

(12) FASCÍCULO DE PATENTE DE INVENÇÃO

(22) Data de pedido: 2007.07.20	(73) Titular(es): TRANSMEDI SA 15 RUE DU BOIS DE LA CHAMPELLE 54500 VANDOEUVRE LES NANCY FR
(30) Prioridade(s): 2006.07.20 EP 06291176	
(43) Data de publicação do pedido: 2009.04.15	(72) Inventor(es): BERNARD BIHAIN FR
(45) Data e BPI da concessão: 2011.09.14 248/2011	(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:

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.

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 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 o estabelecimento do perfil de doentes ou doenças, previsão da gravidade das doenças e avaliação da eficácia de fármacos.

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 adota 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 códons (um códon = 3 nucleótidos) presentes numa determinada ordem no mRNA e 4) pequenos RNAs reguladores referidos como RNA não não codificador (cnRNA).

A sequência de mRNA é então transformada numa linguagem diferente, *i.e.*, a linguagem dos AA por 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.

Foi feita uma observação da leitura molecular errada em indivíduos tendo distúrbios neurodegenerativos, tais como doença de Alzheimer, (Hol Elly M *et al.*: J. Biol. Chem., vol. 278, nº 41, 10 October 2003, pg 39637-39643; Laxminarayana Dama *et al.*, International Immunology, vol. 12, nº 11, November 2000, pg 1521-1529; Fischer David F *et al.*: FASEB Journal, vol. 17, nº 14, November 2003, pg 2014-2024; Van Den Hurk Wilhelmina H *et al.*: J. Biol. Chem. , vol. 276, nº 15, 13 April 2001, pg 11496-11498; Gerez L *et al.*: Neurobiology of Aging, Tarrytown, NY, US Lnk-Doi:10.1016/J. Neurobiolaging 2004.03.011, vol. 26, nº2, 1 February 2005, pg 145-155; W001/40804). Nenhum mecanismo está caracterizado nestas referências.

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.

Em condições de 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 bases 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 de 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.

Sumário do presente invento

O presente invento está relacionado com composições e métodos de utilização, detecção ou alteração da infidelidade da transcrição em células, particularmente em células de mamífero. O invento é particularmente adequado para fins terapêuticos e de diagnóstico, *e.g.*, para detectar e/ou tratar doenças causadas ou associadas a um aumento ou decréscimo do nível de infidelidade da transcrição.

Um objectivo particular deste invento está relacionado com um método de detecção da presença ou fase de um cancro num indivíduo, o método compreendendo a avaliação (*in vitro* ou *ex vivo*) da presença ou taxa de

infidelidade de transcrição numa amostra do referido indivíduo, a referida presença ou taxa sendo uma indicação da presença ou fase de cancro no referido indivíduo.

Um outro objectivo deste invento reside num método de rastreio, identificação ou optimização de um fármaco, o método compreendendo um passo de avaliação se os referidos fármacos alteram (e.g., reduzem ou aumentam) a taxa de infidelidade da transcrição de um gene.

Um outro objectivo deste invento está relacionado com 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 cancro.

Um outro objectivo deste invento reside num método de avaliação da eficácia de um fármaco ou candidato a fármaco, o método compreendendo um passo de avaliação se o referido fármaco altera (e.g., reduz ou aumenta) a taxa de infidelidade de transcrição de um gene de mamífero, tal alteração sendo uma indicação da eficácia do fármaco.

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

O invento também proporciona 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 está relacionado com um método de identificação de biomarcadores de cancro, 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. 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 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 cancros. 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 proliferativo celular num indivíduo.

O 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.).

O 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 pro-

teínas, 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.

É também divulgada neste invento a utilização de um péptido criado por infidelidade da transcrição como definido atrás, como imunogénio. O invento também revela uma composição de vacina compreendendo um péptido compreendendo uma parte de uma proteína criada por infidelidade de transcrição, como definido atrás, e facultativamente um veículo, excipiente e/ou adjuvante adequado.

É igualmente revelado neste invento um método de produção de um anticorpo, o método compreendendo a imunização de um mamífero não humano com um péptido criado pela infidelidade de transcrição, como atrás definido, e recuperação dos anticorpos que se ligam ao referido péptido, ou correspondentes células produtoras de anticorpo. Facultativamente, podem ser produzidos derivados do anticorpo.

Numa realização particular, o invento 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 os derivados do anticorpo.

É também divulgado neste invento 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. 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.

Uma outra descrição deste invento está relacionada 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 está relacionado com 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 descreve 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.

É ainda descrito neste invento 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.

É ainda descrita neste invento 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.

O invento também descreve 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 um aumento ou a uma redução da taxa de infidelidade de transcrição, como seja *e.g.*, distúrbios proliferativos celulares (*e.g.*, cancros), doenças imunes (*e.g.*, doenças

auto-imunes (esclerose múltipla, esclerose amiotrófica lateral (ALS), rejeição de enxertos), envelhecimento, doenças inflamatórias, diabetes, etc., e/ou para produzir, desenhar ou testar fármacos terapeuticamente activos. O invento pode ser igualmente usado para detectar, monitorizar, modular ou atingir a infidelidade da transcrição em quaisquer outros tecidos ou tipos de células, tais como células procarióticas, eucariotas inferiores (e.g., leveduras), células de insectos, células vegetais, fungos, etc.

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

Dentro do contexto do presente invento, 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 infidelida-

de 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 substituição, 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 timi-

dina 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 reflectam 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 anticorpos, 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 transcrição 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 microarranjos 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 deleção de nucleótidos descritas no presente pedido de patente (ver, e.g., Exemplo 5).

Dentro do contexto deste invento, 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 trans-

crição, 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.

Uma realização específica deste invento é uma sonda de ácido nucleico que selectivamente hibrida 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 preferidas deste invento 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 alcóxirribonucleótidos 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.

De acordo com uma outra realização (a qual pode ser usada em combinação com sondas de hibridação), a infidelidade de transcrição é 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 polinucleotídica 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.

Um outro aspecto deste invento também inclui 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. Exemplos específicos de sequências iniciadoras deste invento estão descritos, *e.g.*, na Fig 12.

De acordo com uma outra realização deste invento, 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 eluídos 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 NH₂).

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.

Neste contexto, o invento também está relacionado com qualquer ligando que selectivamente se ligue a um domínio de uma proteína TI. Diferentes tipos de ligandos podem ser considerados, tais como anticorpos específicos, moléculas sintéticas, aptâmeros, péptidos e similares.

Numa realização específica, o ligando é 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 injeção 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. Raven 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 antígeno, particularmente à proteína nativa não contendo o referido domínio). Os anticorpos selectivos para os domínios das proteínas TI permitem a detecção da presença de proteínas contendo tais domínios numa amostra.

Diagnóstico

O presente invento permite 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 farmacogenética, prognóstico, etc.

Num aspecto particular, o invento está relacionado com um método de detecção *in vitro* ou *ex vivo* da presença, ausência, predisposição, risco ou gravidade de cancro 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.

Um objectivo particular deste invento reside num 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.

Uma outra realização deste invento é dirigida a um método de avaliação da resposta ou responsividade 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 está relacionado com 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 do 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 imune 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.

Um objectivo particular deste invento reside num 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.

Uma outra realização do presente invento reside num 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.

Uma outra realização deste invento reside num 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 realizado 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 imunes, 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., anti-corpos, 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, injeções 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 farmacêuticamente aceitável 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 transmissor), 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 descreve 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 imune para reconhecer um padrão proteico específico. Neste contexto, pensamos que a infidelidade da transcrição desempenha um papel na patogénese de doenças imunes. 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 físicoquímicas 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 pro-

teínas 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, desenhamos péptidos específicos tendo a mesma sequência que um domínio resultante de TI. Eles podem ser usados para detectar auto-anticorpos 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 imunes em doentes diagnosticados com qualquer forma de doença (e.g., de cancro) ou para iniciar respostas imunes 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.

Um objectivo particular deste invento está assim relacionado com um método de detecção da presença ou fase de um cancro 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.

Um outro objectivo deste invento está relacionado com 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 no tratamento de um ser humano ou animal, particularmente para o tratamento de cancros.

Um outro objectivo deste invento reside num 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 está relacionado com 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 está relacionado com um método de identificação de biomarcadores de cancro, 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.

Uma outra descrição do invento está relacionada com 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.

Uma outra descrição deste invento está relacionada com 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) KLHTLSAAIYYQQE (VIM) (SEQ ID NO: 2)
 DFLSNK (RPS6) (SEQ ID NO: 3) MYTVEFSVHKNN (RPL7A) (SEQ ID NO: 4) NGS LGDMSDLCT (RPS4X) (SEQ ID NO: 5) ASG (ETH1) (SEQ ID NO: 6) EPSEPSDF (FTL) (SEQ ID NO: 7)
 APSIFPTLPAPKPGTKQPRSPVTALSLHMLLMVSSAPSCGLIQT VSSFTVYIFTL (TPI1) (SEQ ID NO: 8) ARHGRDEEVWHRKSHHFEVQAWAVGGLVCWPRKCHMRSTLISSLD SLLFVIPHRTEAEWV VMEDRRH (AHSG) (SEQ ID NO: 9)
 RSKAYSSVFLFRWCKANTLSKKHKFL (ALB) (SEQ ID NO: 10) GLDSTRALENEMTV (APCS) (SEQ ID NO: 11)
 GARRRPPSRCSE (APOA1) (SEQ ID NO: 12) SVQTI VFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 13)
 IVFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 14) QPDPPSVDKGRVPYSPDPGSD (APOC2) (SEQ ID NO: 15)
 DPPSVDKGRVPYSPD (APOCII) (SEQ ID NO: 16) DLNTPSPPAYPSCCELLGSCNLOGCPCRLKRD S ILSALLPHLMPGPPPGMLASQ (APOC3) (SEQ ID NO: 17)
 PGSTGR LHPLHVTSASLSPTPPPHKDKPINHDKGS (APOD) (SEQ ID NO: 18) TPKPAAMRPHATPCLLPFRSLQRETLSPFPQSSWGGP (APOE) (SEQ ID NO: 19)
 EARVGGNVGSQTQ (AZGP1) (SEQ ID NO: 20) PSVLNTARGPRMPRPPLAPAGREPDLPC (CHI3L1) (SEQ ID NO: 21)
 DVDVAFAPTGASESSSPQDELQPPRESSARHQVTRPQPPGPQLRPASPRSGSCTLTLD SAAHG NRIAPACN (CLU) (SEQ ID NO: 22)
 NVIPLKRKMNTLN (HRG) (SEQ ID NO: 23) TPAARLMWSSNMPYFAQKTAKOMTSSWLQPRFIFLVVN (IGFBP3) (SEQ ID NO: 24)
 GWGVFLLNPMAGGHAPTIISWEERQSWEIDGSHSSLLSLLCLWATLPTPLLSQ (INHA) (SEQ ID NO: 25)
 TGPTHHSPPSPSISTWCLVPVHSVNNKKP (KLK11) (SEQ ID NO: 26) WTPEPLLQPLSHPLPPAHPLGQORL (PKM2) (SEQ ID NO: 27)
 LDGRQSDALHLEAGTWVGI (PLG) (SEQ ID NO: 28) SLPSSSGALSKELGMAQAGCLGLWAQPGCAPSGHGMCGPVCLSLEGDSDSLCS SHMRGPWTL SGGSWAS (SERPINA3) (SEQ ID NO: 29)
 NLRGRAATKVKMGTQMIHEFALVSLAQVVCANHVCLHSSVLPVNLKK (TF) (SEQ ID NO: 30)
 GPAPPRPAPAGPAPPRPAPAALPMGAVFKDTRAPSPPGAPLKMERGLRISVSLGACLGSPSLT PHSLSLSLPLCLLLPVCTIPLPGIKAQGTSGEHYCS (TGFB1) (SEQ ID NO: 31)
 GTSPFVDLKDEGWDFM (TTR) (SEQ ID NO: 32)

Uma outra descrição deste invento reside na utilização de um péptido compreendendo um local de infidelidade da transcrição, 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 descrito atrás, como um imunogénio.

delidade da transcrição, como definido atrás, e facultativamente um veículo, excipiente e/ou adjuvante adequado.

O invento descreve 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 um aumento ou redução da taxa de infidelidade da transcrição, como seja, *e.g.*, distúrbios proliferativos das células (*e.g.*, cancros), doenças imunes (*e.g.*, doenças auto-imunes (esclerose múltipla, ALS) rejeição de enxerto), envelhecimento, doenças inflamatórias, diabetes, etc., e/ou produzir, desenhar ou testar fármacos terapêuticamente activos. O invento pode também ser usado para detectar, monitorizar, modular ou eliminar a infidelidade da transcrição em qualquer outro tipo celular, como sejam células procarióticas, eucariotas inferiores (*e.g.*, leveduras), células de insecto, células vegetais, fungos, etc.

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.*

Os agentes para atingir alvos podem ser igualmente usados como uma molécula terapêutica, *e.g.*, para tratar uma célula ou tecido ou organismo apresentando infidelidade anormal da transcrição. O agente pode ser usado como tal ou, de preferência, acoplado a uma molécula activa, como seja uma molécula tóxica, um fármaco, *etc.*

Uma realização particular deste invento descreve assim 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 sequência 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 imune 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 imune 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 farmacêuticamente aceitável. A composição de vacina pode ser usada para gerar uma resposta imune contra uma célula ou tecido doente expressando proteínas TI, resultando numa destruição da referida célula ou tecido pelo sistema imune. 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.

Uma realização particular deste invento descreve assim um método de produção de um imunogénio que causa, estimula ou inibe uma resposta imune 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 passo (i) do método compreende a identificação de um local de infidelidade da transcrição ou proteína da superfície celular (*e.g.*, um receptor, etc.) ou de uma ácido nucleico. De facto, as proteínas secretadas e as proteínas da superfície celular estão mais sujeitas a uma resposta pelo sistema imune. Exemplos de domínios de tais proteínas TI estão descritos nos exemplos.

Redução da infidelidade da transcrição em sistemas produtores de proteínas recombinantes

Uma revelação particular deste invento reside num 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

Uma divulgação particular deste invento reside num 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 imune 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 imune, determinando portanto a taxa de infidelidade da transcrição nas células circulantes do sistema imune, *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 deverão 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 descritas na secção experimental que se segue, a qual ilustra o invento e não limita o âmbito do presente pedido.

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 eluído com um tampão de baixo teor de sal. A cauda poli(A) do mRNA hibrida então com oligo dT na presença de uma transcriptase 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 hibrida

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 segmento 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 alignment 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 ($\square$) e SNPs biologicamente validados ($\circ$); 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. Dalmaso, 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 restrigente 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 51**).

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 n° 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 $\geq 90\%$ 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 parálogos 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 otimizar 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%) $\approx$ T (32%) $\gg$ C (21%) $\gg$ 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 $A \geq G > T \approx C$.

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 sub-série, 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 $b_0 = A$ (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 $b_0 = T$ (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 envolvidas 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 $b_0 > b-1 > b-2 > b+1 > b-3 > b+2$. Para as substituições $b+1$, a influência relativa da base envolvente seguiu um padrão de $b_0 > b+1 > b+2 > b-3 > b-2$. 2) Na segunda sub-série 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) $b-1 = b_0 = b+1$ (**Fig 7c**). Quando $b-1$ e $b+1$ eram idênticos mas diferentes de b_0 ($n = 339$; **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 selecionadas ao acaso. Ambos os fenômenos 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 triplete 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 imune.

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 possuisse 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 injeções 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 reproduzíveis 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 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 fluorescentes 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 possuísem 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 PSP. 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 eluída 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 eluída 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 dissulfureto.

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 ressuspensos 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 Dezembro de 2011

REIVINDICAÇÕES

1. Um método de detecção da predisposição para ou da fase de um cancro num indivíduo mamífero, o método compreendendo a avaliação *in vitro* ou *ex vivo* da presença ou taxa de:

a) uma ou mais variações da sequência em moléculas de nucleótidos numa amostra do referido indivíduo, as referidas variações sendo seleccionadas entre deleções de bases de nucleótidos (espaços), inserções e alterações de famílias de bases; e/ou

b) uma ou mais modificações da proteína resultando das variações de sequência do passo a); em que as referidas variações de sequência são introduzidas no transcrito primário de RNA durante a transcrição de DNA ou de RNA através de infidelidade da transcrição (YI) e a referida presença ou taxa de variações de sequência e/ou modificações da proteína é uma indicação da predisposição para ou fase de um cancro no referido indivíduo.

2. Um método de rastreio, identificação ou optimização de um fármaco anti-cancro, o método compreendendo um passo de avaliação se o referido fármaco altera a taxa de variações de sequência e/ou modificações de proteína como definido na reivindicação 1.

3. Um método de modificação de um gene de mamífero *in vitro*, o método compreendendo a modificação da se-

quência do gene para reduzir a ocorrência ou taxa de infidelidade da transcrição como definido na reivindicação 1.

4. Um método de avaliação da eficácia de um fármaco anti-cancro ou fármaco candidato, o método compreendendo um passo de avaliação se o referido fármaco altera a taxa de variações de sequência e/ou modificações de proteína como definido na reivindicação 1, tal alteração sendo uma indicação da eficácia do fármaco.

5. O método de qualquer uma das reivindicações 1 a 4, em que as alterações da família de bases é/são seleccionadas entre A→C, T→G, G→T, A→T, A→U, G→C, G→U, T→A, C→A e C→G, de preferência de A→C e T→G.

6. O método de qualquer uma das reivindicações 1 a 5, em que a presença ou taxa de variações de sequência e/ou modificações de proteína introduzidas por infidelidade da transcrição é avaliada usando pelo menos um ligando específico.

7. Um método de identificação de biomarcadores de cancro, o método compreendendo a identificação, numa amostra de um indivíduo tendo um cancro, da presença de variações de sequência e/ou modificações de proteína introduzidas por TI numa proteína alvo ou RNA como definido na reivindicação 1, e, facultativamente, determinação da sequência das referidas variações de sequências e/ou

modificações de proteínas, a referida proteína ou RNA tendo as referidas variações de sequência e/ou modificações de proteína sendo um biomarcador.

8. O método ou utilização de qualquer uma das reivindicações 1 a 7, em que a proteína é uma proteína da superfície celular ou uma proteína secretada, de preferência uma proteína do plasma ou um receptor.

9. O método ou utilização de qualquer uma das reivindicações 1 a 8, em que a proteína compreende uma sequência peptídica seleccionada entre SEQ ID NO:1-32.

10. O método ou uso de qualquer uma das reivindicações 1 a 9, em que a variação da sequência obedece às seguintes regras:

- A substituição de A ocorre preferencialmente quando A é precedido ou seguido de C,
- A substituição de T ocorre preferencialmente quando T é precedido ou seguido de G,
- A substituição de CA ocorre preferencialmente quando C é precedido ou seguido de G ou A e
- A substituição de G ocorre preferencialmente quando G é precedido ou seguido de A.

11. O método ou utilização de qualquer uma das reivindicações 1 a 10, em que o cancro é cancro da mama ou cancro do cólon.

12. Um método da reivindicação 1, para detecção da predisposição para ou fase de um cancro num indivíduo mamífero, o método compreendendo a avaliação *in vitro* ou *ex vivo* da presença ou taxa de:

a) uma ou mais variações de sequência nas moléculas nucleotídicas numa amostra do referido indivíduo, as referidas variações sendo seleccionadas entre as alterações de bases nucleotídicas A→C e T→G; e/ou

b) uma ou mais modificações de proteína resultantes das variações de sequência do passo a); em que as referidas variações de sequência são introduzidas no transcrito primário de RNA durante a transcrição do DNA para RNA por infidelidade da transcrição (TI) e a referida presença ou taxa de variações de sequências e/ou modificações de proteína é uma indicação da predisposição, presença ou fase de um cancro no referido indivíduo.

Lisboa, 12 de Dezembro de 2011

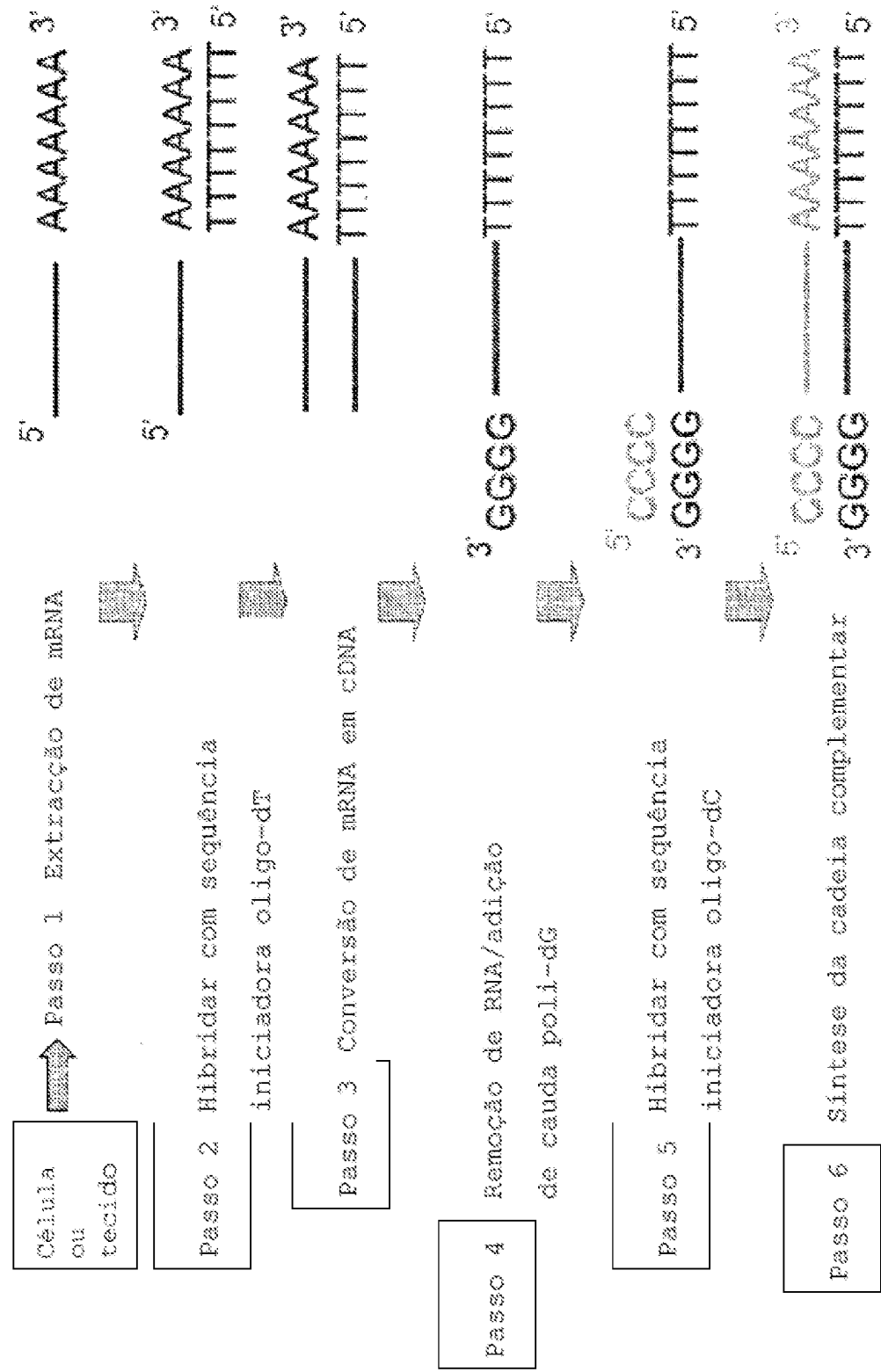


Figura 1a

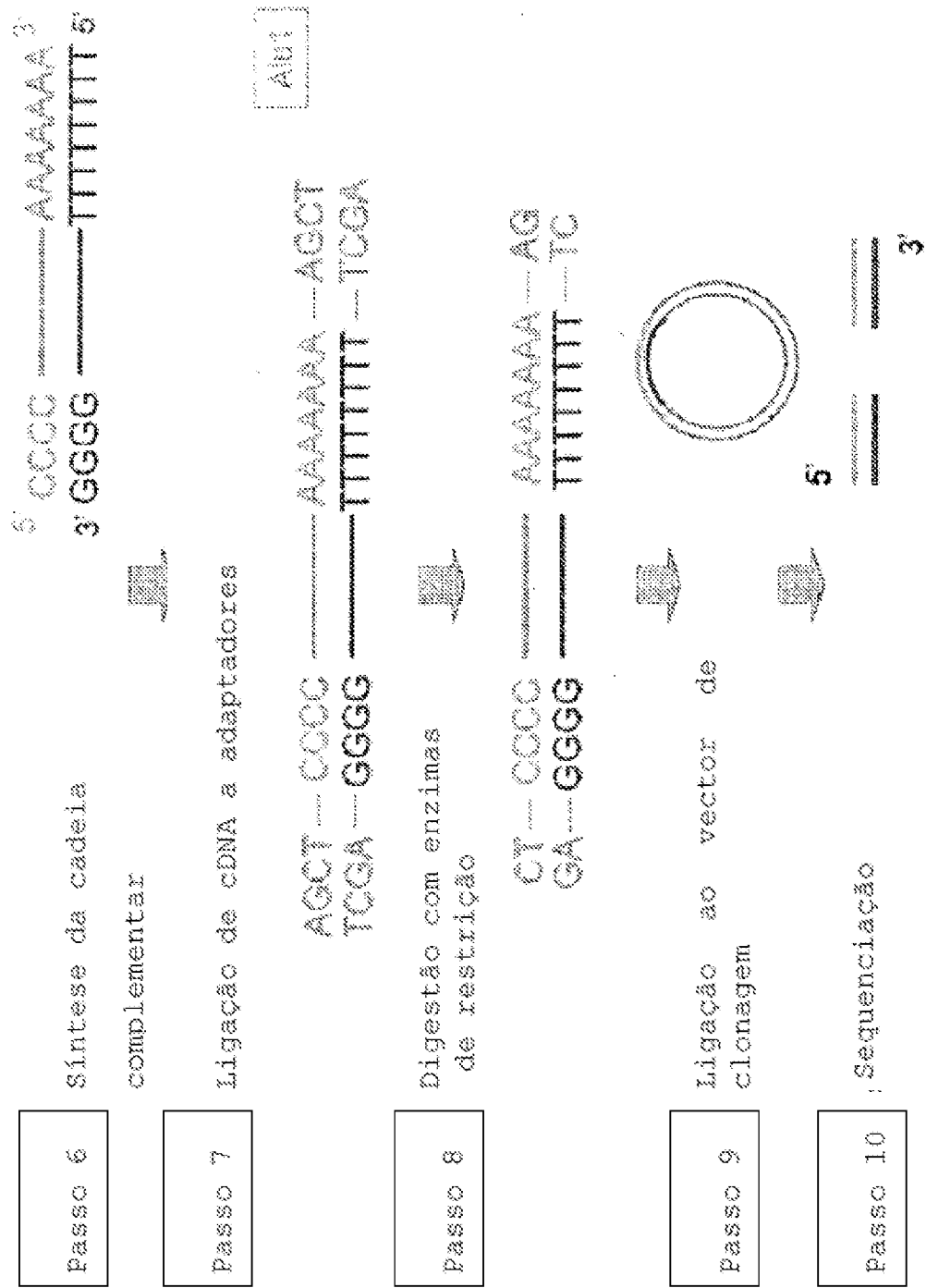


Figura 1b

[illegible]

Figura 2a

>gil56682958|ref|NM_002032.2| Homo sapiens ferritin, heavy polypeptide 1 (FTH1), mRNA

ATAAGAGACCA C AAGCGACCCGTASGGCCAGAGGTTCTTGGCCGAGAGTGGTCGGGGTTTCCCTCTTCACAGTGCCTTGGACGGGAAC
 CCGGCSCTCGTTCCCCACCCCGGCGGCGGCGCCATAGCCAGGCCCTCCGTCAGCTTCTTCACCGGACCCCTCGHACTGGCCGACGAGGCCCC
 CGCGCGCGCTCCAGCGCGGCGCAGGCCACCGCGCGGCGGCGGCGCTCTCTTASTGCGCGGCCATGACGAGACCGCGCTCCACCTGGCAGGT
 GCGCCGAGAACTACCCGCCAGGACTCAGAGGCGCGCCATCAGCGGCGGAGATCAACCTGGAGCTCTAAGCCCTCTTACGTTTACCTGTCCAI
 GTCTTACTACTTTTACCCGGGATGATGTGGCTTTGAGGAGACTTTGCCAATATACTTTCTTCACCAATCTCTCATGAGGAGAGGGAAACATGC
 TGAGGAACTGATGAAGCTGCAGAACCCAGCAGGTGGCCGAATCTTCTCTTCAGGATATCAGGAAACCCAGAGCTGTGTATGACTGAGGAGAG
 CGGGCTGAAATGCCAATGGAGTGTGCHTTACATTTGGGAAAAAATGTGAATCANTCACTACTGTGGAACCTGCACAACTGGCCGCTGACAA
 AATGACCCCCCATTTTGTGTSACTTCATTTGACGACACATTAGCTGAAATGAGCAGGTGAAGCCCATCAAGAGAAATGGGGTGACCCGCTGAC
 CAACCTGGCGCAAGATGGGAGGCGCCCGAATCTGGCTTGGCGGAAATATCTCTTTGACAGCAGCACCCTGGGAAACAGTGTGATATGAAAG
 CTAAAGCCTGGGGCTAATTTCCCCATAGCCCTGGGGTGACTTCCCTGHTCACCCAGGAGTGCATGATGTGTGGGSETTTCTTTTACCT
 TTCTATAGTTGTACCAAAACATCCACTTAAAGTTCTTTGATTTTGTACCAATTCTCTTCAAAATPAAGAAATTTGGGTACCCAGGCTGTGT
 CTTTGAGGTCTTTGGGATGAAATCAGGARAATCTATCAGGCTATCTTCCAGATTCCTTAAAGTGGCGTTGTTCACTTCTTAAATCACACTAAT
 CAAAAAGAAAGGAGTATTTGTATTTATTAACCTCATTAGTTTGGGAGATATACCTAGGTGTGGCTGTCTTGGATTCAATAGAACTA
 AAGGTTCCCGACTCTGAAATCCAGAGTCTGAGTTAATGTTTCCAAATGGCTTCAGTCTTAGCTTTACAGTCTTTTATGAAATAAAGGCCAT
 TAAAGGCTGA

Figura 2b

>gi56682960|ref|NM_000146.3| Homo sapiens fertilin, light polypeptide (FTL), mRNA

GAATTCGGGGTCCCGGGGTCGTCTCTTTTCTTCAACAGTGTGTTGGACGGAACAGATCCGGGGAUTCTCTTCCAGGCTCCGACCGC
CCTCCGATTTCTCTCTGGCTTGGCAAGCTTCGGGACCATCTCTCGGCCATCTCCCTGCTCTCTGGAGCTTCCACACACCGTTTCTGTGCT
TAGCTCTTTCTTGGCAACCAACCAATGAGCTTCCAGATTCTGTCAGAAATTATTCACCGAGCTGAGGGAAGCCGTCGACAGCCCTGGTCAA
TTTGTACCTGCAAGGCTCTTACACCTTACCTCTCTCTGGGCTTCTATTCTGACCTGCGATGATGTGCTCTGGAAGGCTTAGGCACTTC
TTCCGGGAATTGGCCGAGGAGAAAGCGGAGGGCTACGAGCGTCTCTGTAAGATGCAAAACAGCGTGGCGGCGGGCTCTCTCTCCAGG
ACATCARGAAGGCAAGCTGAAGATGAGTGGGTAATAAACCCGAGAGCCCATGAAAGCTGCCATGGCCCTGAGAGAAAGCTGACACAGGC
CCTTTTGGATCTTCATCCCTGCGCTTCTGCGCGGACGGACCCCATCTCTGTGACTTCCCTGGAGATCACTCTCTCTAGATGAGGAAGTG
AAGCTTAICAGAGATGGATGACCACTGAGCCACCTCCACAGGCTGGTGGTGGCCCGAGGCTGGCTGAGGAGATATCTCTCTGAA
GGCTCACTCTCAGCACGACTAAGAGGCTTCTGAGCCCGAGGACTTCTGAGGGCCCTTGGCAAAATAGGGCTCTCTGCTAAGCC
TCTCCCTCCAGACATAGGCAAGCTTTCTTAAGTATCTTACAGGCTTGGACCAATGGGAATAAAGCTTTTGTGATGC

Figura 2c

>gi|83841890|ref|NM_002046.3| Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA

