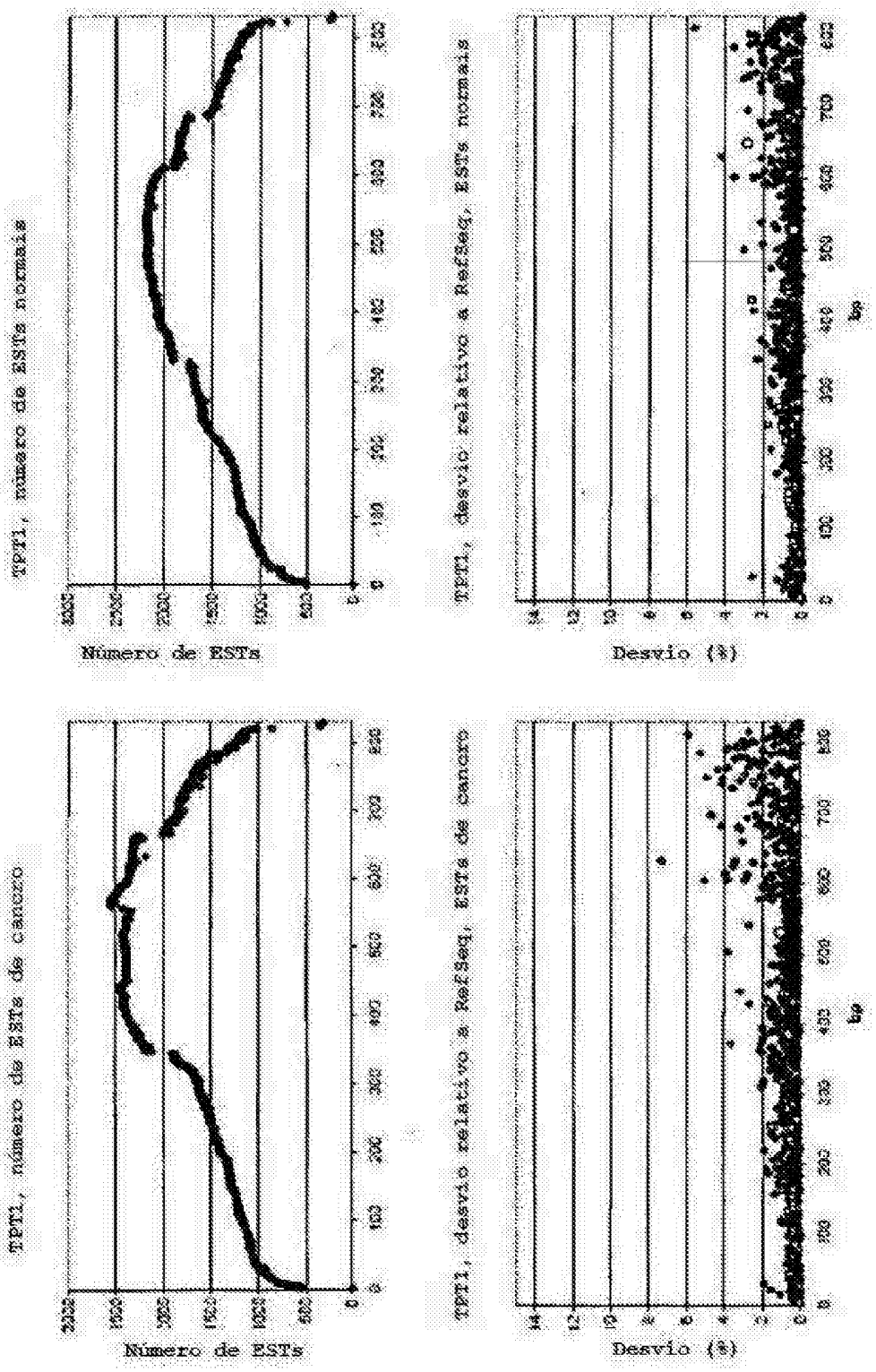


Figura 4d (continuação)

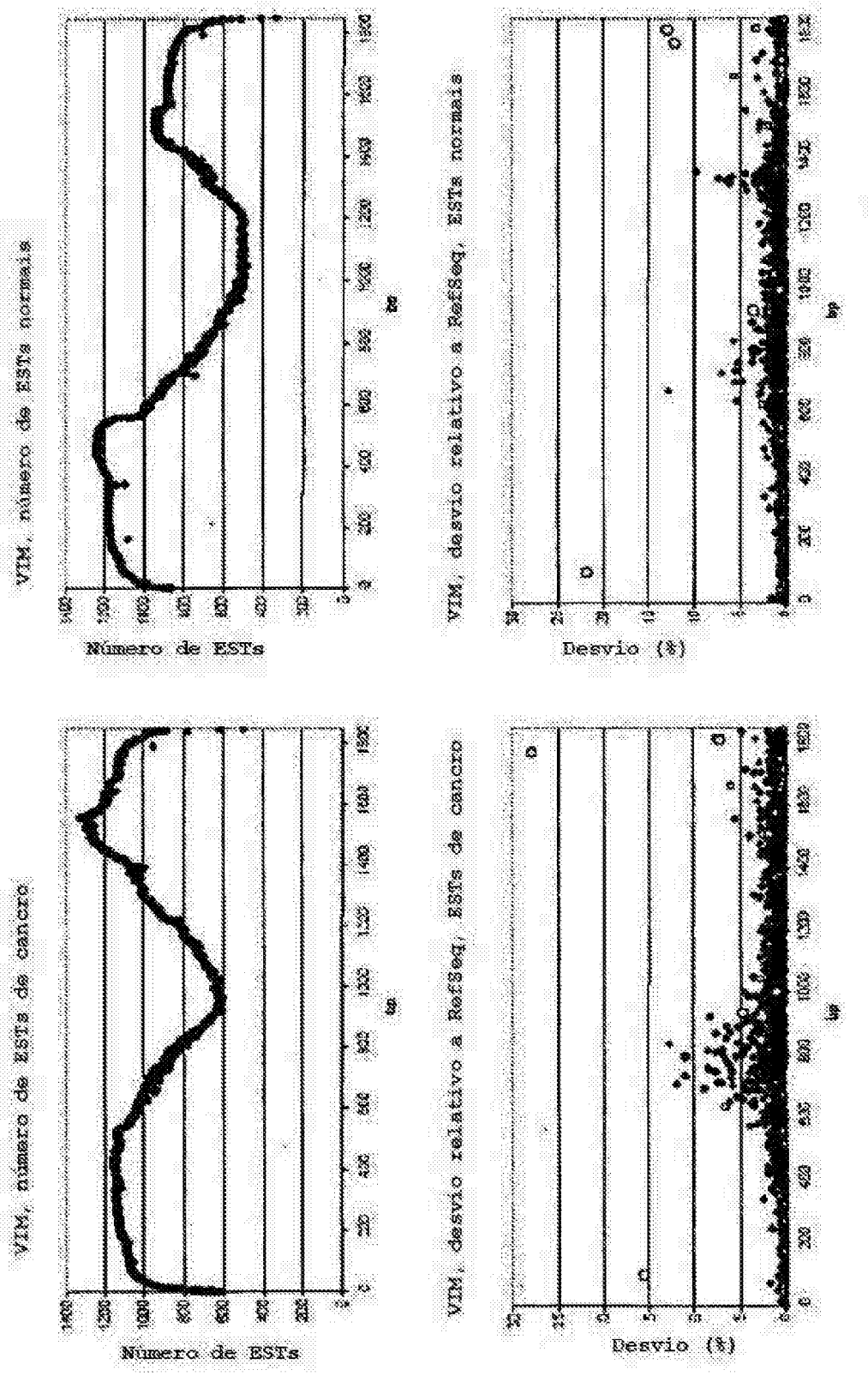


Figura 4d (continuação)

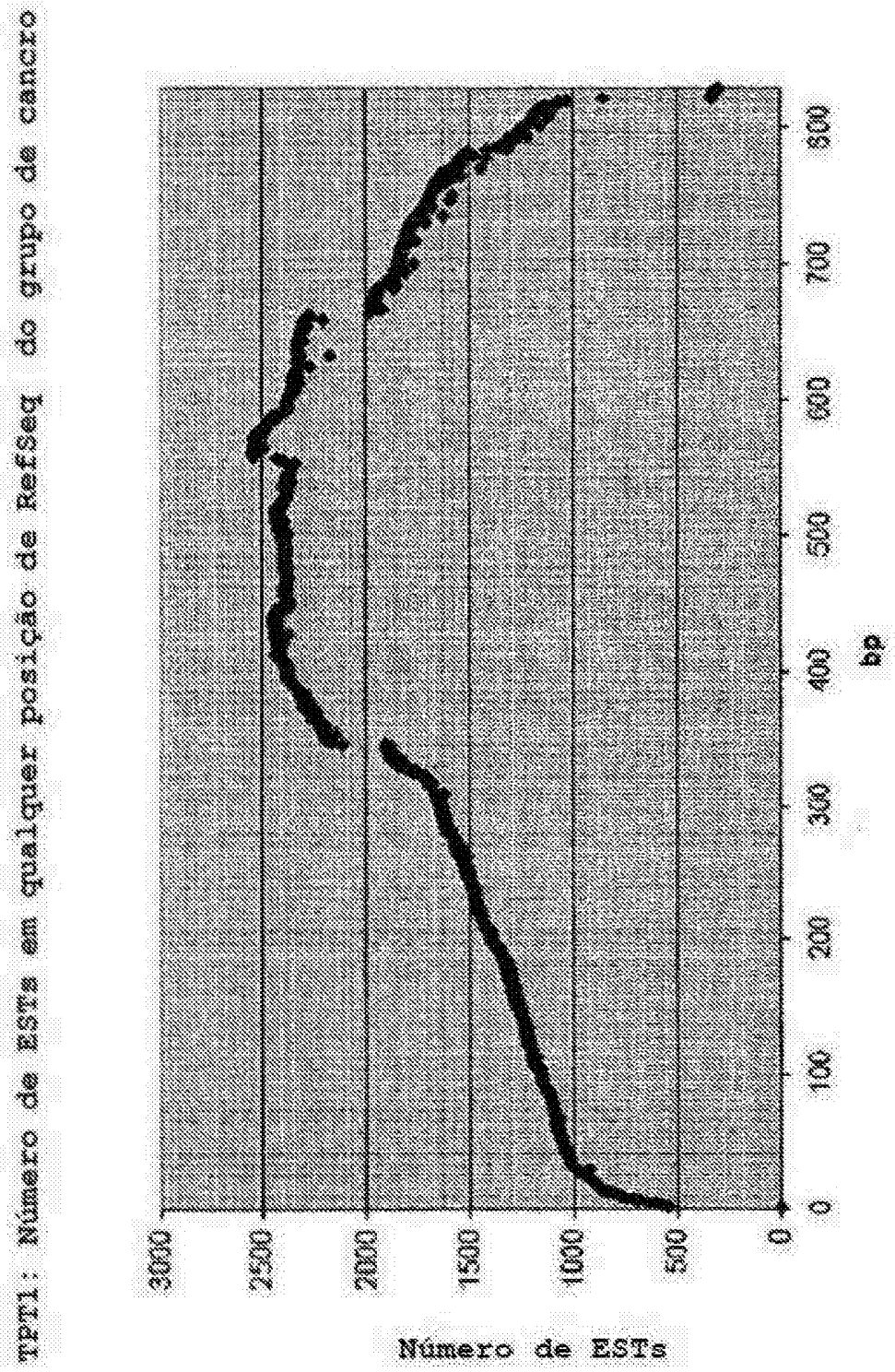

TPT1: Número de ESTs em qualquer posição de RefSeq do grupo de cancro
Número de ESTs
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
bp

Figura 5a

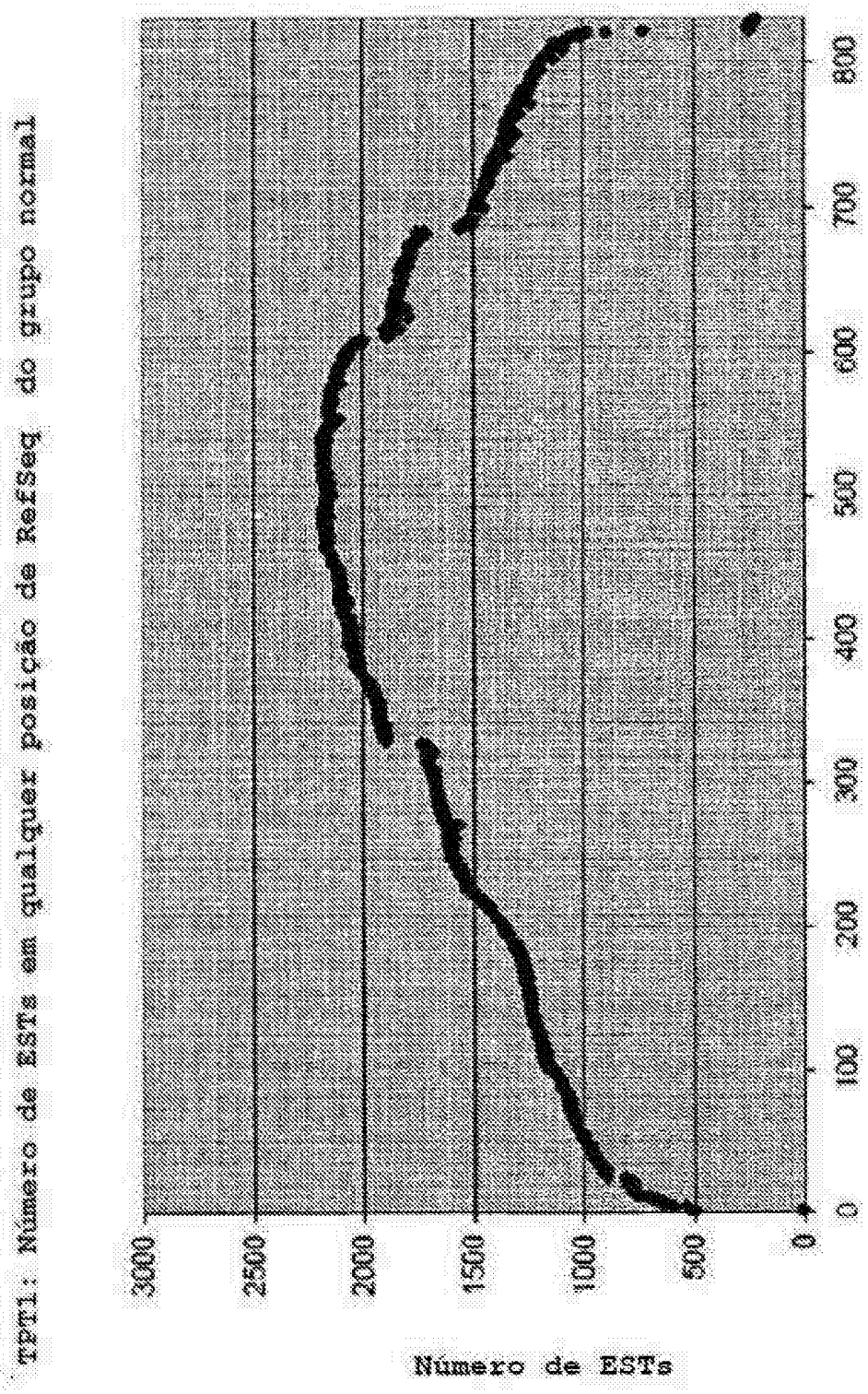

TPTL: Número de ESTs em qualquer posição da RefSeq do grupo normal
Número de ESTs
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
bp

Figura 5b

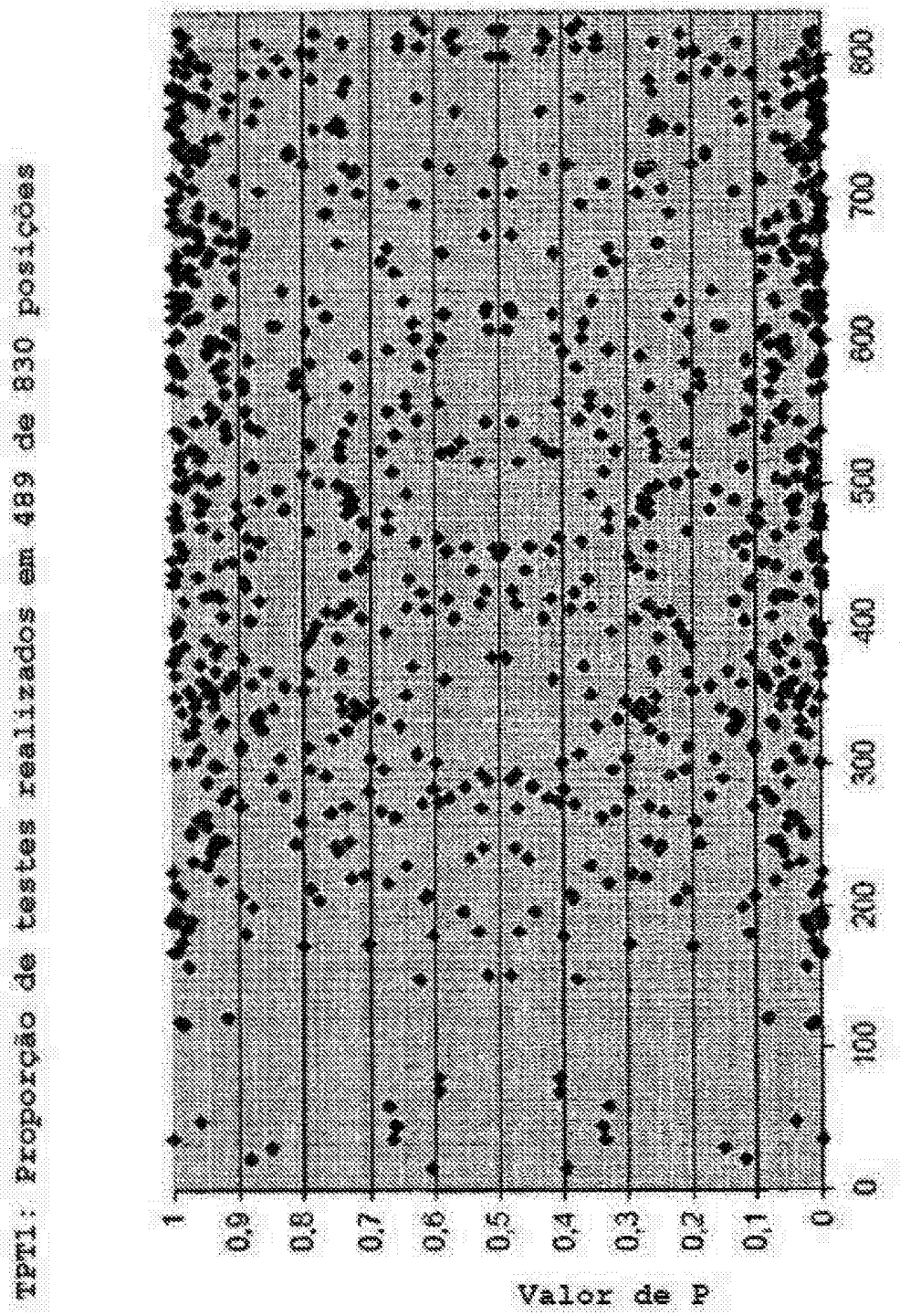

TPTI: Proporção de testes realizados em 489 de 830 posições
Valor de P
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
100
200
300
400
500
600
700
800
bp