AAATTGAGCCCGCAGGCTCCGGCTTCGGTCTCTGCTCTCGACAGTCAAGCCGCACTCTCTTTTGGGTGCGCAGCGCGGCGACAG
TCGCTTAGACACCAATGGGGAGGTGAGGTGGAGTCAACGGATTGTGGTGGTATTGGGGGCGCTGCTCACCAGGCGCTGCTTTTAACTCTG
GTAAAGTGGATATTGTTGCCATCAATGACCCCTTCAATTGACCTCAACTACATGAGTTTACATGTTCCCAATATATGATTCTACCCCATGGCAAA
TTCCATGSCAACCGTCAAGSCTGAGAACGGGAGACTTGTGATCAATCAATGGAATCCCATGACCATCTTCCAGGAGCGAGATCCCTCCCAAAAT
CAAGTGGGGGATGCTGTGGGCTGAGTACGTGTGAGTCCACTGGGCTCTTCAACCACTGGGAGAAAGGCTGGGGCTCATTTTGCAGGGG
GAGCCAAAGGCTCATCATCTCTGTCGCCCCCTCTGTGTGATGCCCCCATGTTCTGTCATGGGTGAAACCATGAGAAATATGACAAACAGCCCTC
AAGATCATCAGCRAATGCTCTGCAACCAACCAACTGCTTAGCAACCTGAGCCCAAGGTCAATCCATGAGAACTTTGGTATCTGTGGAAGGACT
CATGACCAACGCTCCATGCCATCACTGCAACCCGAGAGACGTGGATGAGCCCTCCGGGAAACTGTGAGCTGATGGCCGGGGCTCTCC
AGACATCATCCTGCTCTACTGGGCTGCCAAGGCTGTGAGCAAGCTCATGCTGAGCTGAACGGGAAAGCTTACTGGCATGAGCTTC
CGTGTCCCACTGCCAAGGTGTCAGTGGTGGACCTGACCTGCGCTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAATGTGTGAA
GCAGGGGTGAGAGGCCCCCTGAGGGGCACTCTGAGCTACACTGAGCACCAAGGTGGTCTCTCTGACTTCAACAGCGACACCCACTCTCT
CCACCTTTGAGCGCTGGGGCTTGGCATTTGCCCTCAACGACCACTTTGTCAAGCTCATTTCTCGGTATGACAAACGAATTTGGCTACAGCAAC
AGGATGTTGGAACCTCATGGCCCACTAGGCTCTGAAAGGATAGAGCTCTGAGCCCAAGCCCAAGAGGCAAGAGGCAAGAGAGAG
ACCTCTACTGCTGAGGAGTTCCTGACACACTGAGTCCGCAACCACTGAATCTTCCCTCTCTCAGCTTCCCTATGAGACCCCTTGAAG
AGGGAGAGGGCTTAGGGAGCGCGACCTTGTCTATGTATCCATCAATAAATACCTGTGCTCAACC

Figura 2d

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

TTCCTTTCTCTCTCCGGCCGCCAAGATGCCGARRAGSAAAGAGGGCCAGGGGAAGAGGTGGCTCCGGCCCCAGCTGTCGTGAA
GAAGCAGGAGGCTAAGAAAGTGGTGRAATCCCTGTCTTGAGAAARAGGCTTAAGAAATTTTGGSCATTTGGACAGGACATCCAGGCCAAAGAG
ACCTCAGCCGCTTTGTGSAATAGGCCCCGCTATATCAGCTTTCAGCCTCAGCAGSAGCCATCCTCTATAGCAGCTGAAAGTGGCTCCTG
ATTAAACCAAGTTCAACCCAGGCCCTGGACCCGCCAAACAGCCTACTCAGCTGCTTAAGCTGGCCACAAAGTACAGACCCAGGACAAAGCCAAAG
GACAGAGCAGAGACTGTTGGTCCCGGCGGAGAGAGAGGCTGCTGGCRAAGGGGAGCTCCCAACGAAAGAGACCCACCTGTCTCTCGAGCCAG
GAAATTACACUCCGTCACCACCTTGGTGGAGAGACAGAGAAAGCTCAGCTGGTGGTGGATTGACACAGGACGTGGATCCCATCGAGCTGGTTTC
TTCTTGGCTGCCCCGTGTGTCGTAAATGGGGGTCCCTTACTGCTTATCAGGGGAAGGGCAGAGACTGGAGACCTCTAGTCCACAGSAAAGAC
CTGCACCACTGTGGCCTTCAACACAGGTGAACTCGGAGACACAAGGCCCTTTGGCTAAGCTGGTGGAAAGCTATCAGSACCAATTACAAATG
ACAGATAAGATGAGATCCGCGCTCACCTGGGGTGGCAATGTCTCGGGTCCCTAAGTCTGTGGTCTGTATCGCCAGAGCTCGAARAGGCAAG
GCTAAGAGAACTTGGCCACTAAACTGGGTTAATGTACACTGTTGAGTTTCTGTATCATAAATAATATTCGAATATATACAAATTTTCTCTTC

Figura 2f

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

GGGCGGAGTCAGCTGAAATCCGGCGCGCGCAGTCTCCAGCCGCCAATTCTACGCGCCACCGGAAAGACGGAGGTCCTCTTTCTTACCT
AAGCCAGCCCATGGCTCGTGTGTGCAAGAGAGCATCTGAGCGCGGTGGCGCTCCAAACATTTGGATGCTGGATTAATTGACCCGGTGTG
TTTGGTCTCTGCTCCATCCACCGGTCTCCACAGATTTGAGAGAGTGTCTCTCCCTCATCATTTTCTGAGGAAACAGACTTAATATATGCG
CTGACAGGAGATGAGCTAAGAGAAATTTGCTATGCCAGCGTTCAATTAAATCATGATGCTGAGTCCGAACTGATATATACCTACCCCTGCT
GATTCATGATGATGTCATGTCATGACAGCGGGAGAGAAATTTCTGCTGATCTATGACACCCAGGGTCCCTTTGCTGTACATGCT
ATTACACCTGAGGAGGCCCAAGTACAGTTGTGCAAAAGTGAAGAAAGATCTTTGTGCGCACAAAGGAATCCCTCATCTGCTGACTCAT
GATGCGCGACCATCCGCTACCCCGATCCCTCATGAGGTGAATGATACCATTCAGATTGATTTGGAGACTGAGCAAGATTACTGAT
TTCATCAAGTTGACACTGTTAACTGTGTGTAAGGAGCTGGAGGTGCTTACCTAGGAGAAATTTGGTGTGATCACCACAGAGAGAGG
CACTCTGATCTTTTGGAGGTGTTCAAGGTAAGATGCTCAATGGCAGCAGCTTGGCAGCTCGACTTCGACTTTCACACATTTTGTATTTGGC
AAGGCAACAAACCATGGATTTCCTCTTCCCGGAGGAAGGATATCTGCTCACCATTGCTGAGGAGAGACAAAGACTGAGGAGGTC
AAACAGACAGCTGGGTGAATGAGTCTCCCTGCGCTGACATCTCAGATCTTTGTACGTAATTAAATAATTTGTGGCAGCATTAATAGC

Figura 2g

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

CCCTCTTTCCGTGCGCCGCTCGGAGGGGGTTCCAGCTGCTTCCALMSATGAAGCTGAACATCTCCITCCAGGCTACCTGGGCTGCCAGAAACCTCA
TTGARAGTGGACGATGAACGCAALACCTTCGTRCTTTCTATGAGAAAGCGTAIGSSCCACAGAAAGTTGCTGCTGACGGCTCTGSGTGAAGAAATG
GAAGGGTTATGTGGTCCGNAATCAGTGGTGGGAACGACAPAAACAGGTTTCCCCCTHAAACAGSGGTGTCTTTGACCCCATGSGTCTGTGTCCGC
CTGCTACTGATAGGGGGCATTCCTGTTTACAGACCCAGCCAGAACTGAGAAACAGAAAGACAGAAATCAATTCAGTTCGTGTTGTCATTGTGTCATG
CAAACTGTAGCGGTTCTCAACTTGGTTATTGTATAAAPSSAGAGACGATATTCTCTGACTGACTGATGCTACAGTGCCTGSGCCGCCCT
GSSCCCCAAAGAGCTAGCAGAAATCCGCAAACTTTTTCATCTCTCTAANGAAGAATGATGTCGCCCGCAGTATGTTTAAAGAAAGCCCTTA
AATAAAGAAAGTAAAGAACCTAGGACCAAAACCCCAAGATTCAGGCTCTGTTTCTCCACCTGCTCCCTGCAACACAAACCCGCCCTGTA
TTGCTCTGTGACAGCAGCGTTCGACGHAATAATAAGAGAGAGGCTCCAGAAATATGTTAAACTTTTGGCCACAGAGAAATGAGAGAGCTAA
GGAGAAAGCGCCAGGAAACAAATTCGGAAGAGAGCCAGACTTTCCTCTCTGCGAGCTTCTACTTCTAGTCTGAAATCCCAATCAGAAATAA
GATTTTTCAGTACAAATAAATAGATCAGACTCTG

Figura 2h

ACAACTCGGTGGTGGCCACTGGGCGAACCAGAACTTGGCTGGTACTGGTGGCTGGCTTGGCTTCTGGCAAGCAATTTGTGACAAAGCC
GATATATGGCTGAAATCGAGAAATTGGATTAAGTCGAAACTGAGAAAGCAAGCAAGCAAGAAATACTAGTGGCTTCCAAAGAAACGATATGA
ACAGCGCAAGCAAGCACGCCCAATCTTAATGAGGGGTGGGCCCGCCAAATATGCACTGTGTACATTTCCACACAGCAATGGCTTCTCTTATCTTT
TTAGCTGTATTAACTTTGTAGCAATGCCAAAGAGGTGGATCAAGTTTAAATGACTGTGTCTGGCCCTTTTCACAAATTAAGAAAGCTACTGACAAACGAA
GGCCCGCGTGTGCCCTTTCCCATCTGTCTATCTATCTGGTGGCAGGGAAGGAAAGAACTTGTGCATGTTGGTGAAGGCAAGAAAGTGGGAGGAAAGA
AGTGGGGTGGGACGACACGTGAAATCTAGAGTAAAGCCAGAGCTGGCCCGCAGGGTGTCTGCAGGCTGTAAATCCAGTTTAAATCAGAGTGGCATTT
TTTTTTTGTGTTCAAATGATTTTTHAATATTGGGAATGCCACAATTTTTTTTAAATAGCAATAAATAAGTTTAAAAAACTT

Figura 21

>g|4507668|ref|NM_003285.1| Homo sapiens tumor protein, translationally-controlled 1 (TPT1), mRNA

CCCCCGAGAGGCGCGCTCCGGCTGCAACGGGCGCTCGCTCTGAGATTTCAGAGCTCGTGCTAGCGCGCTGCTGCTCTCCCTTCAGT
CGCATATCATGATTATCTATACCGGGAGCTCATCAGGDAAGCATGAGATGTTCTCCGACATCTACAGATCCGGGAGATCGGGACAGGATT
GTGCTTGGAGGTGAGGGGAGAGATGGTCACTAGGACAGAGGTAACATTGATGATGCTGCTCATIGGTGFAAATGCTCTCGGCTGAGG
CCCCGAGGGGAGGTACCGAAGGCACATTAATCACTGCTGTCGATATTGTCATGAGCCATCACCTGCAGGAAACAGATTTCACAAA
AGAGGCTACAGAAAGTACATCAAGATTACCTGNAATCAATCAAGGGAACTTGAAGAACAGAGACCCAGAGAGTAAAGACCTTT
TATGACAGGGGCTGGAGAACAAATCAGSCHCATGCTTGCCTAATTTCAAAAACIACAGCTTCTTTATTGTTGAAAAACATGAAATGCAAA
TGGCATGGTTGCTCTGTTATGGACIACGCTGAGGAGGCTGTGACCCCATATATGATTTCTTTAAGCATGCTTTAGAAANTGGAAAAATG
TTAACAAATGTGGCAATTATTTTGGATCTATCACCCTGTATCATAACTGGCTTCTGCTTGTCAICACACAAACACCCAGGACTTAAGA
CAATGGGACTGATGTCTATCTTGAAGCTCTTCAATTTATTTTGAAGTTTATTTGAGTGGAGGCTTGTGTTTAAAGAAACACATG
TCAATGTAAGTTGCTAAATAAATGCAATTAACTCATTTGAGAG

Figura 2j

[illegible]

Figura 21

[illegible]

Figura 2m

TTCTGGCATTGGAGAGGCTGCGCTTGGTTGCCACTTCCGCAATCTTCCCGGCTTATTAATCAACACTATCGAAGAGTATAGCGCGCTTAG
TACCAGTTCGGAGAGGTGCGGSCGCCGAGAGTACCGGCTTTGGAATTAACATCGAAGATGCTGTCGCTGGCGGTTTGGTGGGCGCTTGGTCCGCG
CTTCTCCGGCGGGCGGACAGTGGTATTCGGAATAAGCTTTGGGATTCATCTTTGATATGGCTTCAGAGACCTTCCTATGCTCTATCACTCATCT
TTCAAAAGACTGGGAGCTGCTGAGATATCTCTTATCTTISABAGCGTATCTCTGGAGCTGATACCTCTGTTATCTTGAAGAAACTGGGCTT
GCTCTTAAGTATTTGTTGATGTGTTATTCGCGCTACATATGAGTGTGAGGAAATGTTCAAGCGAAGAAATGTGTATGATTTTCTTCAGGCTTAAAG
ATCTCTTGAACTTGGACCTGACATATGTGTGCTTTGCTGTGTTTSSAATGATATAACINATTAAGGAGGAGNATATAGTGAAGAGGAC
AAGAGGCGCATTTGGGACGTTCCAGTTGTTGAGGAGCTGTGTGGCTGTGTATGTTGATGGCCCTTGGTAAATGCTATTTGATGGAAGGTTCCAAAT
GGTTCCAGAGGCGGTAGCGGAGTTGGTGTGAAGGCGGCTGTATCTCTGATCTCTCGAATTCAGTGGCGGACCATGTCAGACGCGCATTAAGAG
CTGTGTGATAGCTTGGTGGCAATTGGTCTGTGGTCAAGCTGACCTGATTTATTTGTTBACCGACAGACTGGGGAAGACCTCAATTGCTATTTGACAC
AATCATATAACCAAGAACGTTTTCATATGATGGATCTTATGAAAGAAAGAGAGCTGTATCTGATTTATTTAGTTTGGCTATTTGGTCAAAAGAGATCTCACT
TTTGGCCAGTTGGTGTGAAGAGCTTACAGATGACATATGATGAAGTACACCATTTGGTGGTGTCTCGGCTAGCGGCTTGGATGCTCTCCCGACCTTC
AAGATACCTTGGCTCTTACTCTGGCTGTTCCATGSGAGAAATATTTTGAAGACAAATGSCARACATGCTTTTGATCACTCATAGACAAATTTATCCAA
ACAGGCTGTGCTTACCGTACATATGTTCTGTGCTTCCGCGACGCGCTTGGTGTGAGAGGCTATGCTGTGCTGATGCTGTGTATGCTATGACCTTC
CGGTTGCTGGAAGAGAGCGCAAAATGAAACGATGCTTTTGTGTGGTCTGCTTGACTGCTTTGCCAGTCAATAGAAACACAGGCTTGGTGTATG
TGTGTGCTTACATTCACAAAGTCAATTCATCACTGACGGACAGATTTCTGTGGAACAGAGATTTCTTACAAAGGATTCGCGGCTTC
AAATTAAGGTTGGTCTGCTGTATCTCGTGTGCGATCTCGCTGGCCAAACCAAGAGGCTATTAAGTAGAGTACAGATTAACATAGAACCTTGAATTTG
CAGTATCTGTGAGGTTGGTGTCTTTTGGCGAGTTGCGTTGTAGACCTGTATGCTGCGACATCAACACATTTTGTATGTGTGGCGTGGCGTCTAAT
CTGTAGTTGCTGAAGCAAGGACAGATATTTCTCCCATGGCTATTTGAAGAACAGATGTGGTGTATCTATATCGGGGTGTAGAGGGGATATCTTTGATAT
ACATGAGCGCCAGCAGATTAACAAAGTTTGAGAAATGCTTTCTGTCTCATGTTCTGTACGCCAGCACCAAGCGCTTGTGTGGAACTATTCAGGCT
GATGTGAAGCATCTCTGAAGACATCTGAATGCAAAAGCTGGAAGAGATTTGTAAACAAATTTCTGTGGCTGGATTTTGAAGCTTAAAGCTCTGTGGGAT
CAACATCTGAATACGAGCTTCAGTTTGTTCATTTGTTCTAGTAAATTTAGTTCCATTTGTGAAGAGGCTTACTCTCTATCTCTCTATATGACAAAGAT
TACATGTGAAGAAATAGAGGTTCCATATATGATATTT

Figura 2n

>gl|5031856|ref|NM_005556.1| Homo sapiens lactate dehydrogenase A (LDHA), mRNA

TGGTGCAGCGCGCTGGCGCGGATTCATTCCACGCGCGCCCGACGCGCGACGTTGCATTCCTGATTCCTTTGGTTCCAG
TCCATATATGCCAACTCTAAGGATCAGCTGATTTATATATCTTCTAAGSAGAGACAGACCCCCAGAAATATACAGTTNTTGGGCTTG
GTCGTGTGGCAATGGCCTGTGGCAATGATATCTTAAATGAGAGACCTGGGAGTGTGACCTGCTCTTGCTGATGATGATATGAA
GGGAGAGATGATGGAATCTGCAACATGGCAGGCTTTTCTCTTAGAATACCAACATTTGTTCTGGCAAGATATATATGTATCTGCAACCTCC
AAGCTGGGTCATTATCAGGCTGGGACAGTCCAGAGAGGAGAGAGCCCTCTTATATTTGGTCCAGCCTAAGCTGAGACATATTTAATTCG
TCATTCCTTATATGTTGTAATACAGCCGAGCTGCGAGTTGCTTATTTGTTTCANAATCCAGTGGATAICTTGACCTACGGTCTTGGAGAT
AAGTGGTTTTCCCAAAACCGTGTATTATGGAGGTGGTTGCCAATCTGSAATTCAGCCCTGATTCGGTTACCTGATGGGGGAAGGCTGGGAGTT
CACCCATTAAAGCTGTATCGGTGGGCTCCTTGGGACATGGCAGATTCAGATTTCCATTTGGCTGTATGGAGTGGATATGATGTTGGTGTCTC
TGAAGACTCTGCACCCAGATTTAAGGACTGTATAAAGATAAAGGATAGTGGAAAGAGGTTCCACAGCGGTGTTTCAGAGTCTCTTATGAGGT
GATCAAACTCAAAAGGCTACAGATCCTGGGCTATTGGACTCTCTCTGTAGCAGATTTTGGCAGAGATATATATGAGAAATCTTASGGGGTGGAC
CCAGTTCCACCATGATTAGGGCTCTTACGGGATTAAGGATCGATGTCTTCTCTAGTGTCTCTTGCATTTTGGGACAGATGGATCTCAG
AOCCTGTCAAGGTGACTCTGACTTCTGAGGAGAGGAGGCTGTTGAGAGAGAGTCCATATACACTTTTGGGGGATCCAAAGAGAGCTGCATTT
TTAAGTCTTCTGATGTTCATATCATTTCCACTCTCTATAGGCTTACACAGGATTTCTAGGTGGAGGTTGTGCTCATGTTSTCTTTTATCTGATCT
GAGATTARAGCAGTATATATTTAAGATGAGACTGGGAAARACATCAACTCTTGAAGTTAGAAATAGAAATGGTTTGTAAATCCACAGCTAT
ATCCTGATGCTGATGATTTATATCTTTGTGTAGTCTTCAACTGGTTAGTGTGAATATGTTCTGCCCACCTCTGAGGCCACCTGCTCATGCT
GTACGTACTGCATTTTGGCCCTTGAAGCCAGGTGGATGTTTACCGTGTGTTATATATTAATCTTCCCTCTTCACTGACATGCTTAGTCCAC
ATTTTTCCTCCAGTGAATCACAATCTGGGATCCAGTGTATATATCAATATCATGCTGTGTGATATATCTTCCAGAGGATCTTATTTTGTG
AAGTATATCATGATGATACATTACCATATATATATGTAT
CCCCCAATAAACCTTGAACAGTC

Figure 2p

>gi|52851448|ref|NM_000365.4| Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA

CCTTCAGCGGCTCGGCTCCAGGACACCTTGGGCGGGGAACTGGAAAGATGAACGGGCGGAAGCAGAGTCIG
GGGAGCTCATCGGCACTCTGAAAGCGCGGCAAGAGTGGCGGCGGACACCGAGGTGGTTTGTGCTGCCCCCTACTGCTATAATCGACTTGGCC
GSCAGAAAGCTAGATCCCAAGATTGCTGTGGCTGGCGCAGAACTGCTACAAATGACATAATGGGCTTTTACTGGGGAGATCAGACCCCTGGCAT
GATCAAGAACTGCGGAGCCACGCTGGTGGTCTCTGAGGCACTCAGAGAGAGAGGCATGTCCTTGGGGAATCAGATGAGCTGATGGGCAAGAA
GTGGCCCATGCTCTGGCAACAGCCACTGGAATATGCGCTGCAATTGGGAGAGAGCTAGATGAAGGGAAAGCTGGCATCACTCAGAAAGGTIG
TTTTGAGACAGACAAAGGTCATCCAGATTAACCTGAAGAACTGGAGCAAGGTCTGCTGSCCTATAGAGCCCTGTGTGGCCATTGGTACTGG
CAAGACTGCCAAGACCCCAACAGGCCCCAGGAATACACAGAGAACTCCGAGAAATGGCTGGAATGCCCAACCTCTCTGATGCGGTGGCTCAAGAC
ACCGGATACATTAAGGAGGCTCTGTGACTGGGAGACCTGCAAGGAGCGCGGCGGCAAGCTGATGATGATGATGATGATGATGATGATGATGATGATG
CTTCCCTCAAGCCCGAATTGCTGGACATCATTAATGCCAATCATGAGCCCCCATCATTTTCCCTAGCCCTTCCGCGCAAGCCAGGGAGCTAA
GCAAGCCAGAGAGCCCACTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCTGTAAGCT
TCTTTTACTGTTTATATCTTCAACCTGTGAAATGGTTGGGACCAAGGCCAAATCCCTTCTCCCACTTACIATAATGGTTGAACTAAAGCTCACCA
AGGTGGCTTCTGCTTGGCTGAGAGATGAAGGCGGTGGTGGATTTGCTCTGCGGTTCCTTGGGCTAGTGGAGGCGCAGAGAGAAACCATC
CTCTCCCTTCTTACACCGGTGAGGCCAAGATCCCTCAGAGAGGCAAGATGCTGACCTCTCCCATGGTCCCGGTGGCTCTGTGCTGTGTATG
TGAAACACCCATGTGAGGCGATTAACCTGGCACTAGG

Figura 2q

TPP1 = Proteína de tumor, controlada a nível da tradução, 1830 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.

TPP1 = Isomerase de triosefosfato 1, 1220 pb, cr 12p13.

Figura 2r

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

[illegible][illegible][illegible]

Fligra 3a

```
Query: 362 aagttacatcaaaagattacatgaataatcaatcaaaagggaaacttgaaggaacagagaccagaa 421
|||||
Sbjct: 372 aagttacatcaaaagattacatgaataatcaatcaaaagggaaacttgaaggaacagagaccagaa 431

Query: 422 aagttaaaaactttttatgacagggggtgcagaaacaaatcaagcatctcttgctaatttc 481
|||||
Sbjct: 432 aagttaaaaactttttatgacagggggtgcagaaacaaatcaagcatctcttgctaatttc 491

Query: 482 aaaaaactaccagttctttattggtgaaacacatgaatccagatggcatggttgcctctattg 541
|||||
Sbjct: 492 aaaaaactaccagttctttattggtgaaacacatgaatccagatggcatggttgcctctattg 551

Query: 542 gactaccgtgaggaatggtgtgaccccatatatgattttcttttaaggaatggttccagaaatg 601
|||||
Sbjct: 552 gactaccgtgaggaatggtgtgaccccatatatgattttcttttaaggaatggttccagaaatg 611

Query: 602 gaaaaatgcttaacaaatgtaggaattatgttggaaactatcacctgtccatcatcaactgacct 661
|||||
Sbjct: 612 gaaaaatgcttaacaaatgtaggaattatgttggaaactatcacctgtccatcatcaactgacct 671

Query: 662 tctgcttgtcatccacacacaccaggaacttaadacaaatgggaactgaatgcatctctgag 721
|||||
Sbjct: 672 tctgcttgtcatccacacacaccaggaacttaadacaaatgggaactgaatgcatctctctgag 731

Query: 722 ctcttcattttattttgactgtgatttattttgagtgaggagcatgttttttaagaaaaaaca 781
|||||
Sbjct: 732 ctcttcattttattttgactgtgatttattttgagtgaggagcatgttttttaagaaaaaaca 791

Query: 782 tgtcatgtgaggtgtctataaaaataaaaatgtcattttaaaactcattttgagag 830
|||||
Sbjct: 792 tgtcatgtgaggtgtctataaaaataaaaatgtcattttaaaactcattttgagag 840
```

Figura 3b

score = 151 bits (82%), expect = 0.0
Identifiers = 928/928 (100%)
plus / plus = plus / plus

[illegible][illegible]

Query: 302 qtcgafatcgtctacgaaccacccacgcggcgaggtttcacaggaggttatcacagg
 | ||||| ||||||| ||||| ||||| ||||| ||||| ||||| |||||
subjct: 318 gtcgaaacttgcgtcaaacaccccacgatcttgtaggaagaaggttatcacagg

Query:	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500
Query:	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500

Figura 3c

```
Query: 482 aaaaactaccagttctttatttggtgaaacacatgaatcccagatggcatggcttgctctattg 541
      |||||
Sbjct: 496 aaaaactaccagttctttatttggtgaaacacatgaatcccagatggcatggcttgctctattg 557

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

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

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

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

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

Figure 3d


```

Query: 422  agagtcataaacctttttatgacaggggctgcagaaacaaatcaagcacatccttgcctaatctc 481
          |||||
Sbjct: 438  agagtcataaacctttttatgacaggggctgcagaaacaaatcaagcacatccttgcctaatctc 497

Query: 482  aaaaacraccaagtctcttattcggctgaaaaacatgaatccagatggcatcgggtgctctattg 541
          |||||
Sbjct: 498  aaaaactaccagtctcttattcggctgaaaaacatgaatccagatggcatcgggtgctctattg 557

Query: 542  gactaccgtgagggatggtcttattcggcccatatatgatcttcttaaggatcgttcttagaatac 601
          |||||
Sbjct: 558  gactaccgtgagggatggtcttattcggcccatatatgatcttctttaaaggatcgttcttagaatac 617

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

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

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

Query: 782  tgcctatgtagggtgtcttaaaaataaataatgcattttaaactcattttgagag 830
          |||||
Sbjct: 798  tgcctatgtagggtgtcttaaaaataaataatgcattttaaactcattttgagag 846

```

Figure 3f

Score = 564 bits (303), Expect = e-158
 Identities = 314/318 (98%), Gaps = 1/318 (0%)
 Strand = Plus / Plus

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

Figura 39

>gi110224966|gb|BE771308.1| QV1-FT0082-210600-249-812 FT0082 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 aaaaagagcctacaaagaa-gtaccatca-aagattacatcgaataccaatcaaaaggggaactt 403
|||||
Sbjct: 6 aaaaagagcctacaaagaaagacgtacatcagttgattacatg-aattcaattcaaaaggggaactt 64

Query: 404 gaagaacacagagaccagaaagaggtaaaaccctttctatgacacaggggctgcagaaacaaatccaaag 463
|||||
Sbjct: 65 gaagaacacagagaccagaaagaggtcaaaaccctttctatgacacaggggctgcagaaacaaatccaaag 124

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

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

Query: 584 aaggacggttttagaaaatggaaaaaatgttaacaaatgtggcaattattctggatctatccac 643
|||||
Sbjct: 245 aaggacggttttagaaaatggaaaaatgttaacaaatgtggcaattattcttggatctatccac 304

Query: 644 ctgtcaccataaactggcttct 664
|||||
Sbjct: 305 ctgtcaccataaactggcttct 325

Figura 3h

```
>g110246833|gb|6E814599.1| NR0-6N0070-070800-033-b04_1 6N0070 Homo sapiens cDNA, mRNA
sequence.
Length = 140

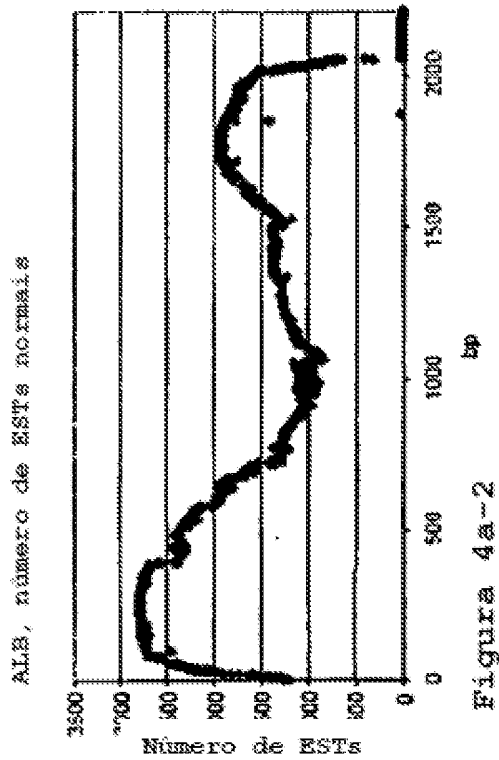
Score = 259 bits (140), Expect = 2e-066
Identities = 140/140 (100%)
Strand = Plus / Plus

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

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

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

Figura 3i



ALB, desvio relativo a RefSeq, cancerous ESTs

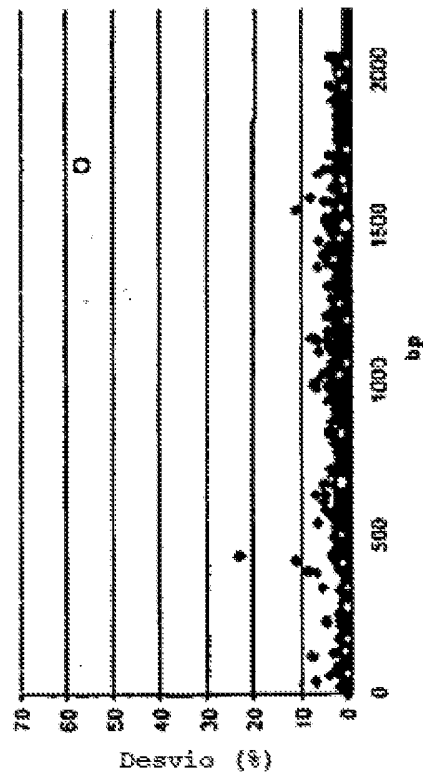
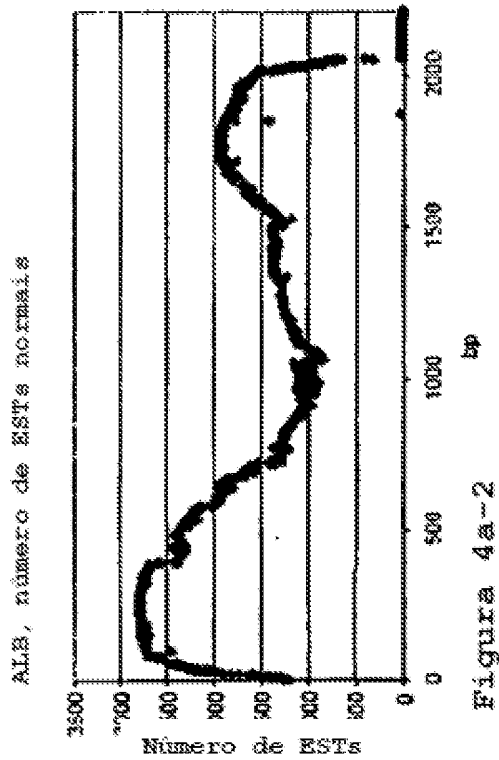


Figura 4a-3



ALB, desvio relativo a RefSeq, ESTs normais

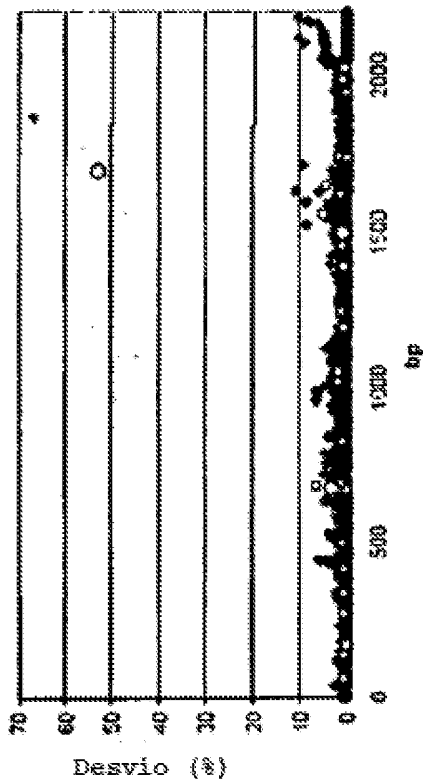


Figura 4a-4

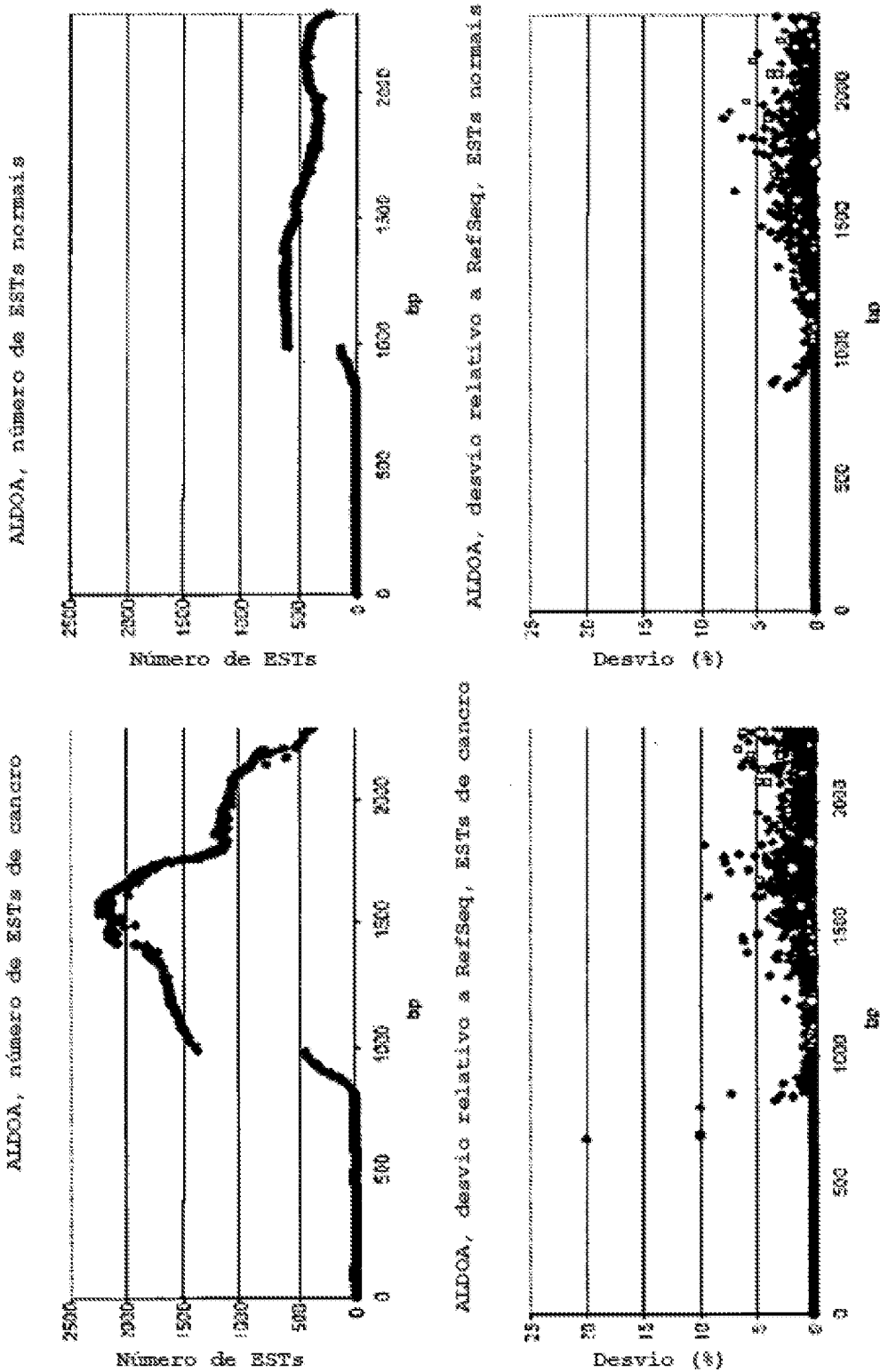


Figura 4a (continuação)