Figura 5c

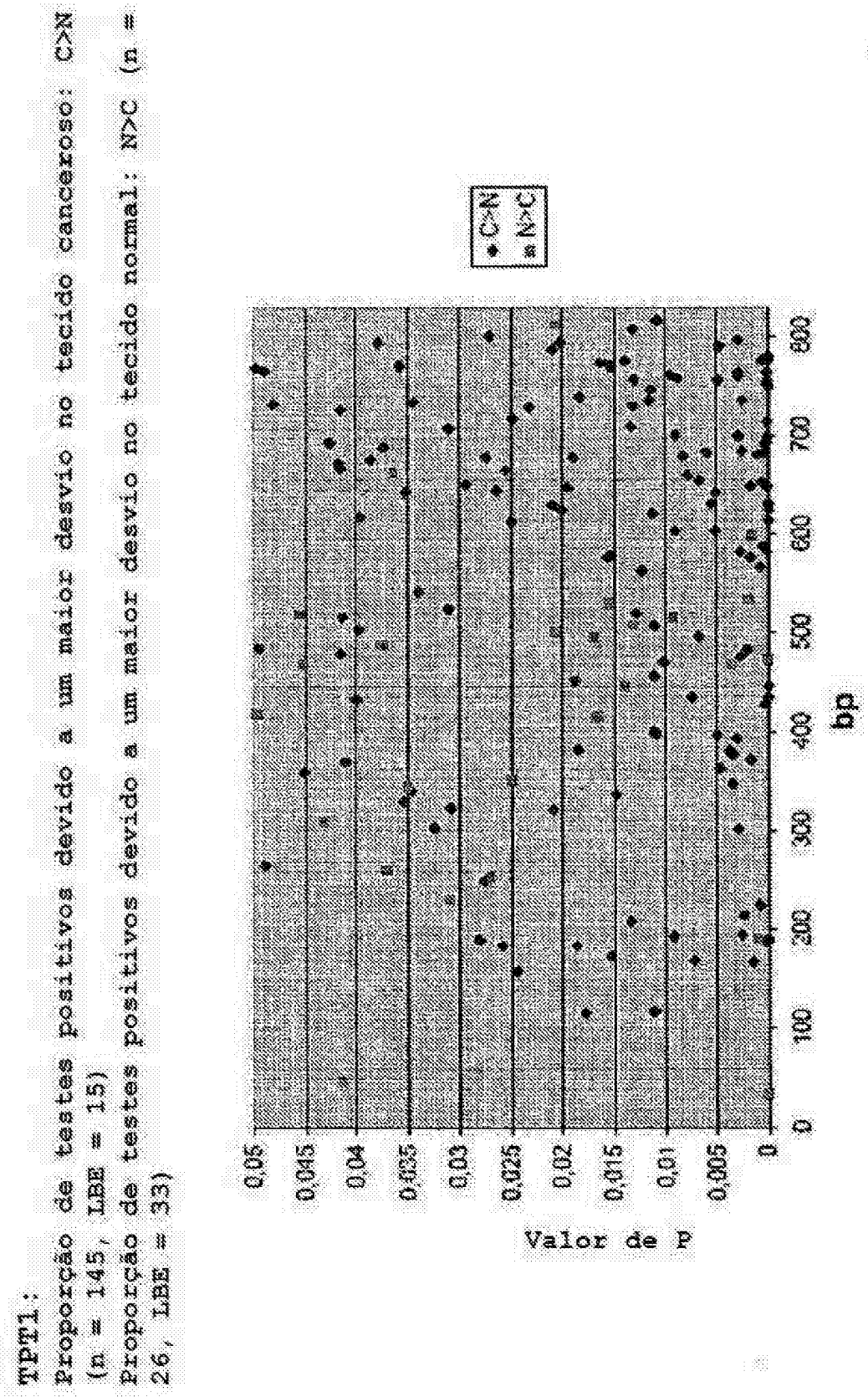

TPT1:
Proporção de testes positivos devido a um maior desvio no tecido canceroso: C>N (n = 145, LBE = 15)
Proporção de testes positivos devido a um maior desvio no tecido normal: N>C (n = 26, LBE = 33)
C>N
N>C
Valor de P
0,05
0,045
0,04
0,035
0,03
0,025
0,02
0,015
0,01
0,005
0
0
100
200
300
400
500
600
700
800
bp

Figura 5d

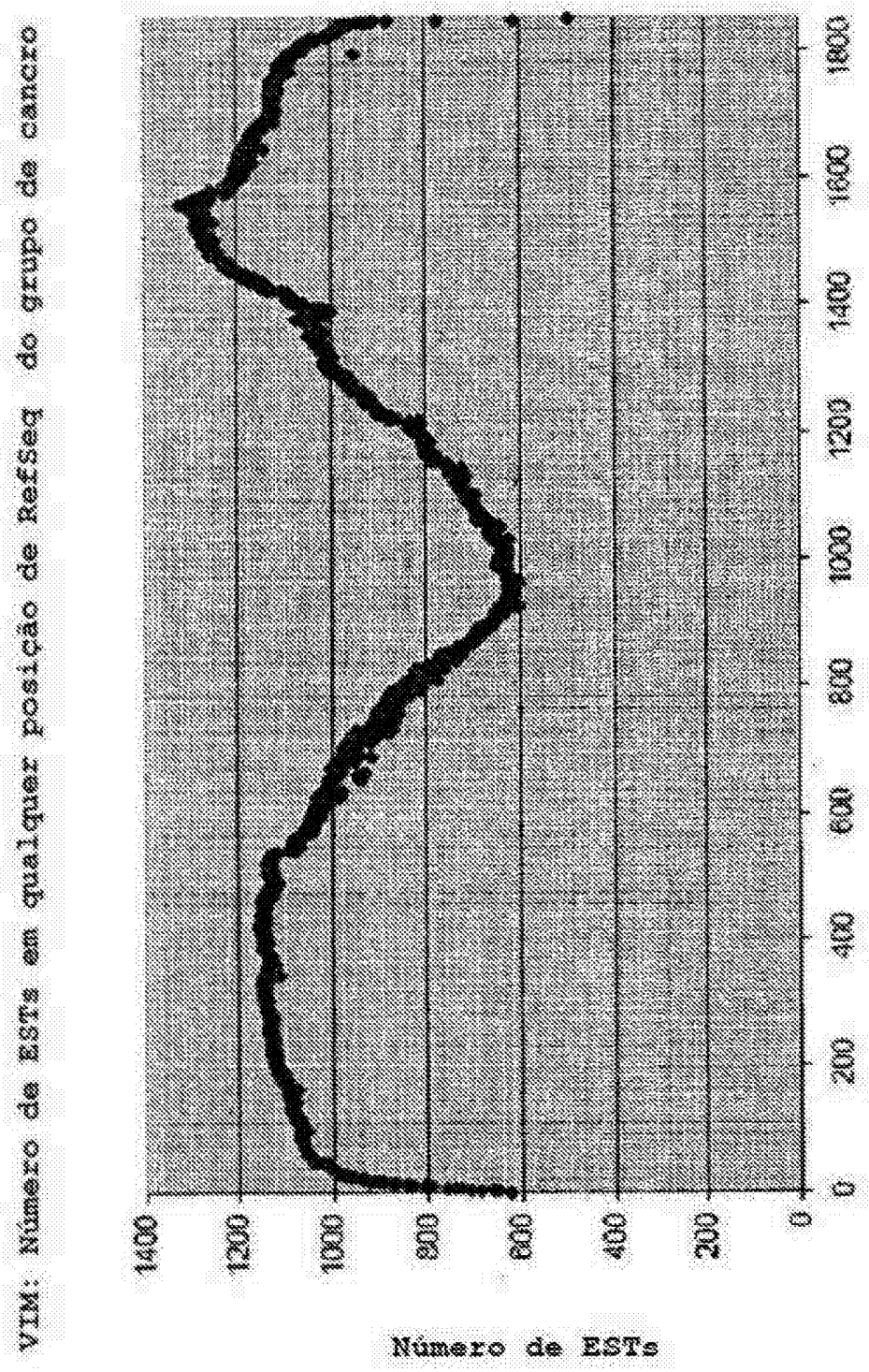

VIM: Número de ESTs em qualquer posição de RefSeq do grupo de cancro
Número de ESTs
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600
1800
bp

Figura 5e

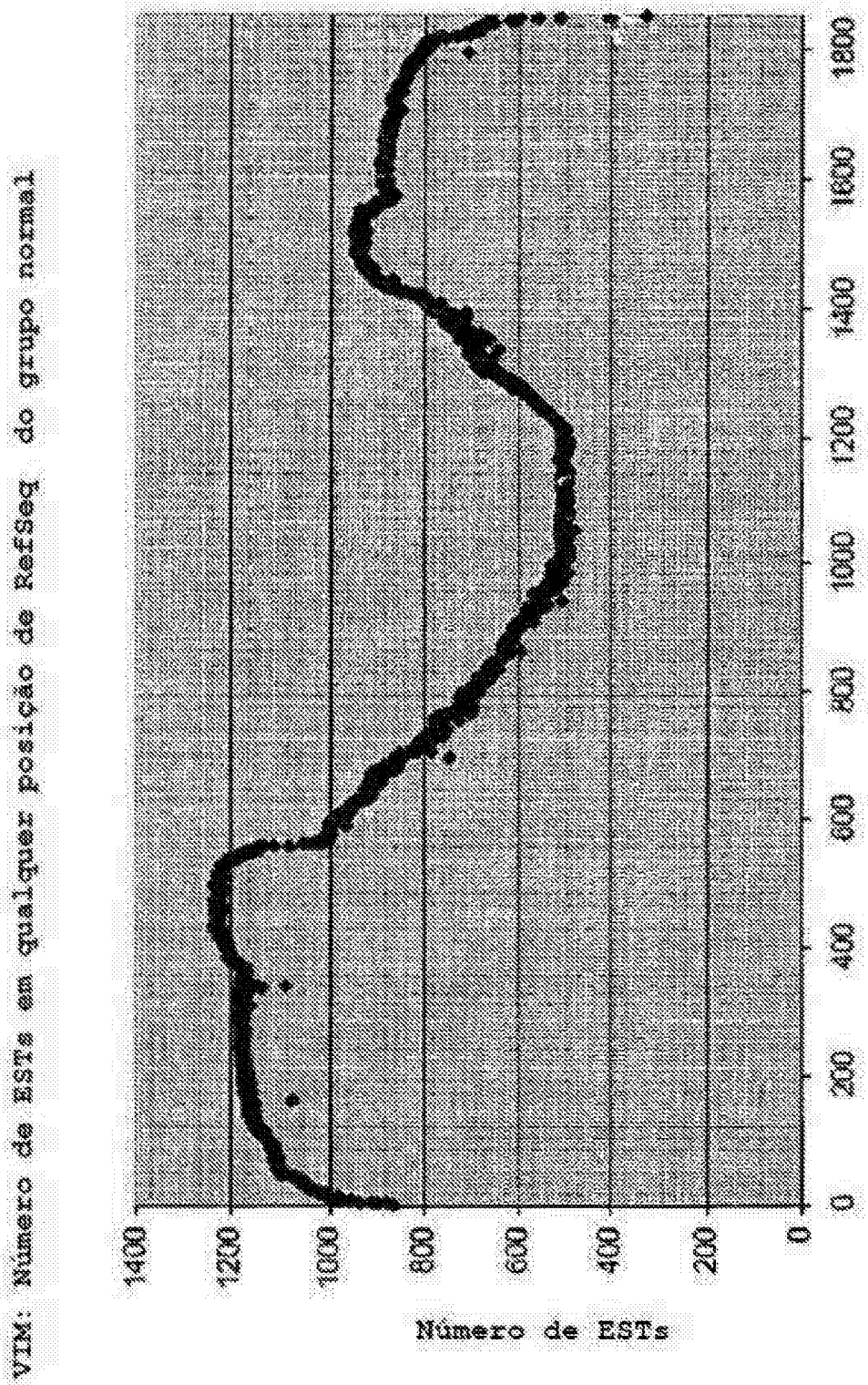

VIM: Número de ESTs em qualquer posição de RefSeq do grupo normal
Número de ESTs
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600
1800
bp

Figura 5E

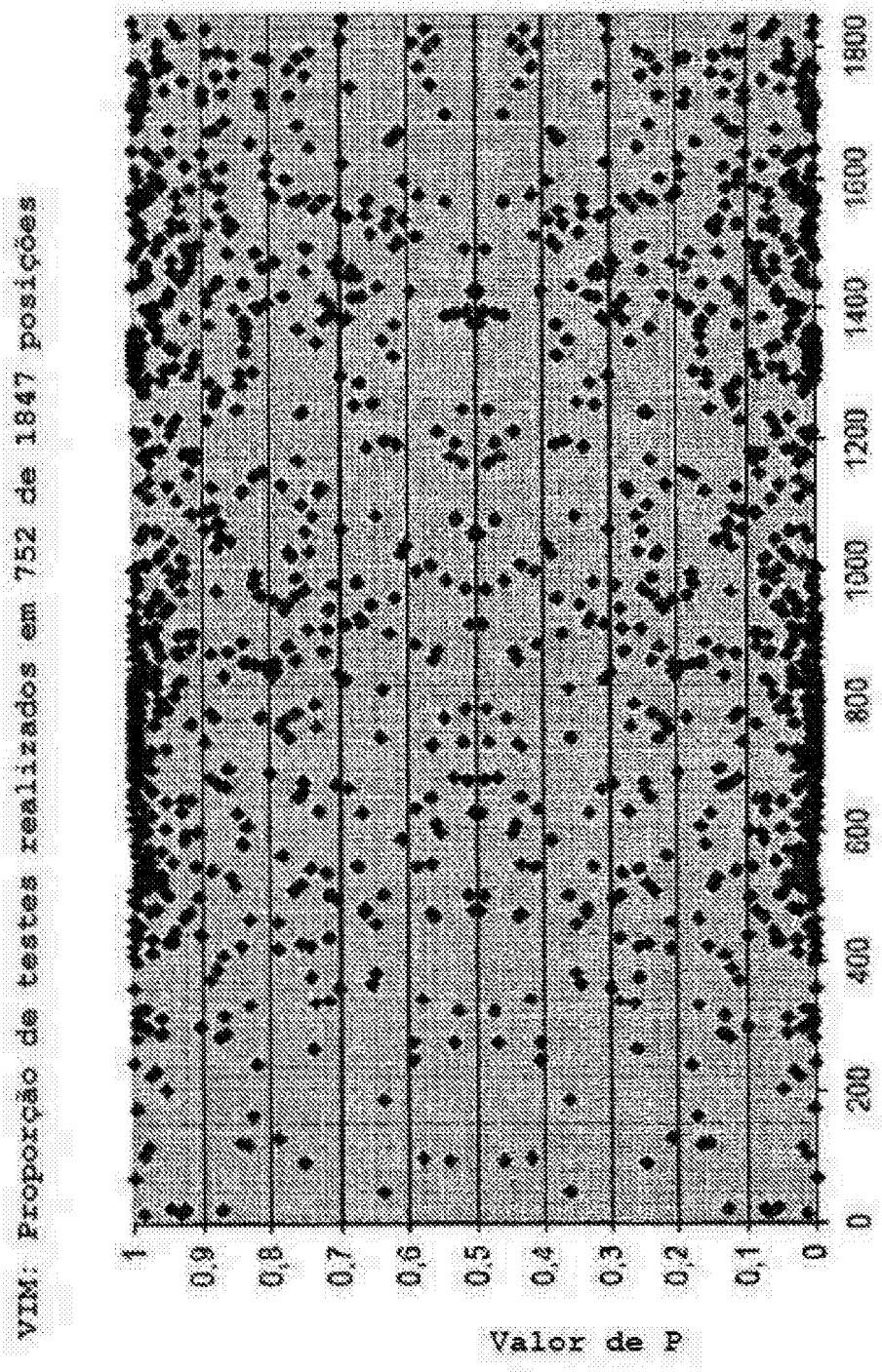

VIM: Proporção de testes realizados em 752 de 1847 posições
Valor de P
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
200
400
600
800
1000
1200
1400
1600
1800
bp

Figura 5g

VIM:

Proporção de testes positivos devido a um maior desvio no tecido canceroso: C>N (n = 269, LBE = 24)

Proporção de testes positivos devido a um maior desvio no tecido normal: N>C (n = 78, LBE = 50)

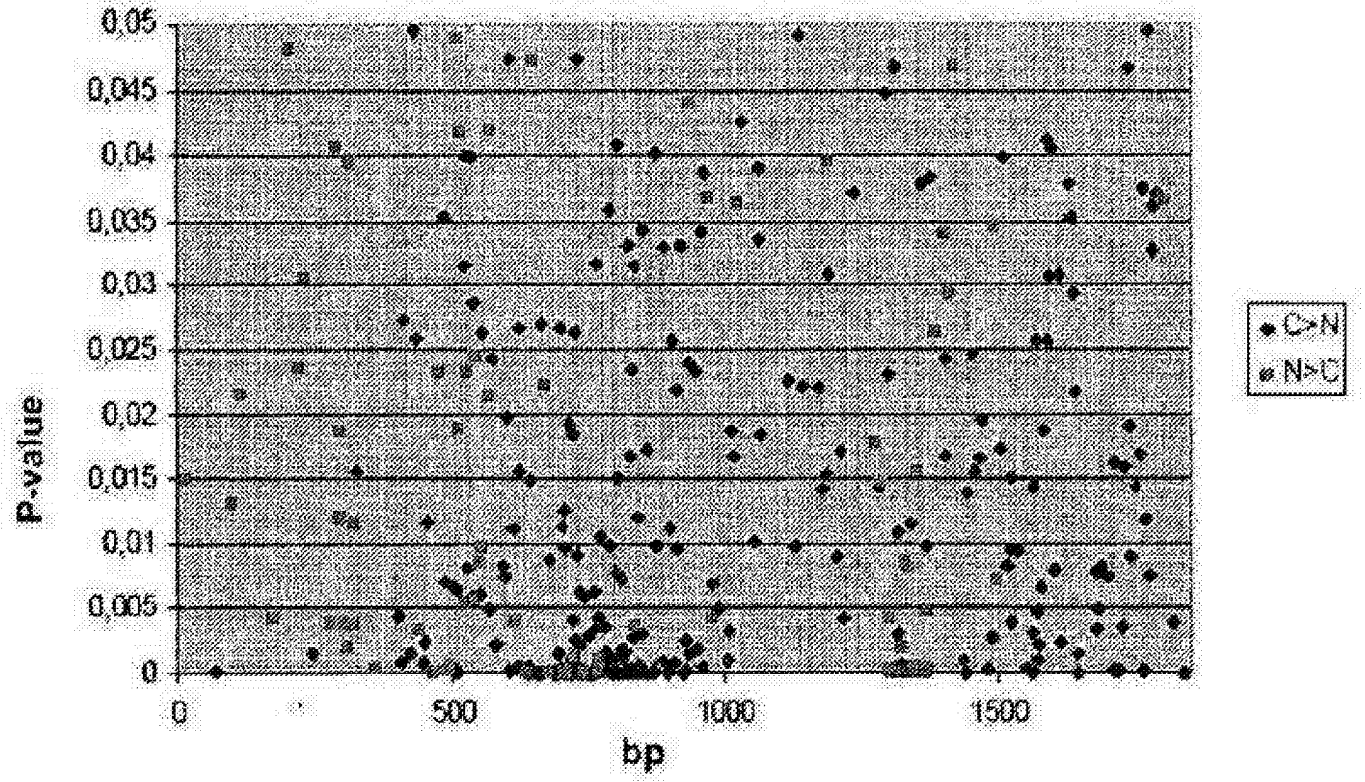


Figura 5h

Análise do teste de proporções: resultados

Gene	Número de testes de proporções	% de cobertura	C>N	LBE	N>C	LBE
ALB	194	9	38	10	56	8
ALDOA transcript1	336	30	81	11	35	21
ATP5A1 transcript1	239	38	123	3	6	20
CALM2	374	23	69	16	40	21
ENO1	614	27	186	18	31	42
FTH1	592	71	226	20	67	38
FTL	678	56	118	31	86	36
GAPDH	812	91	311	23	61	57
HSPA8 transcript1	513	59	163	13	18	37
LDHA	181	9	70	4	10	13
RPL7A	337	35	103	11	33	22
RPS4X	371	45	124	11	27	25
RPS6	432	19	86	17	40	25
TMSB4X	352	19	70	16	74	17
TPI1	379	31	99	14	47	23
TPT1	489	26	145	15	26	33
VIM	752	57	269	24	78	50

Figura 5i

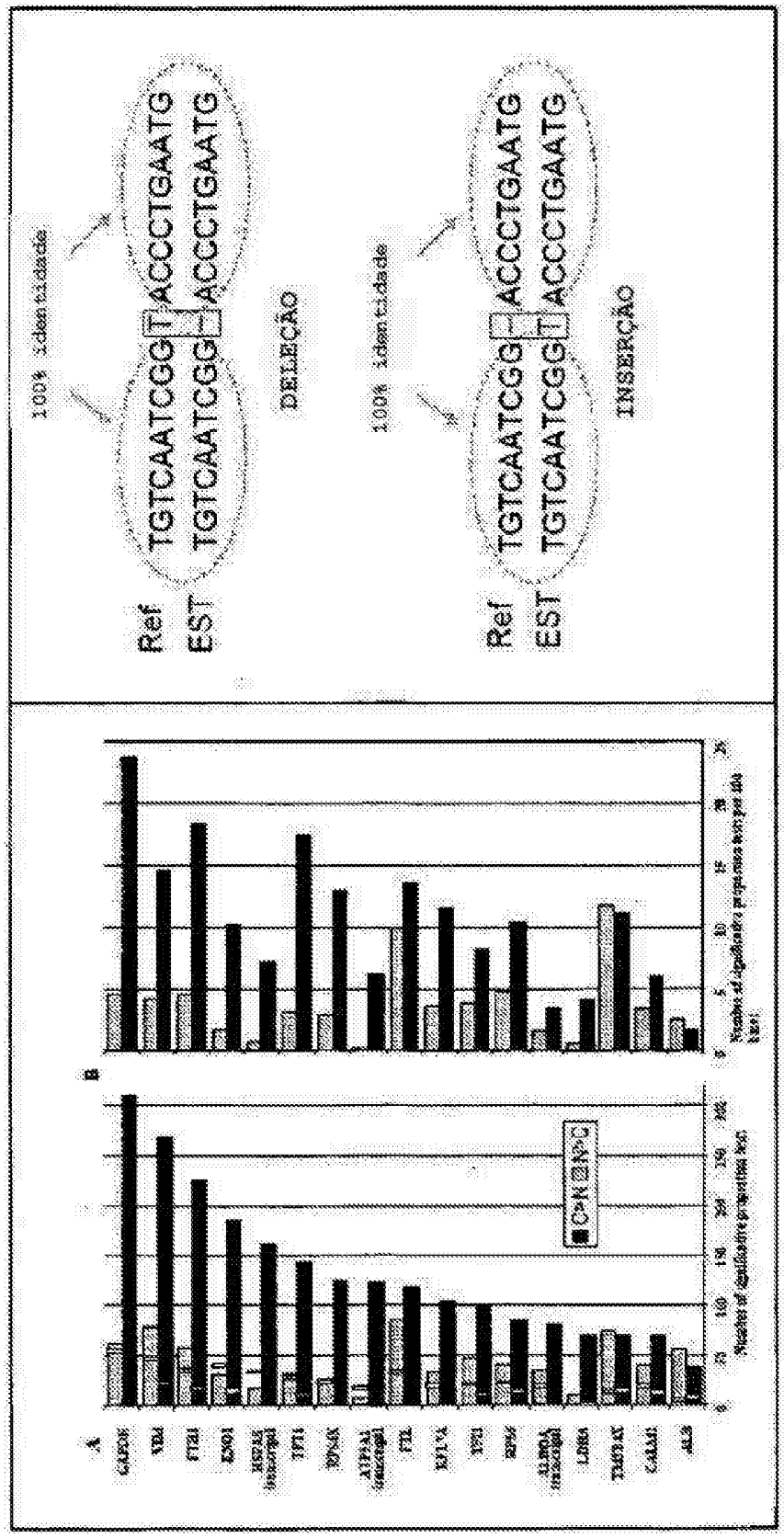

100% identidade
Ref TGTCAATCGGTACCCTGAATG
EST TGTCAATCGG-ACCCTGAATG
DELEÇÃO
100% identidade
Ref TGTCAATCGG-ACCCTGAATG
EST TGTCAATCGGTACCCTGAATG
INSERÇÃO
A
B

Figura 5j

Análise do teste de proporções: resultados (deleções e inserções)

DELEÇÕES						
genes	Número de testes de proporções	C>N	LBE	N>C	LBE	C>N / N>C
ALB	17	1	1	13	0	0,1
ALDOA	34	9	1	7	1	1,1
ATP5A1	24	16	0	0	2	
CALM2	50	14	2	4	2	3,5
ENO1	41	29	0	5	3	5,8
FTH1	67	46	0	4	5	11,5
FTL	100	42	2	9	7	4,7
GAPDH	99	59	2	17	7	3,5
HSPA8	30	27	0	0	2	
LDHA	14	6	0	3	0	2,0
RPL7A	45	26	0	1	3	26,0
RPS4X	35	29	0	1	3	29,0
RPS6	45	24	0	1	3	24,0
TMSB4X	28	16	0	4	1	4,0
TPI1	30	17	0	1	2	17,0
TPT1	43	27	0	4	3	6,8
VIM	26	11	1	9	1	1,2
Total	728	336	9	83	45	4,8

INSERÇÕES						
genes	Número de testes de proporções	C>N	LBE	N>C	LBE	C>N / N>C
ALB	9	0	0	9	0	0,0
ALDOA	5	1	0	1	0	1,0
ATP5A1	9	6	0	0	0	
CALM2	72	10	3	10	3	1,0
ENO1	40	14	0	2	3	7,0
FTH1	114	54	1	3	3	18,0
FTL	188	25	9	33	9	0,8
GAPDH	137	46	4	18	9	2,6
HSPA8	14	7	0	2	1	3,5
LDHA	2	0	0	0	0	
RPL7A	56	12	1	1	3	12,0
RPS4X	29	9	1	2	1	4,5
RPS6	44	12	1	1	3	12,0
TMSB4X	22	5	1	6	0	0,8
TPI1	26	6	1	5	1	1,2
TPT1	41	7	1	4	2	1,8
VIM	27	10	1	3	1	3,3
Total	835	225	24	100	45	2,3