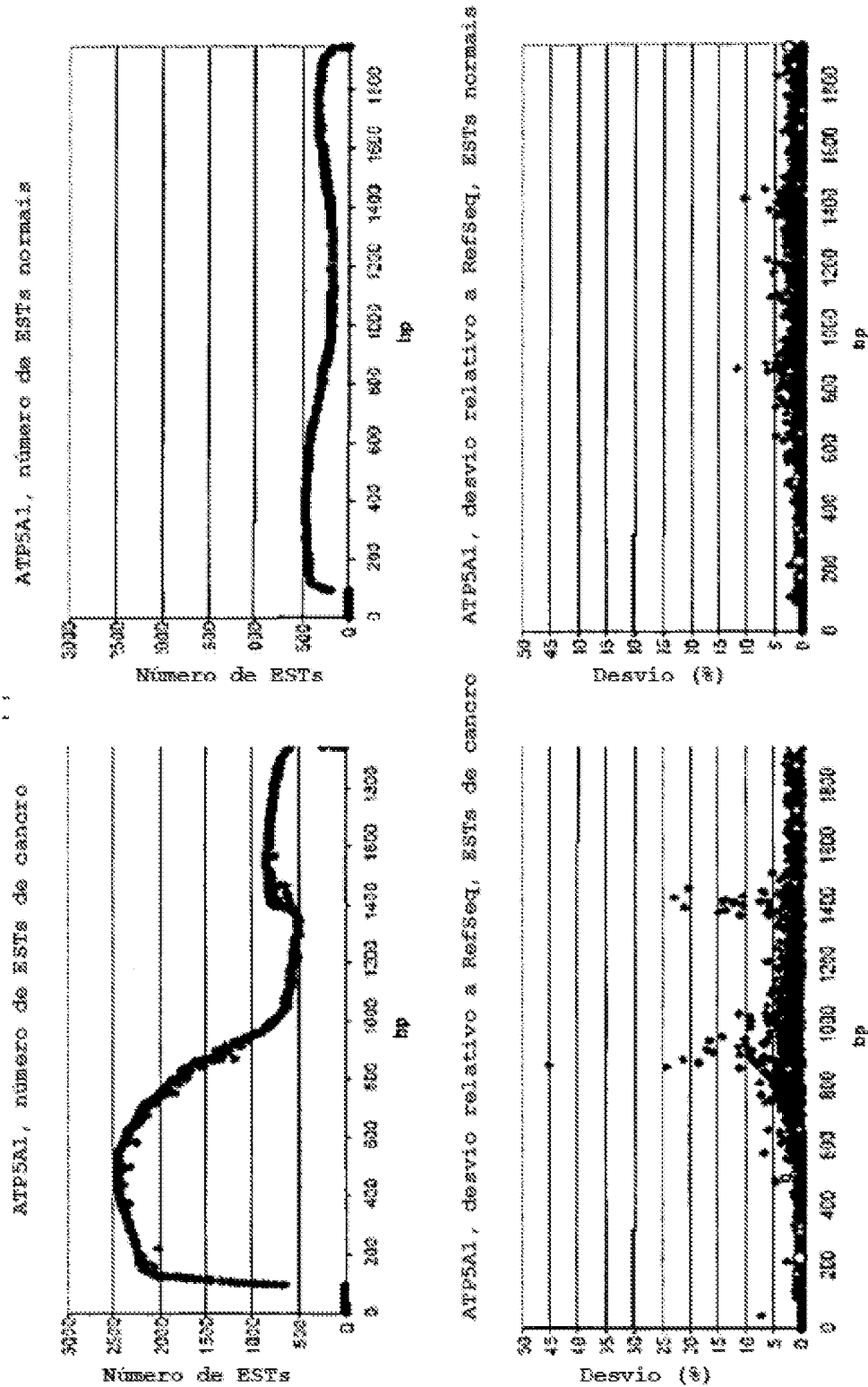


Figura 4a (continuação)

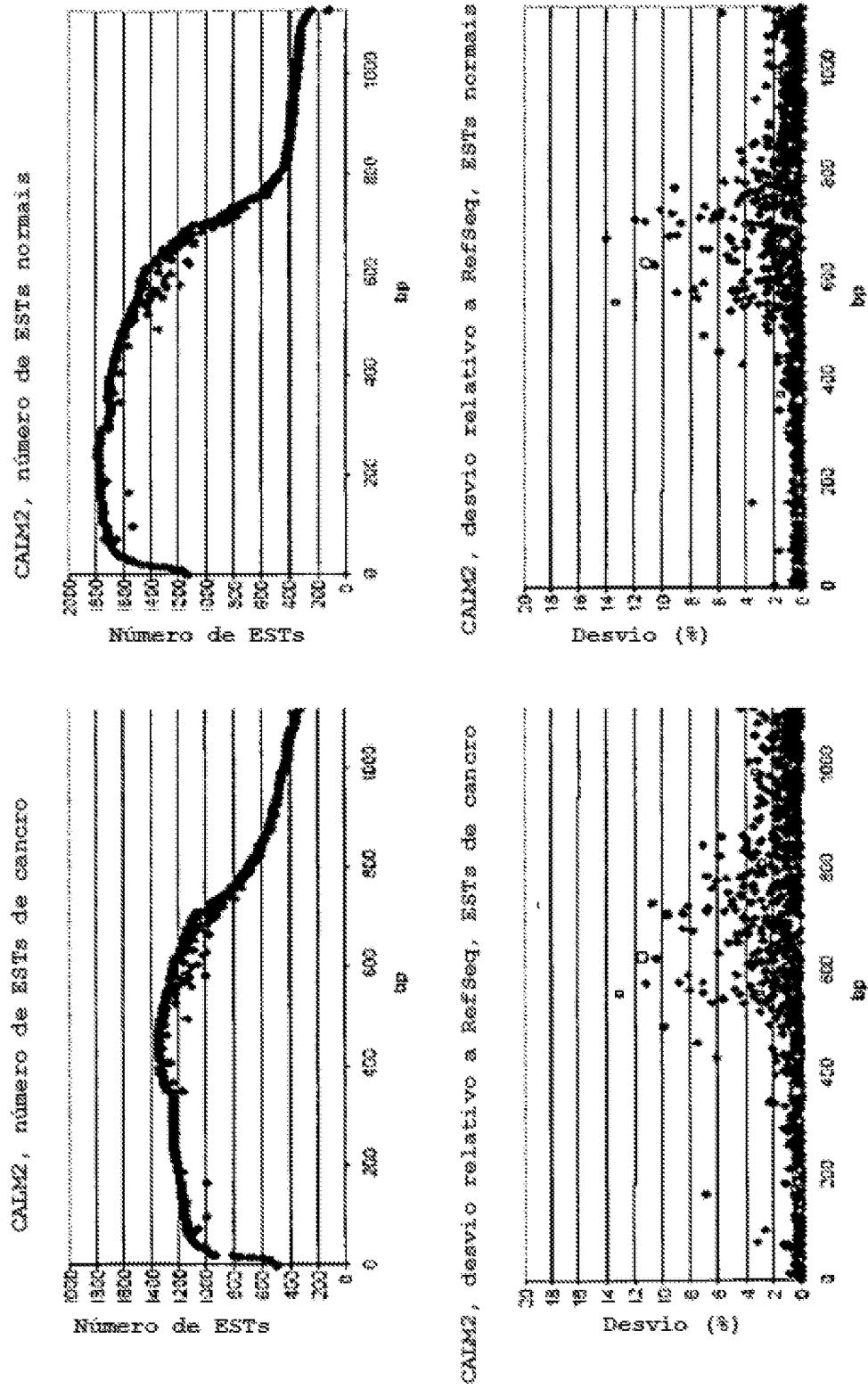


Figura 4a (continuação)

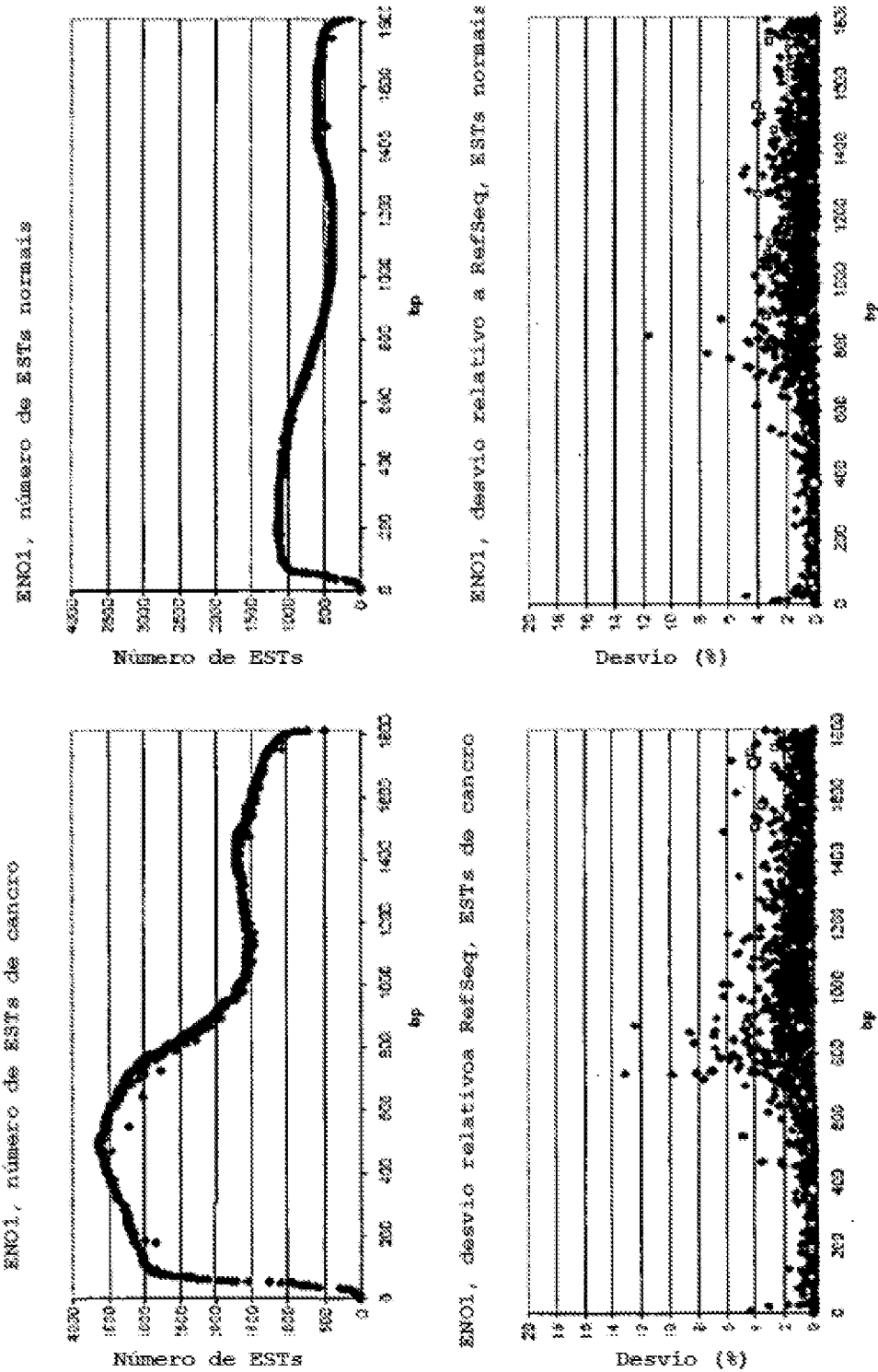


Figura 4b

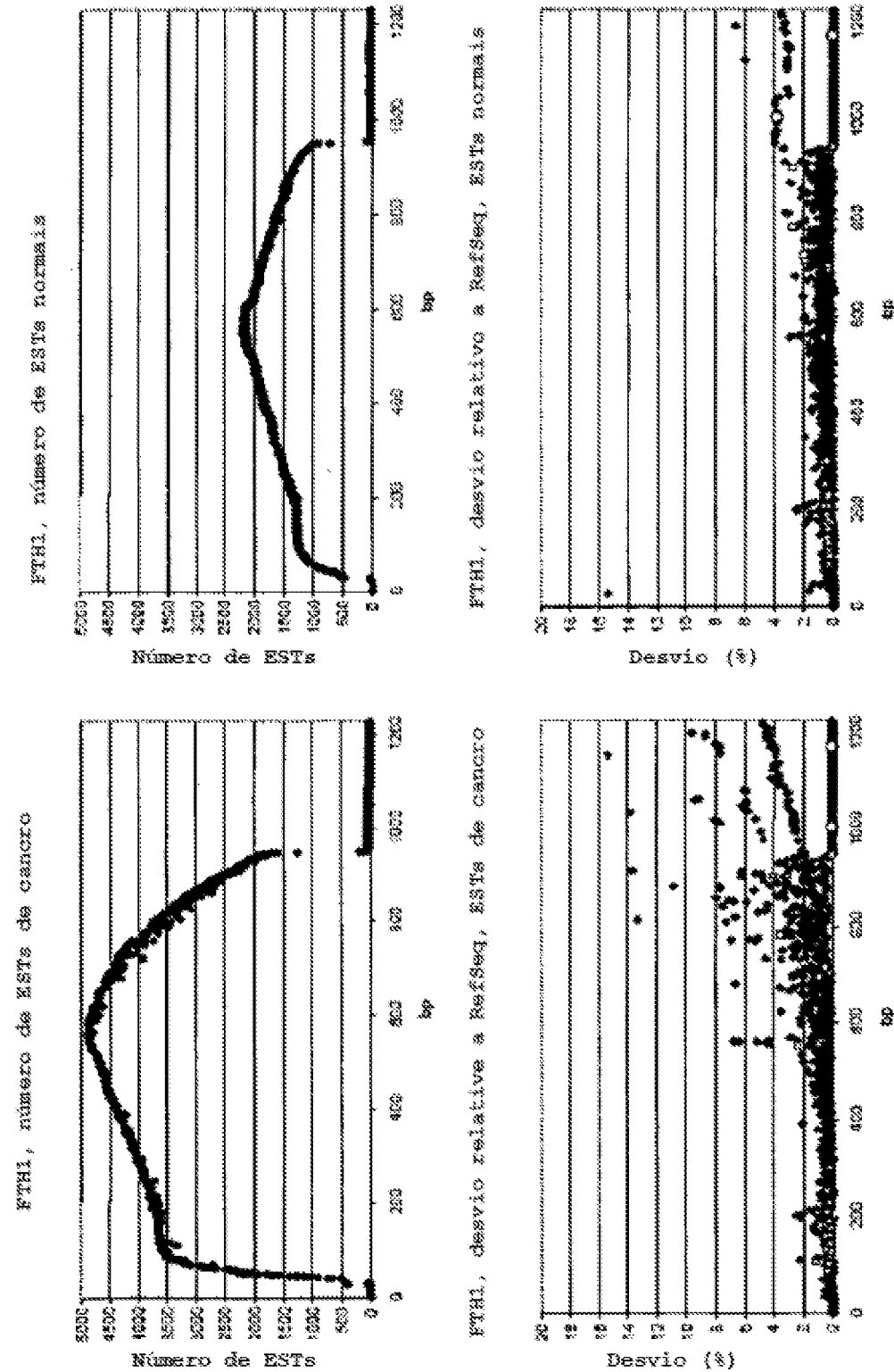


Figura 4b (continuação)

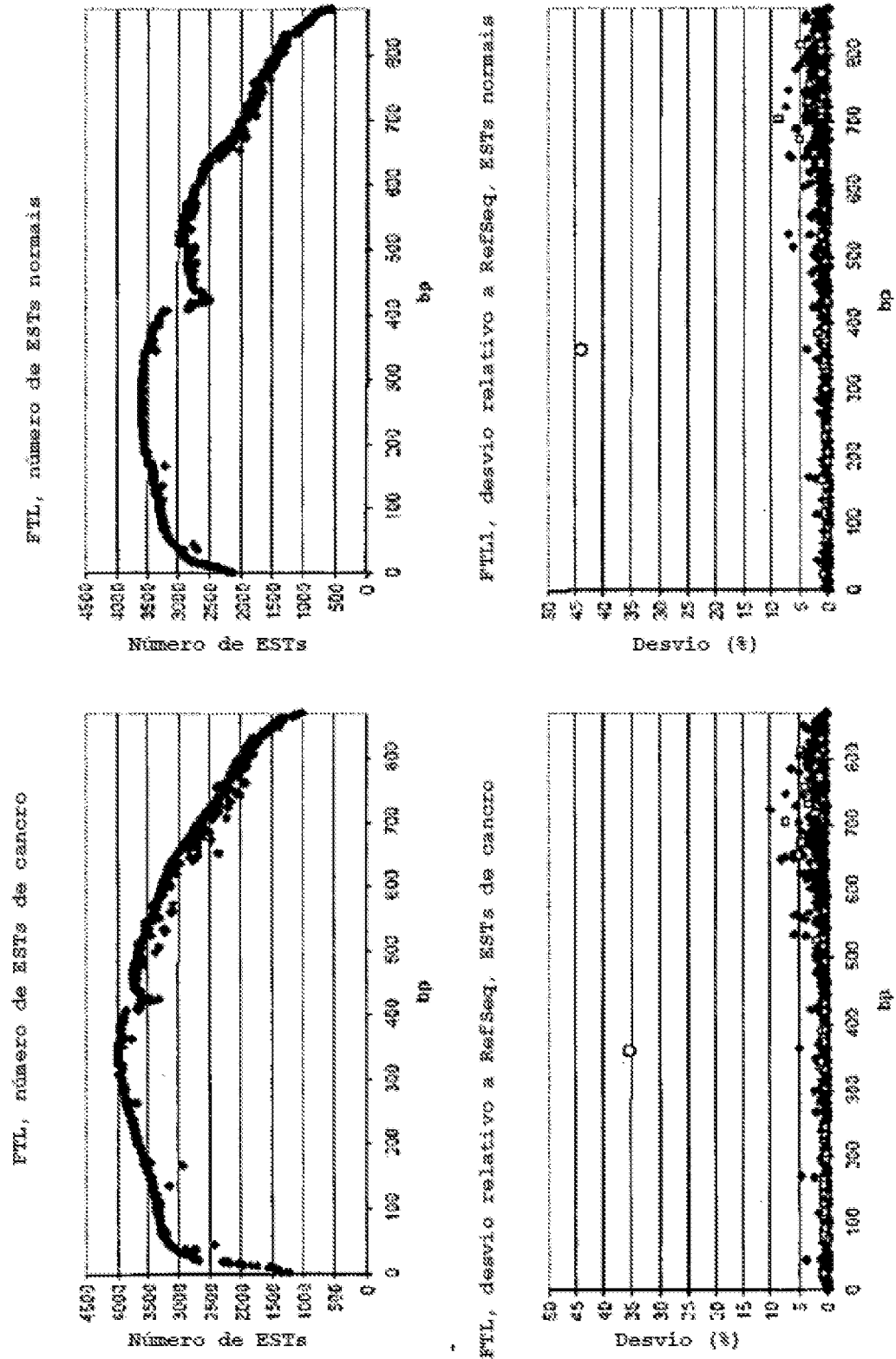


Figura 4b(continuação)

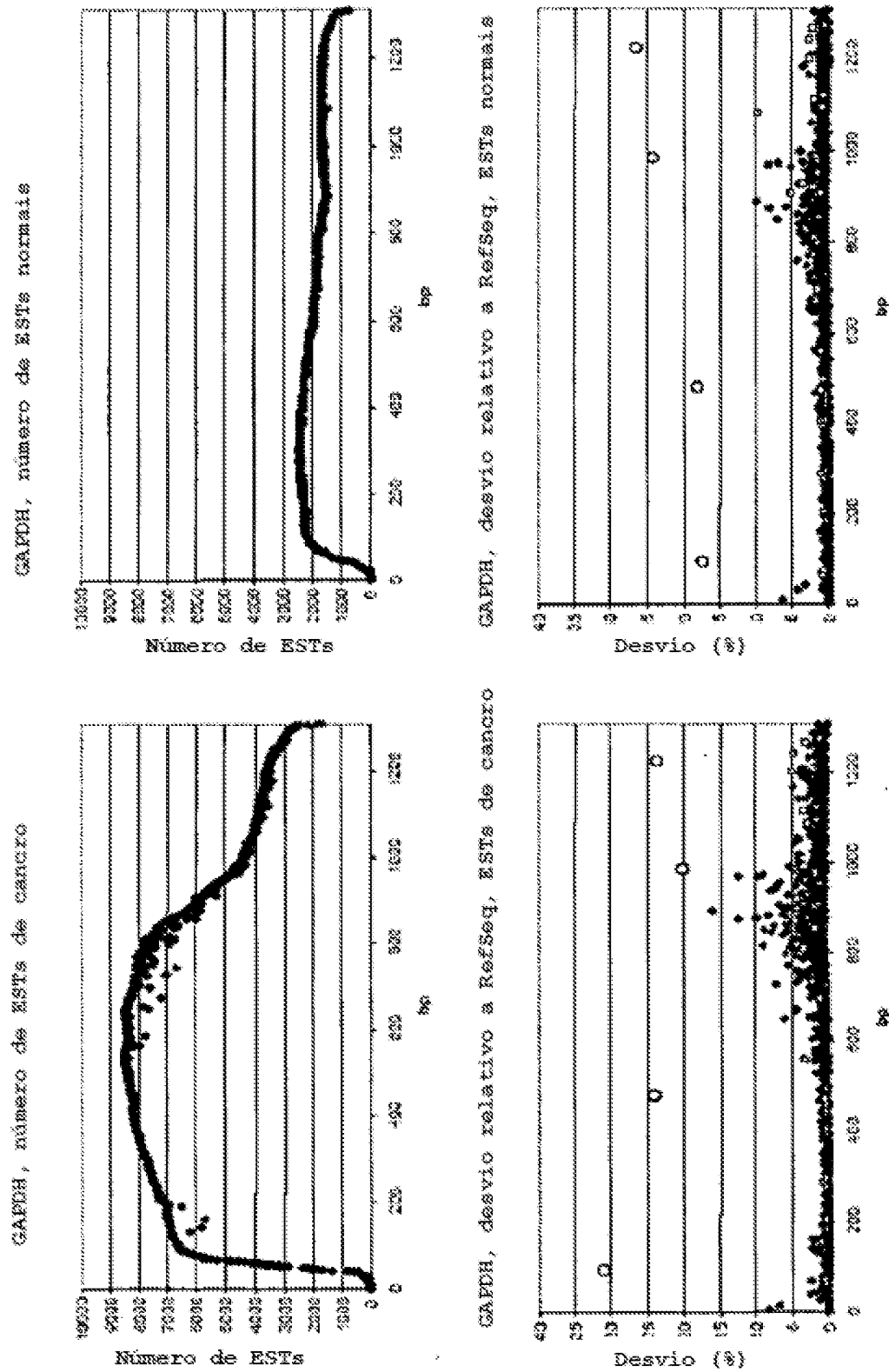


Figura 4b (continuação)

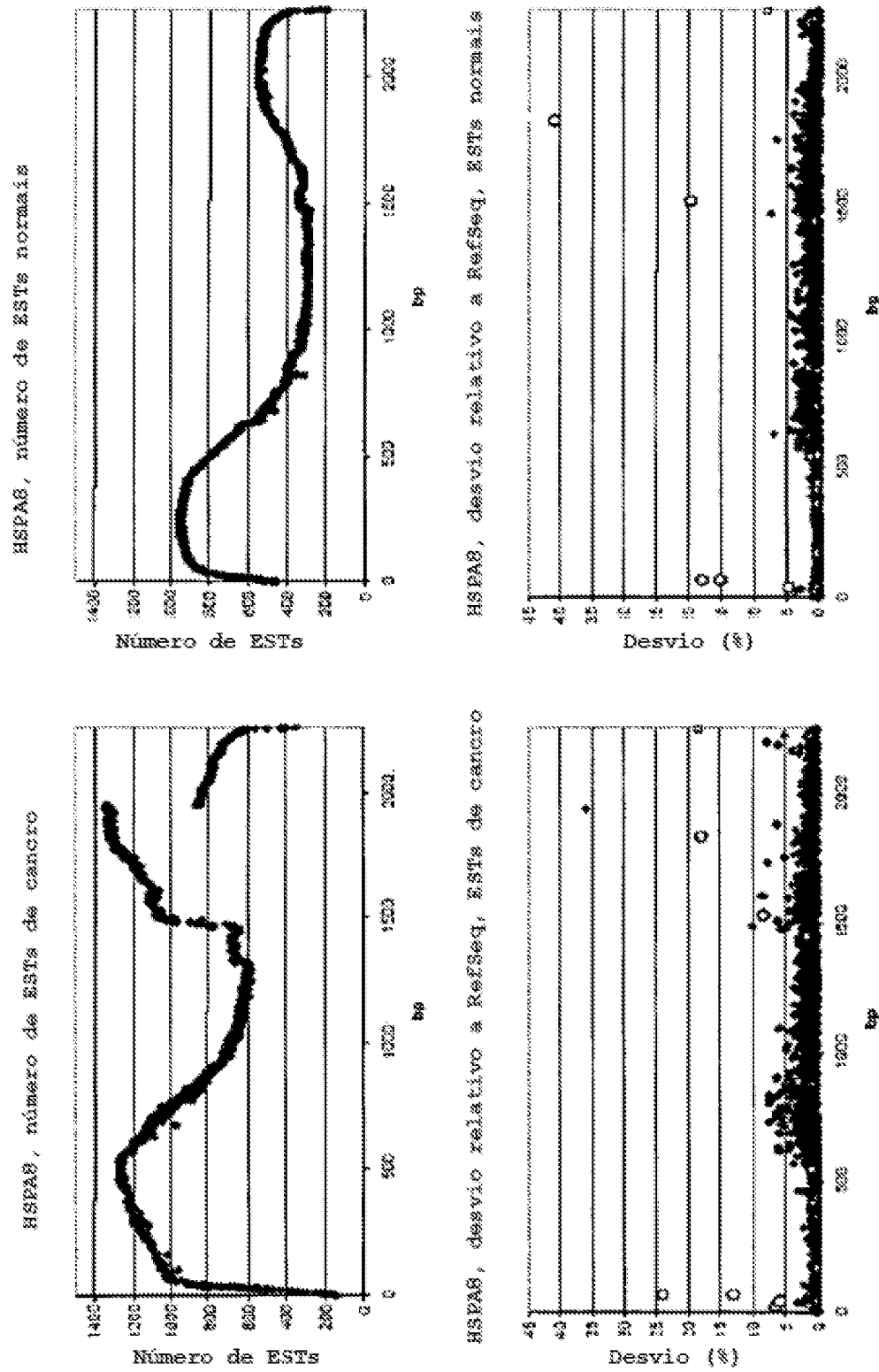


Figura 4c

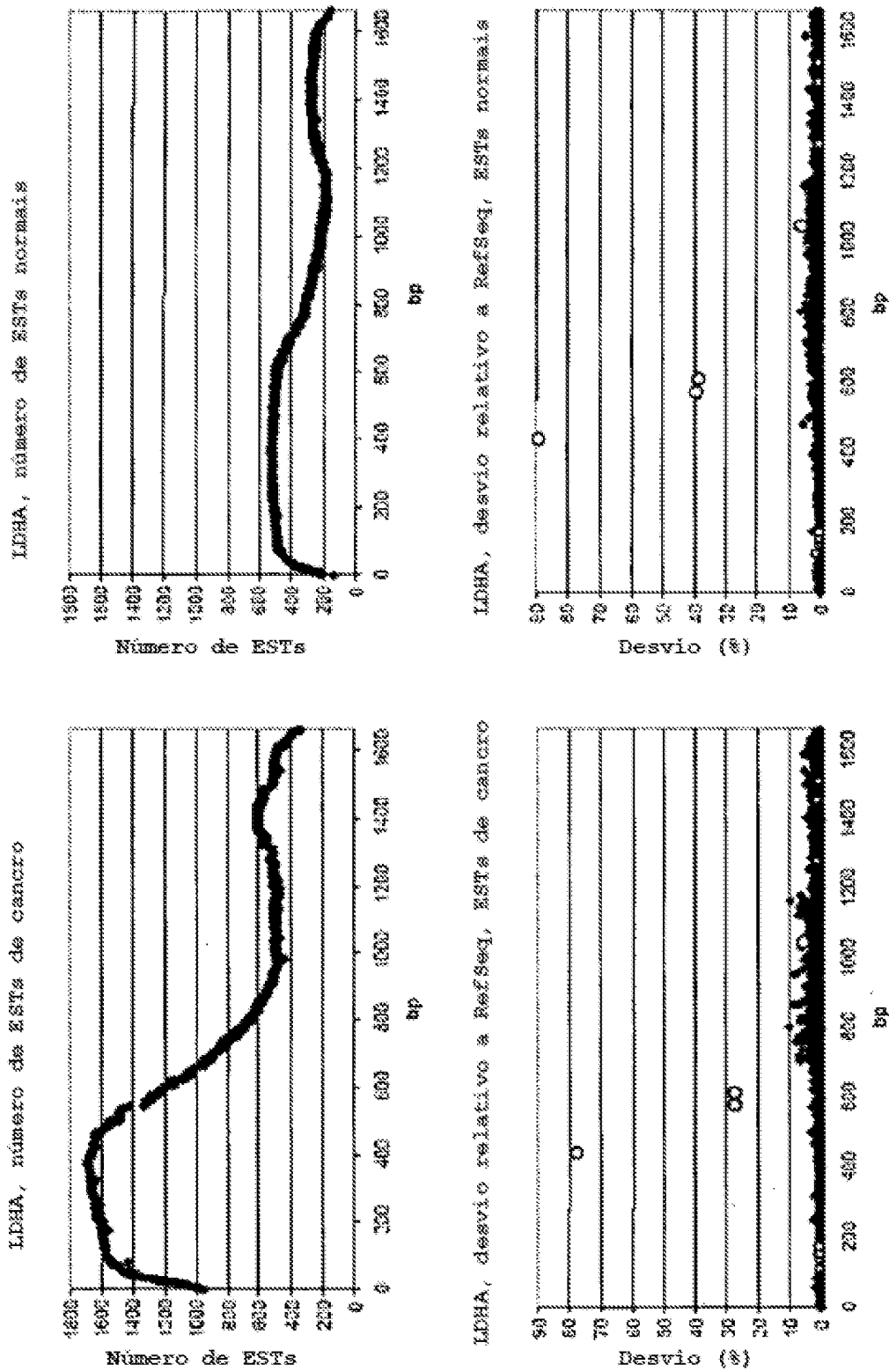


Figura 4c (continuação)