Figura 5k

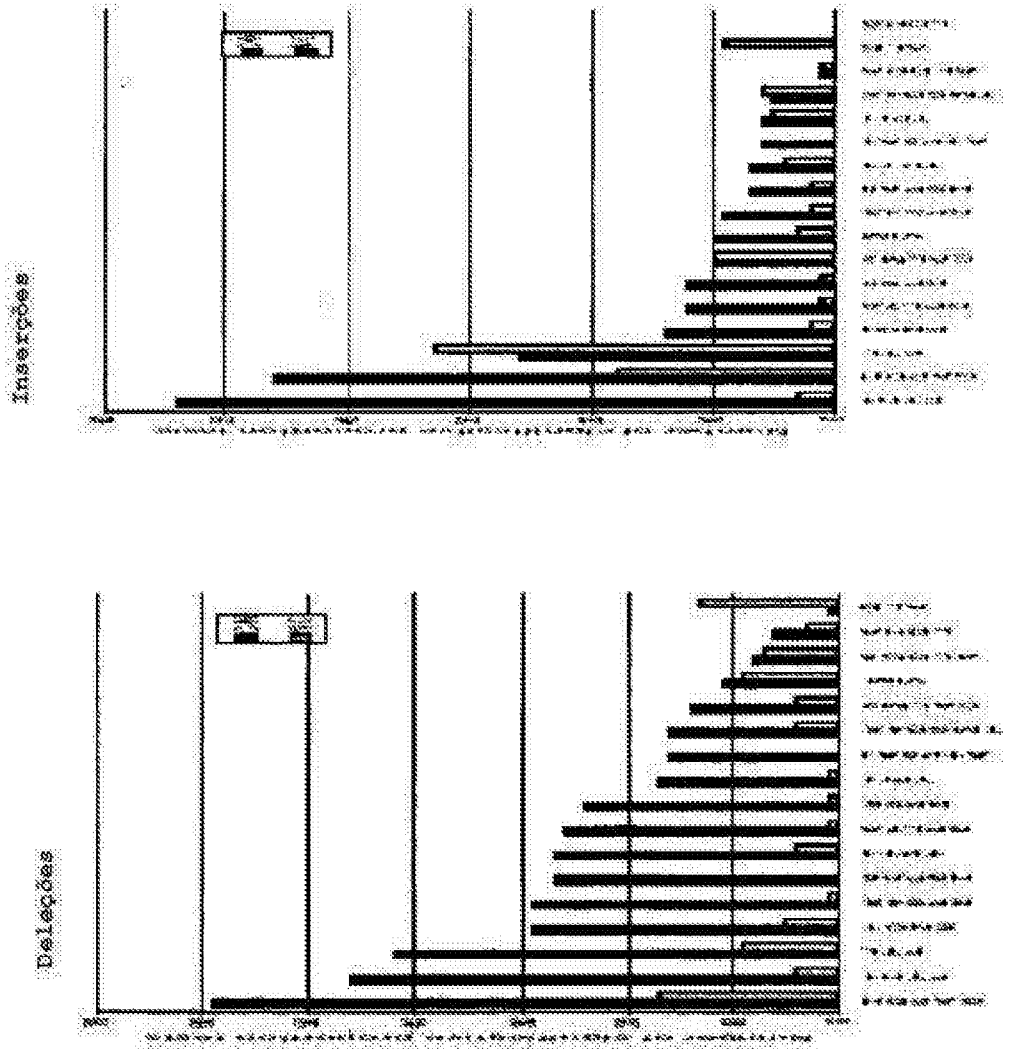


Figura 51

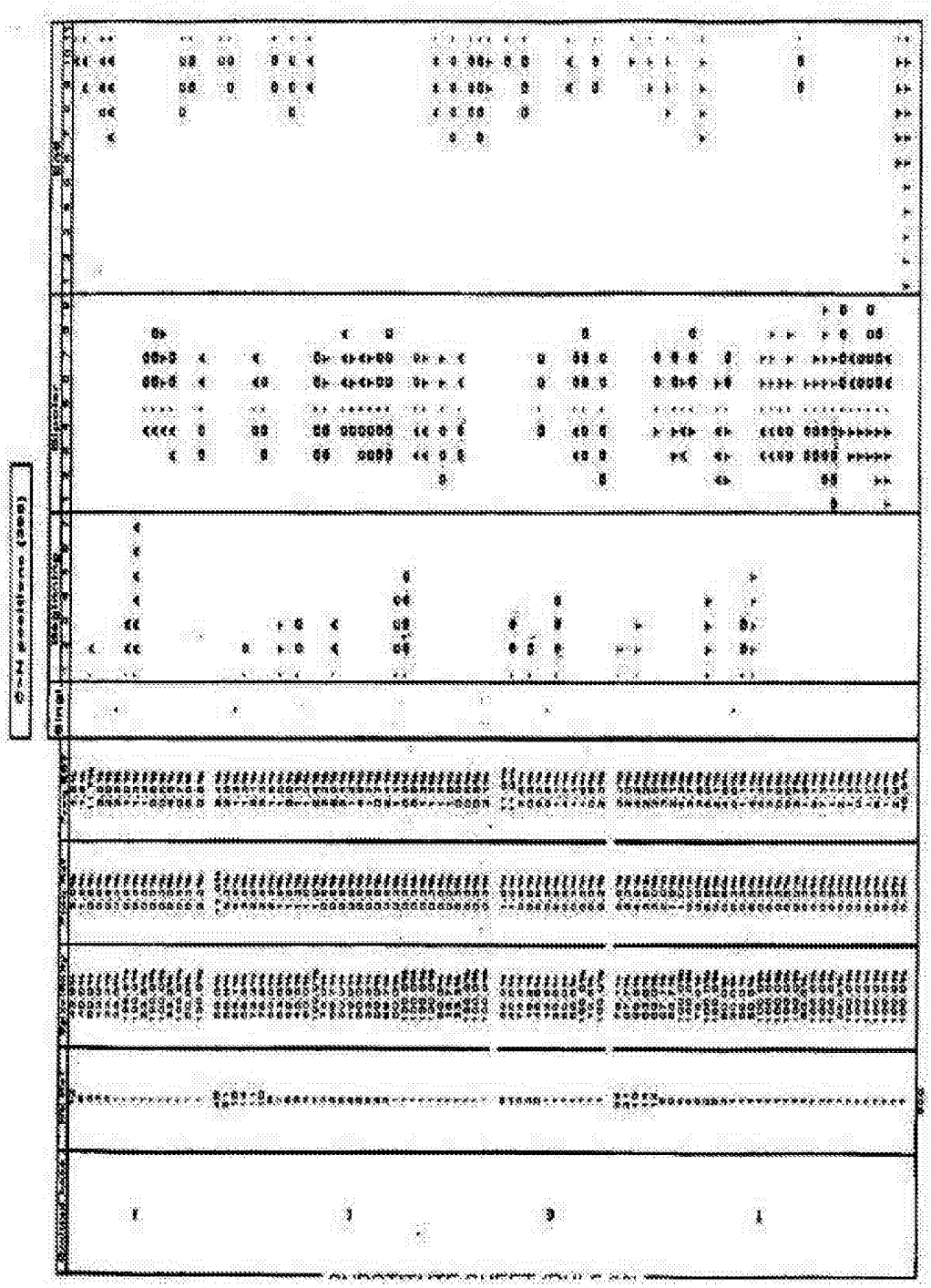


Figura 5m

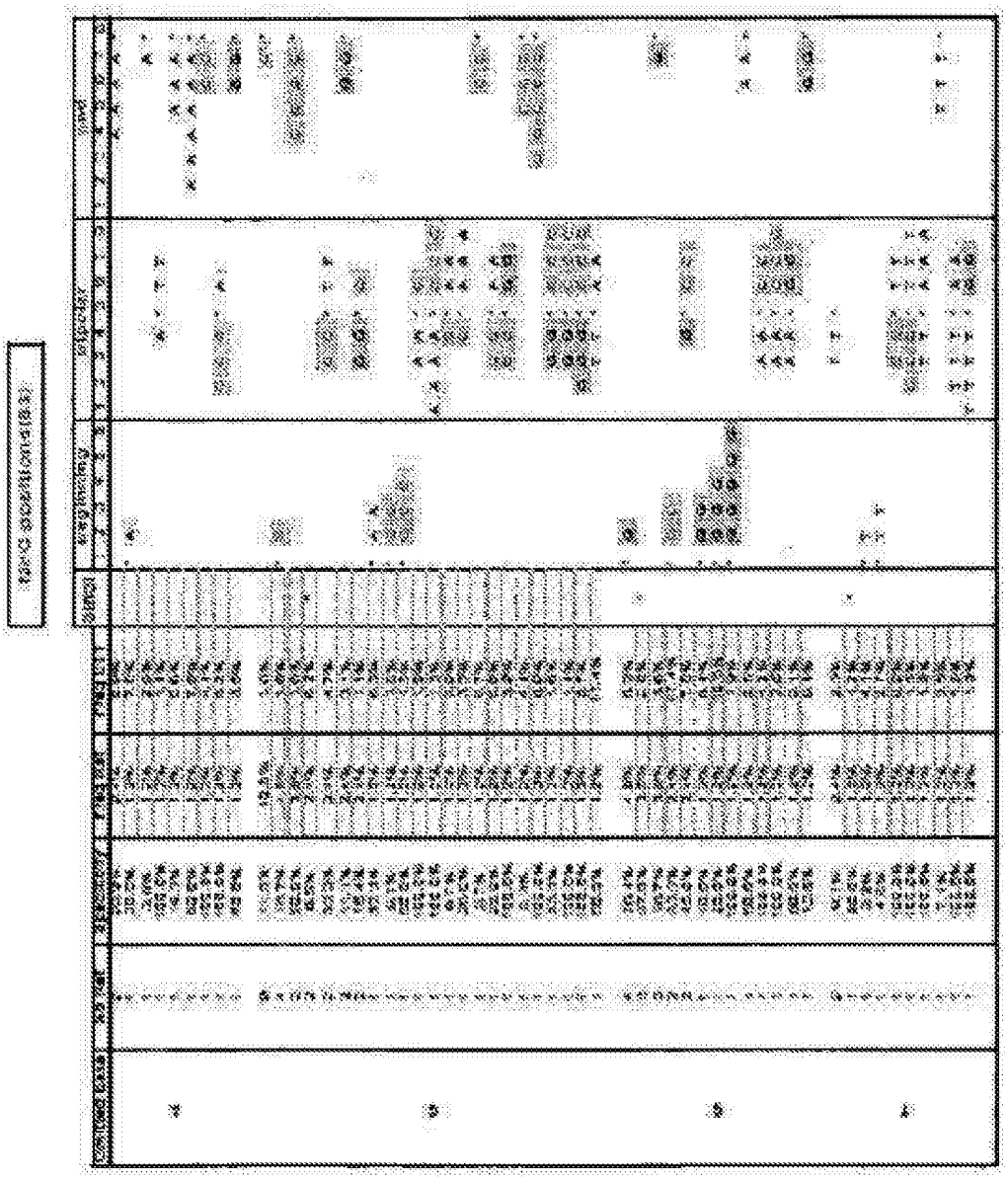


Figura 5n

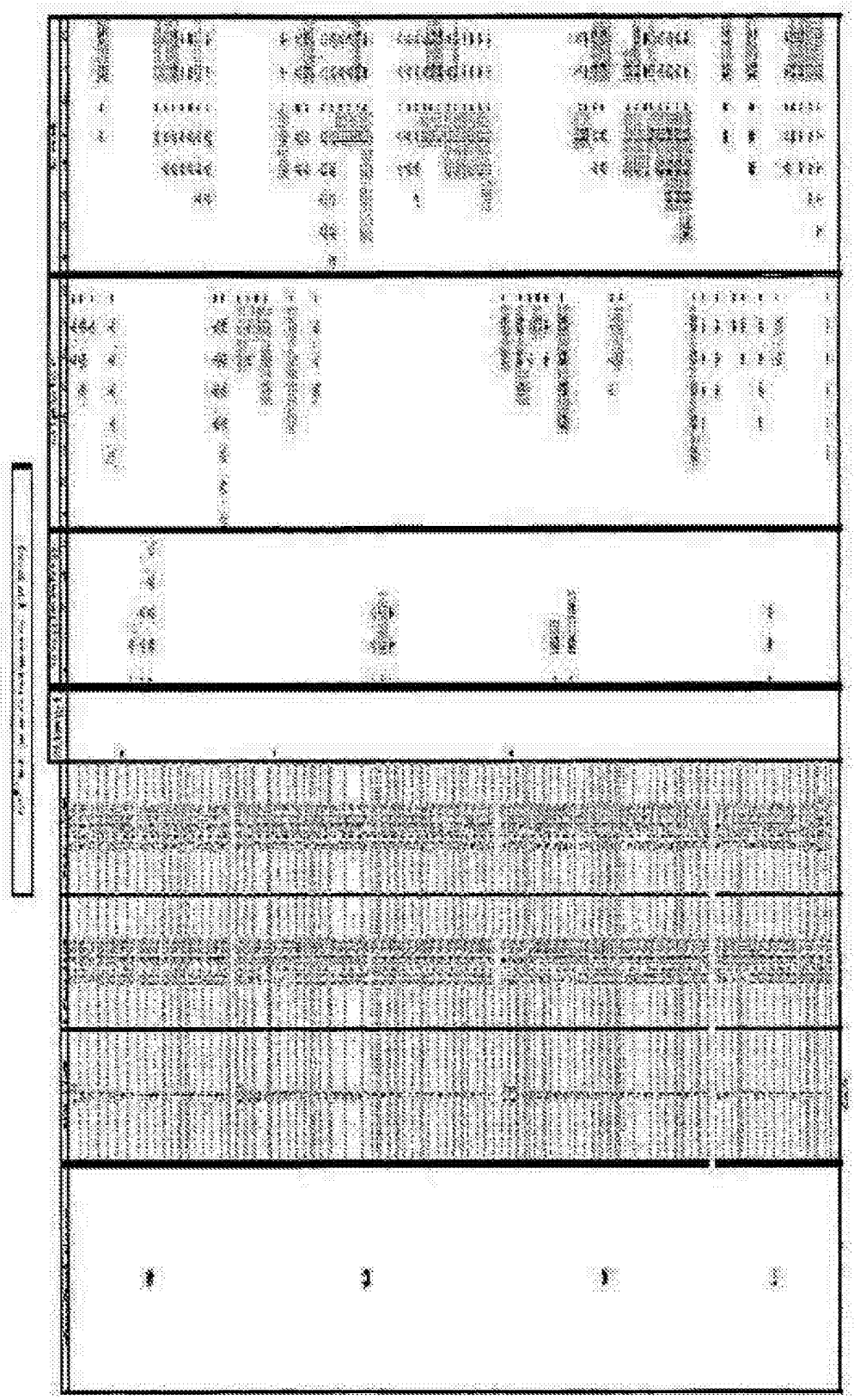


Figura 5o

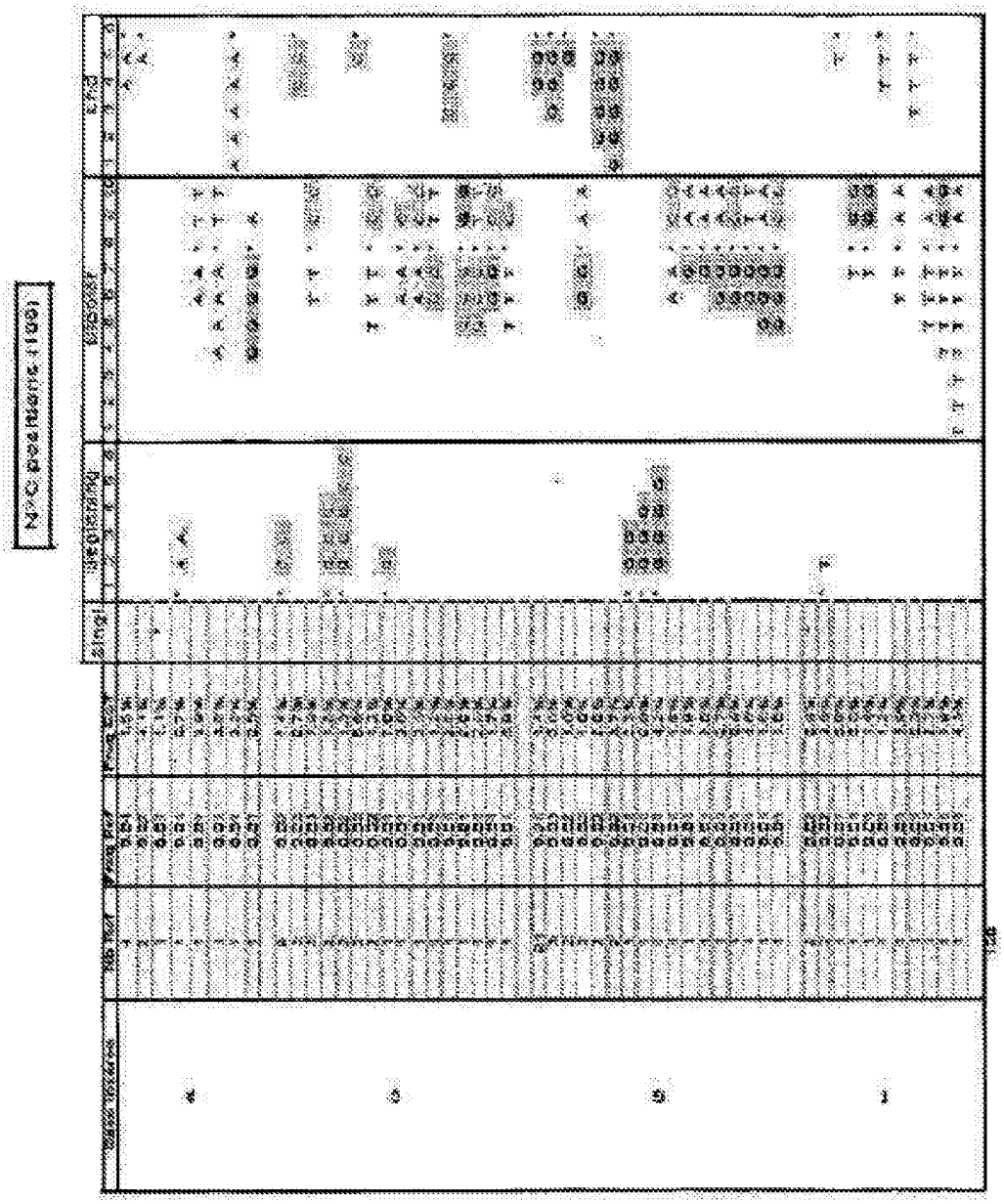


Figura 5p

Figura 6a

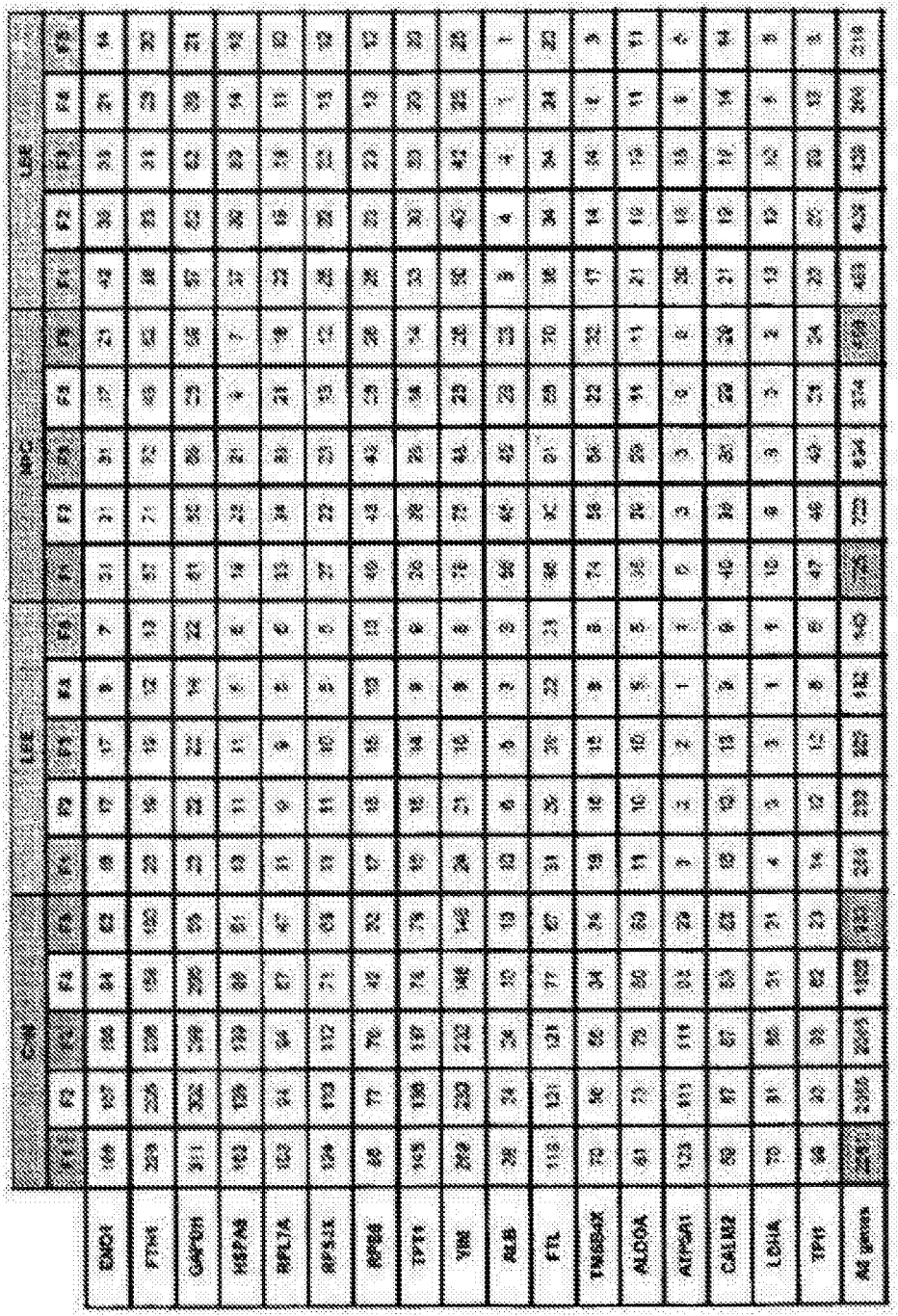


Figura 6b

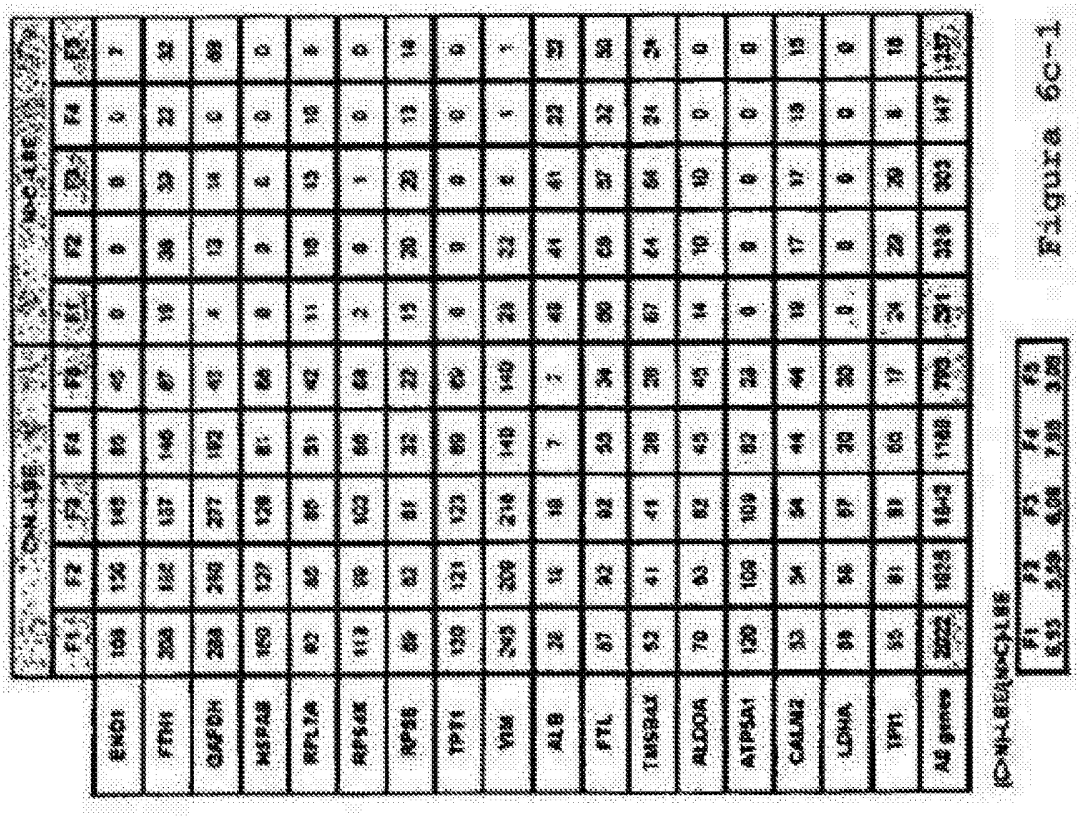


Figura 6c-1

	CNR	LBE	KPC	LBE	(CNR)/(KPC)	(CNR)/(LBE)
F1	2281	259	725	488	3,15	8,95
F2	2065	230	722	429	2,86	9,59
F3	2065	223	664	428	2,98	9,03
F4	1300	132	374	268	3,48	7,95
F5	803	140	456	219	2,85	3,03

Figura 6c-2

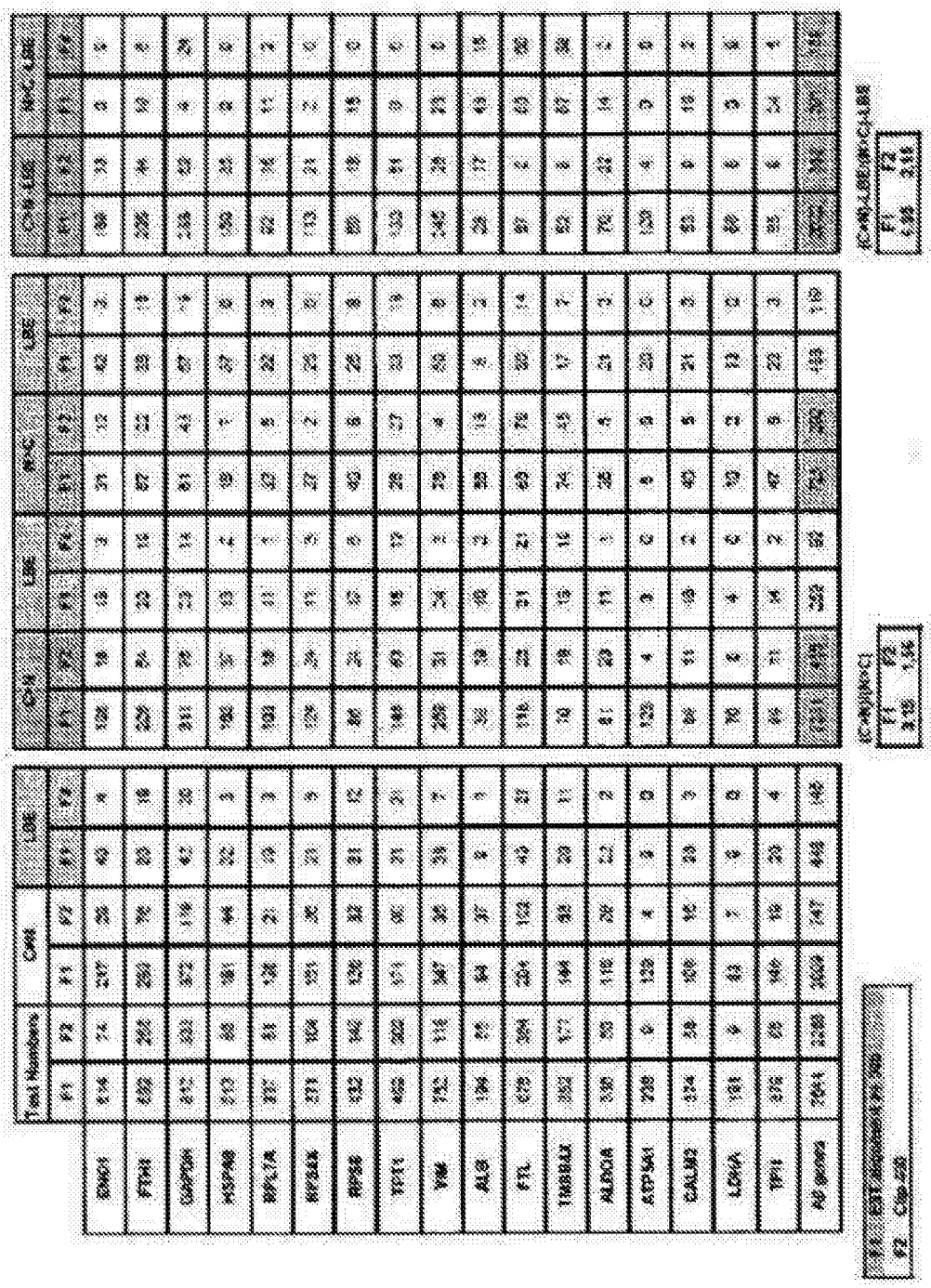


Figura 6d

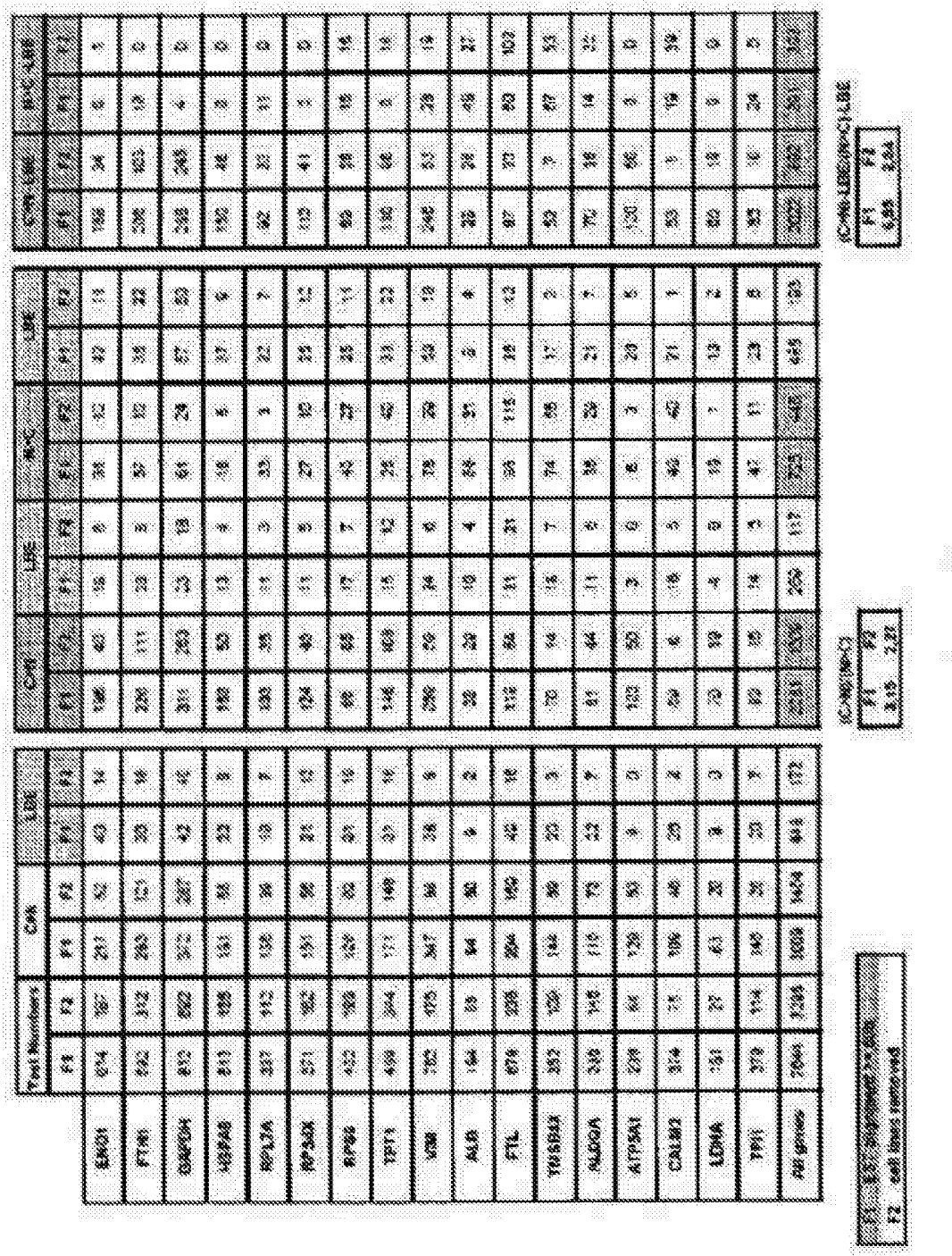
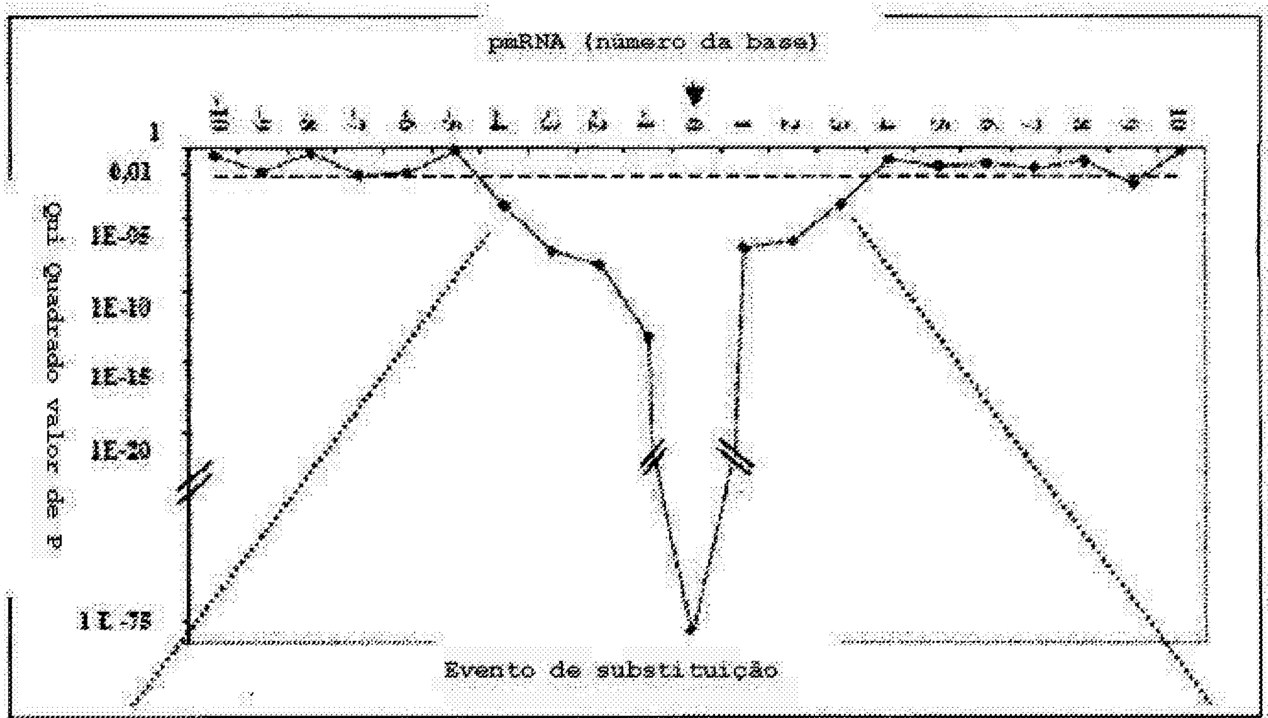


Figura 6e



	b-4		b-3		b-2		b-1		b0		b+1		b+2		b+3	
	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH
A	28	24	24	27	25	27	30	22	33	23	29	25	27	26	28	24
T	13	25	28	23	26	23	32	27	32	22	22	24	24	16	24	24
C	24	25	25	24	31	26	23	26	21	26	21	25	20	24	22	26
G	25	26	23	26	28	24	25	25	14	29	28	26	29	24	26	26

Figura 7a

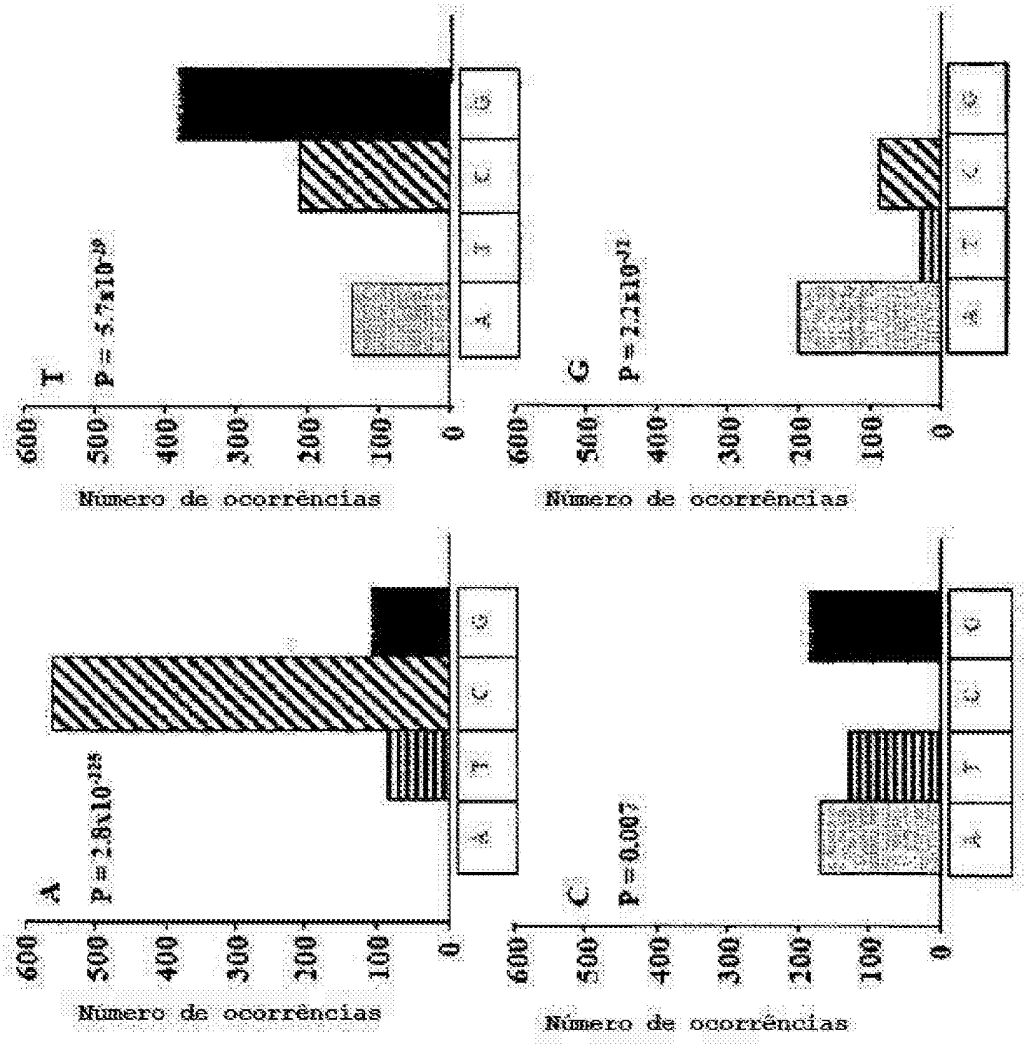

T
P = 5.7x10-39
G
P = 2.2x10-32
A
P = 2.8x10-235
C
P = 0.007
600
500
400
300
200
100
0
Número de ocorrências
A
T
C
G

Figura 7b

A. Todos os desvios (n=2281)

afectada \ substituição	A	T	C	G	%
A		85	560	107	33
T	139		210	393	32
C	169	130		185	21
G	199	26	86		14
%	22	11	37	30	

B. Desvios para b-1 (n=799)

afectada \ substituição	A	T	C	G	%
A		41	201	42	30
T	28		58	83	21
C	77	41		126	30
G	75	9	20		13
%	23	11	33	31	

C. Desvios para b+1 (n=630)

afectada \ substituição	A	T	C	G	%
A		24	52	37	21
T	27		45	177	47
C	36	48		15	19
G	40	11	16		13
%	20	16	21	43	

D. Desvios nem para b-1 nem para b+1 (n=347)

afectada \ substituição	A	T	C	G	%
A		7	148	8	12
T	42		50	9	16
C	17	10		23	12
G	7	2	24		4
%	19	5	64	12	

E. Desvios para b-1=b+1 (n=339)

afectada \ substituição	A	T	C	G	%
A		5	52	12	20
T	11		12	95	35
C	25	30		17	21
G	71	1	8		24
%	31	11	21	37	

F. Desvios dentro do triplet (n=266)

afectada \ substituição	A	T	C	G	%
A		8	107	8	46
T	31		47	20	37
C	12	1		4	6
G	6	5	17		11
%	19	5	64	12	

Figura 7c

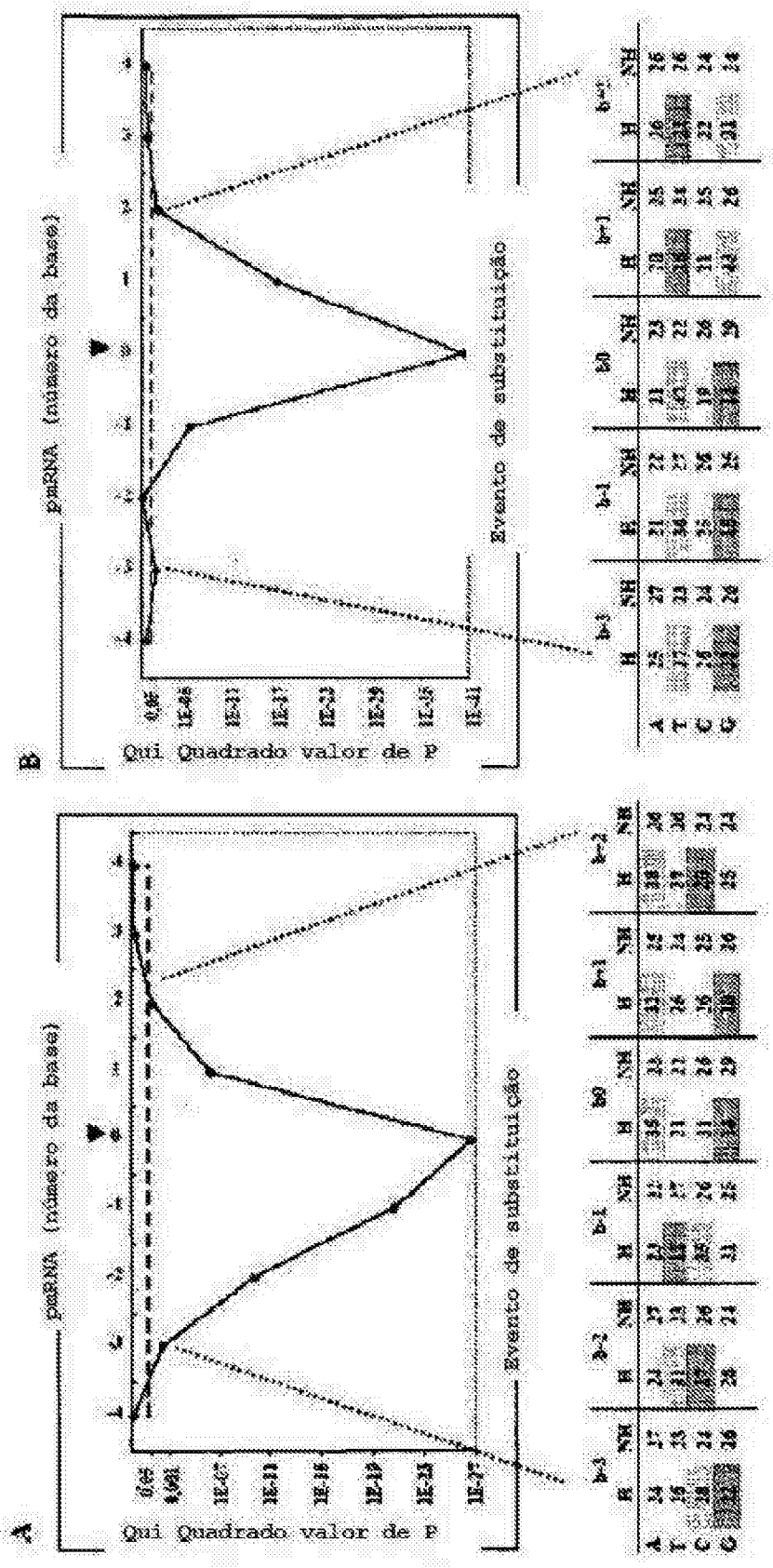

A
B
Qui Quadrado valor de P
Evento de substituição
pmRNA (número da base)

Figura 7d

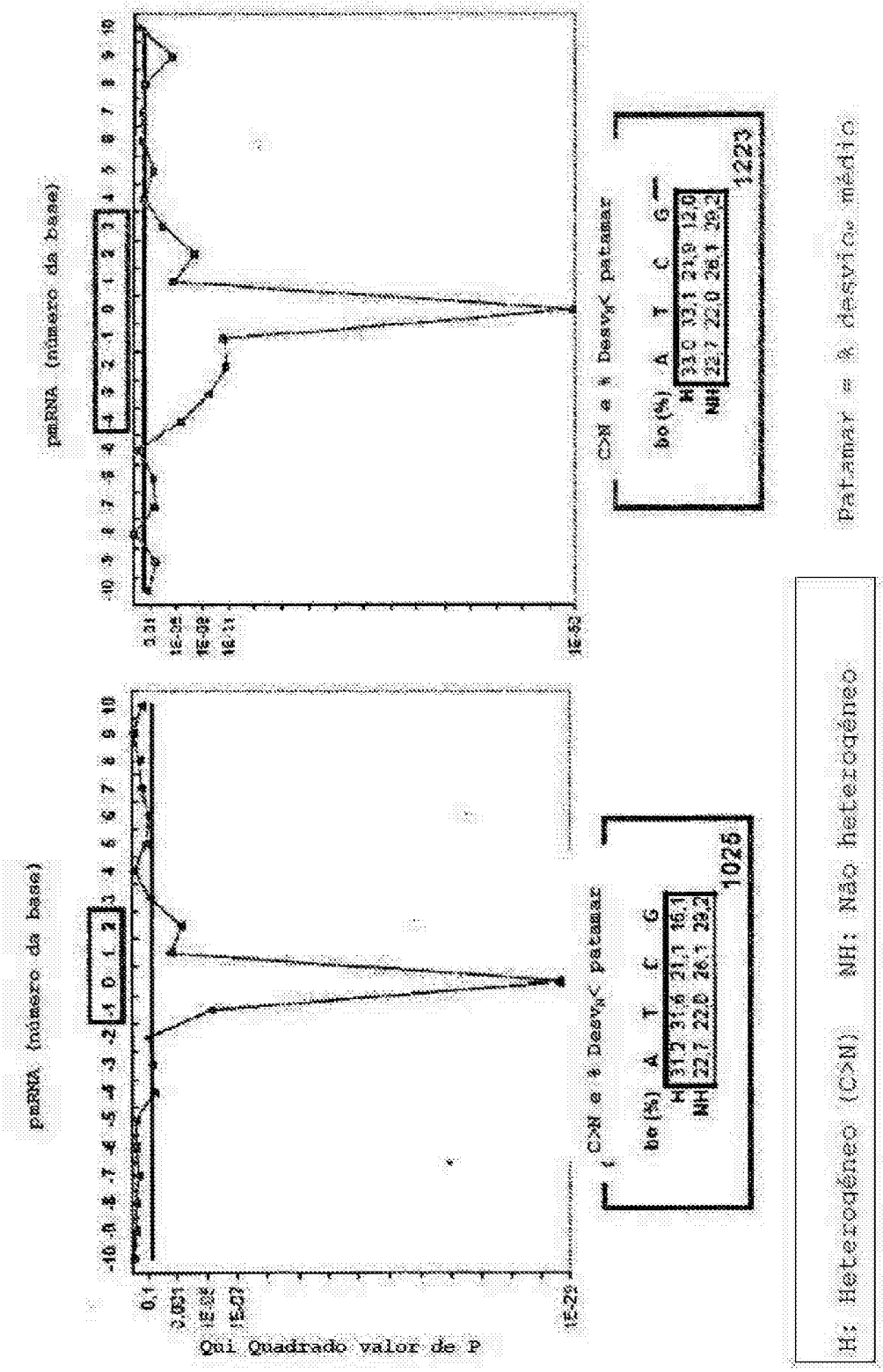

paRNA (número da base)
Qui Quadrado valor de P
C>N e % Desv< patamar
bo (%) A T C G
H 33.0 32.1 21.9 12.9
NH 22.7 22.0 28.1 28.2
1223
1025
H: Heterogêneo (C>N) NH: Não heterogêneo
Patamar = % desvio médio