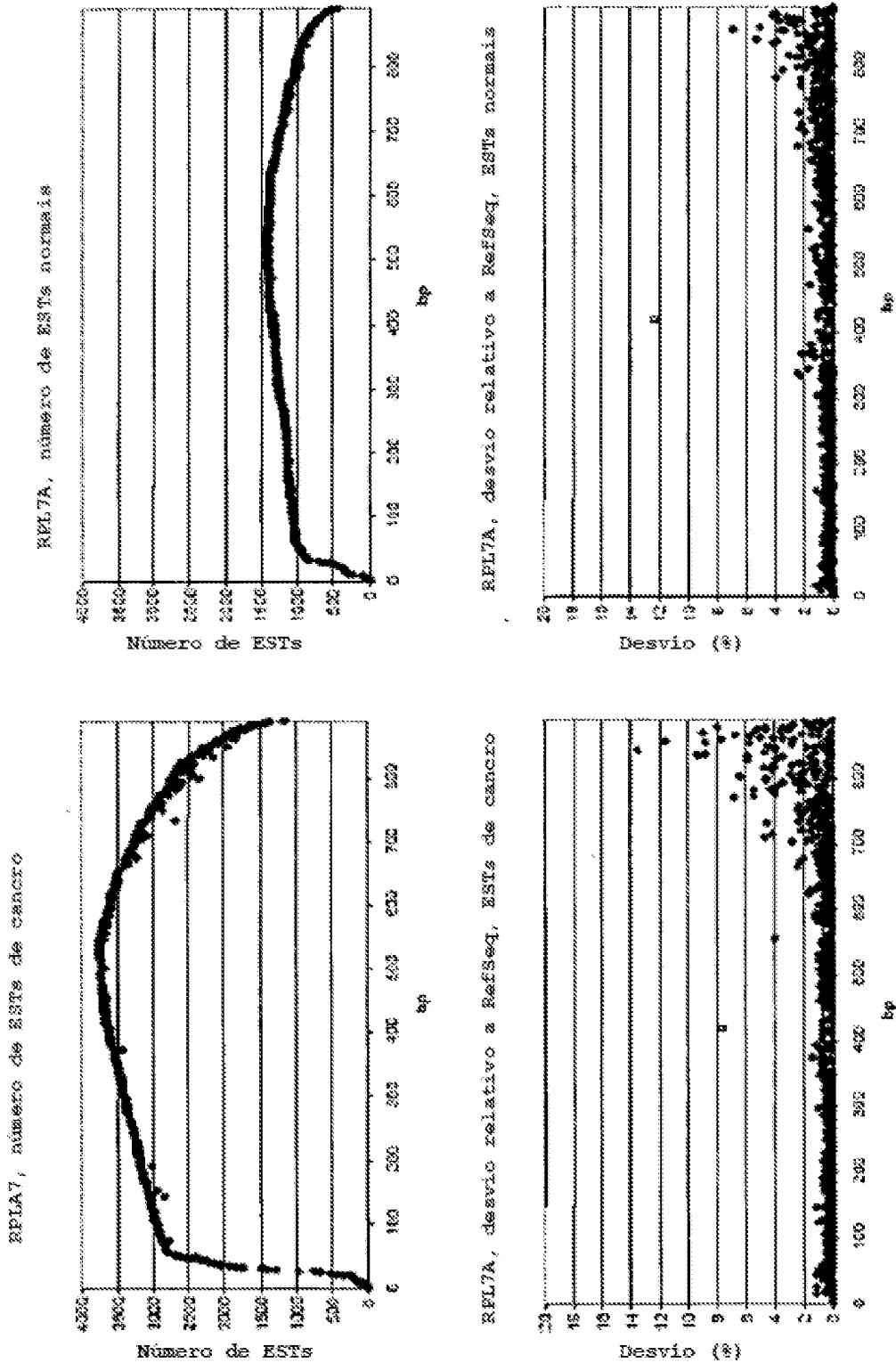


Figura 4c (continuação)

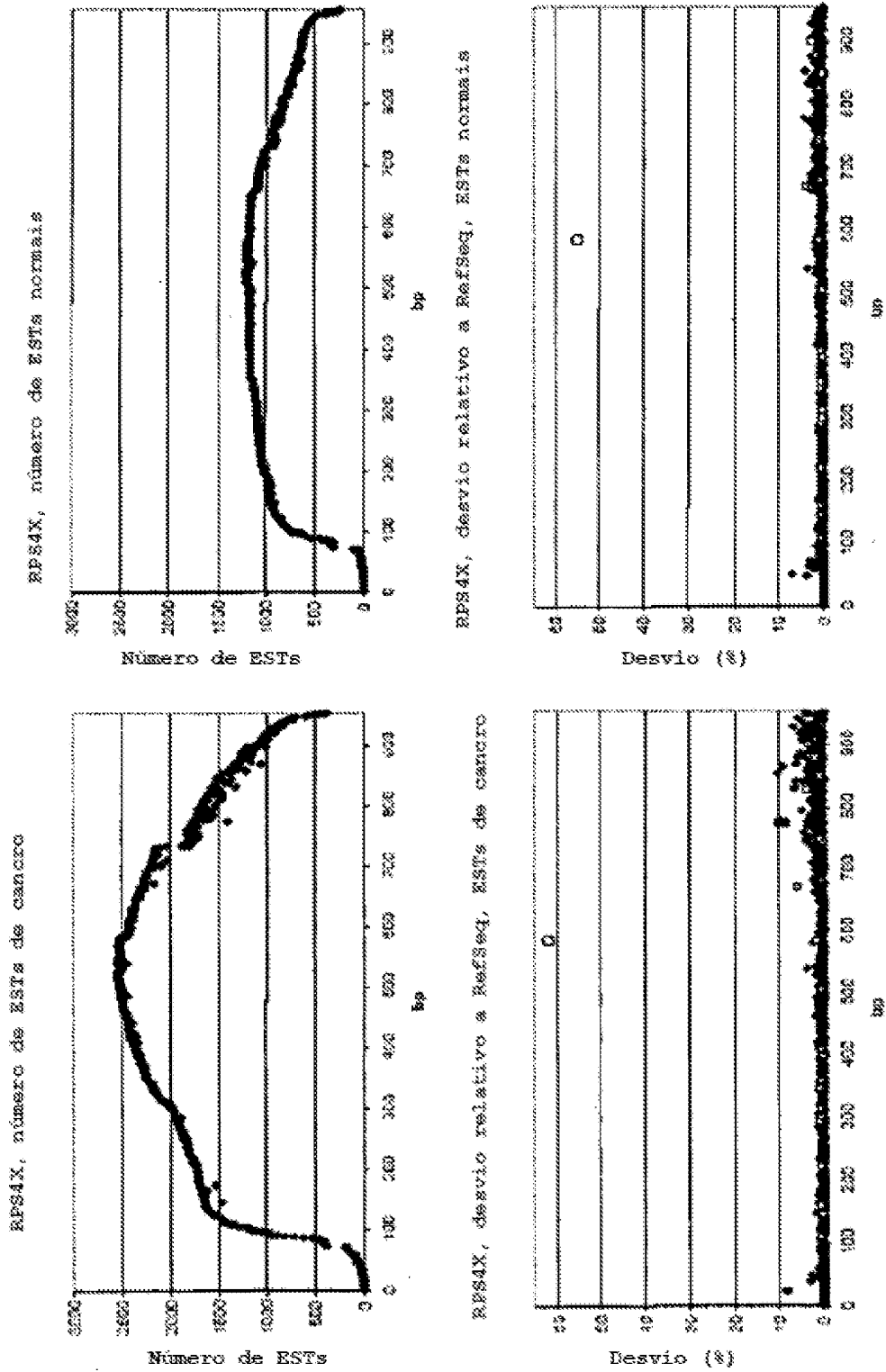


Figura 4c (continuação)

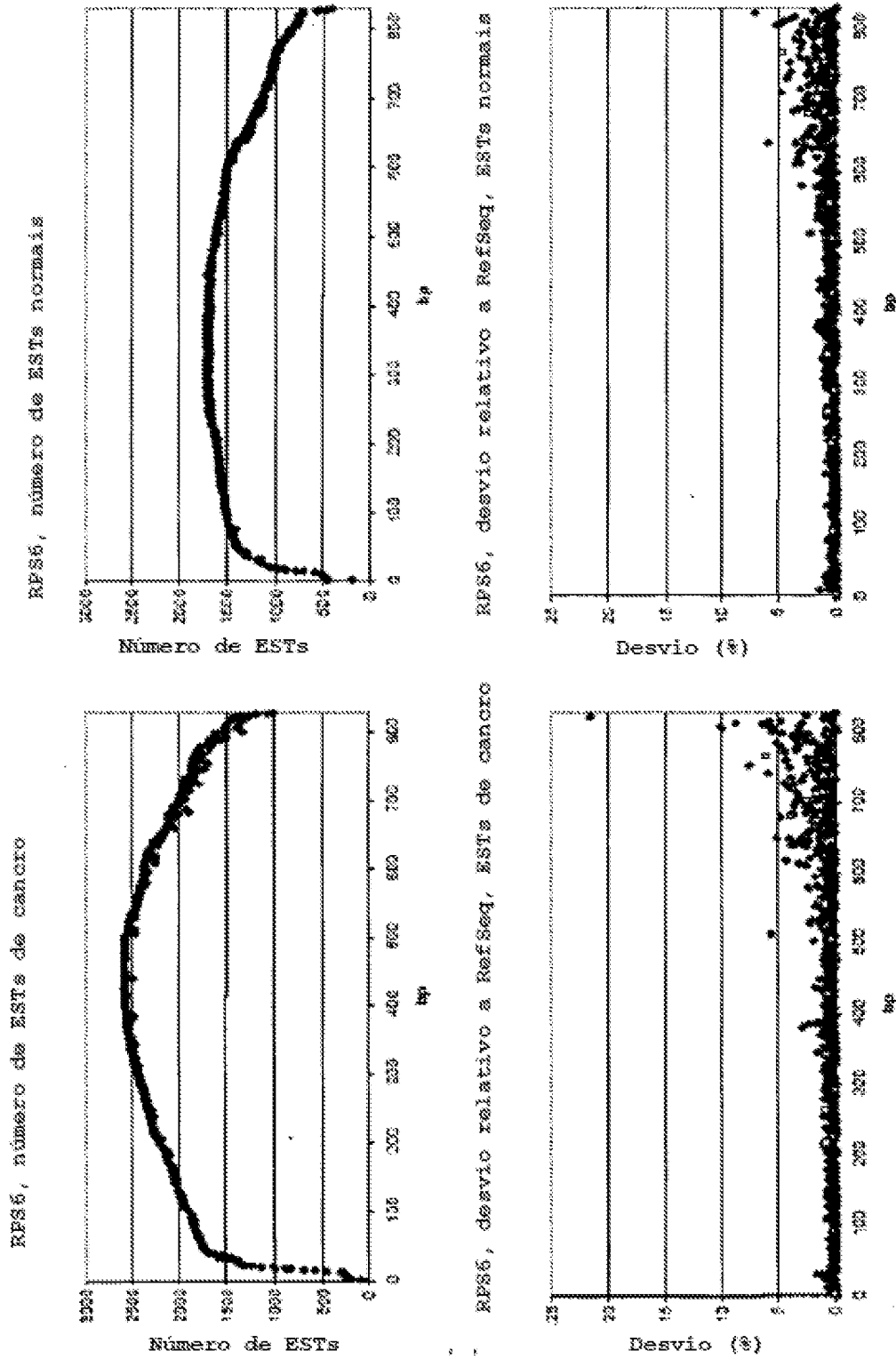


Figura 4d

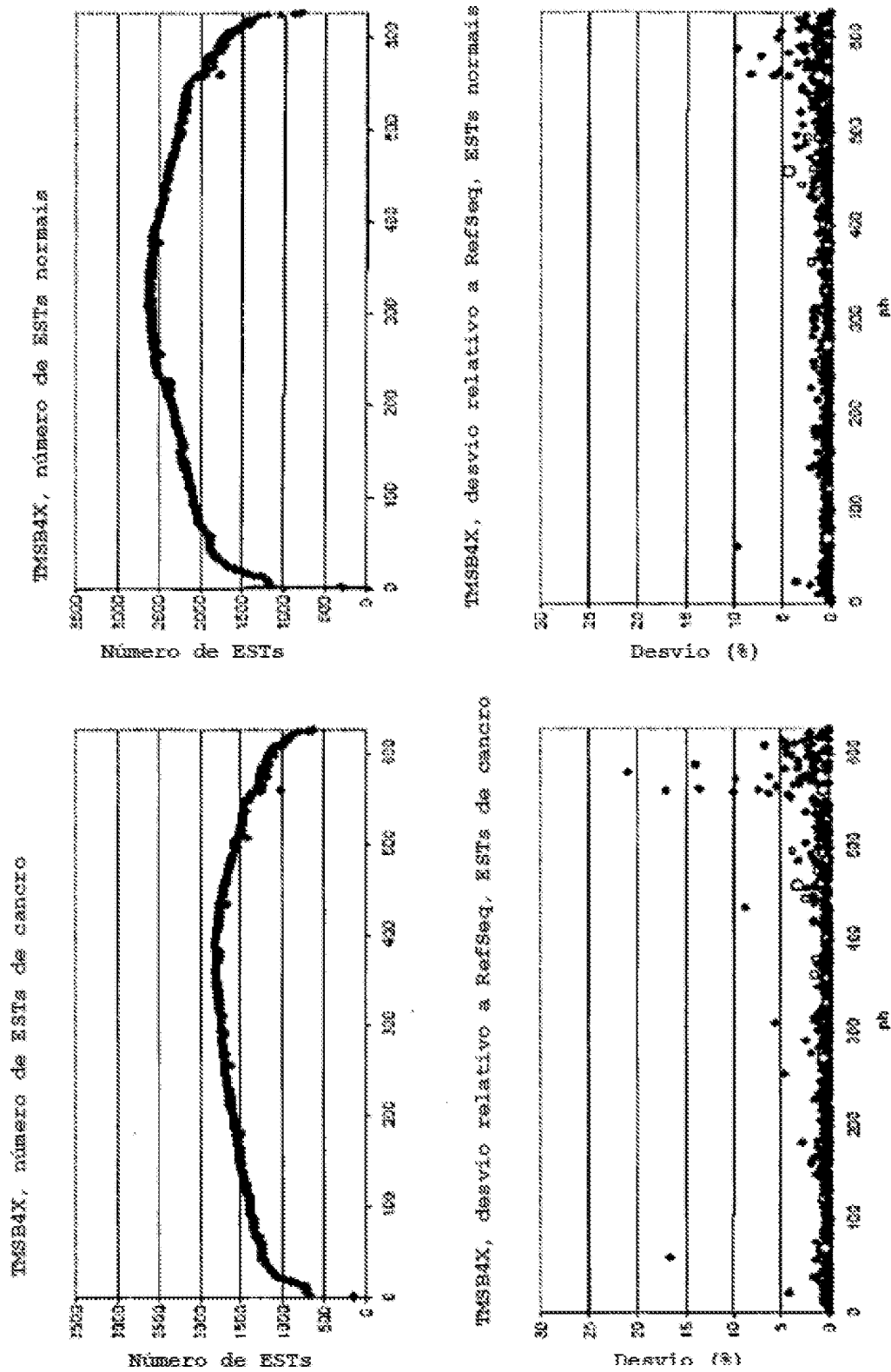


Figura 4d (continuação)

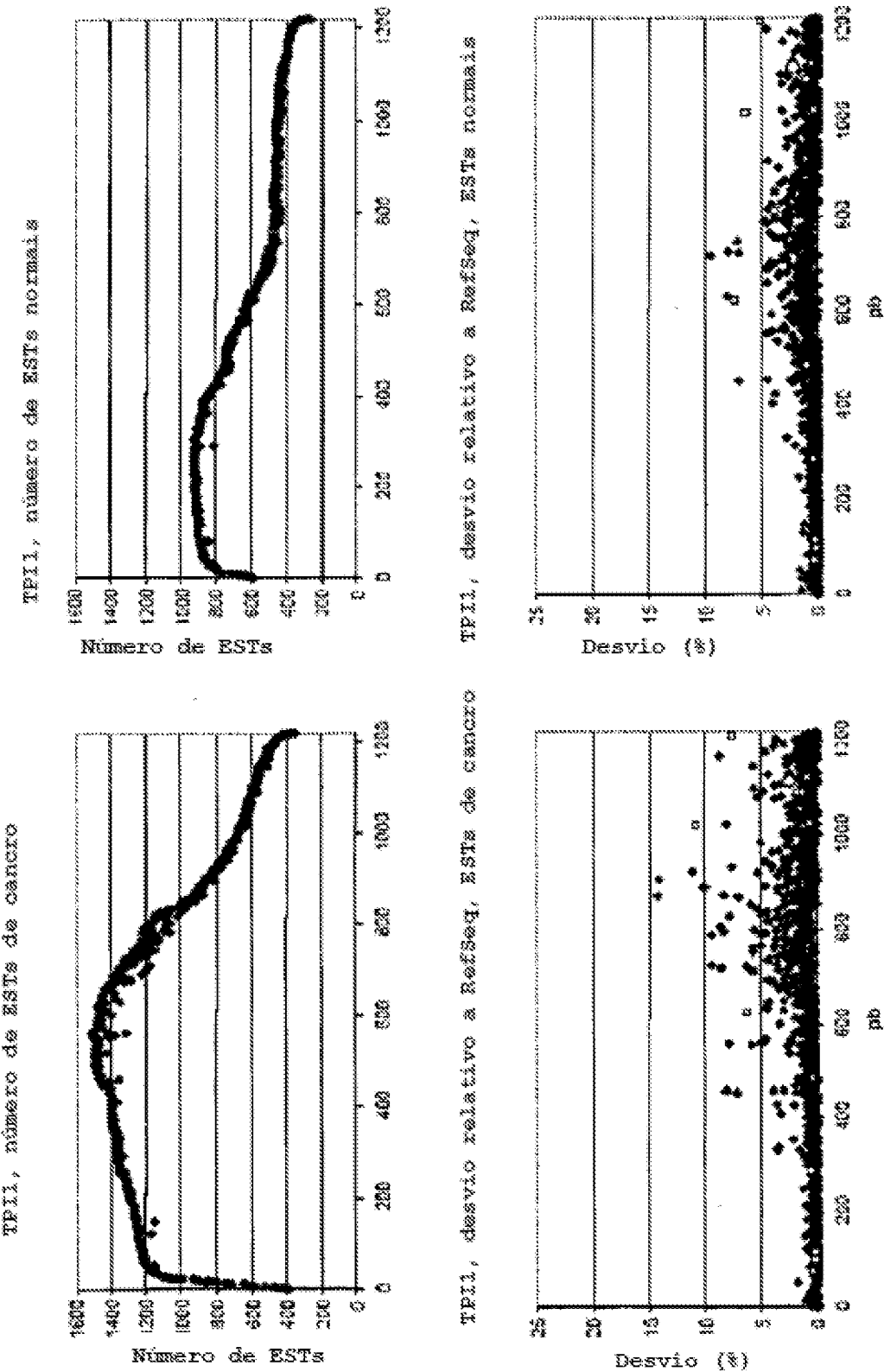


Figura 4d (continuação)

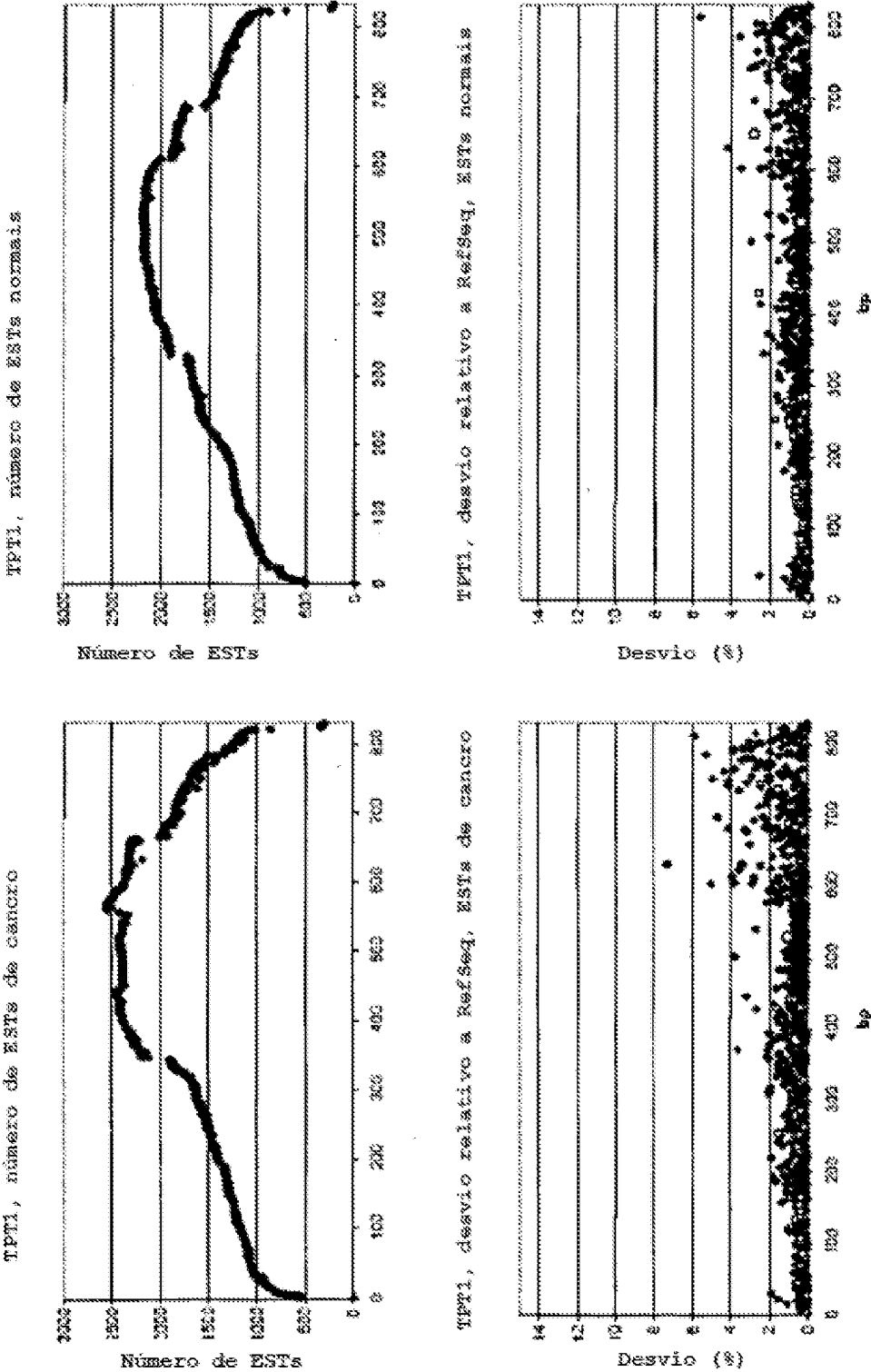


Figura 4d (continuação)

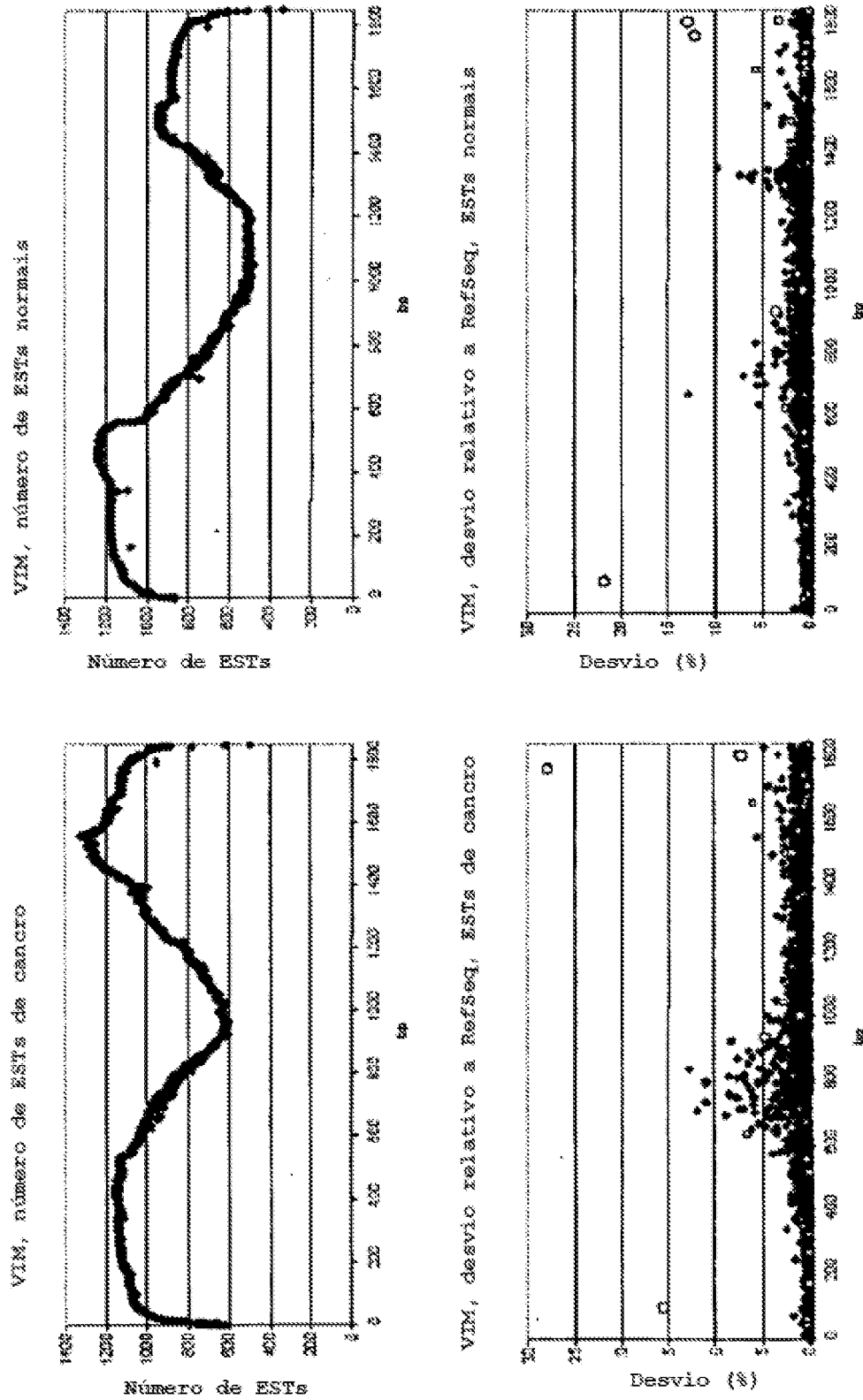


Figura 4d (continuação)

TP1: Número de ESTs em qualquer posição de RefSeq do grupo de cancro

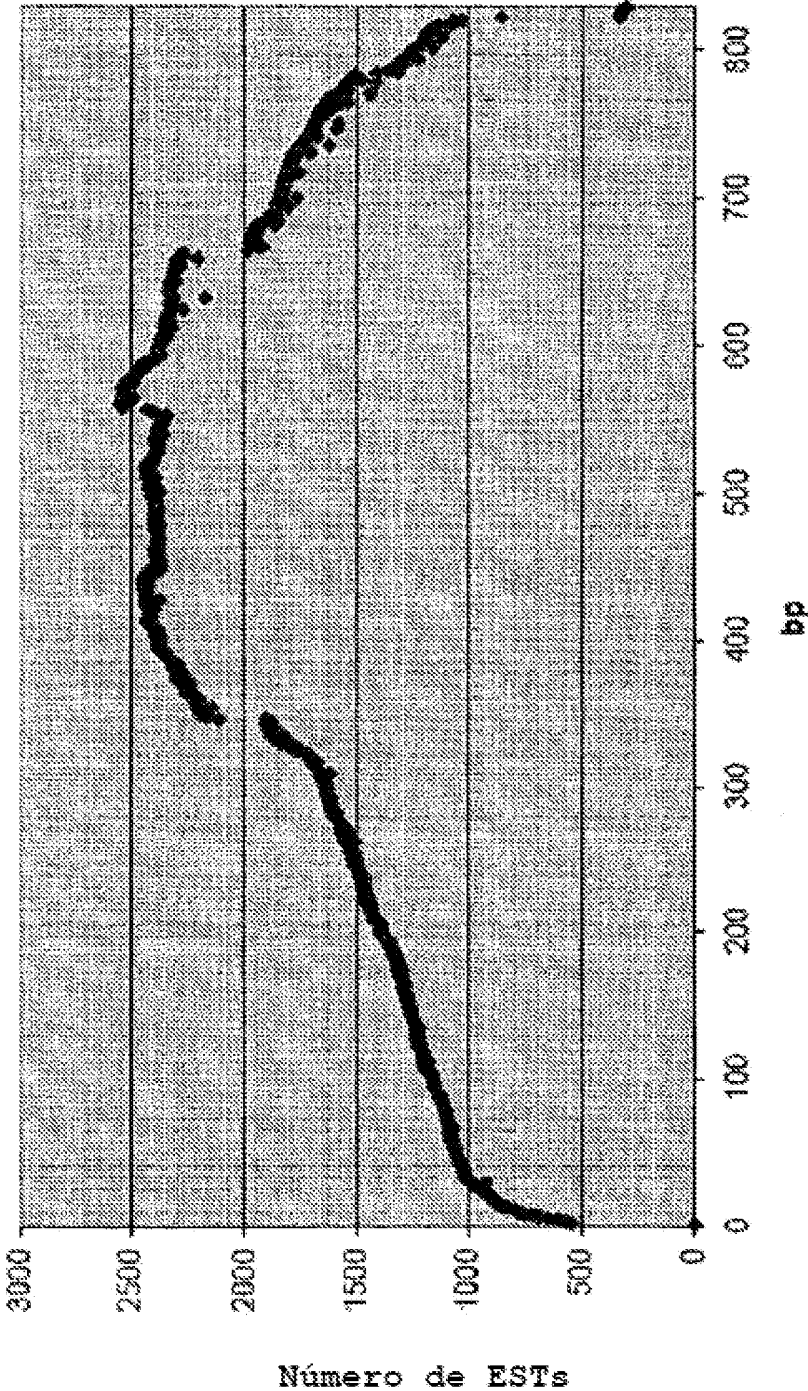


Figura 5a

TPP1: Número de ESTs em qualquer posição de RefSeq do grupo normal

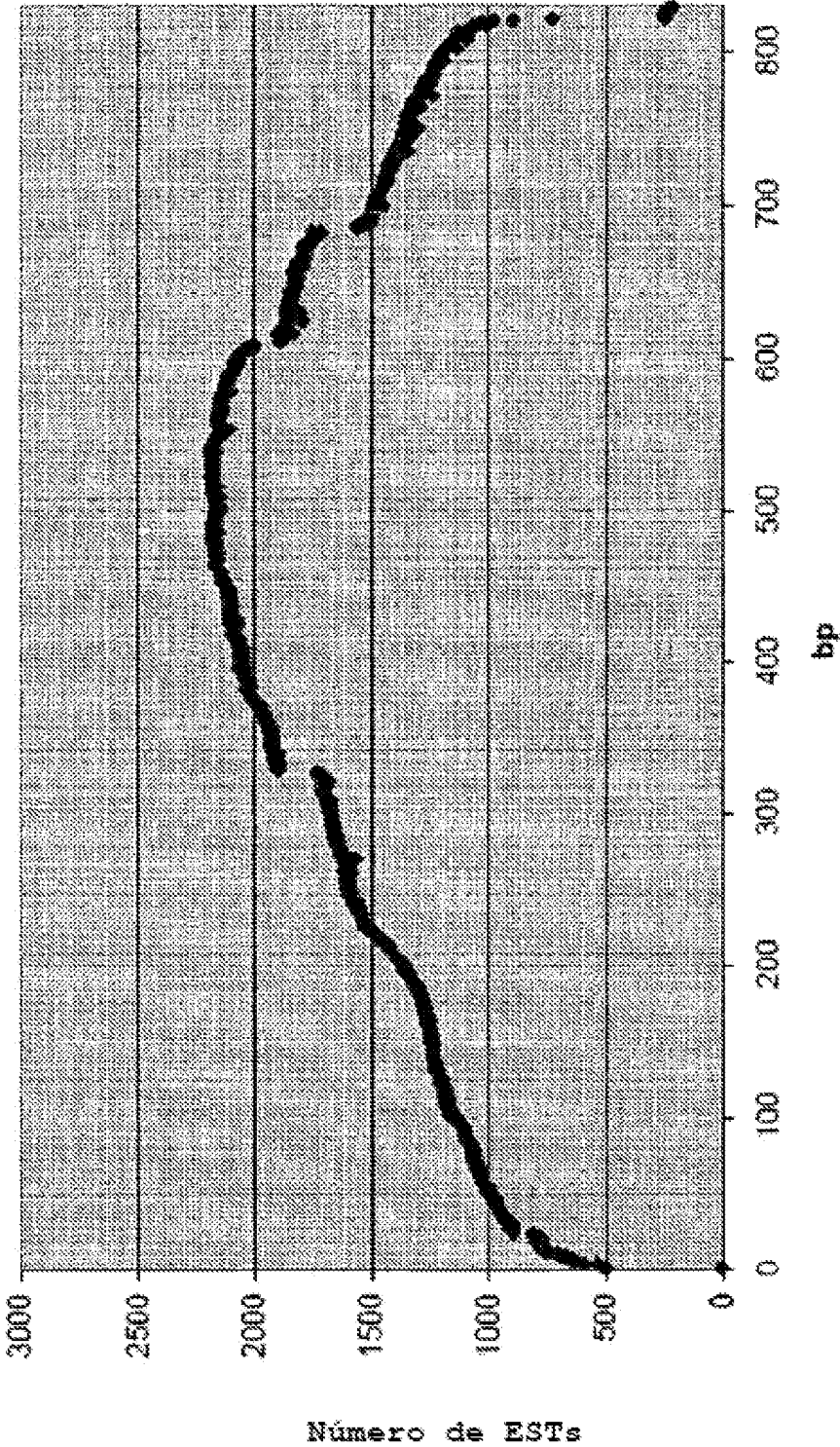


Figura 5b

TP1: Proporção de testes realizados em 489 de 830 posições

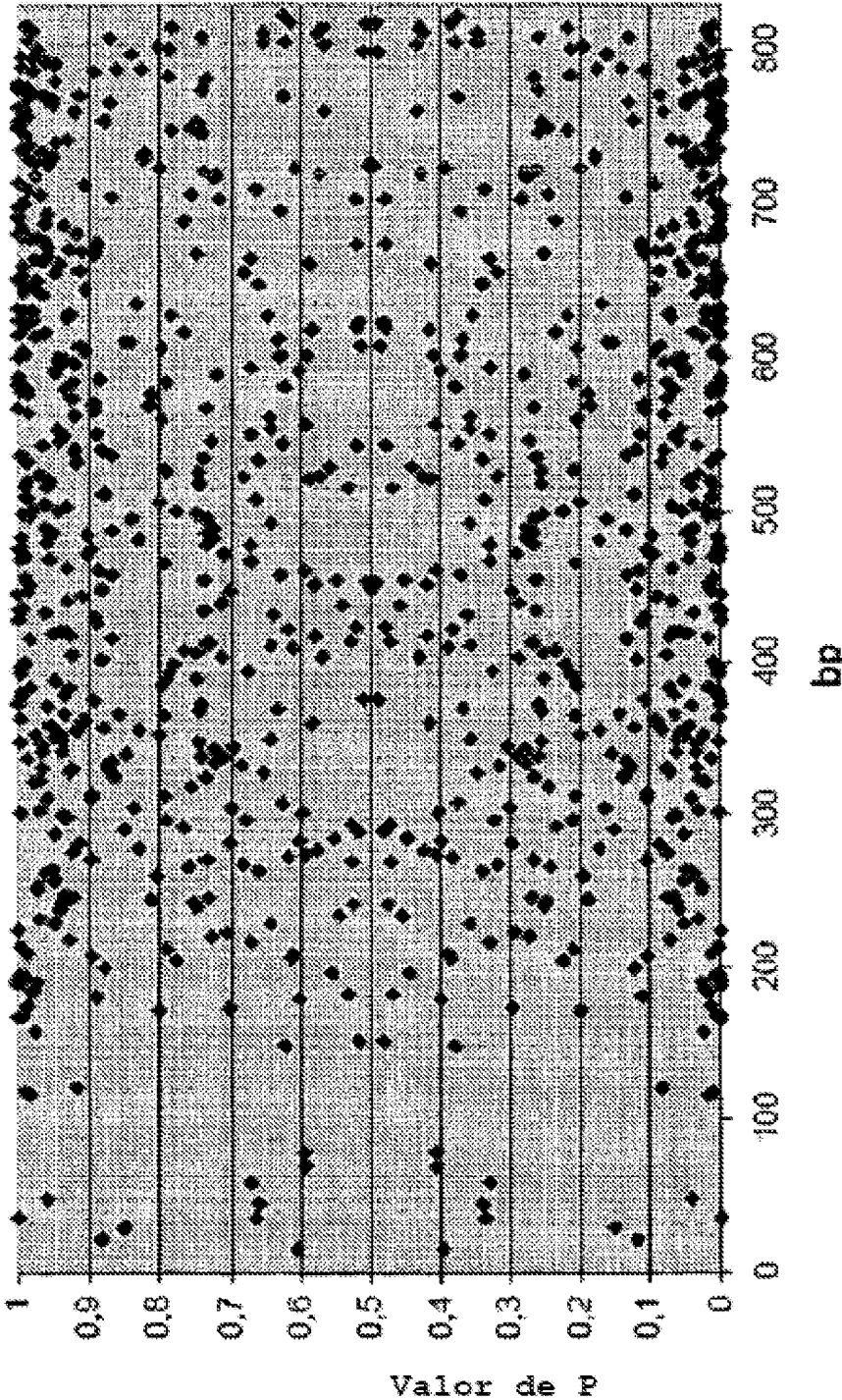


Figura 5c

TPW1:

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)

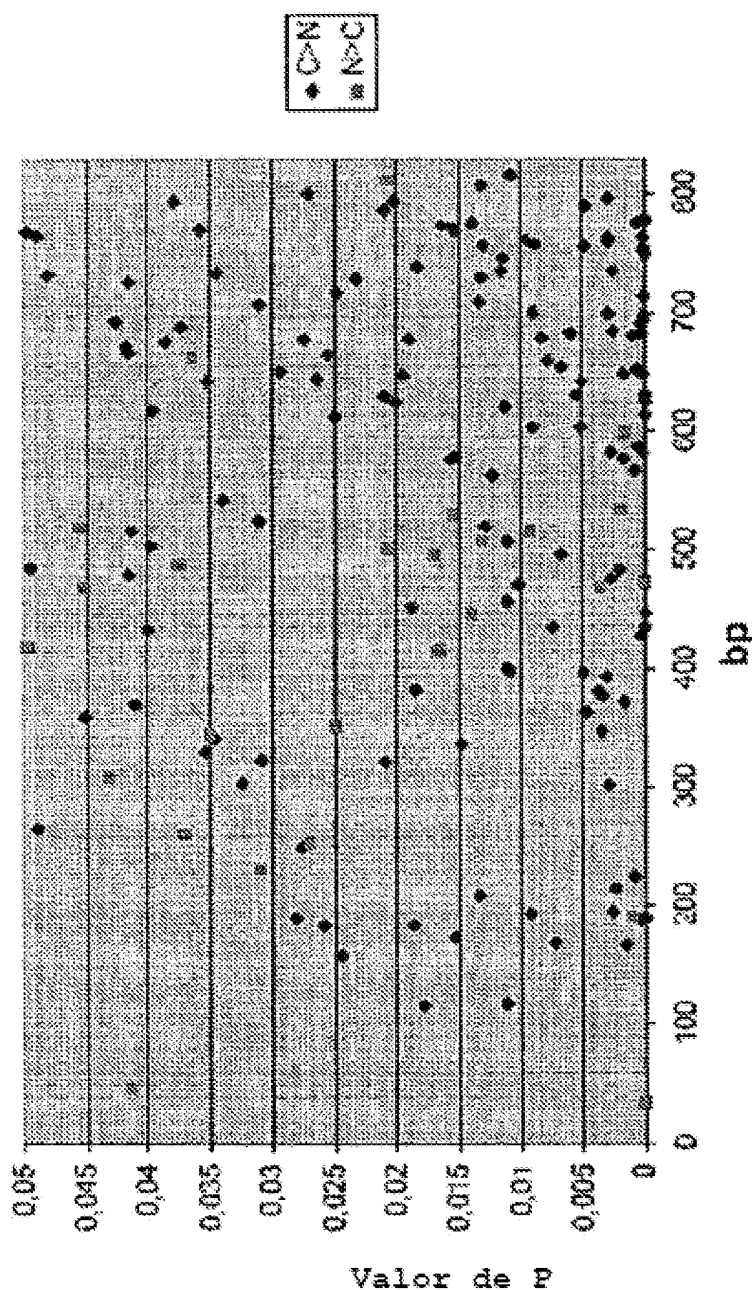


Figura 5d

VIM: Número de ESTs em qualquer posição de RefSeq do grupo de cancro

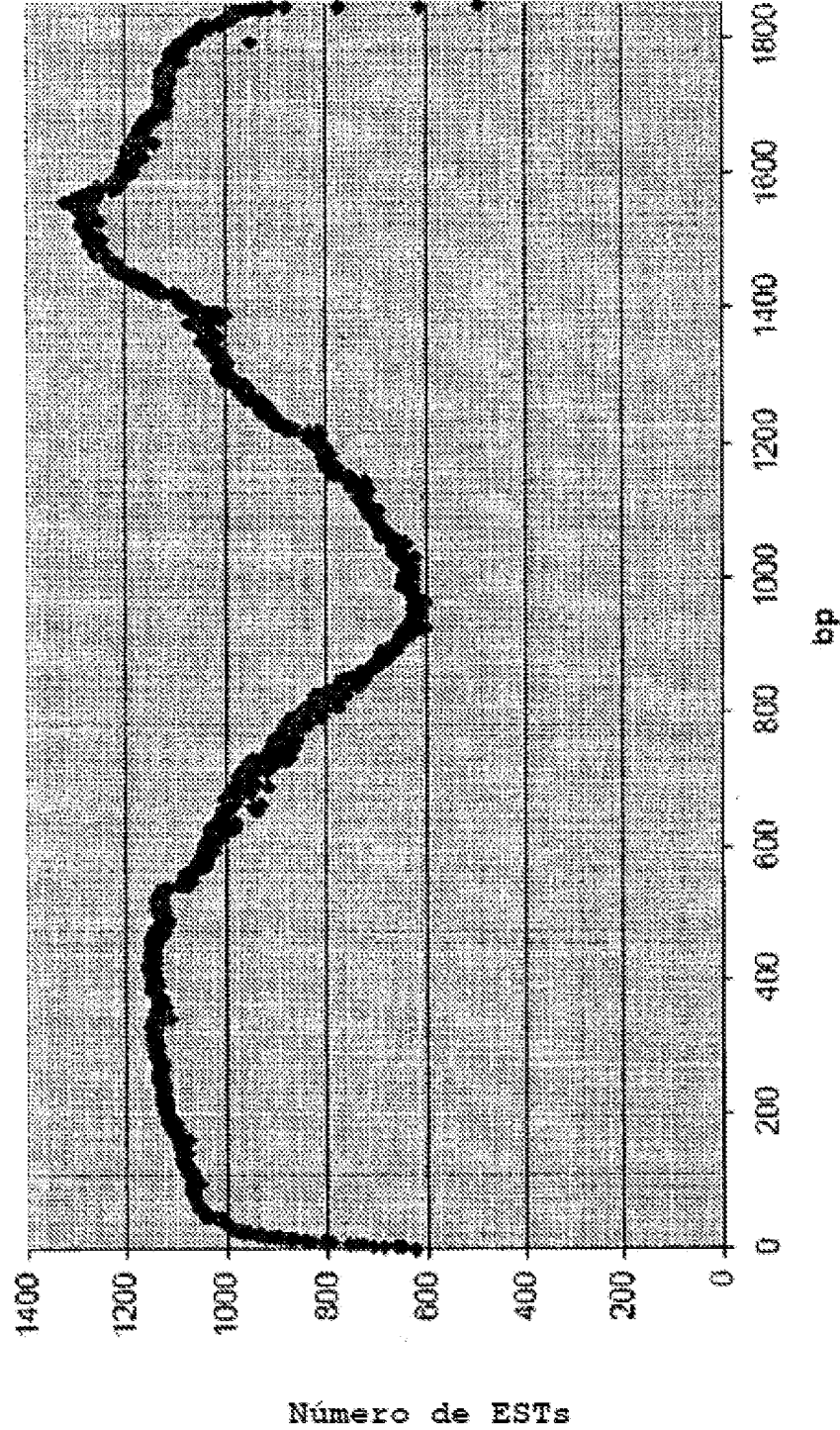


Figura 5e

VIM: Número de ESTs em qualquer posição de RefSeq do grupo normal

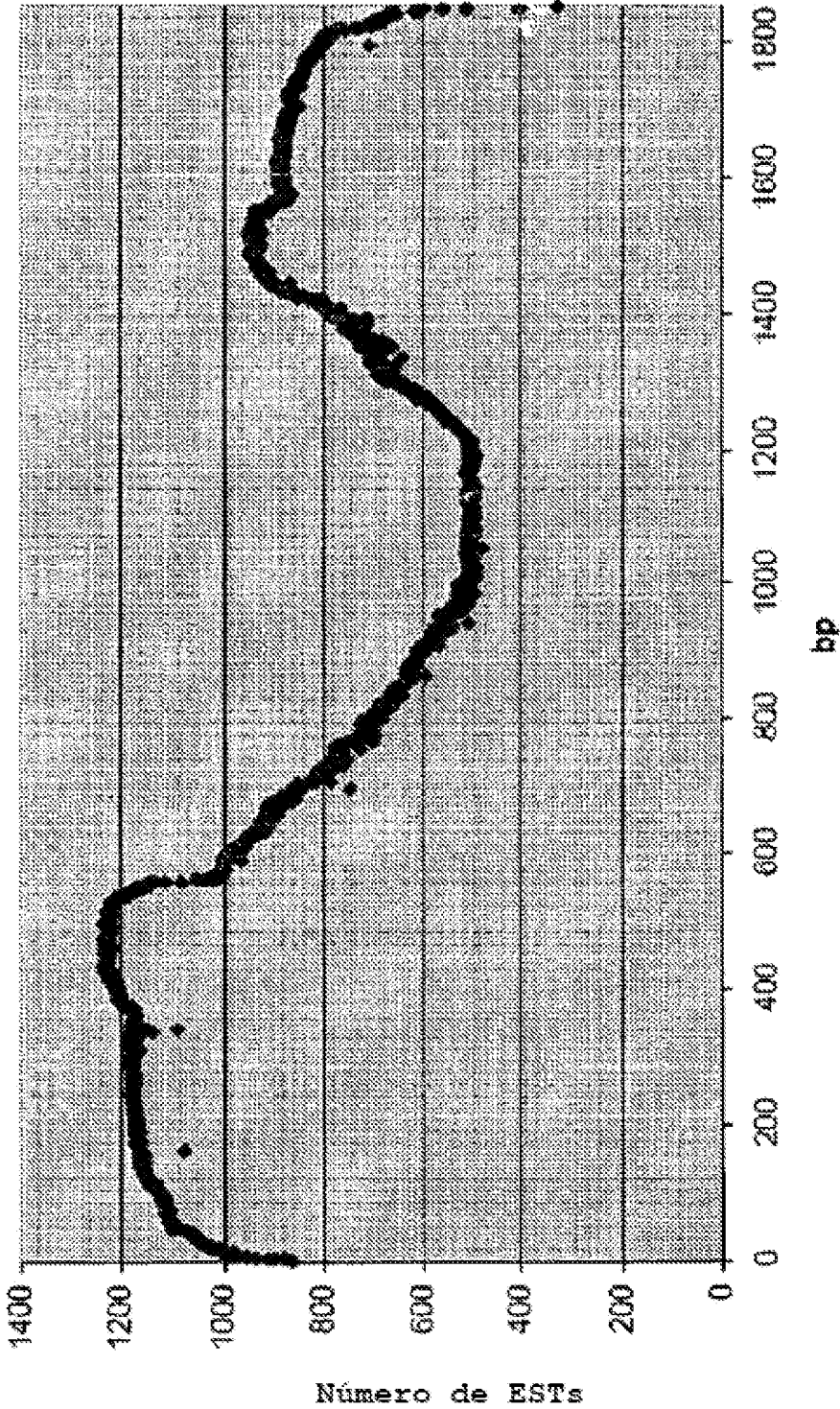


Figura 5f

VIM: Proporção de testes realizados em 752 de 1847 posições

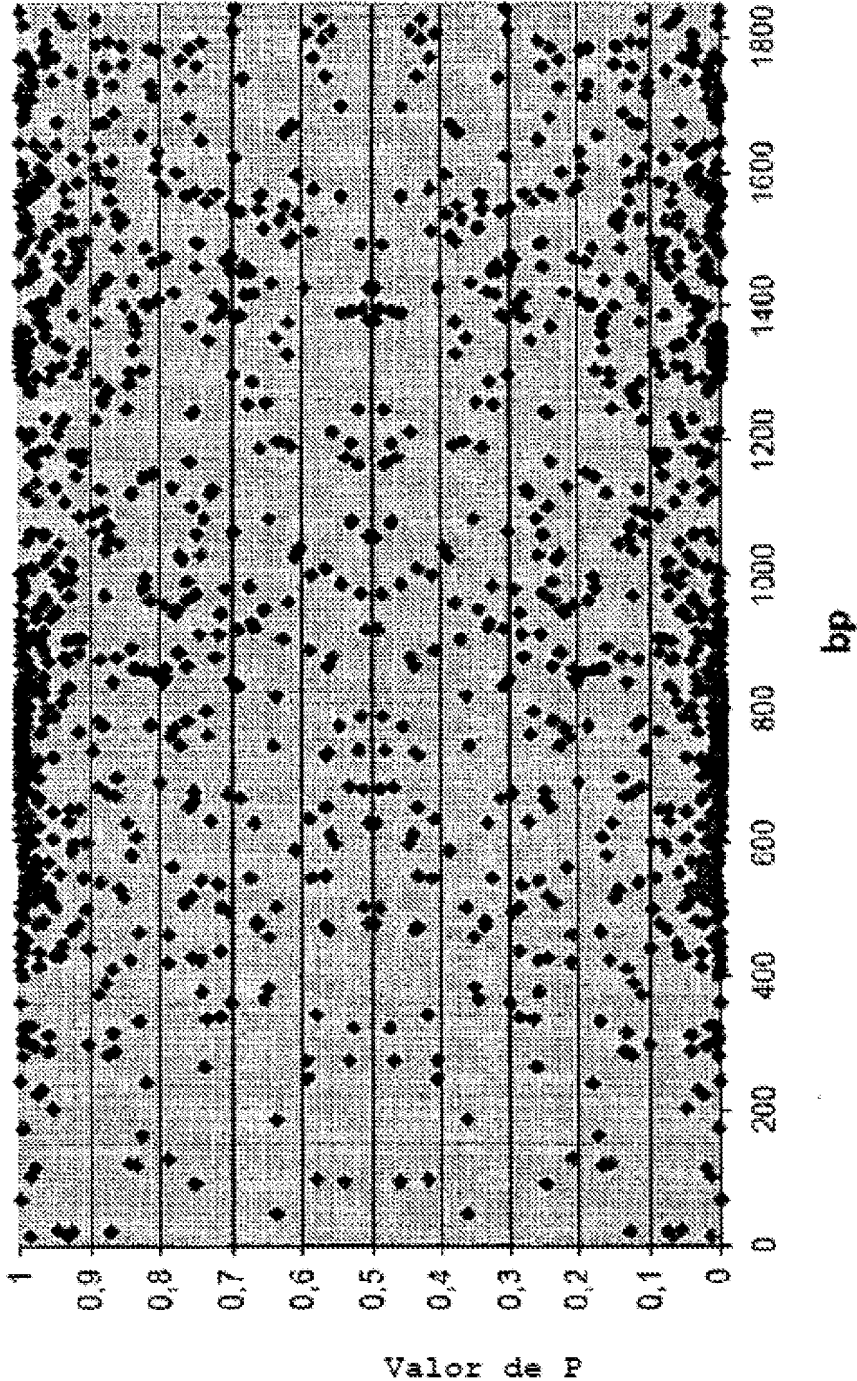


Figura 5g

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)

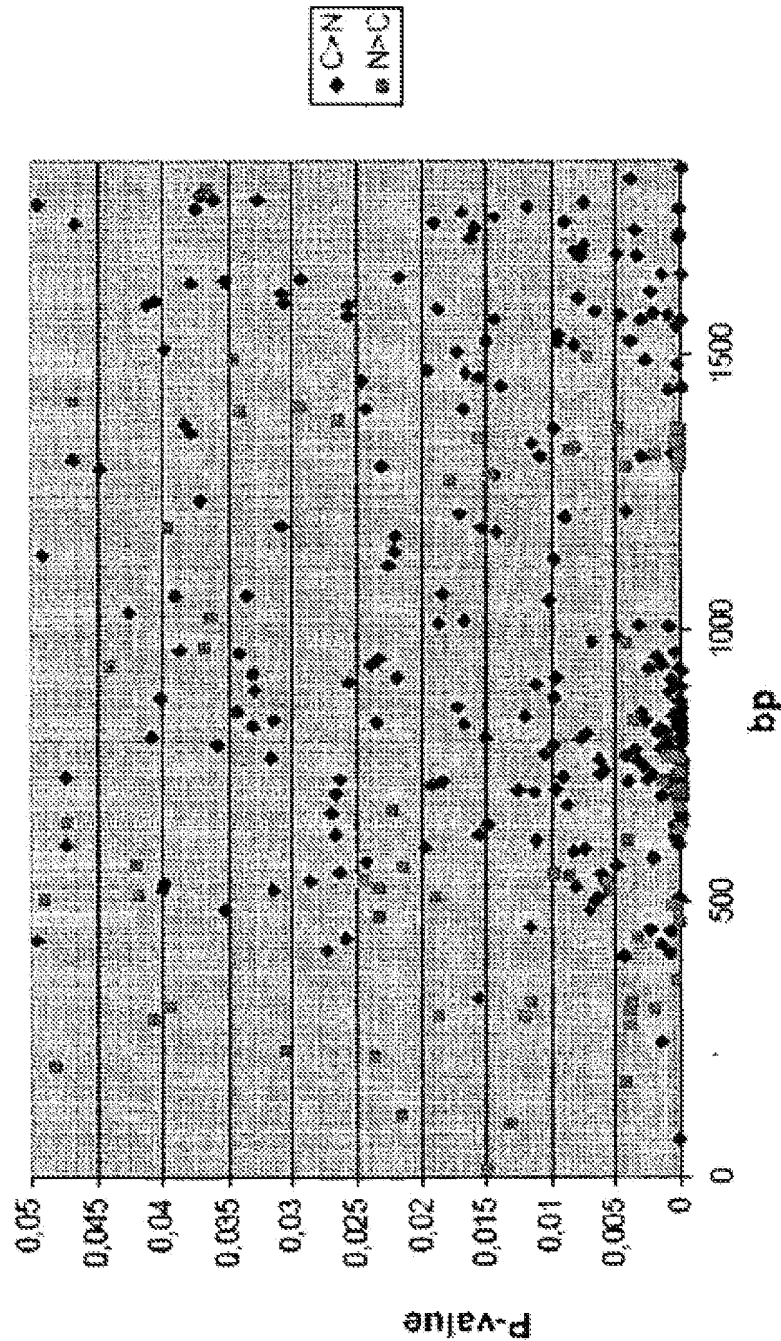


Figura 5h

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	336	30		81		11	35	21
transcript1								
ATP5A1	238	38		123		3	6	20
transcript1								
CALM2	374	23		69		16	40	21
ENO1	614	27		186		18	31	42
FTH1	592	71		226		20	57	38
FTL	678	56		118		31	86	36
GAPDH	812	91		311		23	61	57
HSPA8	513	59		163		13	18	37
transcript1								
LDHA	181	9		70		4	10	13
RPL7A	337	35		103		11	33	22
RPS4X	371	45		124		11	27	25
RPS6	432	19		88		17	40	25
TMSB4X	352	19		70		18	74	17
TPH1	379	31		99		14	47	23
TPT1	489	26		145		15	26	33
VIM	752	57		269		24	78	50

Figura 5i

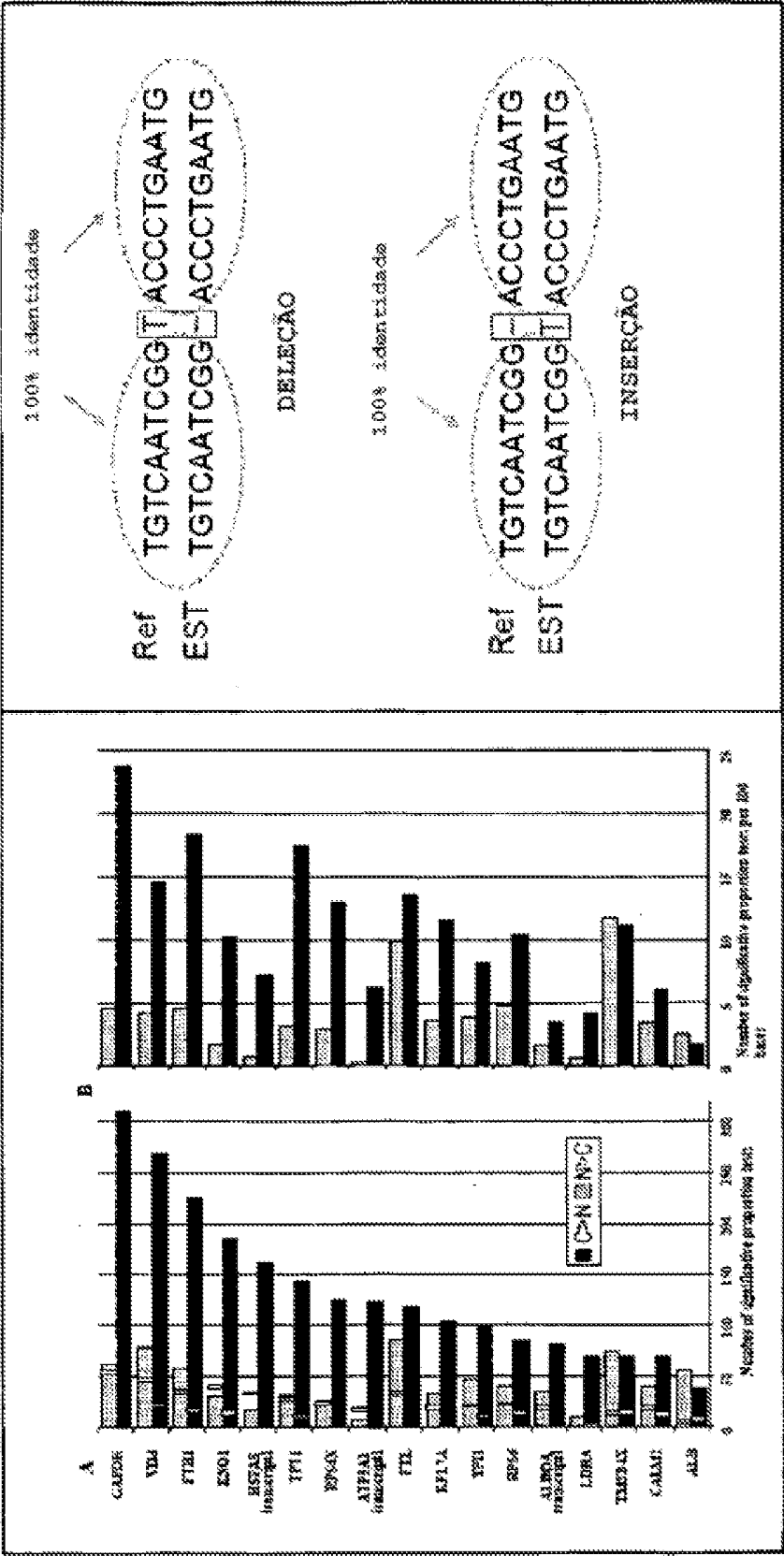
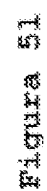


Figura 5j

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

DELEÇÕES					INSERÇÕES				
Número de testes					Número de testes				
genes					genes				
de proporções					de proporções				
>N	LBE	N>C	LBE	C>N / N>C	>N	LBE	N>C	LBE	C>N / N>C
ALB	17	1	13	0	0.1	9	0	9	0.0
ALDOA	34	8	1	7	1	5	1	0	1.0
ATP5A1	24	16	0	0	2	9	0	0	
CALM2	50	14	2	4	2	72	3	10	3
ENOT	41	29	0	5	3	40	0	2	3
FTIH1	67	46	0	4	5	114	1	3	9
FTL	100	42	2	9	7	188	9	33	9
GAPDH	99	59	2	17	7	137	4	18	9
HSPA8	30	27	0	0	2	14	0	2	1
LDHA	14	6	0	3	0	2	0	0	0
RPL7A	45	26	0	1	3	56	1	1	3
RPS4X	35	29	0	1	3	29	1	2	1
RPS6	43	24	0	1	3	44	1	1	3
TMSB4X	28	16	0	4	1	22	1	6	0
TPH	30	17	0	1	2	26	1	5	1
TPT1	43	27	0	4	3	41	1	4	2
VIM	26	11	1	9	1	27	1	3	1
Total	728	396	9	83	45	835	24	109	45
				4.8					2.3

Figura 5k



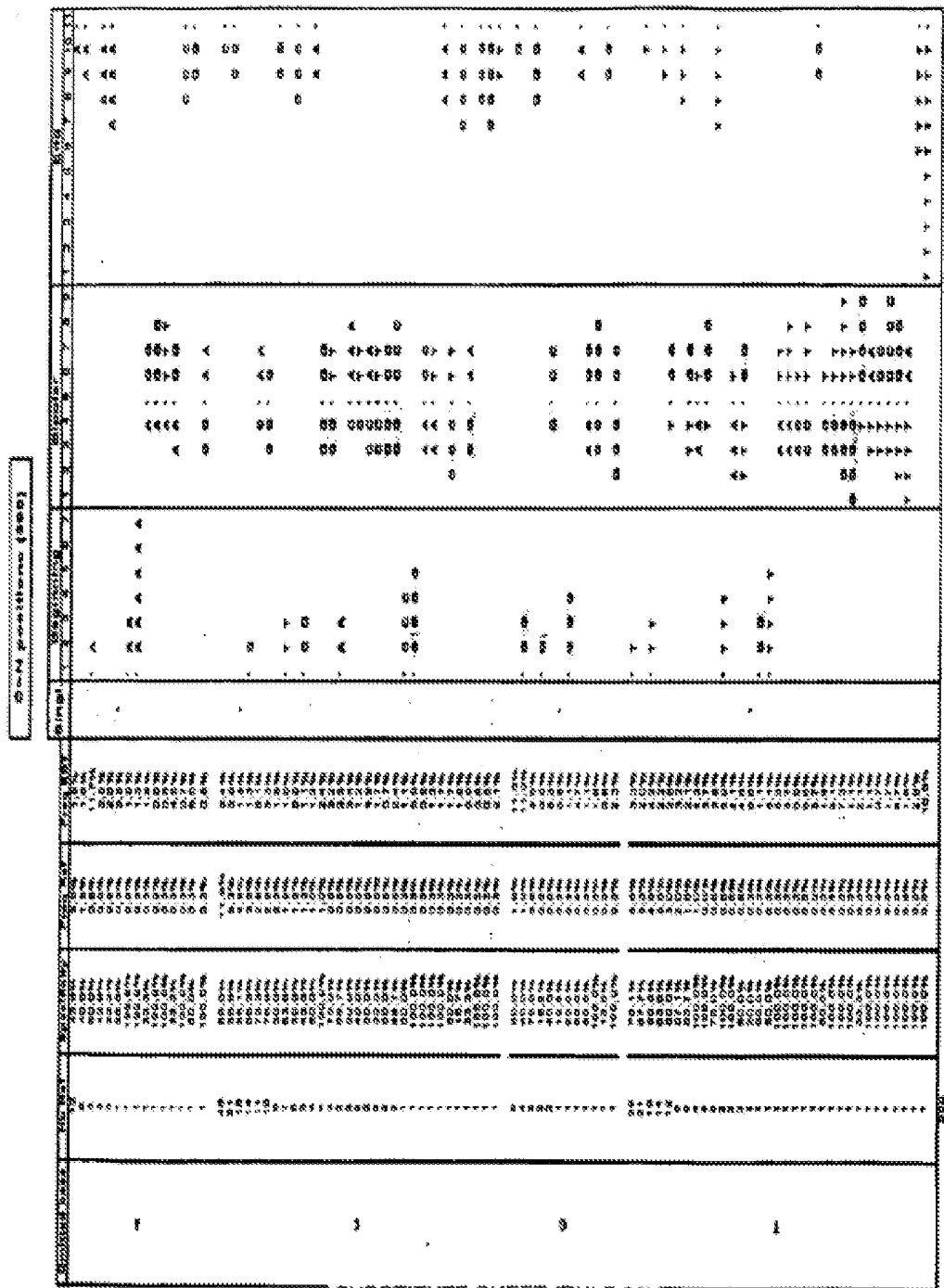


Figura 5m

4		3		2		1		0		-1		-2		-3		-4		-5		-6		-7		-8		-9		-10		-11		-12		-13		-14		-15		-16		-17		-18		-19		-20		-21		-22		-23		-24		-25		-26		-27		-28		-29		-30		-31		-32		-33		-34		-35		-36		-37		-38		-39		-40		-41		-42		-43		-44		-45		-46		-47		-48		-49		-50		-51		-52		-53		-54		-55		-56		-57		-58		-59		-60		-61		-62		-63		-64		-65		-66		-67		-68		-69		-70		-71		-72		-73		-74		-75		-76		-77		-78		-79		-80		-81		-82		-83		-84		-85		-86		-87		-88		-89		-90		-91		-92		-93		-94		-95		-96		-97		-98		-99		-100		-101		-102		-103		-104		-105		-106		-107		-108		-109		-110		-111		-112		-113		-114		-115		-116		-117		-118		-119		-120		-121		-122		-123		-124		-125		-126		-127		-128		-129		-130		-131		-132		-133		-134		-135		-136		-137		-138		-139		-140		-141		-142		-143		-144		-145		-146		-147		-148		-149		-150		-151		-152		-153		-154		-155		-156		-157		-158		-159		-160		-161		-162		-163		-164		-165		-166		-167		-168		-169		-170		-171		-172		-173		-174		-175		-176		-177		-178		-179		-180		-181		-182		-183		-184		-185		-186		-187		-188		-189		-190		-191		-192		-193		-194		-195		-196		-197		-198		-199		-200		-201		-202		-203		-204		-205		-206		-207		-208		-209		-210		-211		-212		-213		-214		-215		-216		-217		-218		-219		-220		-221		-222		-223		-224		-225		-226		-227		-228		-229		-230		-231		-232		-233		-234		-235		-236		-237		-238		-239		-240		-241		-242		-243		-244		-245		-246		-247		-248		-249		-250		-251		-252		-253		-254		-255		-256		-257		-258		-259		-260		-261		-262		-263		-264		-265		-266		-267		-268		-269		-270		-271		-272		-273		-274		-275		-276		-277		-278		-279		-280		-281		-282		-283		-284		-285		-286		-287		-288		-289		-290		-291		-292		-293		-294		-295		-296		-297		-298		-299		-300		-301		-302		-303		-304		-305		-306		-307		-308		-309		-310		-311		-312		-313		-314		-315		-316		-317		-318		-319		-320		-321		-322		-323		-324		-325		-326		-327		-328		-329		-330		-331		-332		-333		-334		-335		-336		-337		-338		-339		-340		-341		-342		-343		-344		-345		-346		-347		-348		-349		-350		-351		-352		-353		-354		-355		-356		-357		-358		-359		-360		-361		-362		-363		-364		-365		-366		-367		-368		-369		-370		-371		-372		-373		-374		-375		-376		-377		-378		-379		-380		-381		-382		-383		-384		-385		-386		-387		-388		-389		-390		-391		-392		-393		-394		-395		-396		-397		-398		-399		-400		-401		-402		-403		-404		-405		-406		-407		-408		-409		-410		-411		-412		-413		-414		-415		-416		-417		-418		-419		-420		-421		-422		-423		-424		-425		-426		-427		-428		-429		-430		-431		-432		-433		-434		-435		-436		-437		-438		-439		-440		-441		-442		-443		-444		-445		-446		-447		-448		-449		-450		-451		-452		-453		-454	
---	--	---	--	---	--	---	--	---	--	----	--	----	--	----	--	----	--	----	--	----	--	----	--	----	--	----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	-----	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--