Figura 7e

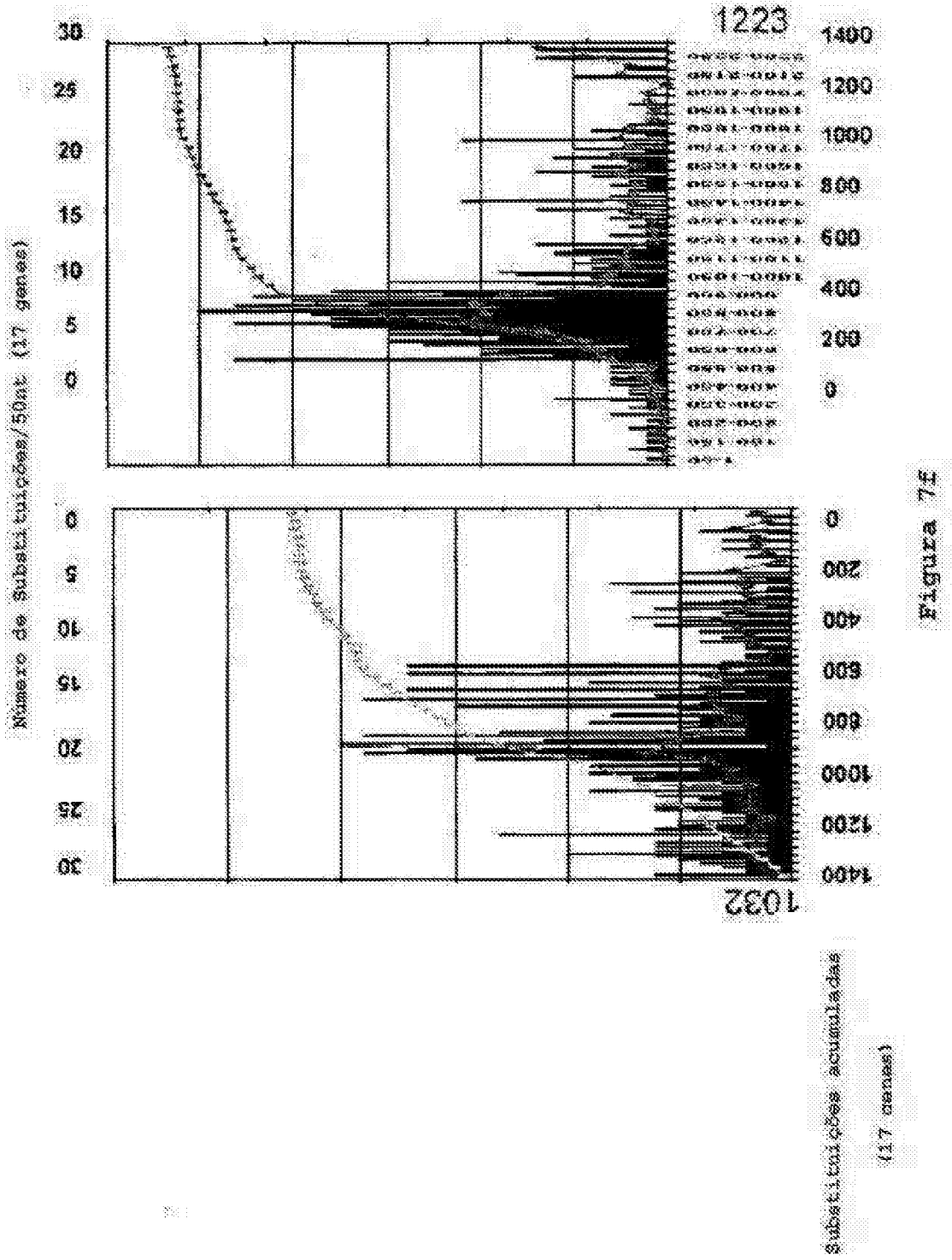
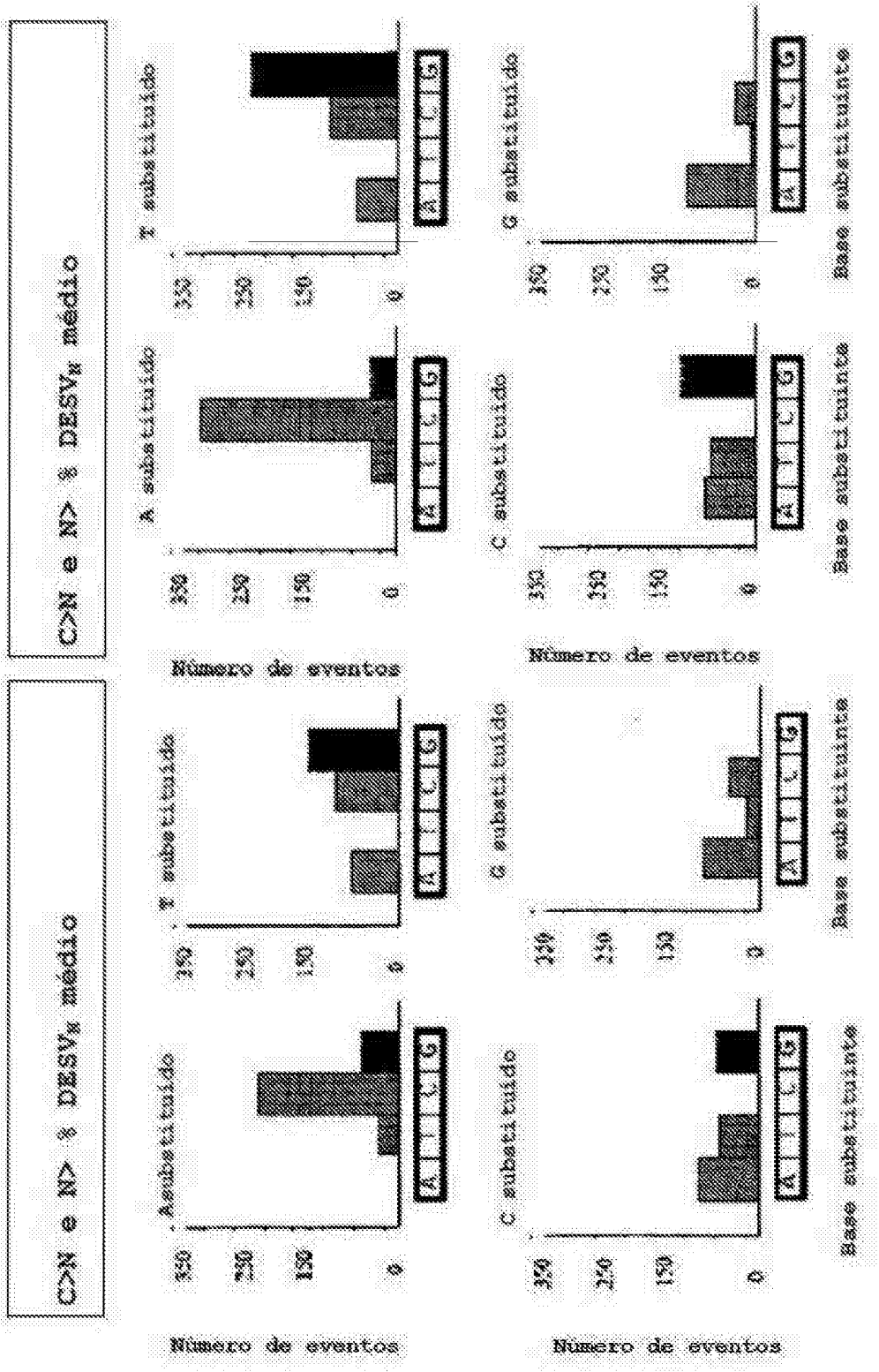

1223
1032
Número de Substituições/50nt (17 genes)
Substituições acumuladas (17 genes)

Figura 7F


C>N e N> % DESV$_N$ médio
C>N e N> % DESV$_M$ médio
A substituído
T substituído
C substituído
G substituído
Número de eventos
Base substituinte
A
T
C
G
350
250
150
0

Figura

Figura 8a

Figura 8b

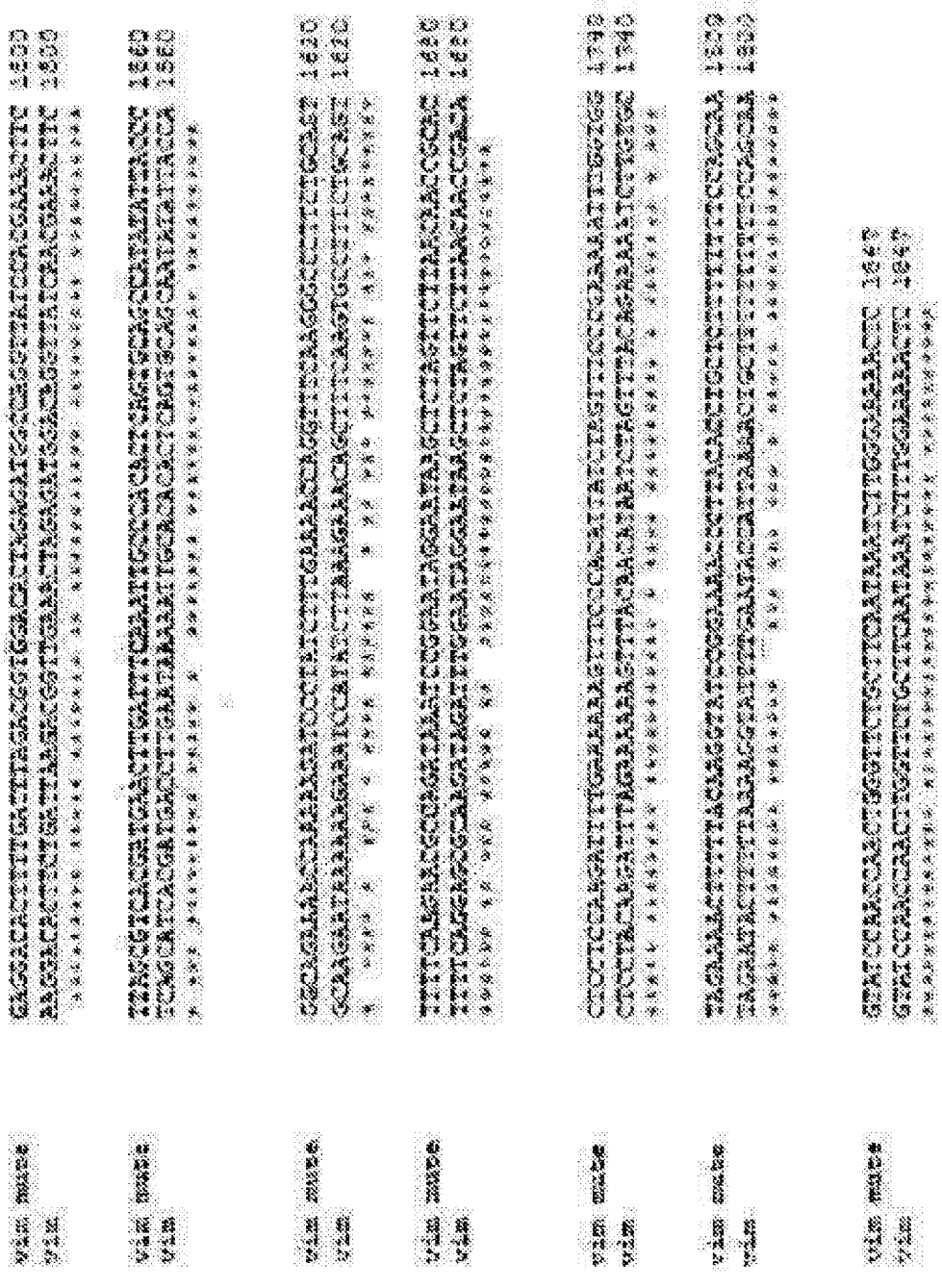


Figura 8c

Figura 8d

Estudo do impacto codificante das substituições (cancro > normal) em 17 genes		
Número de substituições C>N substituições	2281	
Numero de substituiçõee C>N em CDS	1548	
Impacto codificante %	67,8%	
Estudo da região CDS	Número de substituições	%
Substituições silenciadoras	369	23,8%
Impacto codificante (modificação de AA mas da mesma família)	393	26,4%
Impacto codificante (modificação da família de AA)	744	48,1%
Substituições sem sentido	24	1,6%
Modificações de codões de paragem canónicos	18	1,2%
Substituições que originam um codão ATG (MET)	33	2,1%

Figura 9

proteínas normais
Péptidos adicionais de proteínas cujo codão de paragem é afectado normais

Figura 10a

proteínas normais
Peptídos adicionais de proteínas cujo codão de paragem é afectado normais

Figura 10b

(DHPLC = Cromatografia líquida de alta resolução
Princípio:
amplificação do cDNA por PCR
Desnaturação pelo calor
Arrefecimento lento
homoduplexes e
heteroduplexes
formação de
Fusão 80% (55-60°C)
cDNA normal
cDNA mutado

Figura 11a

Injecção da amostra na coluna de
DHPLC:
Heteroduplexes eluem antes
dos homoduplexes
Heteroduplexes homoduplexes
absorvância
tempo

Figura 11b

Source : Theru A. Sivakumaran et al., Current Science, 84 : 291 – 296, 2003

Figura 11c

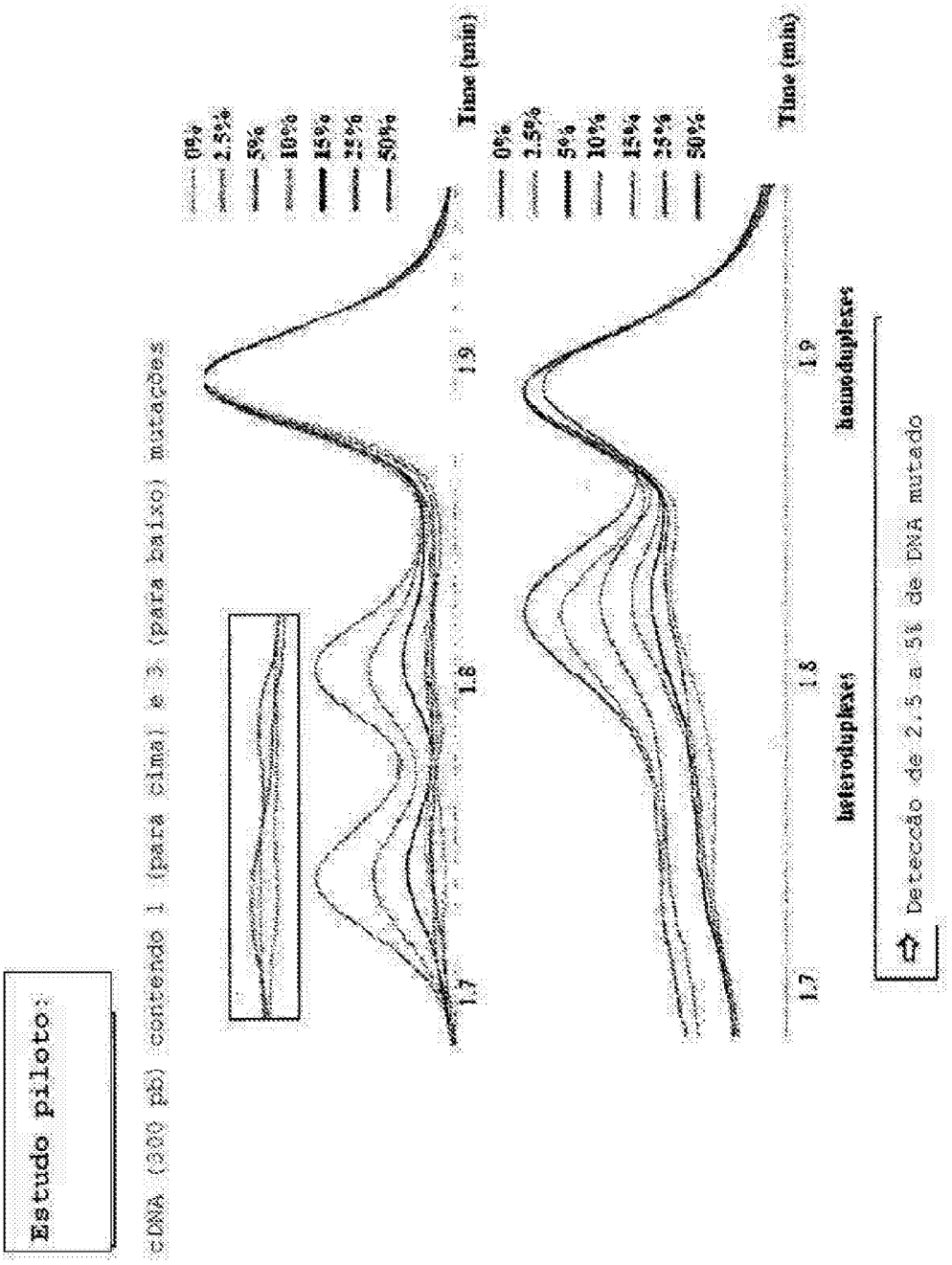

Estudo piloto:
cDNA (300 pb) contendo 1 (para cima) e 3 (para baixo) mutações
0%
2.5%
5%
10%
15%
25%
50%
Time (min)
1.7
1.8
1.9
heteroduplexes
homoduplexes
Detecção de 2.5 a 5% de DNA mutado

Figura 11d

Figura 12a

Figura 12b

Figura 12c

Símbolo	Nome	Expasy	Plasma Proteome database	
		proteín		mRNA
		SP accession	PPD accession	PPD accession
A2M	Alfa-2-macroglobulina	P01023	NP_000005	NM_000014
AHSG	Alfa-2-HS-glicoproteína	P02765	NP_001613	NM_001622
ALB	Albumina	P02768	AAH00236	BC000232
APCS	Componente amilóide P	P02743	NP_001630	NM_001639
APOA1	Apolipoproteína A-I	P02647	NP_000030	NM_000039
APOA2	Apolipoproteína A-II	P02652	NP_001634	NM_001643
APOC1	Apolipoproteína C-I	P02654	NP_001635	NM_001645
APOC2	Apolipoproteína C-II	P02655	NP_000483	NM_000474
APOC3	Apolipoproteína C-III	P02656	NP_000031	NM_000040
APOD	Apolipoproteína D	P05090	NP_001638	NM_001647
APOE	Apolipoproteína E	P02649	NP_000032	NM_000041
AZGP1	Zn-alfa2-glicoproteína	Q3KKQ4	CAA42438	X59768
B2M	Beta-2-microglobulina	P61769	NP_004039	NM_004048
C1S	Componente do complemento C1, subcomponente s	P09871	NP_963869	NM_201442
C3	Componente do complemento 3	Q5LL30	NP_000055	NM_000064
C7	Componente do complemento 7	P10643	NP_000578	NM_000587
		P12830	BAA00957	AB021105
		P80761	NP_001731	NM_001719
CHI3L1	chitinase 3-like 1	P36222	NP_001267	NM_001276

F

Figura 13a

Símbolo	Nome	Expasy	Plasma Proteome database	
		proteína		mRNA
		SP accession	PPD accession	PPD accession
CLEC3B	Domínio 3 de lectina tipo C, membro B	P05452	NP_003269	NM_003278
CLU	Clusterina	P10909	NP_001822	NM_001831
CP	Ceruloplasmina	P00450	NP_000087	NM_000096
CRP	Proteína C reactiva	P02741	AAH20766	BC020766
EDN1	Endotelina 1	P05305	NP_001946	NM_001955
EGFR epidérmico	Receptor do factor de crescimento	P00533	NP_958441	NM_201284
F13A1	Factor de coagulação XIII, polipéptido A1	P00488	NP_000120	NM_000129
F13B	Factor de crescimento XIII, polipéptido B	P05160	NP_001985	NM_001994
FGA	Fibrinogénio cadeia alfa	P02671	NP_000499	NM_000508
FGB	Fibrinogénio cadeia beta	Q32065	NP_005132	NM_005141
FGC	Fibrinogénio cadeia gama	P02679	NP_068566	NM_021870
GPI	Isomerase de glucose fosfato	P06744	NP_000166	NM_000175
GSTP1	Glutationa S-transferase p1	P09211	NP_000843	NM_000852
HDLBP (VIGF)	cDNA FLJ45936 fis, altamente semelhante a	Q03591	BAC28155	AK127833
		P00739	NP_005134	NM_005143
HPX	hemopexin	P02790	NP_000604	NM_000613
HRG	histidine-rich glycoprotein	P04196	NP_000403	NM_000412
IGFBP3	insulin-like growth factor binding protein 3	P17936	NP_000589	NM_000598
IGJ	immunoglobulin J polypeptide	P01591	NP_653247	NM_144646

Figura 13b

Símbolo	Nome	Expasy	Plasma Proteome database	
		proteína		mRNA
		SP accession	PPD accession	PPD accession
ITIH4	Inibina alfa			
KLKB1	Peptidase 11 relacionada com calicreína	P05111	NP_002182	NM_002191
LDHA	Lactato desidrogenase A	Q5JBX7	NP_659166	NM_144567
LGALS3BP	Lectina, ligação a galactosídeo, solúvel, proteína de ligação 3	P00338	NP_005557	NM_005566
		Q08380	NP_005558	NM_005567
LRG1	Alfa-2-Glicoproteína 1 rica em leucinas	P02750	NP_683704	NM_052972
OSM1	Oncostatina M precursor	P02763	NP_000590	NM_000607
PMK2	Quinase 1 de Homo sapiens	P14618	NP_872270	NM_182470
PLG	Quinase de piruvato, músculo	P00747	NP_000292	NM_000301
PON1	Plasminogênio	P27169	NP_003457	NM_003466
PROC	Paraoxonase 1	P04070	NP_000303	NM_000312
PSP	Proteína C	P20742	NP_002855	NM_002864
RBP4	Proteína da zona de gravidez	P02753	P02753	BC020633
SAA1	Proteína de ligação ao retinol 4	P02735	NP_000322	NM_000331
SERPINA1	Amilóide sérica A1	P01009	NP_000286	NM_000295
SERPINA3	Inibidor da peptidase serpina, grupo A	P01011	NP_001076	NM_001085
SERPINA5	Inibidor da peptidase serpina, grupo A, membro 3	P05154	NP_002663	NM_002674
TF	Inibidor da peptidase serpina, grupo B, membro 1	P02787	NP_001054	NM_001063
TFRC	Transferrina	P02786	NP_003225	NM_003234
[illegible]	Receptor da transferrina	P01137	NP_000551	NM_000560
[illegible]	[illegible]			

Figura 13c

Símbolo	Nome
TIMP 1	inibidor da metalopeptidase 1
TTR	Transtiretina
VTN	vitronectina

Expasy	Base de dados Plasma Proteome	
proteínas		mRNA
SP accession	PPD accession	PPD accession
P01033	NP_003245	NM_003254
P02766	NP_000362	NM_000371
P04004	NP_000629	NM_000638

Figura 13d

Selecção de 3 proteinas do plasma:

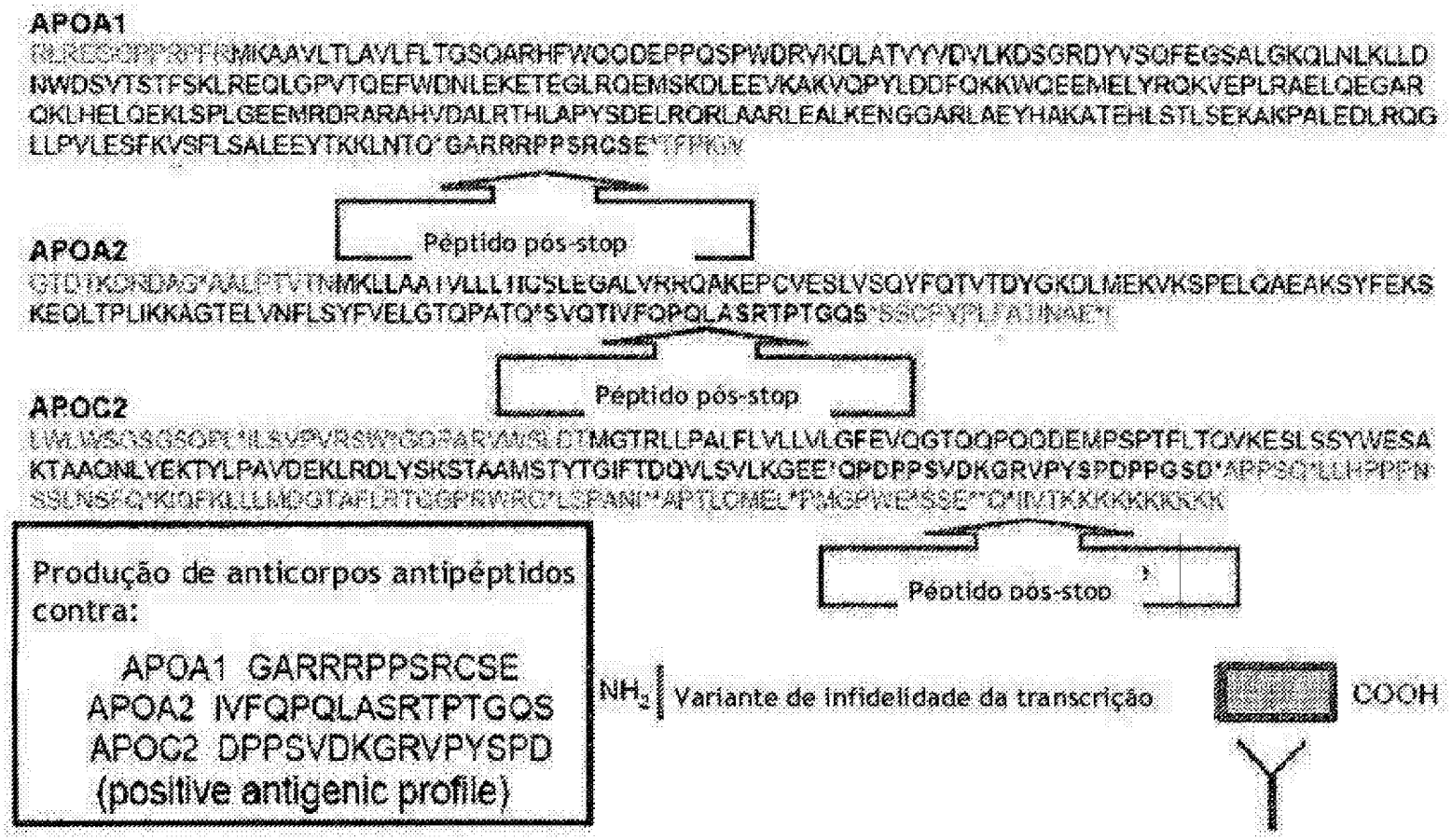
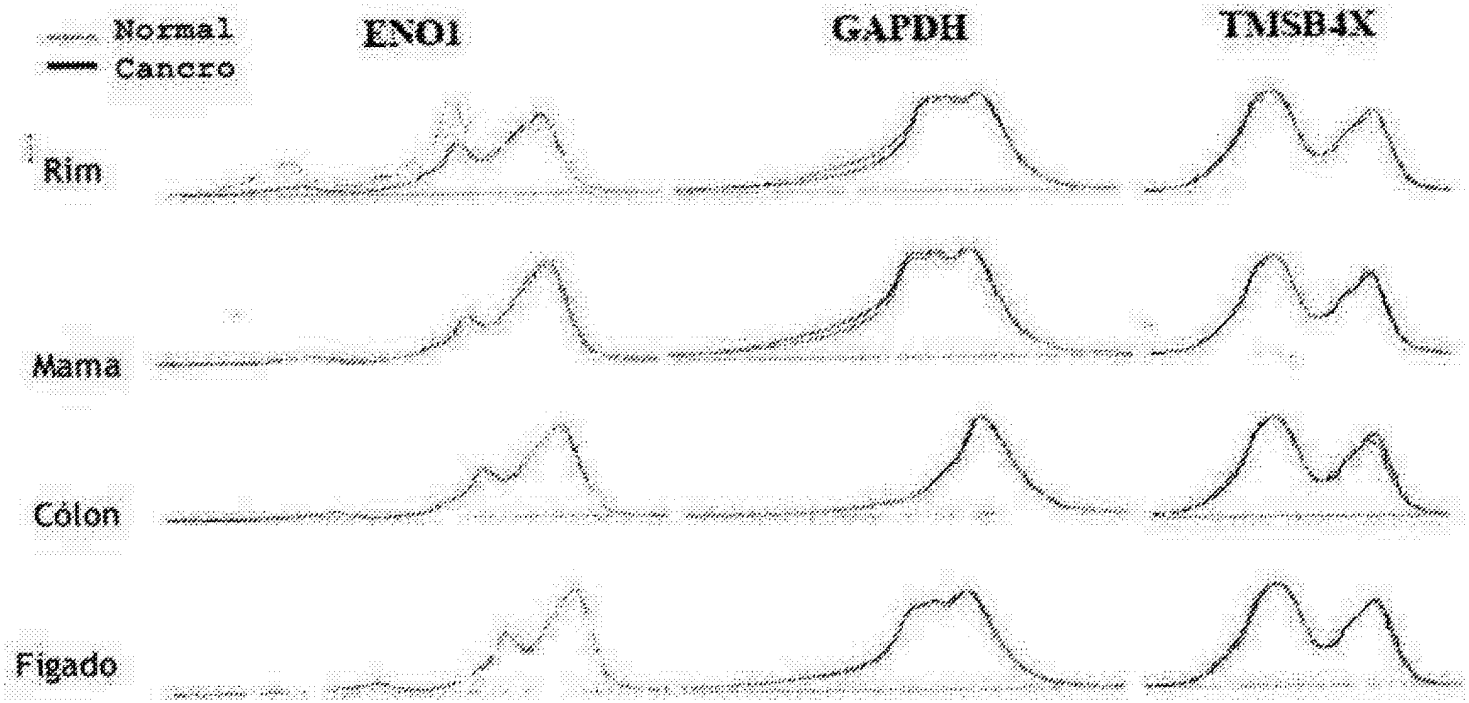


Figura 14



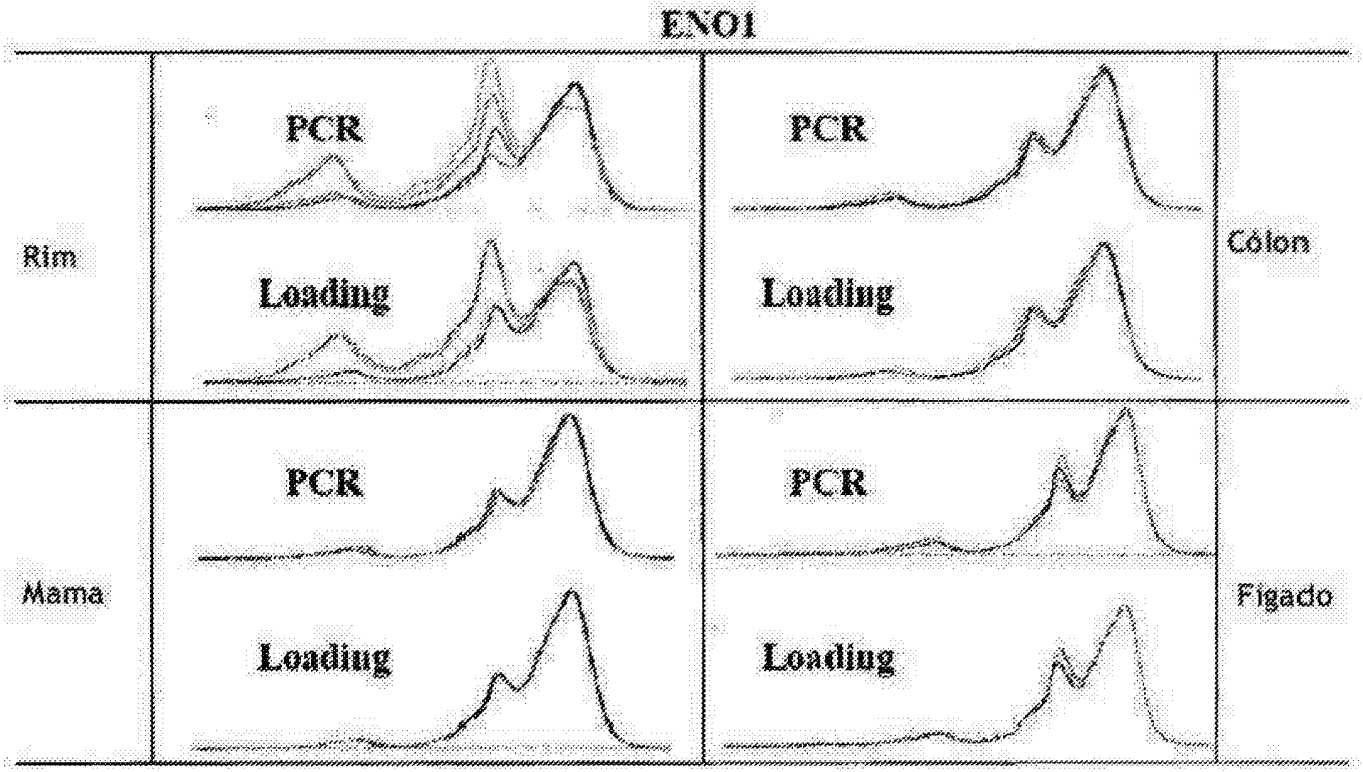
DHPLC profiles of tumor tissue v.
cDNA from cancerous patients (Bio
TMSB4X genes and the *pfx* polym
represented in Green and the cance
system (Transgenomic). The tempe
TMSB4X gene. Curves were visuali

Perfis de DHPLC de tecido tumoral versus tecido adjacente normal por PCR para os genes ENO1, GAPDH e TMSB4X

cDNA de doentes cancerosos (Biochain Inc.) foram amplificados por PCR usando oligonucleótidos complementares dos genes ENO1, GAPDH e TNSB4X e a polimerase pfx. cDNA foi obtido a partir de tecidos de rim, mama, cólon e fígado. O perfil de DHPLC normal está representado a verde e o perfil de cancro está representado a preto. As

Figure 15a

Figura 15a



Different PCR and Loading DHPLC pro

Two different PCR products were prepare

DHPLC profile is represented in Green (

duplicate is represented in : Red for Non

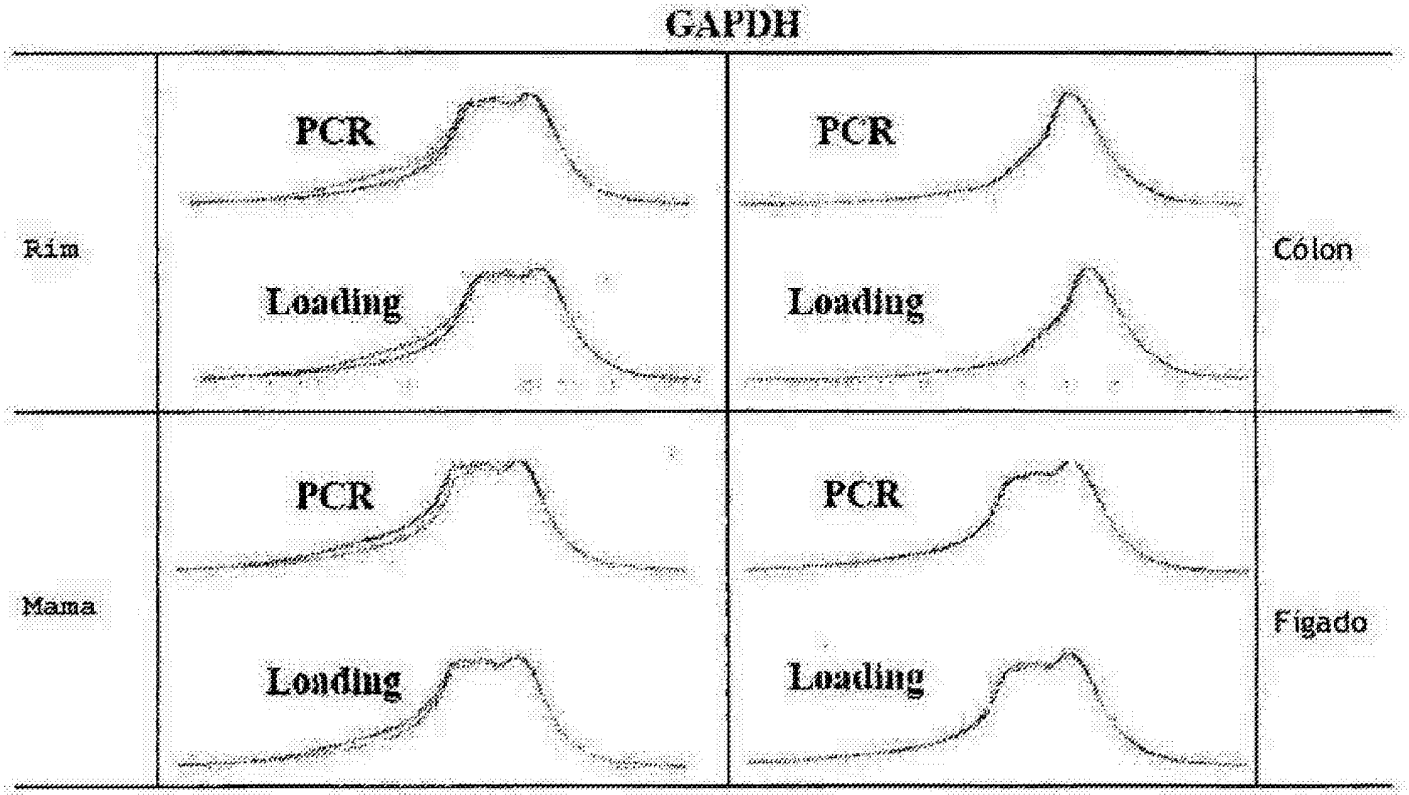
Normal and Light Blue for Cancerous. Cur

Diferentes perfis de DHPLC para PCR e aplicação ("loading") de tecidos tumorais versus tecido normal adjacente para o gene ENO1

Dois produtos de PCR diferentes foram preparados a partir de 4 tecidos e cada produto de PCR foi injectado duas vezes no sistema de DHPLC. O perfil normal de DHPLC está

Figure 15b

Figura 15b



Different PCR and Loading DHPLC profi Diferentes perfis de DHPLC para PCR e aplicação ("loading") de tecidos
Two different PCR products were prepared tumorais versus tecido normal adjacente para o gene GAPDH
DHPLC profile is represented in: Green an Dois produtos de PCR diferentes foram preparados a partir de 4 tecidos e cada produto
duplicate is represented in : Red for Norm de PCR foi injectado duas vezes no sistema de DHPLC. O perfil normal de DHPLC está
Normal and Light Blue for Cancerous. Curve

Fi~~~ 15~

Figura 15c

Sequência A: CCGCTCAGCAT
100%A-0%B
95%A-5%B
85%A-15%B
70%A-30%B
50%A-50%B
Sequência B: CCGCAGCGCAT
0%A-100%B

Figura 16

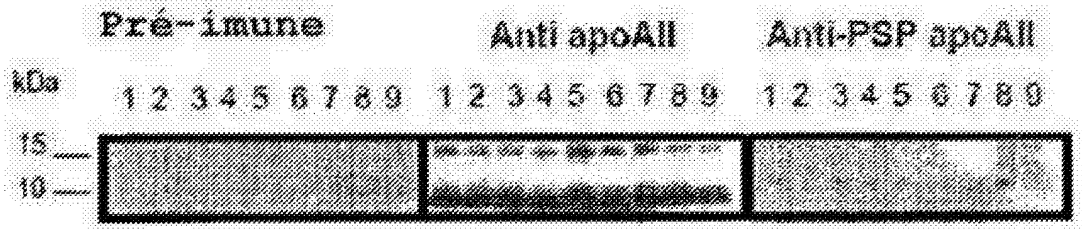


Figura 17a

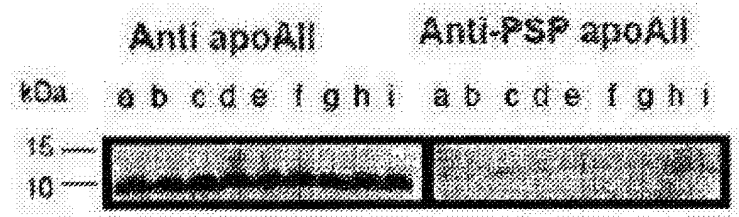


Figura 17b

Plasma from patients with prostate polyacrylamide gels under reducing nonfat dry milk (1h, room temperatu commercially available goat anti-hur (1/100, right panel). After 1 h incu appropriate conjugated second antibo

Plasmas de diferentes doentes com cancro da próstata (a) ou com cancro na fase metástica (b) foram aplicados em géis de SDS 4-12% poliacrilamida em condições redutoras. Após transferência para membranas PVDF e bloqueio em TBST contendo 5% de leite em pó desnatado (1h, temperatura ambiente). As transferências Western foram realizadas usando soro pré-imune de coelho (diluição 1/100, painel esquerdo),

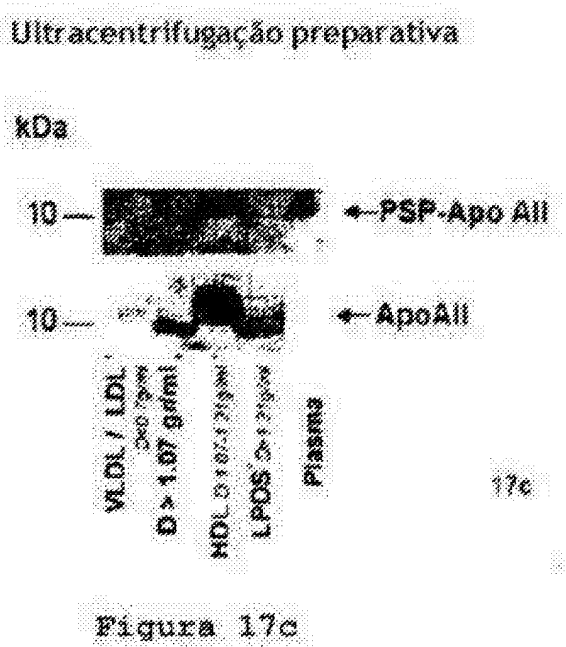


Figura 17c

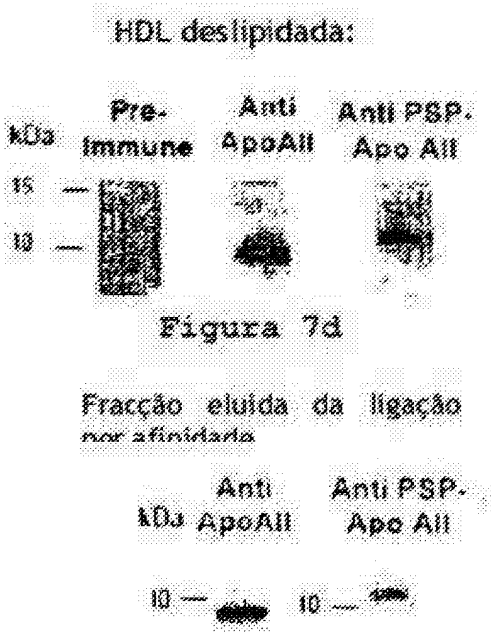


Figura 7d

c. HDL (d 1,07-1,21 g/ml) foi preparado por ultracentrifugação diferencial de um conjunto de plasmas de 10 doentes com cancro na fase metástica. As diferentes fracções recuperadas após cada passo foram dialisadas contra NaCl 0,15M, pH 7,4 contendo 0,01% EDTA. As transferências Western foram realizadas nas fracções indicadas com anticorpo comercial anti-apoAII (painel inferior) ou com anticorpo anti-PSP (painel superior) usando as mesmas diluições indicadas nas Figuras 17a e 17b. Os géis foram sobrecarregados, o que explica a distorção e as bandas largas. A transferências Western foi igualmente realizada no plasma original ou num volume correspondente do soro deficiente em lipoproteina (d>1,21 g/ml). d. HDL purificado foi dialisado contra 0,01% EDTA, pH 7,4, congelado a -80°C e depois liofilizado. A deslipidação de HDL liofilizado foi então efectuada usando clorofórmio-metanol arrefecido em gelo (2:1, v/v, 5 ml por 20 mg de proteína HDL). Uma segunda deslipidação de HDL liofilizada foi então realizada usando éter dietílico, seguida de

c. HDL (d1,07-1,21 g/ml) was
The different fractions recover
performed on the indicated fi
using the same dilutions as i
Western blot was also perform
d. Purified HDL was dialyzed
then performed using ice-cold
then followed by centrifugatio
N2(g). The pellet was redisso
using preimmune serum, con
carried out to isolate the PSP
Western blotting using both the