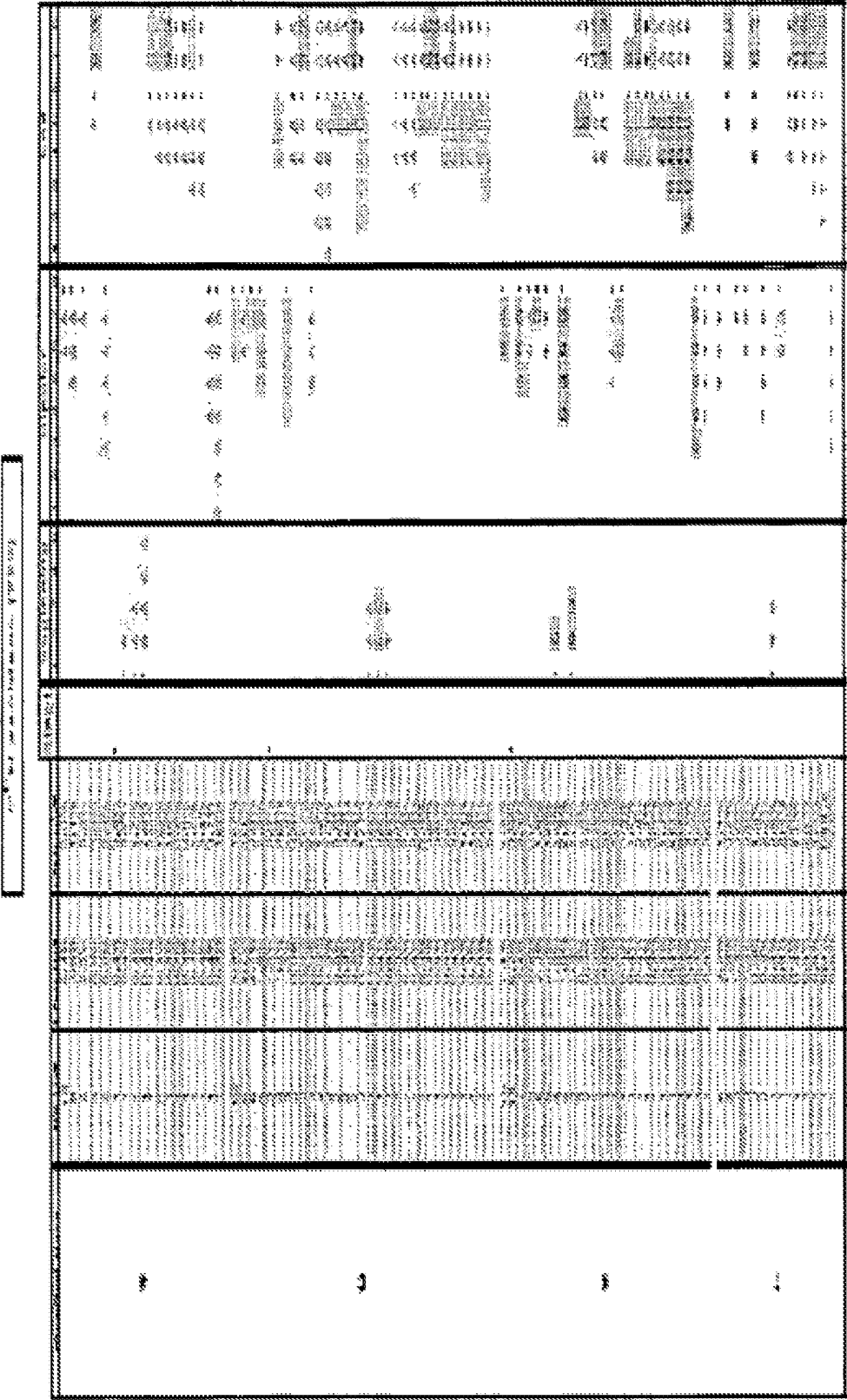


Figure 50

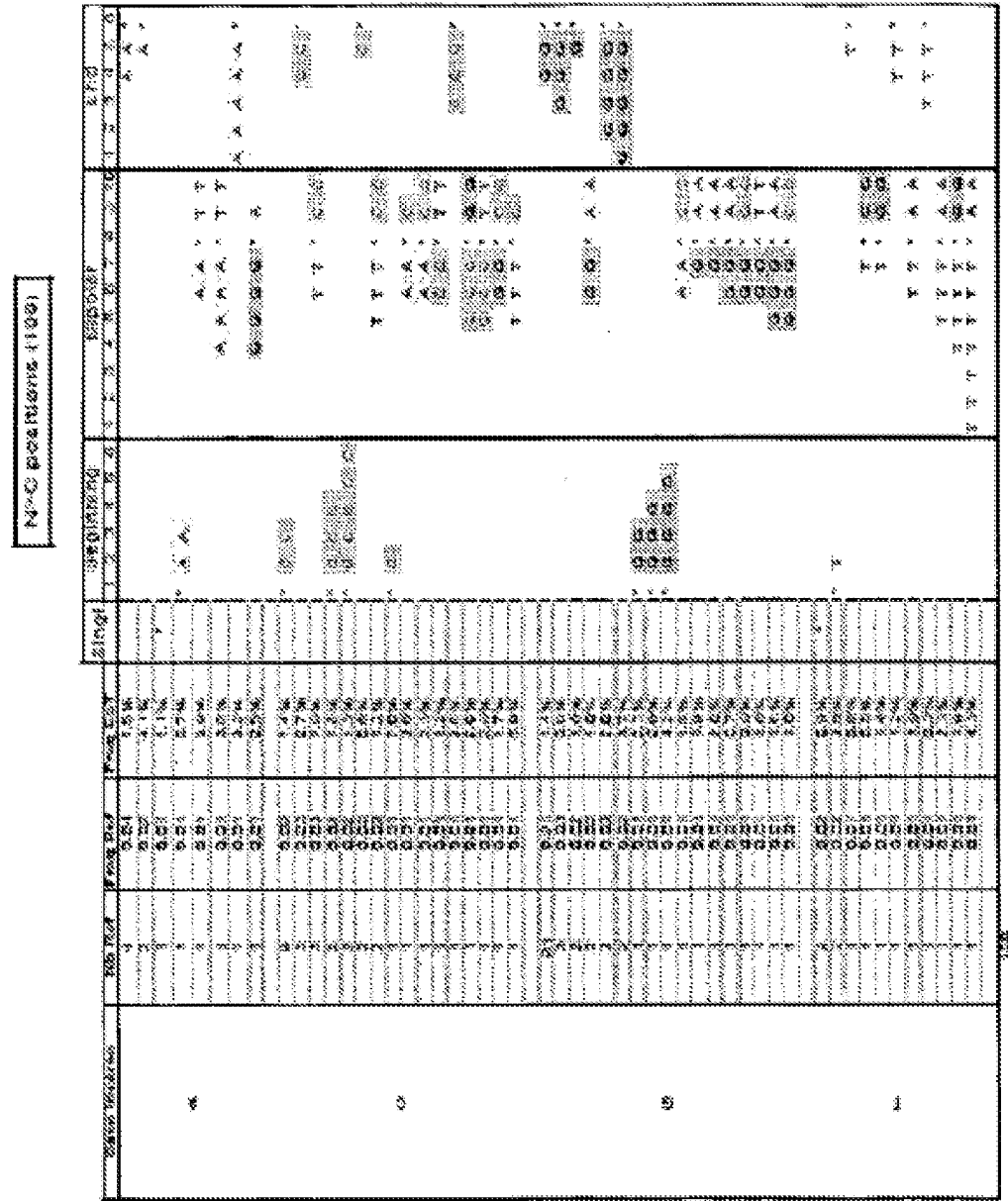


Figure 5p

	Test Numbers								CAN								LBE							
	F1	F2	F3	F4	F5	F6	F7	F8	F1	F2	F3	F4	F5	F6	F7	F8	F1	F2	F3	F4	F5	F6	F7	F8
ENO1	514	599	557	321	220	217	193	185	111	32	40	37	27	18	14	14								
FTH1	522	535	532	380	235	232	275	276	202	152	30	25	25	15	13	13								
GAPDH	512	752	745	524	440	372	361	325	252	154	42	37	27	25	28	28								
HSP48	513	423	415	197	160	181	151	133	75	25	32	23	23	11	13	13								
RPL7A	337	254	281	175	94	135	125	127	75	22	15	14	14	5	2	2								
RPS12	371	340	238	183	170	161	137	135	54	75	21	12	12	15	13	13								
RPS6	432	357	357	230	225	125	123	115	58	58	31	23	23	11	13	13								
TPST1	435	452	423	350	300	171	154	155	52	33	31	23	23	23	23	23								
VIM	752	544	525	327	337	347	355	253	174	574	26	31	30	15	15	15								
ALB	194	117	117	43	49	54	53	55	33	33	3	4	4	1	1	1								
FTL	575	544	544	454	444	304	213	212	133	127	42	40	44	33	31	31								
TMSB4X	352	303	332	153	403	144	124	124	55	33	30	15	15	15	15	15								
ALDOA	335	237	237	174	174	115	152	152	51	51	22	12	12	11	11	11								
ATPSA1	235	211	211	122	58	125	114	115	53	25	5	5	5	4	4	4								
CALM2	374	255	335	245	245	150	153	135	52	52	25	23	23	15	15	15								
LDHA	151	580	550	75	75	53	55	55	33	33	5	7	7	4	4	4								
TPH1	375	337	335	325	150	145	142	142	53	47	20	15	15	15	15	15								
All genes	7544	5152	5701	4140	3759	2524	2757	2759	1574	1353	449	151	372	220	220	220								

Figura 6a

EN1: EN1 sequence, 500 bp
F1: 500 bp EN1 signal (mean)
F2: 500 bp EN1 signal (mean)
F3: 500 bp EN1 signal (mean)
F4: 500 bp EN1 signal (mean)
F5: 500 bp EN1 signal (mean)
F6: 500 bp EN1 signal (mean)
F7: 500 bp EN1 signal (mean)
F8: 500 bp EN1 signal (mean)

	C-M					LEE					M-C					LBE				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
EMOH	164	167	188	84	82	18	17	17	9	7	31	31	31	17	21	42	38	39	21	14
FTM	229	235	236	186	183	23	19	19	12	13	67	71	72	48	82	38	38	33	29	20
GAPDH	311	322	339	206	85	23	22	22	14	22	81	82	88	29	66	87	82	82	38	21
HSPA6	183	189	120	88	81	13	11	11	5	6	16	22	21	9	7	37	30	30	34	12
RPL7A	103	94	64	87	47	11	6	9	9	6	33	34	33	21	18	22	19	13	11	13
RPS3L	124	133	112	71	83	11	11	10	8	6	37	22	23	13	12	26	26	22	19	12
RPS8	88	77	76	48	32	17	16	16	13	13	40	43	43	29	38	28	28	29	19	12
TPT1	145	138	137	74	74	16	16	14	9	9	26	28	28	34	14	33	30	30	20	20
YMG	269	233	232	148	148	24	21	16	9	8	76	79	43	23	28	50	43	42	25	26
ALB	38	24	24	19	19	10	9	9	2	3	68	46	45	23	23	8	4	4	1	1
FTL	118	121	121	71	87	31	26	26	23	23	86	92	91	29	70	36	34	34	24	23
TMSB4X	70	90	56	34	24	19	16	16	9	8	74	58	66	22	32	17	14	34	8	9
ALDOA	81	73	78	50	53	11	10	10	8	8	36	36	39	11	11	21	19	19	11	11
ATPSA1	153	111	111	81	29	3	2	2	1	1	9	3	3	0	0	20	18	18	8	8
CALM2	89	87	87	53	53	13	13	13	9	9	40	34	38	29	29	21	19	19	14	14
LON4	70	81	80	31	31	4	3	3	1	1	16	8	9	3	2	13	12	12	8	8
TPT1	98	88	88	82	23	14	12	12	8	8	47	46	48	21	24	20	20	20	12	12
Ad genes	2381	2382	2005	1300	132	249	233	225	132	140	329	722	624	374	489	438	428	428	368	218

(C-M(N-C))

F1	F2	F3	F4	F5
3.15	3.86	2.98	3.18	2.85

Figura 6b

	C>N-LBE					N>C-LBE				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
BN01	109	150	142	80	40	0	0	0	0	7
FTN1	308	122	127	140	87	19	38	32	22	32
QAPDH	288	280	277	182	43	4	13	14	0	88
HSPAB	150	127	128	51	86	0	0	0	0	0
RPL7A	80	85	85	51	42	11	18	15	18	6
RPL4X	113	99	102	88	28	2	0	1	0	0
RPS6	84	82	81	32	22	15	20	20	13	14
TPY1	130	121	123	89	69	0	0	0	0	0
YH8	243	208	216	140	140	23	22	6	1	1
ALB	28	18	18	7	7	48	41	41	22	22
FTL	87	92	92	53	34	50	59	57	32	50
TMSB4X	52	41	41	29	29	67	64	64	24	24
ALDOA	78	83	82	45	45	14	10	10	0	0
ATP5A1	129	109	109	82	23	0	0	0	0	0
CALM2	53	54	54	44	44	18	17	17	15	15
LDNA	88	88	87	30	30	0	0	0	0	0
TPY1	95	91	91	55	17	24	29	29	6	18
All genes	2022	1835	1842	1188	703	231	324	303	147	257

(C>N)-LBE(N>C)-LBE

F1	F2	F3	F4	F5
6.33	5.59	6.08	7.55	3.09

Figura 6c-1

	C>N	LBE	N>C	LBE	(C>N)/(N>C)	(C>N)-LBE	(N>C)-LBE
F1	2281	259	723	489	3.15	6.95	
F2	2065	230	722	429	2.86	5.59	
F3	2065	223	694	428	2.98	6.08	
F4	1300	132	374	268	3.46	7.55	
F5	933	140	456	218	2.05	3.03	

Figura 6c-2

	Test Numbers				CmH		LBE		CmN		LBE		N-C		LBE		CmN-LBE		N-C-LBE					
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2				
END1	814	74	237	25	49	4			182	18	18	3	31	12	42	3	188	13	3	2				
FTN1	892	263	253	78	20	18			225	54	20	18	87	21	38	13	235	44	18	6				
GA9PDH	840	323	372	118	42	26			311	78	23	14	81	43	57	13	135	42	4	24				
HSP98	843	58	183	44	32	3			183	37	13	2	18	7	37	6	158	35	3	2				
RPL7A	337	51	136	21	19	3			103	98	11	1	32	5	22	3	22	16	11	2				
RP3AX	371	104	153	35	21	3			124	24	11	3	27	2	25	6	113	21	2	6				
RP36	432	140	126	32	31	12			86	26	17	6	40	6	28	8	89	18	15	0				
TP11	452	303	573	80	31	21			148	89	15	12	28	27	33	14	133	31	0	0				
YMA	752	118	347	35	33	7			288	31	24	3	78	4	50	8	345	38	23	0				
ALB	194	87	64	37	8	1			35	19	10	2	55	18	5	2	38	17	43	16				
FTL	878	284	104	102	40	27			118	23	21	21	68	70	38	14	89	2	53	82				
TNBBAX	252	177	144	33	20	11			19	68	15	18	24	45	17	2	52	3	87	52				
ALDOA	330	83	118	28	22	2			81	23	11	1	35	5	21	3	70	22	14	2				
ATP5A1	238	6	123	4	8	0			123	4	3	0	5	9	23	0	128	4	3	0				
CALM2	374	58	106	16	28	3			84	11	18	2	40	8	21	2	83	8	19	2				
LDHA	181	6	83	7	6	0			76	6	4	6	19	2	13	0	88	6	3	0				
TP11	375	83	140	16	20	4			84	11	14	2	47	5	23	3	85	8	24	5				
All genes	7841	2283	3039	747	448	145			7341	127	253	22	731	250	483	119	3000	341	307	14				
																					(C-N)-LBE(N-C)-LBE			
																					F1	F2		
																					6.95	2.15		
																					(C-N)-C1			
																					F1	F2		
																					3.15	1.56		
																					F1: EST Alignment (N=500)			
																					F2	C1e-420		

Figura 6d

	Test Numbers				CAN		LBE				CAN				LBE				N+C				CAN LBE				N+C LBE			
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2				
ENOT	014	187	217	52	43	14							188	46	18	3	21	12	12	11			188	34	0	1				
FTN1	592	312	283	121	83	18							238	111	23	3	53	10	38	22			238	103	12	0				
GNPDM	812	582	372	287	42	40							311	283	53	18	61	24	57	53			333	243	4	0				
NSPAB	513	125	151	55	22	9							182	52	13	4	18	8	37	9			182	48	0	0				
RPL7A	337	412	136	26	18	7							103	25	11	3	33	3	22	7			82	33	11	0				
RP34X	271	182	151	56	21	12							124	48	11	5	27	10	15	12			112	41	2	0				
RP56	432	185	126	82	25	10							98	68	17	7	43	27	38	11			88	38	15	16				
TPT1	439	384	171	149	37	18							146	163	15	12	28	40	31	22			190	68	3	16				
VAB	782	175	347	88	35	9							238	56	34	6	76	26	53	10			248	83	23	18				
ALB	184	58	94	80	9	2							28	29	10	4	68	21	9	4			28	28	49	27				
FTL	678	325	294	159	49	18							118	54	21	21	84	116	38	12			87	33	80	103				
TM5B4X	252	129	146	36	23	3							78	14	18	7	74	88	17	2			52	7	37	53				
ALDOA	339	148	118	73	22	7							81	44	11	5	38	28	21	7			70	38	14	22				
ATPSAT	238	88	136	53	9	0							103	60	3	0	6	1	28	5			120	46	0	0				
CALB2	374	15	106	46	25	2							29	8	18	5	45	40	21	1			53	5	19	39				
EDNA	151	27	63	20	9	3							33	18	4	0	19	1	13	2			88	18	0	0				
TF11	379	114	148	26	23	7							82	10	14	3	47	11	33	0			85	10	24	5				
All genes	7044	2385	3809	1404	448	172							231	103	269	117	723	948	488	183			3037	882	251	323				

IC-MP(N+C)

F1 F2
3.15 3.27

IC-MP(N+C)

F1 F2
6.99 2.84

IC-MP(N+C)LBE

F1 F2
6.99 2.84

F1 F2
3.15 3.27

F2 cell lines removed

Figure 6e

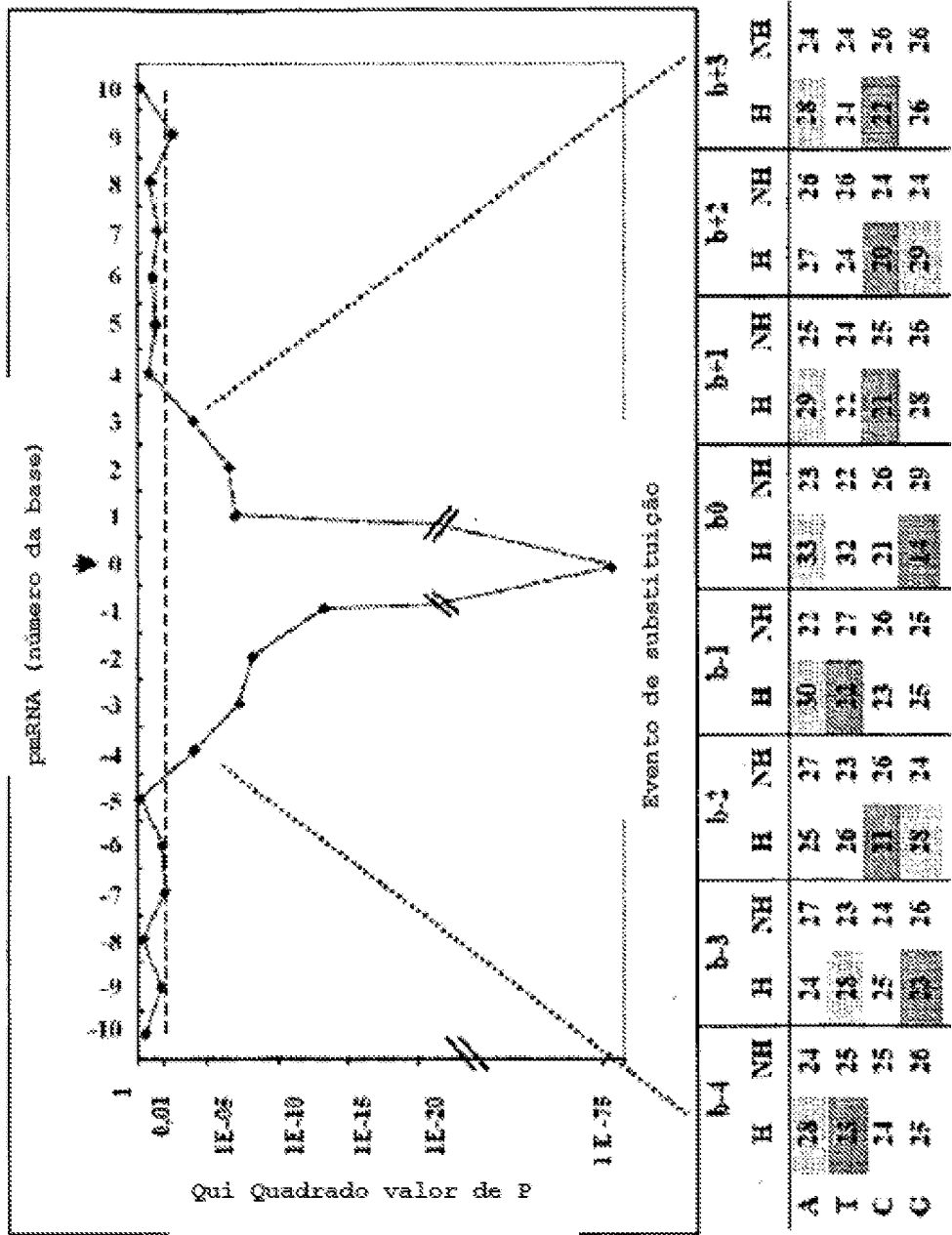


Figura 7a

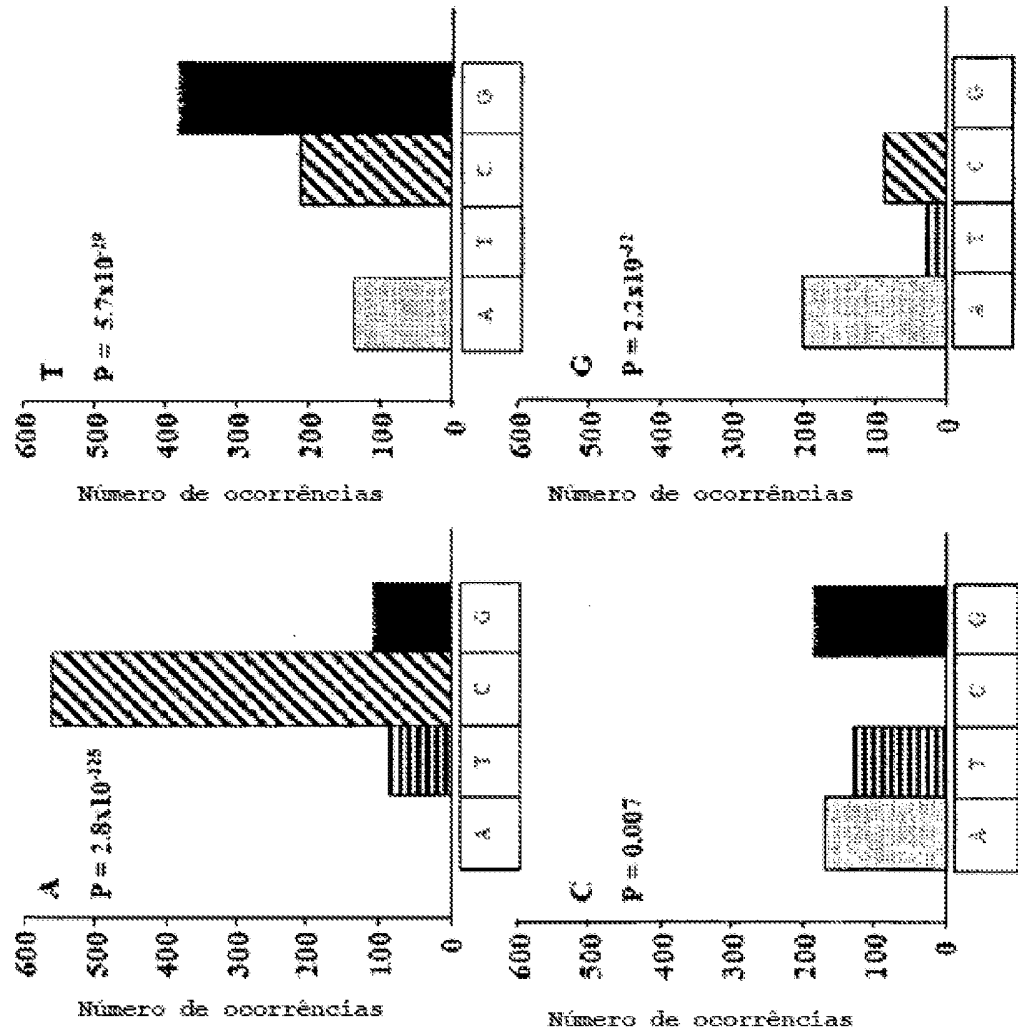


Figura 7b

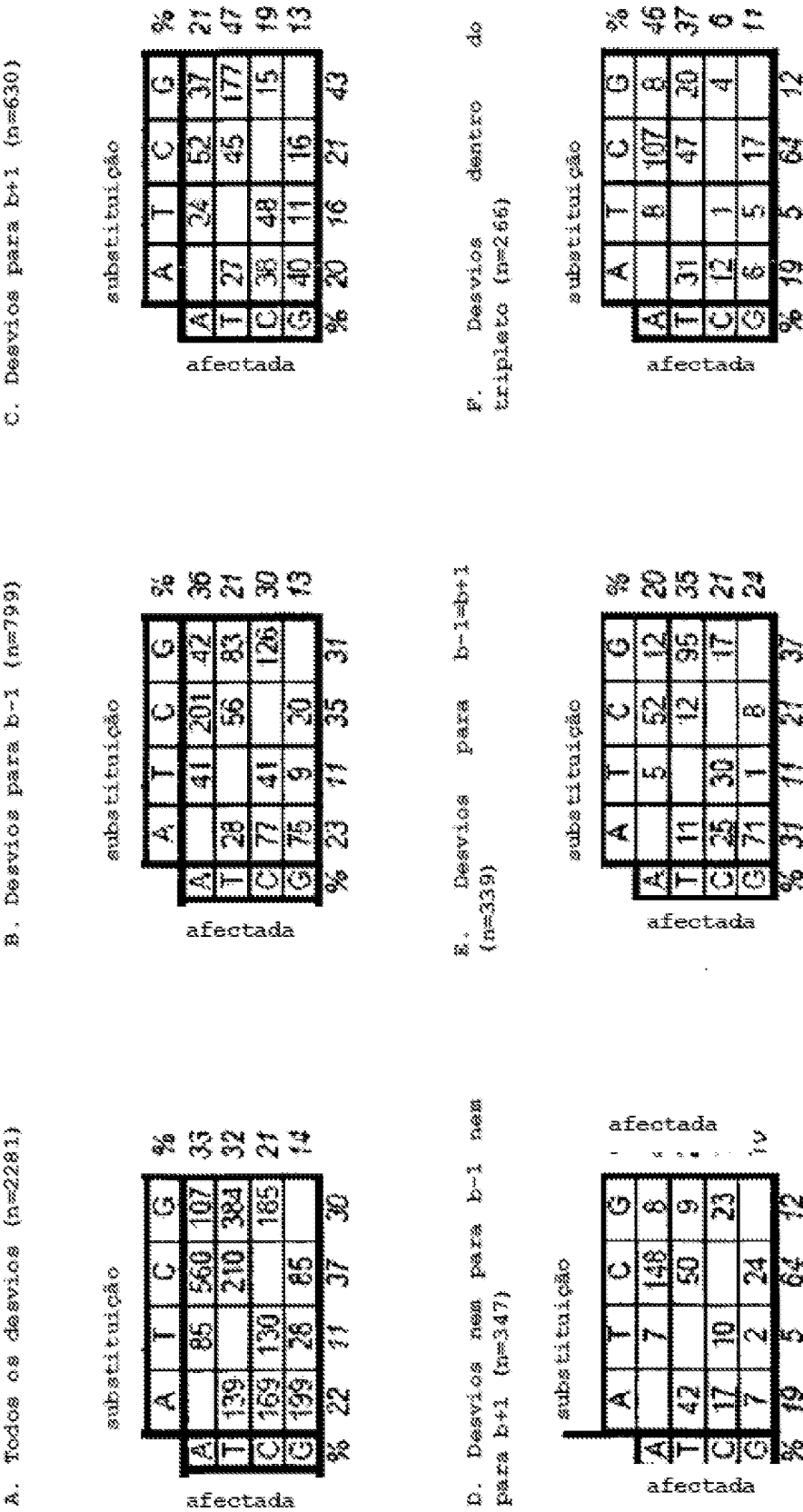


Figura 7c

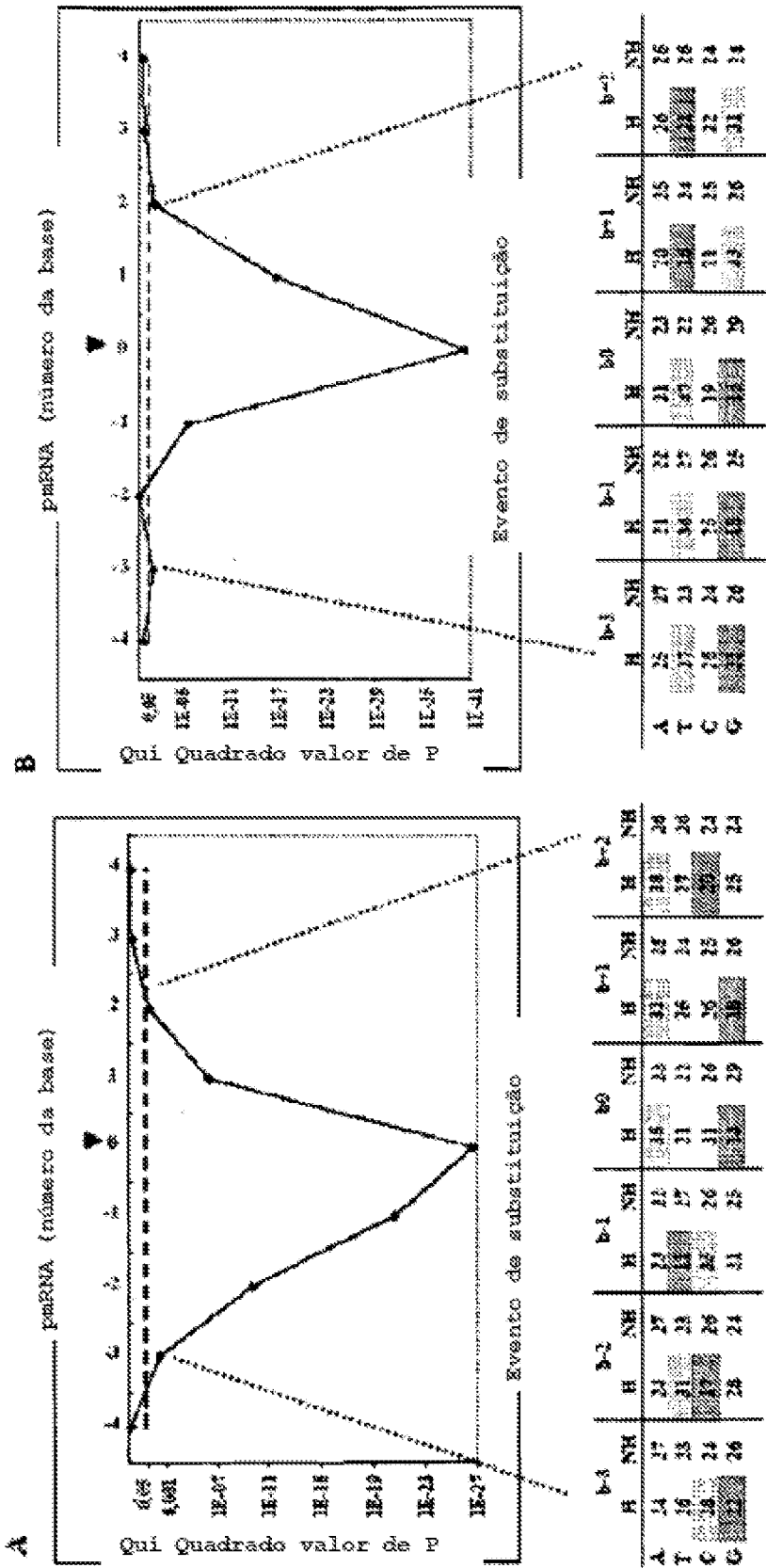
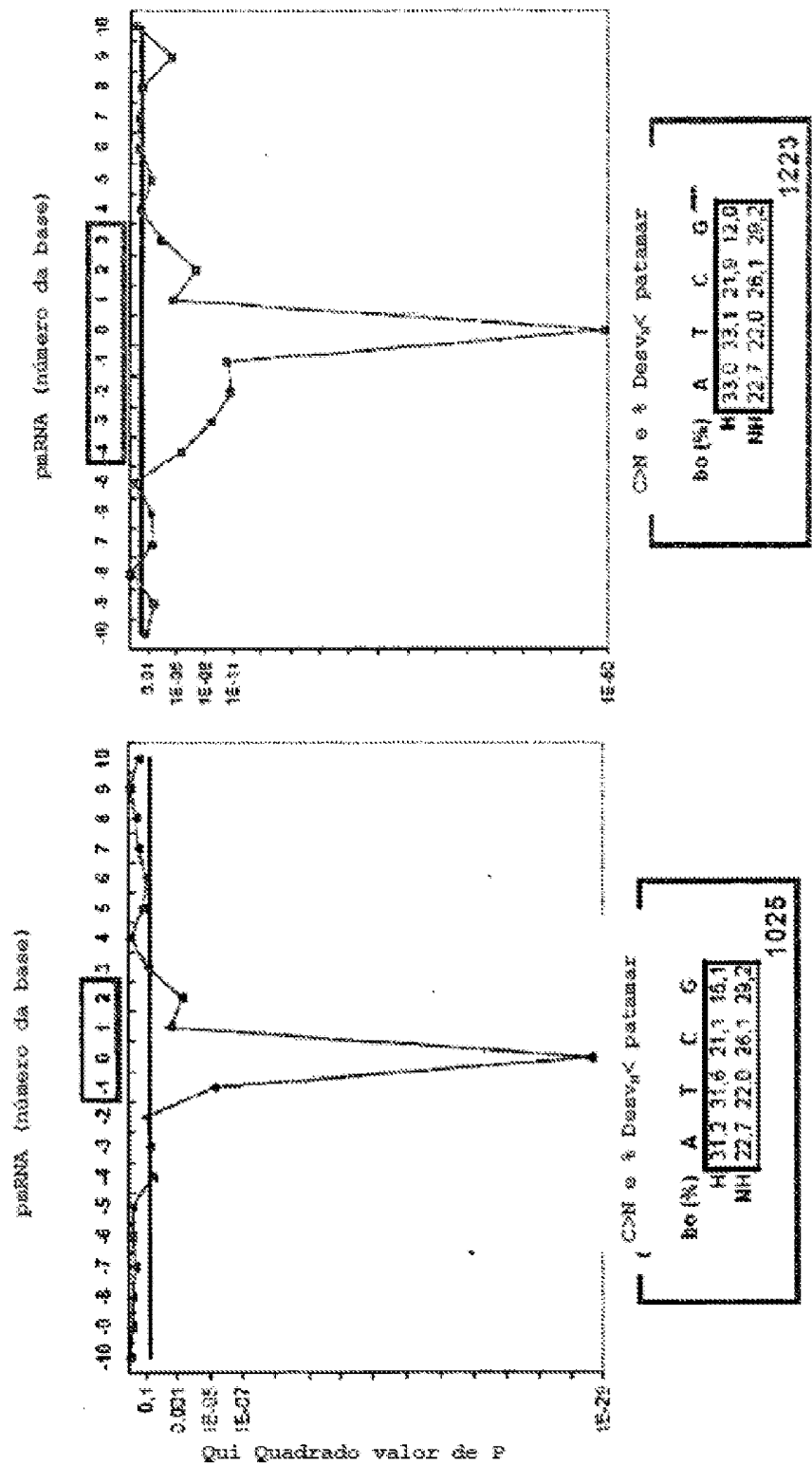