Figura 17f

Similar experiment was performed as for Foi realizada uma experiência semelhante à da Fig a. Com excepção de ser

Name	mRNA id	PSP	Decision
AHSG	NM_001622.1	[illegible]	YES
ALB	BC023299	[illegible]	YES
APCS	NM_001639.2	[illegible]	YES
APOA1	NM_000039.1	[illegible]	YES
APOA2	NM_001643.1	[illegible]	YES
APOC2	NM_000483.3	[illegible]	YES
APOC3	NM_000040.1	[illegible]	YES
APOD	NM_001647.2	[illegible]	YES
APOE	NM_000041.2	[illegible]	YES
AZGP1	X59766.1	[illegible]	YES
CHI3L1	NM_001276.1	[illegible]	YES
CLU	NM_001831.2	[illegible]	YES
HRG	NM_000412.2	[illegible]	YES
IGFBP3	NM_000598.4	[illegible]	YES
INHA	NM_002191.2	[illegible]	YES
KLK11	NM_144947.1	[illegible]	YES
PKM2	NM_182470.1	[illegible]	YES
PLG	NM_000301.1	[illegible]	YES
SERPINA3	NM_001085.4	[illegible]	YES
TF	NM_001063.2	[illegible]	YES
TGFB1	NM_000660.3	[illegible]	YES
TTR	NM_000371.1	[illegible]	YES

Figura 18a

normal proteins
additional peptide of proteins whose STOP codon is affected

Figura 18b

>P02649|APOE Apolipoprotein E
MKVLWAALLVTFLAGCQAKVEQAVETEPEPELRQQTEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRALMDETMKELKAY
KSELEEQLTPVAEETRARLSKELQAAQARLGADMEDVCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKRLAVYQAGAR
EGAERGLSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMGSRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQAR
LKSWFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVPSDNH*TPKPAAMRPHATPCLLPPRSLQRETLSPPQPSSWGGP*
>Q5XKQ4|AZGP1 Alpha-2-glycoprotein 1, zinc-binding
MWASMSRMLPVLLSLLLLLGPAVPQENQDGRYSLTYIYTGLSKHVEDVPAFQALGSLNDLQFFRYNSKDRKSQPMGLWRQVEGMEDWKQDSQ
LQKAREDIFMETLKDIVEYYNDSNGSHVLQGRFGCEIENNRSSGAFWKYYYDGKDYIEFNKEIPAWVPFDPAAQITKQKWEAEPVYVQRAKAYLE
EECPATLRKYLKYSKNILDRQDPPSVVVTSHQAPGEKKKLKCLAYDFYPGKIDVHWTRAGEVQEPELRGDVLHNGNGTYQSWVVVAVPPQDTAP
YSCHVQHSSLAQPLVVPWEAS*EARVGGMVGSQTQ*
>P36222|CHI3L1 Chitinase-3-like protein 1
MGVKASQTGFVVLVLLQCCSAYKLVCYYTSWSQYREGDGSCFPDALDRFLCTHIIYSFANISNDHIDTWEWNDVTLYGMLNTLKNRNPNLKTLLS
VGGWNFGSQRFSKIASNTQSRRTFIKSVPPFLRTHGFDGLDLAWLYPGRRDKQHFTTLIKEMKAEFIKEAQPGKKQLLLSAALSAGKVTIDSSYDIA
KISQHLDFISIMTYDFHGAWRGTTGHHSPLFRGQEDASPDRFSNTDYAVGYMLRLGAPASKLVMGIPTFGRSFTLASSETGVGAPISGPGIPGRFT
KEAGTLAYYEICDFLRGATVHRTLGQQVPYATKGNQWVGYDDQESVKSKVQYLKDRQLAGAMVWALDLDDFQGSFCGQDLRFPLTNAIKDALA
AT*PSVLHTARGPRMPRPPLAPAGREPDHLPC*
>P10909|CLU Clusterin
MQVCSQPQRGCVREQSAINTAPPSAHNAASPGGARGHRVPLTEACKDSRIGGMMKTLLLFVGLLLTWESGQVLGDQTVSDNELQEMSNQGSK
YVNKEIQNAVNGVKQIKTLIEKTNEERKTLLSNLEEAKKKKEDALNETRESETKLKELPGVCNETMMALWEECKPCLKQTCMKFYARVCRSGSGL
VGRQLEEFLNQSSPFYFWMNGDRIDSLLENDRQQTHMLDVMQDHFSRASSIIDELFQDRFFTREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLM
PFSPYEPLNFHAMFQPFLEMIHEAQQAMDIHFHSPAFQHPPTEFIREGDDDRTVCREIRHNSTGCLRMKDQCDKCREILSVDCSTNNPSQAKLRR
ELDESLQVAERLTRKYNELLKSYQWKMLNTSSLLEQLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDVPSGVTEVVVKLFDSDPITVTVP
VEVSRKNPKFMETVAEKALQEYRKKHREE*DVDVAFAPTGASESSSPQDELQPPRESSARHQVTRPQPPGPQLRPASPRSGSGTLTLDSAAHG
KNRIAPACN*
>P04196|HRG Histidine-rich glycoprotein
MKALIAALLLITLQYSCAVSPTDCSAVEPEAEKALDLINKRRRDGYLFQLLRIADAHLDRVENTTVYYLVLDVQESDCSVLSRKYWNDCEPPDSRRP
SEIVIGQCKVIATRHSHESQDLRVIDFNCTTSSVSSALANTKDSPVLIDFFEDTERYRKQANKALEKYKEENDDFASFRVDRIERVARVRGGEGTGY
FVDFSVRNCPRHHFPRHPNVFGFCRADLFYDVEALDLESPKNLVINCEVFDPQEHENINGVPPHLGHPFHWGGHERSSTTKPPFKPHGSRDHHH
PHKPHEHGPPPPPDERDHSHGPPLPQGPPPLLPMSCSSCQHATFGTNGAQRHSHNNNSSDLHPHKHHSHEQHPHGHHPHAHHPHEHDTHRQ
HPHGHHPHGHHPHGHHPHGHHPHGHHPHCHDFQDYGPCDPPPHNQGHCCHGHGPPPGHLRRRGPGKGPRPFHCRQIGSVYRLPPLRKGEV
LPLPEANFPSFPLPHHKHPLKPDNQPFPQSVSESCPGKFKSGFPQVSMFFTHTFPK*KVPLKRKMNNTLN*

Figura 18c

>P17936|IGFBP3 Insulin-like growth factor-binding protein 3
MQRARPTLWAAALTLLVLLRGPPVARAGASSAGLGPVVRCEPCDARALAQCAPPPAVCAELVREPGCGCCLTCALSEGQPCGIYTERCGSGLR
CQPSPDEARPLQALLDGRGLCVNASAVSRLRAYLLPAPPAPGNASESEEDRSAGSVESPSVSSTHRVSDPKFHPLHSKIIIIKKGHAKDSQRYKVD
YESQSTDTQNFSSESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQPLPGYTTKGKED
VHCYSMQSK*TPAARLMVSSNNPYFAQ*QTAKDMTSSWLQPRHIPLFWN*
>P05111|INHA Inhibin alpha chain
MVLHLLLFLLLTPQGGHSCQGLELARELVLAKVRALFLDALGPPAVTREGGDPGVRRLPRRHALGGFTHRGSEPEEEEDVSQAILFPATDASCED
KSAARGLAQEAEEGLFRYMFRPSQHTRSRQVTSAQLWFHTGLDRQGTAASNSSEPLLGLLALSPGGPVAVPMSLGHAPPHWAVLHLATSALSLL
THPVLVLLLRCPLCTCSARPEATPFLVAHTRTRPPSGGERARRSTPLMSWPWSPSALRLLQRPPEEPAAHANCHRVALNISFQELGWERWIVYP
PSFIFHYCHGGCGLHIPPNLSLPVPGAPPTPAQPYSLLPGAQPCCAALPGTMRPLHVRTTSDGGYSFKYETVPNLLTQHCACI*SW*SVFLLNPMA
GGHAPTISWEERQSWEDQSHSSLLSLLCLWATLPTPLLSQ*
>Q9UBX7|KLK11 Kallikrein-11
MQRLRWLRDWKSSGRGLTAAKEPGARSSPLQAMRILQLILLALATGLVGGETRIIKGFECKPHSQPWQAALFEKTRLLCGATLIAPRWLLTAAHCL
KPRYIVHLGQHNLQKEEGCEQTRTATESFPHPGFNNSLPNKDHRNDIMLVKMASPVSITWAVRPLTLSSRCVTAGTSCLISGWGSTSSPQLRLPH
TLRCANITIIEHQKCENAYPGNITDTMVCASVQEGGKDSCQGDSGGPLVCNQSLQGIISWGQDPCAITRKPGVYTKVCKYVDWIQETMKNN*TQP
THRSPSPSISTWCLYPVHSVNKKP*
>P14618|PKM2 Pyruvate kinase isozymes M1/M2
MSKPHSEAGTAFIQTQQLHAAMADTFLEHMCRLDIDSPPITARNTGIICTIGPASRSVETLKEMIKSGMNVARLNFSHGTHEYHAETIKNVRTATES
FASDPILYRPVAVALDTKGPEIRTGLIKGSGTAEVELKKGATLKITLDNAYMEKCDENILWLDYKNICKVVEVGSKIYVDDGLISLQVKQKGADFLVTE
VENGGSLGSKKGVNLPGAAVDLPAVSEKDIQDLKFGVEQDVDMVFASFIRKASDVHEVRKVLGEKGKNIKIISKIENHEGVRRFDEILEASDGIMVA
RGDLGIEIPAEKVFLAQKMMIGRCNRAGKPVICATQMLESMIKKPRPTRAEGSDVANAVLDGADCIMLSGETAKGDYPLEAVRMQHLIAREAEAA
MFHRKLFEELVRASSHSTDLMEAMAMGSVEASYKCLAAALIVLTESGRSAHQVARYRPRAPIIAVTRNPQTARQAHLYRGIFPVLCKDPVQEAWA
EDVDLRVNFAMNVGKARGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP*WTPEPLLQPLSHPLPPAHPLGQQRL*
>P00747|PLG Plasminogen
MEHKEVVLLLLLFLKSGQGEPLDDYVNTQGASLFSVTKKQLGAGSIEECAAKCEEDEEFTCRAFQYHSKEQQCVIMAENRKSSIIIRMRDVVLFEK
KVYLSECKTGNGKNYRGTMSKTKNGITCQKWSSTSPHRPRFSPATHPSEGLEENYCRNPDNDPQGPWCYTTDPEKRYDYCDILECEEECMHC
SGENYDGKISKTMSGLECQAWDSQSPHAHGYIPSKFPNKNLKKNYCRNPDRELRPWCFTTDPNKRWELCDIPRCTTPPPSSGPTYQCLKGTGE
NYRGNVAVTVSGHTCQHWSAQTPHTHNRTPENFPCKNLDENYCRNPDGKRAPWCHTTNSQVRWEYCKIPSCDSSPVSTEQLAPTAPPELTPV
VQDCYHGDGQSYRGTSSTTTTGKKCQSWSSMTPHRHQKTPENYPNAGLTMNYCRNPDADKGPWCFTTDPSVRWEYCNLKKCSGTEASVVAP
PPVVLLPDVETPSEEDCMFGNGKGYRGKRATTVTGTPCQDWAAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDVGGPWCYTTNPRKLYDYCD
VPQCAAPSFDCGKPQVEPKKCPGRVVGGCVAHPHSWPWQVSLRTRFGMHFCGGTLISPEWVLTAAHCLEKSPRPSSYKVILGAHQEVNLEPHV
QEIEVSRLFLEPTRKDIALLKLSSPAVITDKVIPACLPSPNYVVADRTECFITGWGETQGTFGAGLLKEAQLPVIENKVCNRYEFLNGRVQSTELCAG
HLAGGTDSCQGDSGGPLVCFEKDKYILQGVTSWGLGCARPNKPGVYVRVSRFVTWIEGVMRNN*LDSRQSQALTTWVGI*

Figura 18d

>P01011|SERPINA3 Alpha-1-antichymotrypsin
MERMLPLLALGLLAAGFCPAVLCHPNSPLDEENLTQENQDRGTHVDLGLASANVDFAFSLYKQLVLKAPDKNVIFSPLSISTALAFLSLGAHNTTLT
EILKGLKFNLTETSEAEIHQSFQHLLRTLNQSSDELQLSMGNAMFVKEQLSLLDRFTEDAKRLYGSEAFATDFQDSAAAKKLINDYVKNGTRGKITD
LIKDLDSQTMMVLVNYIFFKAKWEMPFDPQDTHQSRFYLSKKKWVMVPMMSLHHLTIPYFRDEELSCTVVELKYTGNASALFILPDQDKMEEVEA
MLLPETLKRWRDSLEFREIGELYLPKFSISRDYNLNDILLQLGIEEAFTSKADLSGITGARNLAVSQVVHKAVLDVFEEGTEASAATAVKITLLSALVE
TRTIVRFNRPFLMIIVPTDTQNIFFMSKVTNPKQA*SLPSSSGALSKELGMQAGCLGLWAQPGPCAPSGHGMQGPVCLSLEGDSDSLCSSHMHR
GPWTLQSQGSVAS*
>P02787|TF Serotransferrin
MRLAVGALLVCAVLGLCLAVPDKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTLDAGLVYDAYLAPNN
LKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQL
CQLCPGCGCSTLNQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCHLAQVPSHTVVARSMGGKEDLIW
ELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGFLKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKPVKWCALSHHERLKCDE
WSVNSVGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGKCGLVPVLAENYNKSDNCEDTPEAGYFAVAVVKKSASDLTWDNLKGKKSCH
TAVGRTAGWNIPMGLLYNKINHCRFDEFFSEGCAPGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAFRCLVEKGDVAFVKHQTVPQNTGG
KNPDPWAKNLNEKDYELLCLDGTRKPVEEYANCHLARAPNHAVVTRKDKEACVHKILRQQQHLFGSNVTDCSGNFCLFRSETKDLLFRDDTVCL
AKLHDRNTYEKYLGEEYVKAVGNLRKCSTSSLLEACTFRRP*NLRGRAATKVKMGTQMHEFALVSLAQVVCANHVCLHSSVLPCVLNKK*
>P01137|TGFB1 Transforming growth factor beta-1
MPPSGLRLLPLLLPLLWLLVLTPGRPAAGLSTCKTIDMELVKRKRIEAIRGQILSKLRLASPPSQGEVPPGPLPEAVLALYNSTRDRVAGESAEPEP
EPEADYYAKEVTRVLMVETHNEIYDKFKQSTHSIYMFFNTSELREAVPEPVLLSRAELRLLRLKLKVEQHVELYQKYSNNSWRYLSNRLLAPSDSP
EWLSFDVTGVVRQWLSRGGEIEGFRLSAHCSCDSRDNTLQVDINGFTTGRRGDLATIHGMNRPFLLLMATPLERAQHLQSSRHRRALDTNYCFS
STEKNCCVRQLYIDFRKDLGWKWIHEPKGYHANFCLGPCPYIWSLDTQYSKVLALYNQHNPGASAAPCCVPQALEPLPIVYYVGRKPKVEQLSN
MIVRSCKCS*GPAPPRPAPACPAPPRPAPAALPMGAVFKDTRAPSPPGAPLKMERGLRISVSLGACLGSPSLTPPHSHSLSLPLCLLLPVCTIPLP
GIKAQGTSGEHYCS*
>P02766|TTR Transthyretin
MASHRLLLLCLAGLVFVSEAGPTGTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTTEEEFVEGIYKVEIDTK
SYWKALGISPFHEHAEVVFTANDSGPRRYTIAALLSPYSYSTTAVVTNPKE*GTSPPVDLKDEGWDFM*

Figura 18e

Previsões

	M	nM	Total	
M	315	132	**447**	70.47%
nM	**705**	1725	**2428**	71.05%
Total	**1020**	**1855**	2875	
	TP 30.88%	TN 92.99%		70.83%

Observação

TP: Verdadeiro positivo; TN: Verdadeiro negativo

Figura 19

REFERÊNCIAS CITADAS NA DESCRIÇÃO

Esta lista de referências citadas pelo requerente é apenas para conveniência do leitor. A mesma não faz parte do documento da patente Europeia. Ainda que tenha sido tomado o devido cuidado ao compilar as referências, podem não estar excluídos erros ou omissões e o IEP declina quaisquer responsabilidades a esse respeito.

Documentos de patentes citadas na Descrição

- US 6329147 B
- US 4979330 A
- WO 02077286 A
- US 6120992 A

Literatura que não é de patentes citada na Descrição

- **GOTT, J. M. ; EMESON, R. B.** *Annu Rev Genet*, 2000, vol. 34, 499-531
- **MAAS, S. ; RICH, A.** *Bioessays*, 2000, vol. 22, 790-802
- **NISWENDER, C. M.** *Cell Mol Life Sci*, 1998, vol. 54, 946-64
- **EISENBERG, E. et al.** *Nucleic Acids Res*, 2005, vol. 33 (14), 4612-7
- **RUIZ et al.** *Clinical cancer Research*, 2004, vol. 10 (8), 2560-2567
- **PAN ; WEISSMAN ; LIU et al.** *PNAS*, 2002, vol. 99 (14), 9346-9351
- *Anal Biochem.*, 01 September 2006, vol. 356 (1), 117-24
- Fundamental Immunology. **PAUL, W.E.** Fundamental Immunology. Raven Press, 1989, 176
- **HARLOW et al.** Antibodies: A laboratory Manual. CSH Press, 1988
- **WARD et al.** *Nature*, 1989, vol. 341, 544
- **KOHLER-MILLSTEIN ; GALFRE, G. ; MILSTEIN, C.** *Methods Enz.*, 1981, vol. 73, 1
- Remington's Pharmaceutical Sciences. Mack Publishing
- Basic Local Alignment Search Tool, Zheng Zhang, A greedy algorithm for nucleotide sequence alignement search. *J Comput Biol*, 2000
- **SHERRY, S. T. et al.** *Nucleic Acids Res*, 2001, vol. 29, 308-11
- **C.DALMASSO ; P.BROET.** *Journal de la Societe Française de Statistique*, 2005, vol. 146 (1-2
- **BRULLIARD M. et al.** *PNAS*, 01 May 2007, vol. 104 (18), 7522-7527
- **SJOBLOM, T. et al.** *Science*, 2006, vol. 314, 268-74
- **THOMAS, R. K et al.** *Nat Med*, 2006, vol. 12, 852-5