Figura 7d



H: Heteroquêneo (C>N) NH: Não heteroquêneo

Figura 7e

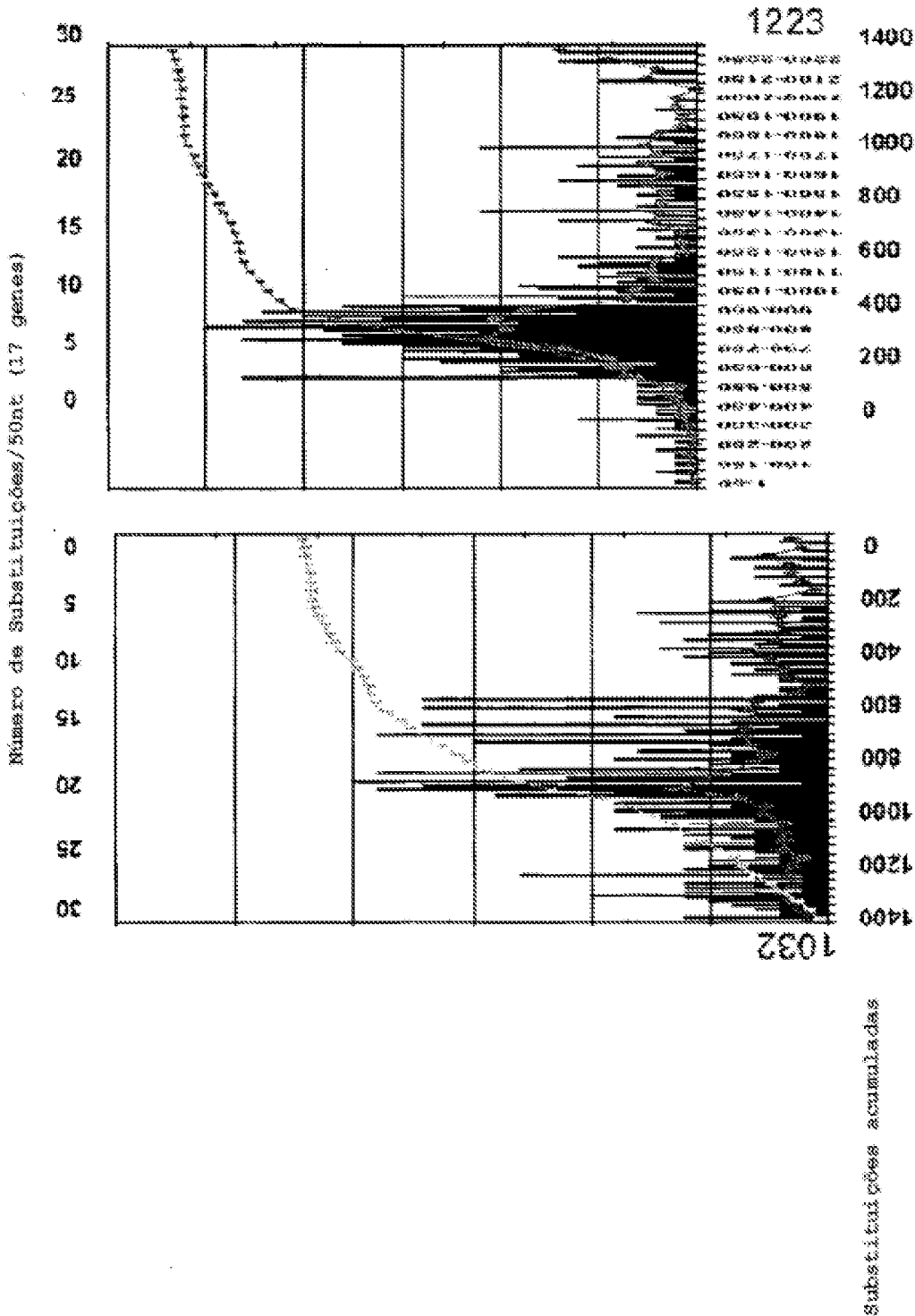
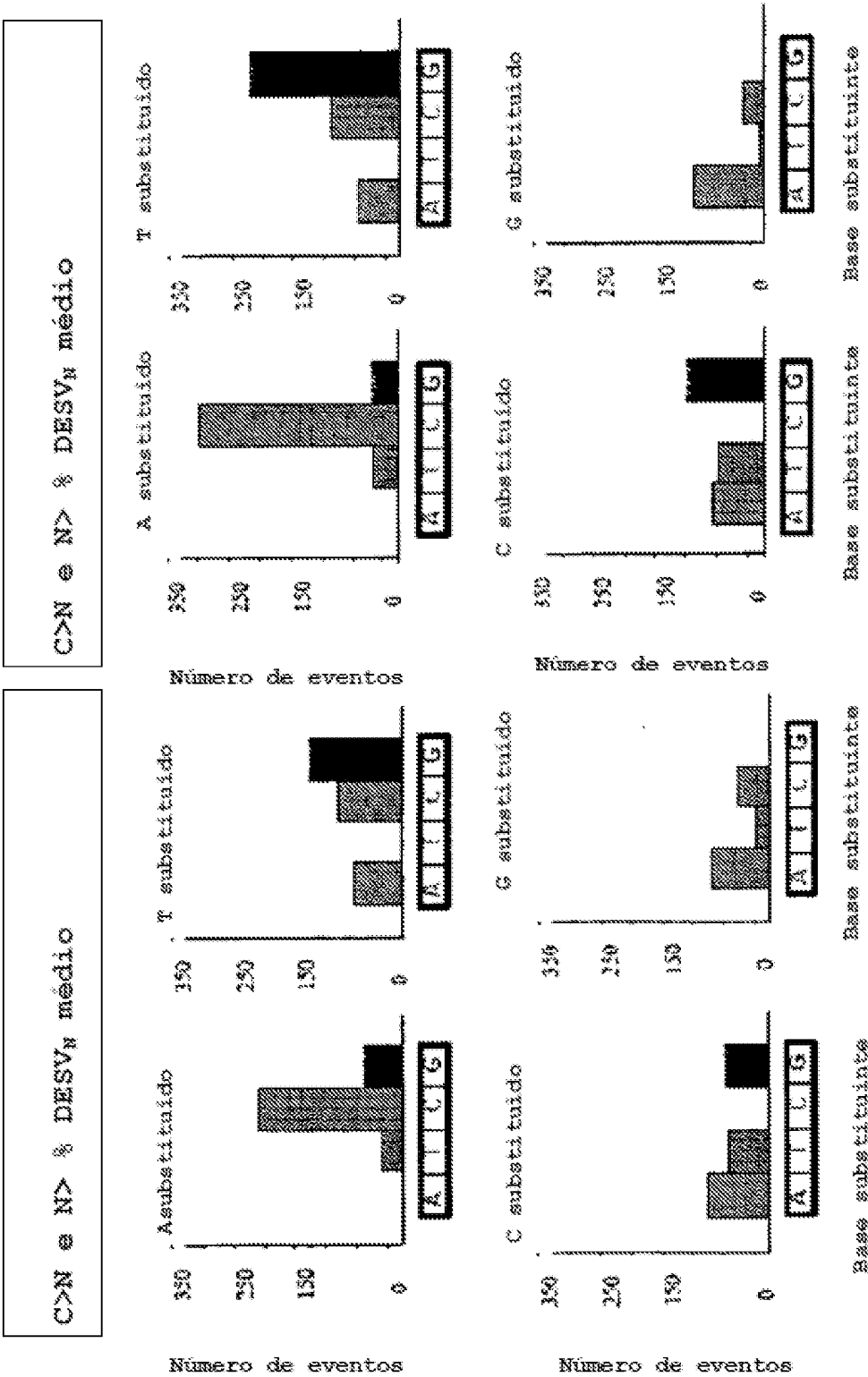


Figura 7f



Figura

-PREF. MAR. 1987 BASES, 30E17CB32

protein variant, 1847 bases, 3387CB32

Figure 8a

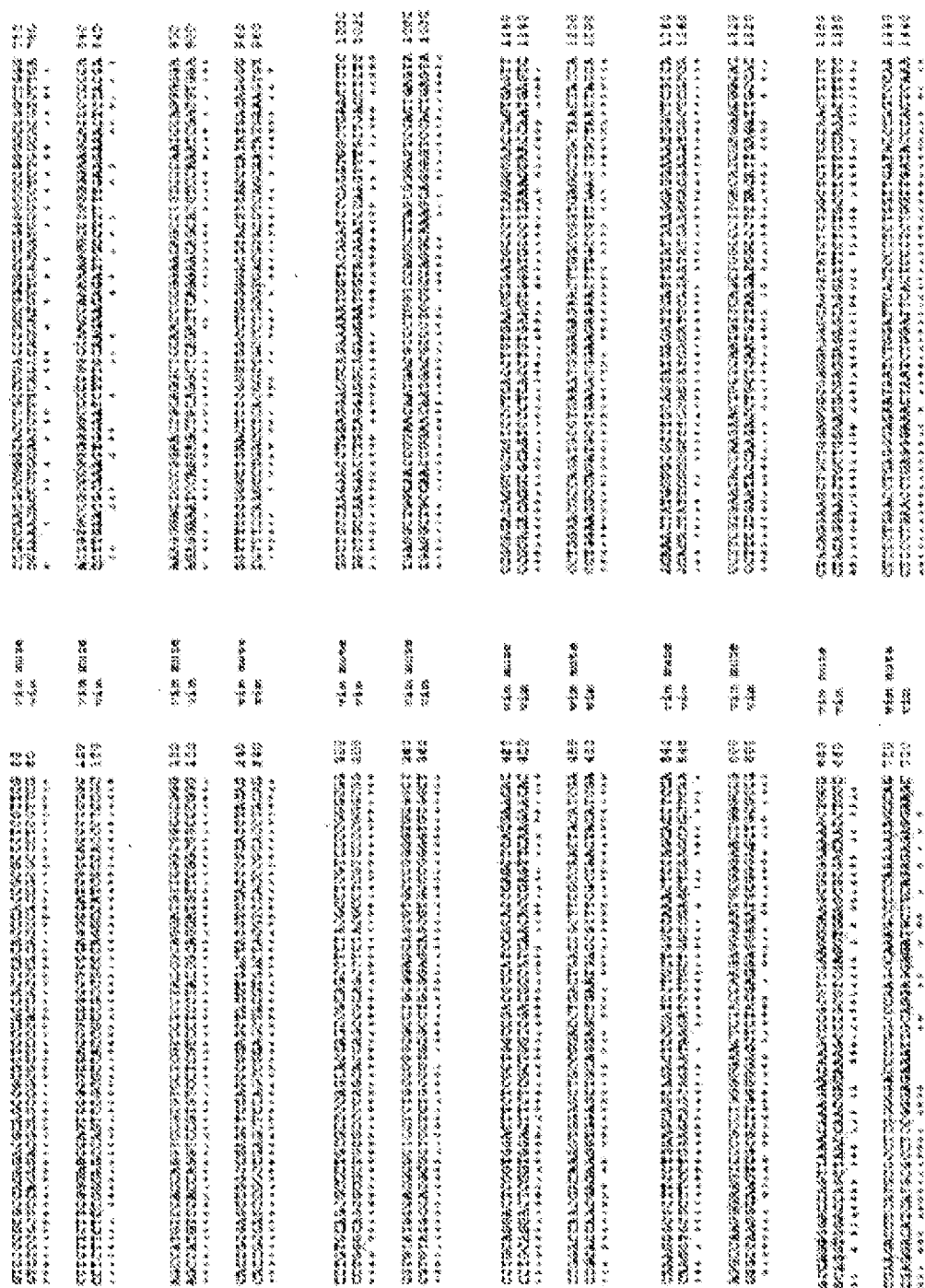


Figura 8b

Estado do impacto codificante das substituições (cancro > normal) em 17 genes		
Número de substituições C>N substituições	2281	
Número de substituições C>N em CDS	1548	
Impacto codificante %	67,9%	
Estudo da região CDS	Número de substituições	%
Substituições silenciadoras	369	23,8%
Impacto codificante (modificação de AA mas da mesma família)	383	26,4%
Impacto codificante (modificação da família de AA)	744	49,1%
Substituições sem sentido	24	1,6%
Modificações de códons de paragem canónicos	18	1,2%
Substituições que originam um codão ATG (MET)	33	2,1%

Figura 9

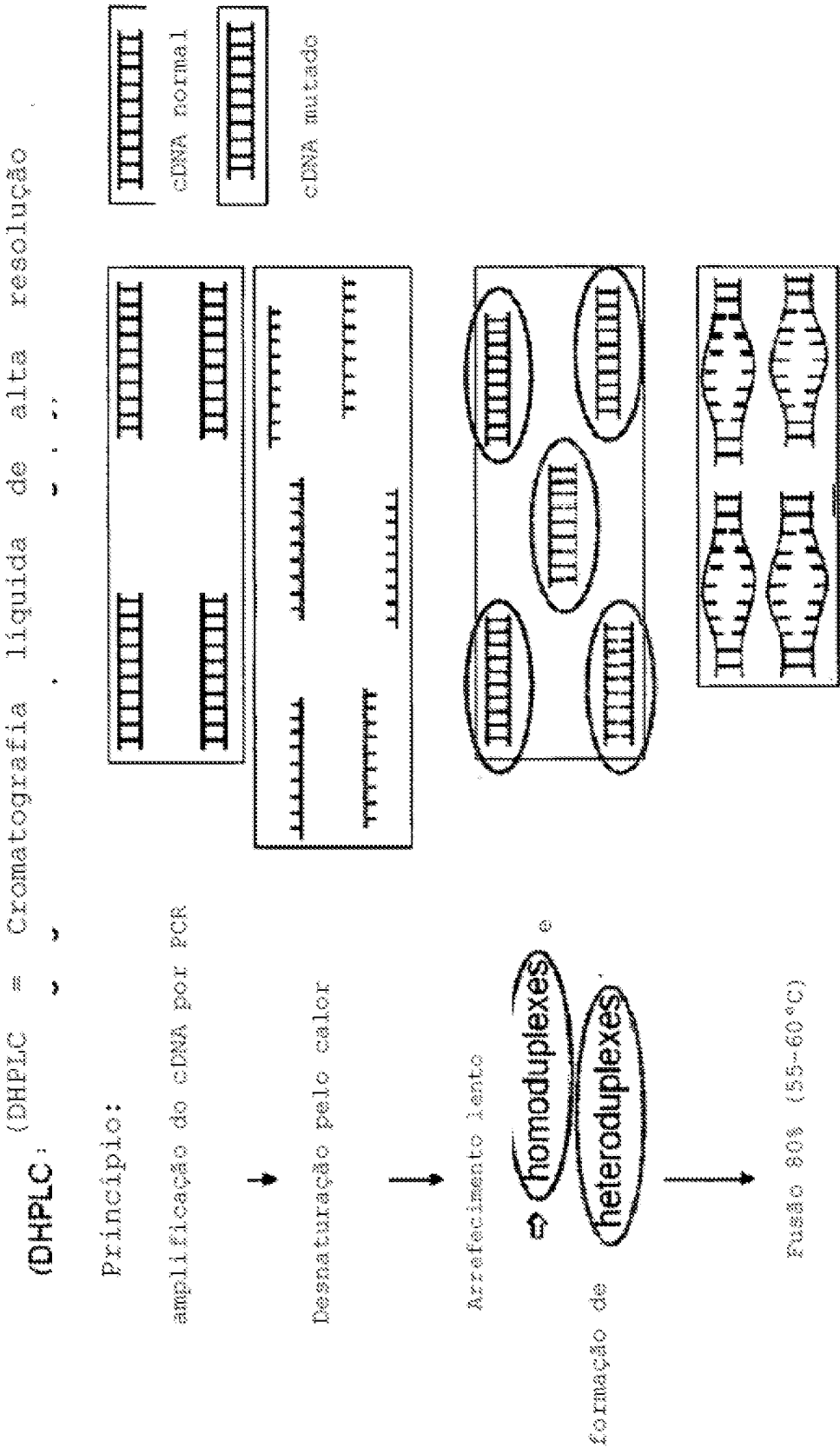


Figura 11a

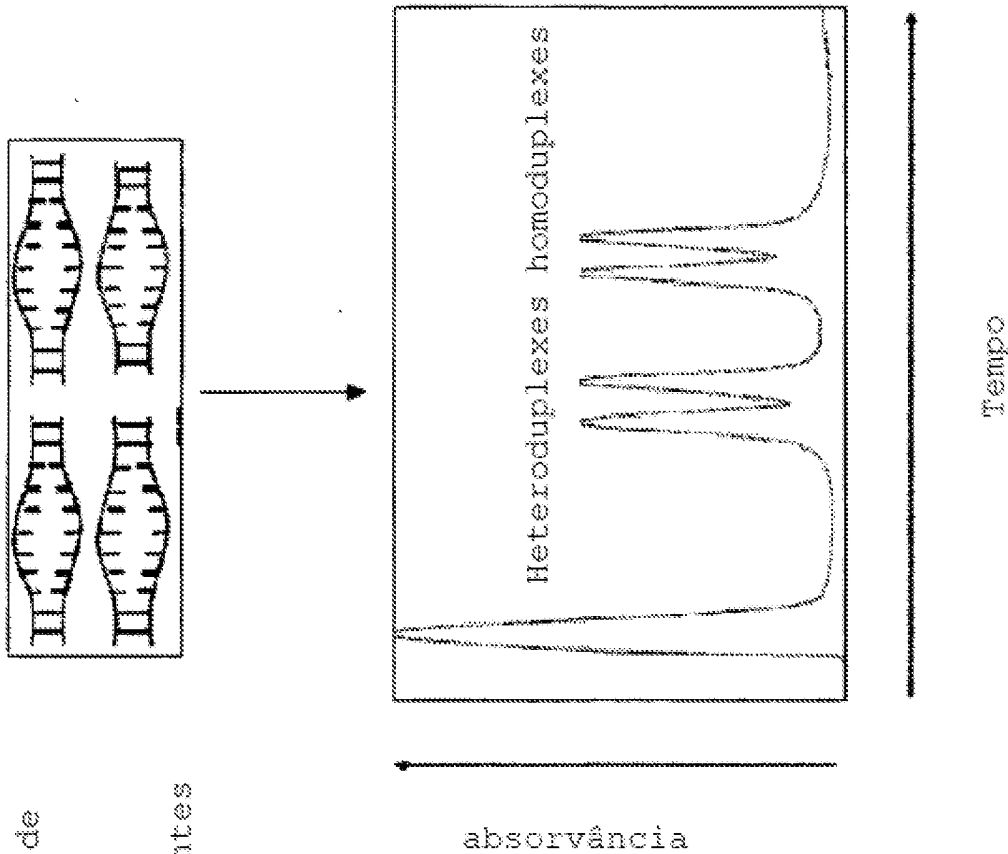


Figura 11b

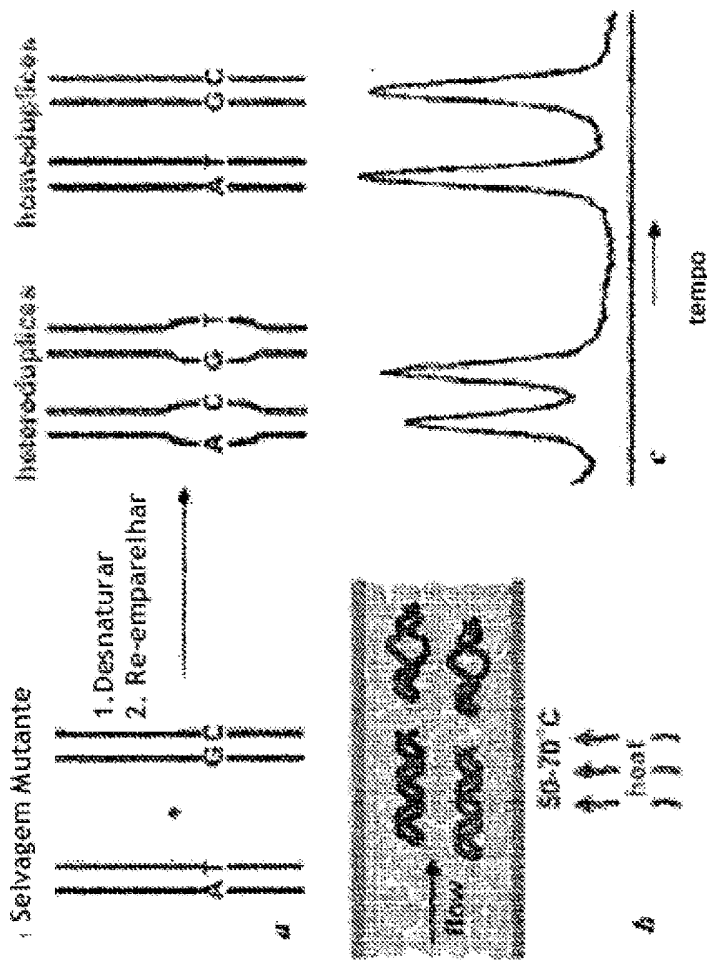


Figure 2. Principle of mutation detection by DHPLC. *a*, DHPLC. Typically two channels, sources as a mixture of PCR products denatured at 85°C for 3 min. and re-annealed over 30 min by gradual cooling from 65 to 55°C prior to analysis. In the presence of mismatch, the only are the original homoduplexes formed, but also the source and indicate strands of either homoduplex form heteroduplexes; *b*, heteroduplexes denature more extensively at the analysis temperature (ranges from 50 to 70°C) and are eluted earlier than the homoduplexes in the IEVA Sep column; *c*, Corresponding chromatographic pattern shows four different peaks belonging to four different species of DNA

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

Figura 11c

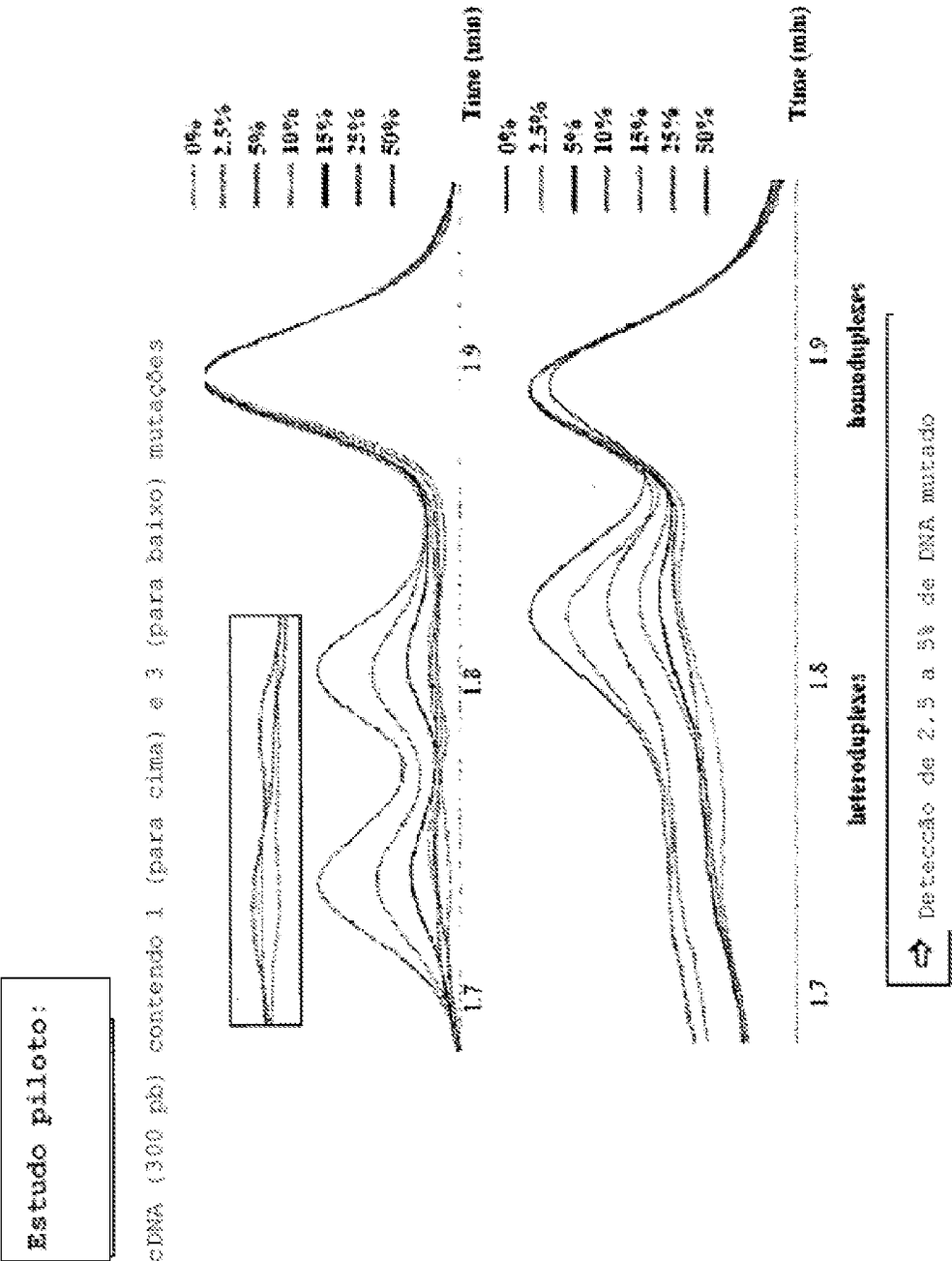


Figura 11d



Figura 12a


```

5' GATCAATTAAAGACCTGTTG 3'      218-318
5' TTTTATTCATAAATAAAGAATGGC 3'    597-614
Amplification product length : 304

GC% : 42.0%
Suggested hybridization temperature (TL) : 51.5
Primer strand + :
Primer strand - :
Primer :      5' GCATCAATTAAAGACCTGTTG 3'
              |||||
Target : 312 : 5' GCATCAATTAAAGACCTGTTG 3'
score : 151

TM : 70.3 / 42.3 / 50,
Primer :      3' GCACAATTTTTTAATCATGAATAAA
          |||||
Target : 587 : 5' GCACAATTTTTTAATCATGAATAAA
          |||||
score : 155

TM : 64.4 / 42.3 / 50, 5

```

doi.org/10.1101/2019.02.11.092110 | Home | rapamycin | beta-4, X-linked | T3SH4Y | mRNA

[illegible]

Figura 12c

Símbolo	Nome	Expaty		Plasma Proteome database	
		protein		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	AAH39236	BC039236	
APCS	Componente amilóide P	P02743	NP_001630	NM_001639	
APOA1	Apolipoproteína A-I	P02647	NP_000030	NM_000039	
APOA2	Apolipoproteína A-II	P03652	NP_001634	NM_001643	
APOC1	Apolipoproteína C-I	P02654	NP_001636	NM_001645	
APOC2	Apolipoproteína C-II	P02655	NP_000493	NM_000474	
APOC3	Apolipoproteína D	P02656	NP_000031	NM_000040	
APOC4	Apolipoproteína E	P03690	NP_001638	NM_001647	
AZGP1	2n-alfa2-glicoproteína	P02649	NP_000032	NM_000041	
B2M	Beta-2-microglobulina	Q5XKQ4	CAA42438	X59766	
C1S	Componente do complemento 1,	P61789	NP_004039	NM_004048	
subcomponente 3		P08871	NP_063850	NM_201442	
C3	Componente do complemento 3	Q6LDJ0	NP_000055	NM_000064	
C7	Componente do complemento 7	P10643	NP_000576	NM_000587	
		P12830	EAA88957	AB025106	
		P00751	NP_001701	NM_001710	
		P36222	NP_001267	NM_001276	
CH3L1	chitinase 3-like 1				

F
Figura 13a

Símbolo	Nome	Plasma Proteome database			
		Exposy	proteins	miRNA	
		SP accession	PPD accession	PPD accession	
CLEC3B	Domínio 3 de lectina tipo C, membro B	P05452	NP_003265	NM_003279	
CLU	Clusterina	P10909	NP_001822	NM_001831	
CP	Ceruloplasmina	P03450	NP_000387	NM_000395	
CRP	Proteína C reactiva	P02741	AAH20766	BC020766	
EDN1	Endotelina 1	P05305	NP_001946	NM_001955	
EGFR	Receptor do factor de crescimento	P05533	NP_963441	NM_201284	
EGF	Factor de coagulação XIII, polipeptido A1	P03466	NP_000120	NM_000125	
F13B	Factor de crescimento XIII, polipeptido B	P05160	NP_001935	NM_001934	
FGA	Fibrinogénio cadeia alfa	P02671	NP_000499	NM_000503	
FGB	Fibrinogénio cadeia beta	Q02265	NP_005132	NM_005141	
FGC	Fibrinogénio cadeia gama	P02679	NP_003536	NM_021870	
GPI	Isomerase de glucose fosfato	P05744	NP_000166	NM_000175	
GSTP1	Glutathione S-transferase pi	P08211	NP_000943	NM_000952	
HPLEP	cdNA FLJ42936 fi, altamente semelhante a	Q00341	BACN7153	AK127833	
HPLEP		P00738	NP_005134	NM_005143	
HPX	hemopexin	P02750	NP_000604	NM_000613	
HRG	histidine-rich glycoprotein	P04196	NP_000403	NM_000412	
IGFBP3	insulin-like growth factor binding protein 3	P17936	NP_000569	NM_000583	
IGJ	immunoglobulin J polypeptide	P01591	NP_953297	NM_144646	

Figura 13b

Símbolo	Nome	Plasma Proteome database		
		proteins		miRNA
		SP accession	PPD accession	PPD accession
INH1	Inibina alfa	P05111	NP_002162	NM_002191
KLK11	Peptidase 11 relacionada com calicreina	Q8UBX7	NP_653196	NM_142917
LDHA	Lactato desidrogenase A	P00308	NP_005557	NM_005586
LGALS3BP	Lectina, ligação a galactósido, solúvel, proteína de	Q08390	NP_005558	NM_005557
ligação 3		P02760	NP_433704	NM_023972
LRG1	Alfa-2-Glicoproteína 1 rica em leucinas	P02763	NP_000598	NM_000607
ORM1	Orosomucóide I de Homo sapiens	P14618	NP_072270	NM_182470
PKM2	Cinase de piruvato, músculo	P00747	NP_000292	NM_000301
PLG	Plasminogénio	P27165	NP_000437	NM_000446
PONI	Paraoxonase I	P04070	NP_000303	NM_000312
PROC	Proteína C	P20742	NP_002855	NM_002864
PZP	Proteína da zona de gravidez	P02762	P02763	BC030633
RBP4	Proteína de ligação ao retinol 4	P02735	NP_000322	NM_000331
SAA1	Amlóide sérica A1	P01009	NP_000366	NM_000265
SERPINA1	Inibidor da peptidase serpina, grupo A	P01011	NP_001076	NM_001085
SERPINA3	Inibidor da peptidase serpina, grupo A, membro 3	P00164	NP_002066	NM_002074
SERPINA5	Inibidor da peptidase serpina, grupo B, membro 1	P02727	NP_001054	NM_001063
TF	Transferrina	P02726	NP_000225	NM_000234
TFRC	Receptor da transferrina	P01137	NP_000651	NM_000660
membro 1				

Figura 13c

Exposy		Base de dados Plasma Proteome	
proteinas		mRNA	
SP accession	PPD accession	PPD accession	PPD accession
P01033	NP_003245	NP_003245	NM_003254
P02766	NP_000362	NP_000362	NM_000371
P04004	NP_000529	NP_000529	NM_000638

Símbolo Nome

TIMP I TIMP inibidor da metalopeptidase 1

TTR Transtirretina

VTN vitronectina

Figura 13d

Seleção de 3 proteínas do plasma:

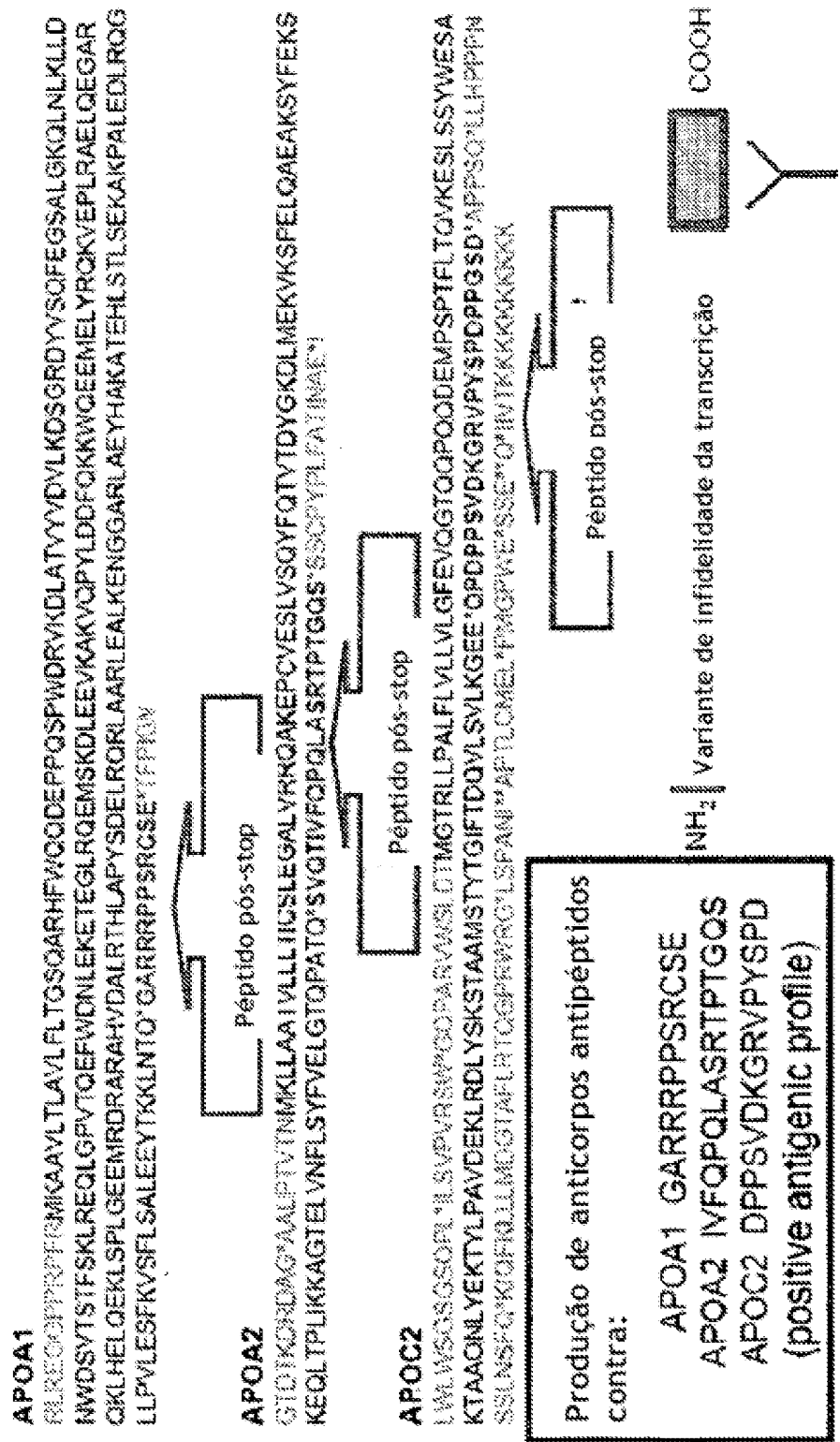


Figura 14

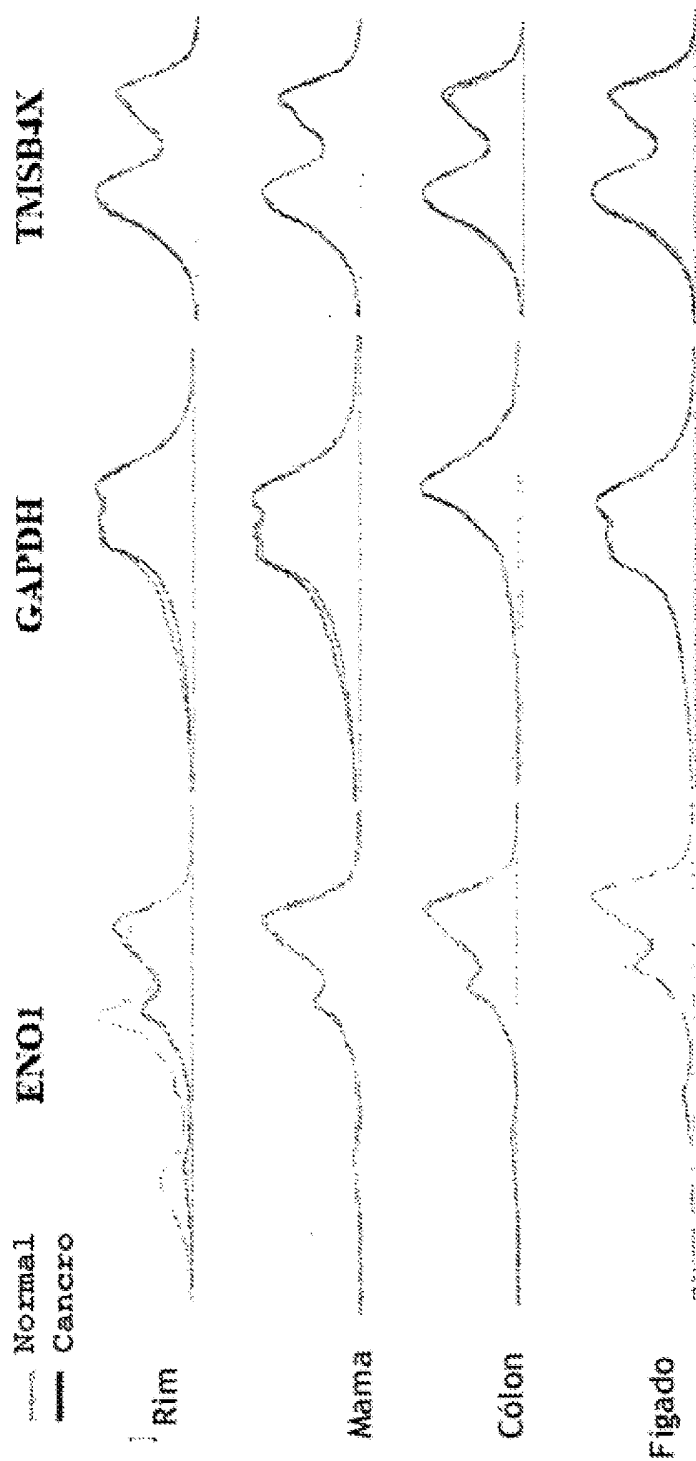
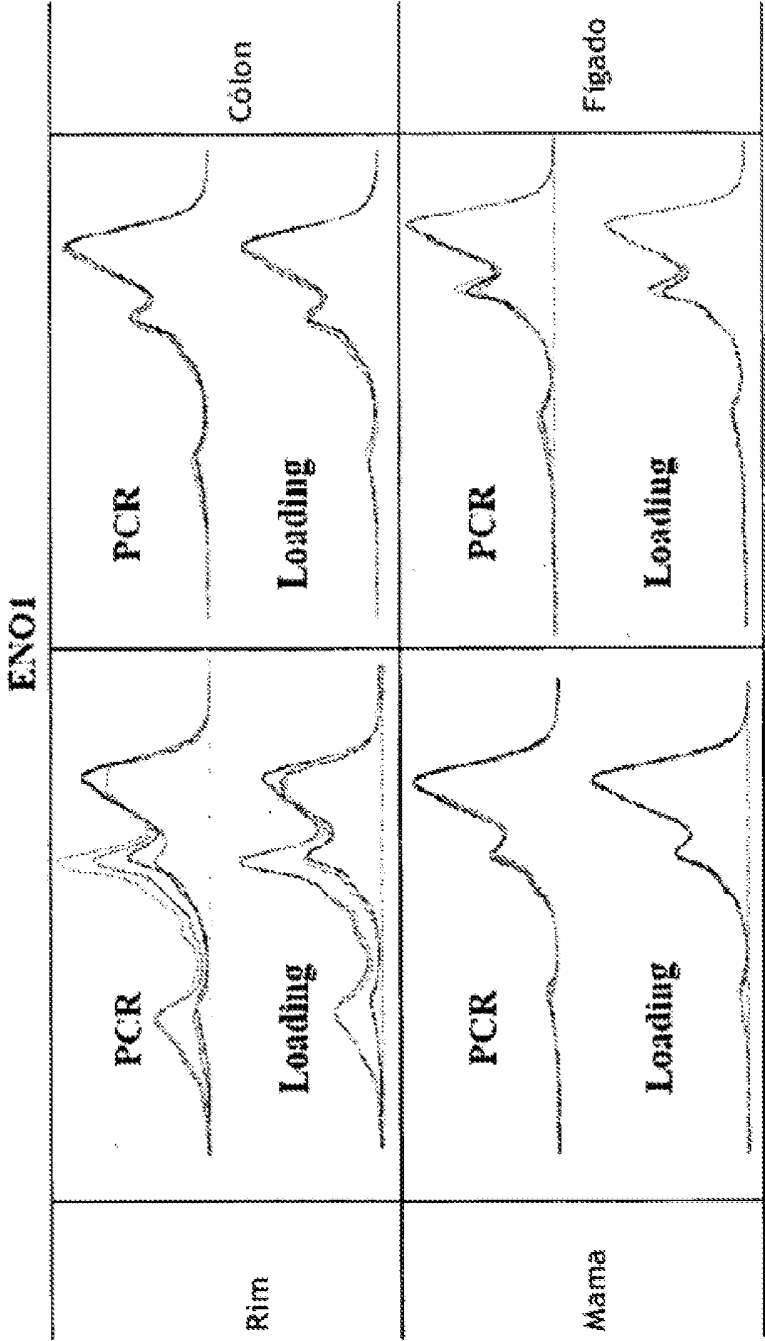


Figure 15a

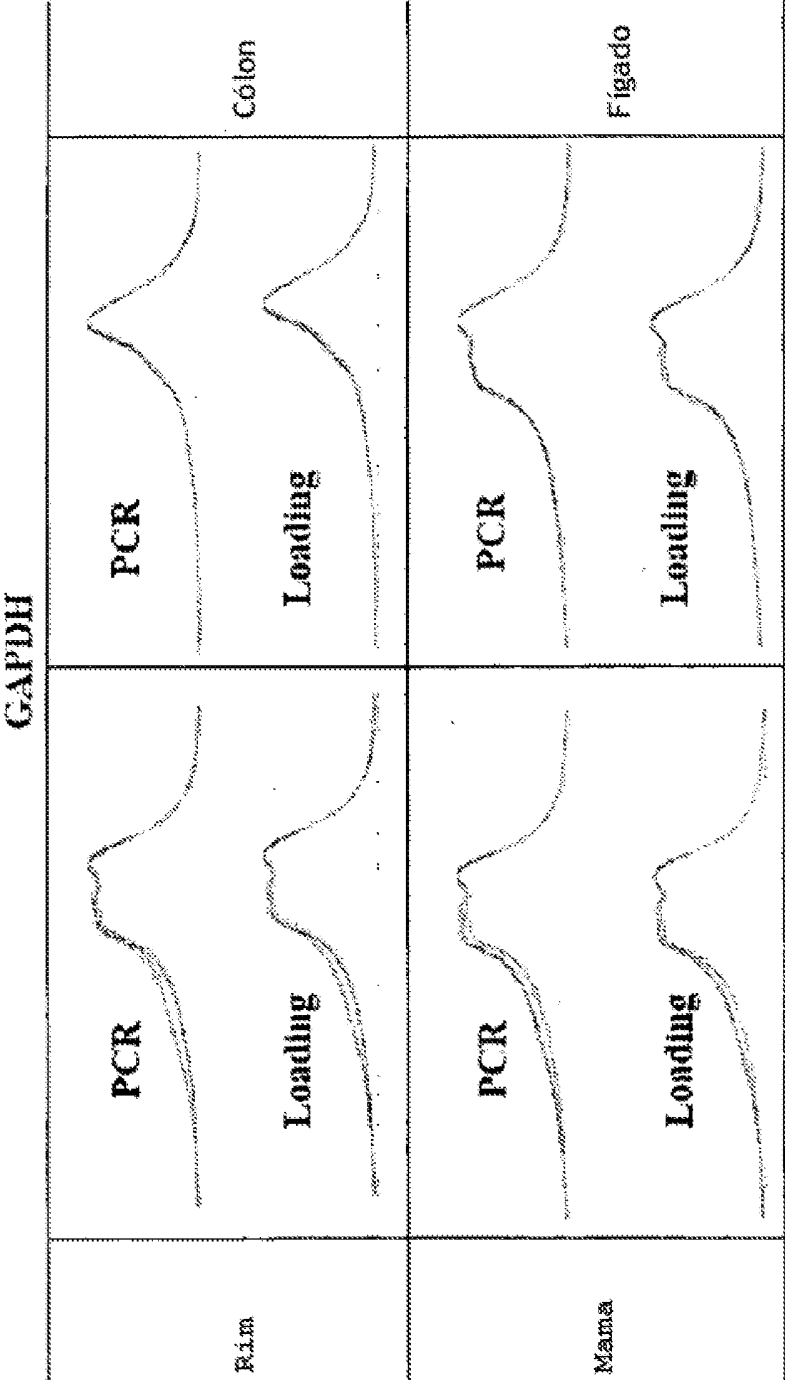
Figura 15a



Different PCR and Loading DHPLC pro
Two different PCR products were prepared versus tissue normal adjacent for the gene ENO1
DHPLC profile is represented in Green ; Dois produtos de PCR diferentes foram preparados a partir de 4 tecidos e cada produto
duplicate is represented in : Red for Non de PCR foi injectado duas vezes no sistema de DHPLC. O perfil normal de DHPLC está
Normal and Light Blue for Carcinomas. Car

Figure 15b

Figura 15b



Different PCR and Loading DHPLC profile Differentes perfis de DHPLC para PCR e aplicação ("loading") de tecidos
Two different PCR products were prepared tumourais 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. Curva, para o sistema de DHPLC para PCR e aplicação ("loading") de tecidos

Figura 15c
Figura 15c

Sequência A: CCGCTCAGCAT

Sequência B: CCGCAGCGCAT

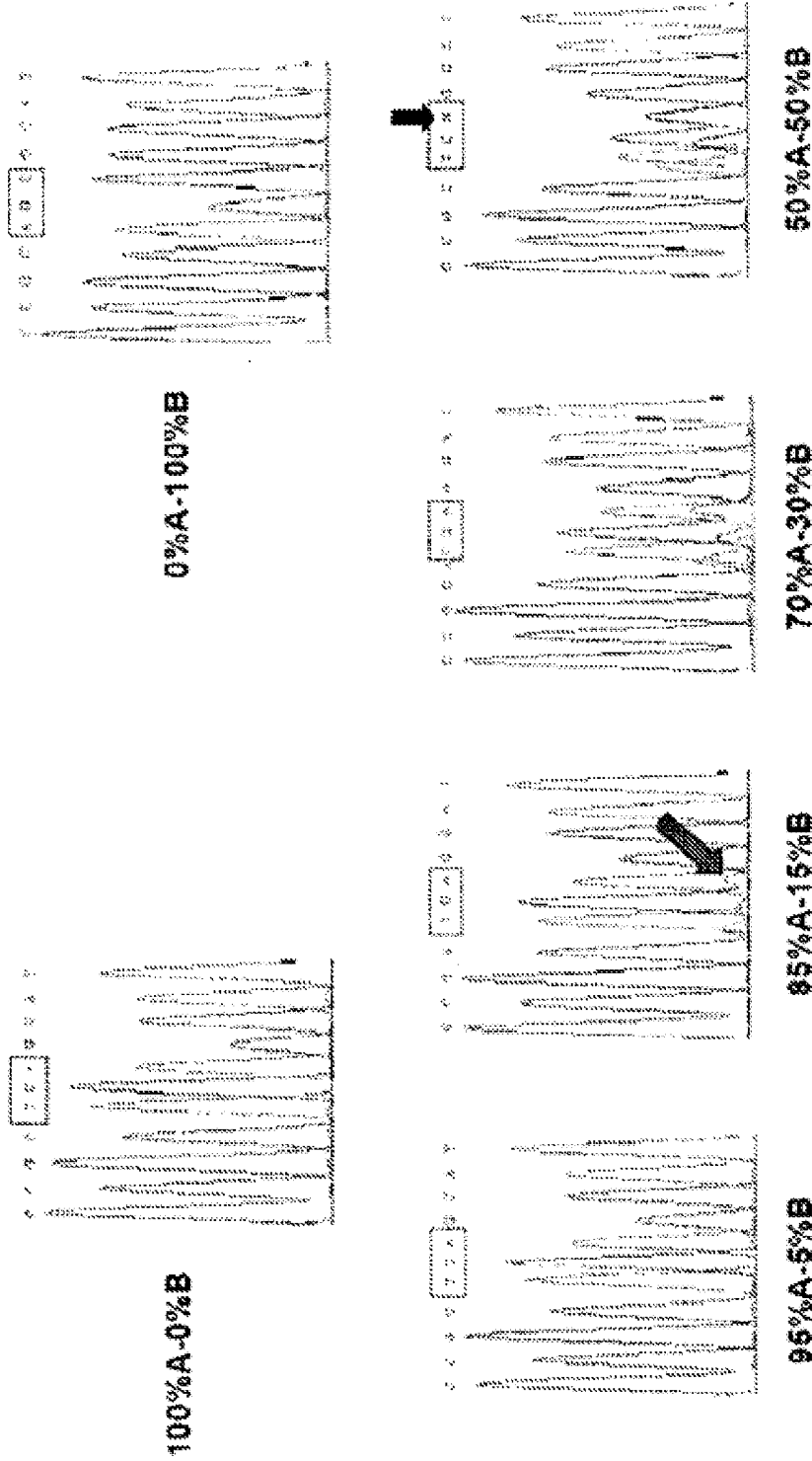


Figura 16

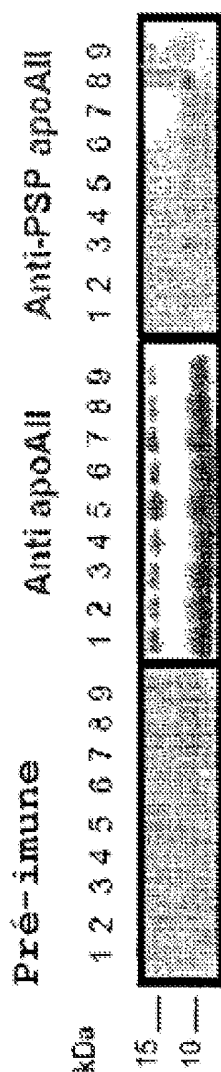


Figura 17a

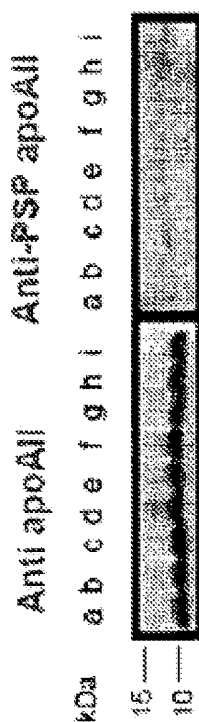
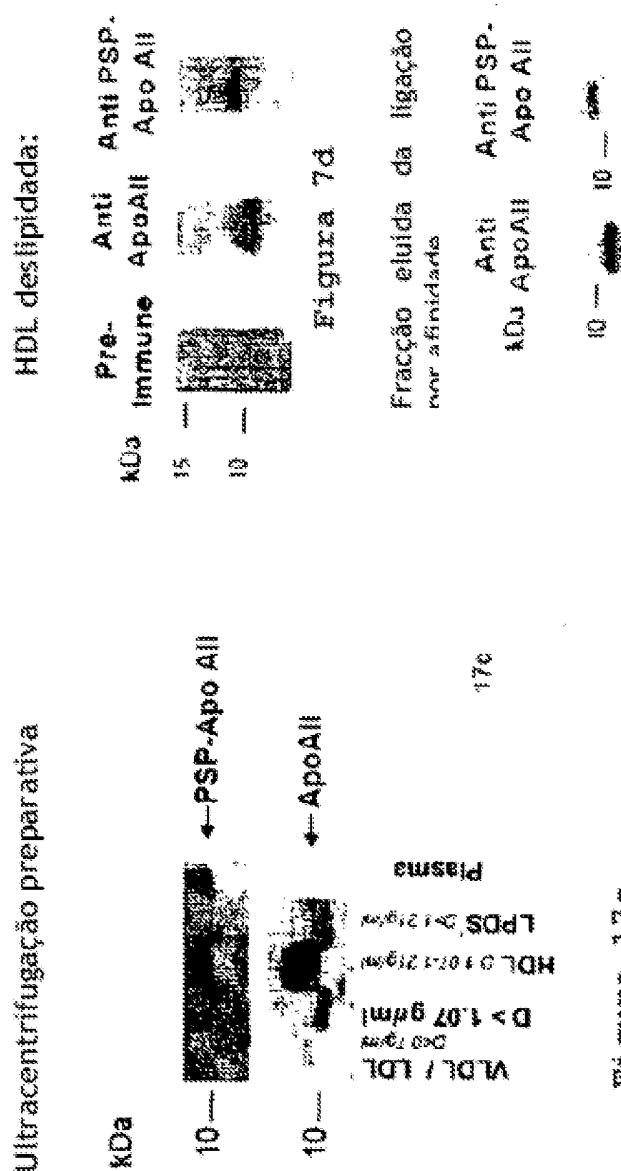


Figura 17b

Plasma from patients with prostate cancer was applied to polyacrylamide gels under reducing conditions at 4°C for 1 h, room temperature for 1 h, and then commercially available goat anti-human IgG (1:100, right panel). After 1 h incubation, appropriate conjugated second antibody was applied to each assay.



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 metastática. 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 lipoproteína ($d > 1,21$ 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 cloroformio-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

e. HDL (d 1,07-1,21 g/ml) was prepared by ultracentrifugation differential of a set of plasmas from 10 cancer patients in the metastatic phase. The different fractions recovered after each step were dialysed against NaCl 0,15M, pH 7,4 containing 0,01% EDTA. The Western transfers were performed on the indicated fractions using the same dilutions as indicated with commercial anti-apoAII (lower panel) or anti-PSP (upper panel) using the same dilutions indicated in Figures 17a and 17b. The gels were overexposed, which explains the distortion and the wide bands. The Western transfers were also performed on the original plasma or on a volume corresponding of the lipoprotein deficient serum ($d > 1,21$ g/ml). d. Purified HDL was dialysed against 0,01% EDTA, pH 7,4, frozen at -80°C and then lyophilized. The lipid removal of lyophilized HDL was then performed using ice-cold chloroform-methanol (2:1, v/v, 5 ml per 20 mg of HDL protein). A second lipid removal of lyophilized HDL was then performed using diethyl ether, followed by

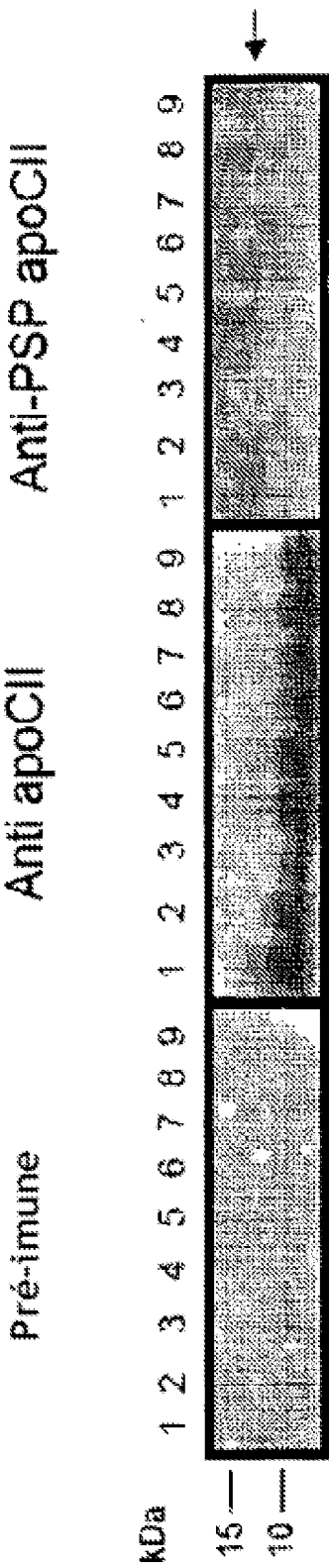


Figura 17f

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

Name	mRNA id	5'UTR	Decision
AHEG	NM_001323.1	APHOSEEVNHRHSHHFVGVANWVNGELUCWPRKCHPCTLSLGLPYPNFTGATNWWWVAFDZRH	YES
ALB	B0023235	RSLAVESVFLFACVCAITLSH-HPL	YES
APCS	NM_001338.2	GLDSTRALQEMTV	YES
APCA1	NM_000328.1	QARPPPPRQSE	YES
APCA2	NM_001341.1	SVQTNFQPLAERTFTQDS	YES
APC2	NM_000333.3	QPPPPSVQNGRPRVDFPPSSD	YES
APC3	NM_000340.1	DLNTPFPAYPSCGLLGGALGGCPRLKRDGLSALLPMLPSPPPGALAEQ	YES
APOD	NM_001347.2	PQSTGRLHPLKNTSLSLPTPPPPRNDQPRHNDQES	YES
APCE	NM_000341.2	SPKPPMRSHKATLCPLPPRSQRGTLSPFPSSGNGGP	YES
A20P1	X55762.1	SARVGGNVGDTG	YES
C4B1.1	NM_001372.3	PSVLTARPSPRMRPPLAFAGHEPDRPC	YES
CLU	NM_001331.2	QVQNAFPTGASESSGCELOPPRESARNGVTRPQPPSPQNRKAPPSGCTLTGSAHGVNRHAPACR	YES
HPS	NM_000433.2	KVPTLRMMNTLN	YES
ISFPP3	NM_000328.4	TPAARLVGSGMAPYFAQTAFQNTSSWLOPFFPLPVH	YES
IN4A	NM_000310.2	QVGGVLLNPMAGGCHAPTEHNEERQSVNGDGSBSLSLLCLWATLFTALLSQ	YES
KLK14	NM_146347.5	TGPTHEPSPFSSTWCLVPVNSVANSR	YES
PMAC	NM_152470.1	WTPPELLOPLSHPLPRAPLQOORL	YES
PLG	NM_000331.3	LDGPPGCMATKLENGTAVSR	YES
SERPP1A3	NM_001325.4	SLPSSGQALSHGLQAGAGCLGLWADPPQAPAGGAGWSPVQSLIEFVZDELQSSHARPPPTLQSGGMAAS	YES
TF	NM_001323.2	NLRJRAATNVKMGTDGHEFALVSLQVVCANPVCLHSSVALPQVLRK	YES
TGFB1	NM_000340.3	QMAPPPAPAPAPPPAPALAHGAAGTGTAPSPQAPLKLERGIRHSLGACLOSPLTFPHSHSLPLCLLL PVCTPLPQSGRGGTSGEHCYS	YES
TTR	NM_040371.3	STSPFQVQCEGNDPM	YES

Figure 18a

>P02765JAH9G Alpha-2-HS-glycoprotein
MKSLVLLQLAQLWGHSAFHGPGGLYRPNQDDPETEEAALVADYINQNLPWGYKHTLNGIDENKVPQPSGELFEIEDILETTCHVLDPTP
VARCSVRLKEHAVEGDCDFOLLKLDGKFSVAYAKDSSPDSAEVRKVCQDCPLAPINDTRVYHAAKAAALAFNAONNGSNFOLIEISRAQL
VPLPSTVTEFTVSGTDCYAKEATEAAKCNLLAEKQYGFCAKATLSEKLGAEVAVCTVFCQPVTSQPSQPEGANEAVPTPWVDPDAFPPSPFLG
APGLPAGSPDDSHVLLAAPPGHCLHRAHYDLRHTFMGVVSLGSPSSGEVSHPRKTRTVQPSVGAAGAPVWPPQGRIRHFXVARGIGRDEEV
VARRKHSSEFVCAWAWVMSLVQVNPQKCHMRSLUSSLSLLPLVPHRTAEKVVVYMFDRRH*

>P02768JAlbumin
KKGWVTFISLLFLFSAYSGVFRFRAHKSEVAHPRFKOLGEENFKALVIAFAGYLOQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGD
KLCVTATLRETYGEMADCCAKGEPRNECFLOHKKDDNPNLPRLYRVEDVMCTAFHONEETFLKYLYVEARRHPYFYAFELLFFAKRYKAAAFTE
CCQAAADKAOLLPKLDELROEGKASSAKORLKCASLOKFGERAFKAWAVARLSQRFPAKFAEVEYSKLVTDLTKVHTECCCHGDLLECAADRQALA
KYICENQDSISSKLKECCENPLEKSHCIAEYENDCMPADLPSLAOFVESKDYKQVYAEAKDVFLOFMELYEARRHPDYSVLLRLAKTYETILE
KCCAAADPHECYAKVDFEEKPLVEEPONLIKONCELFEOUGEYKFNALLVRYTKVPOVSTPLVEVSRNLGKVGSKCKCKHPEAKRNPQCAEDY
LSVVLNOLCVLHEKTPVSDRVTKGCTESLVNRRPQFSALEVDETYWPKFNAETFTFHADICTLSEKRDIKKOTALVELVKHKPKATKEOLKAYM
DDFAAPVEKCKKADDKETQFAEEGKKTCCCKSSQLRLTSHLKASOPTMRIRERK^{QSSKAYSSVLEFMYOKANTLSQKHFL}*

>P02743JAPCS Serum amyloid P-component
MKKPLLWISVLTSLLEAFATDLSGKVVFPRRESYTDHVNLIITPLEKPLQNFILCFRAYSOLSRAYSLSFYNTQGRONELLVYKERVGEYSLYIGRH
KYTSKVIKFPAPVHICVSWESSGIAEPWINGTFLVKGLRQGVYEAQPKVILGCEQDSYGCKFORSQSFYGEIGDLYMWDVSLPPENILSAYQ
GTLPLANILDVQALNYEIRGYVHKPLVWVYVTL^{ESTNALLENEMTY}*

>P02647JAPOA1 Apolipoprotein A-I
MKAAMLTAVLFTGSOARHFVQDDEPQSPWDRVKOLATVYVVLKDSGRDYYSQFEGSALGKQLNLKLLDNWDSVTSTFSKLREQLGPVTO
EFWDNLEKETEGLRQEMS^{KOLEEMVQKVCPLYDDFQKKWQDEEMELYROKVEPLRAELQEGARQKLHELOEKLSPLGEENRDRARAHVDAALRT}
HLAPYSDELRORLAARLEALKENOGARLAEYHAKATEHLSTLSEKAKPALEDLRQGLLPVLESFKVSFLSALEEYTKMLNTQ^{GARRRQPPRRQSE*}

>P02652JAPOA2 Apolipoprotein A-II
MKLLAATVLLLTICLEGALYRROAKEPOVESLYSQVFTVTDYGKOLMEKVKSPELQAEAKSYFEKSKEQLTPLIKKAGTELVNFLSVFVELGTQF
ATQ^{SVQITVWFOQLASRTTQGS*}

>P02655JAPOC2 Apolipoprotein C-II
MGTALLPALFLVLLVGFEDGTQDQDDEMPSPITFLIQVKESLS^{SSYWIESAKTAQNLVEKTYLPAVDEKLRQLYSKSTAA MSTYTIGIFTDQYLS}
VLKQEE^{QFQPTSMKGRQFYSPDPQSL*}

>P02658JAPOC3 Apolipoprotein C-III
MOPRVLLVYALLALLASARASEAEDASLLSFMQGYMKHATKTAKQALSSVQESQVAQQAQGWVTDGFS^{SLKDYWSTVKDKFSEFWELQFEVRP}
TSAAVA^{YDNLTPSPAYPSDELQSDMLQADQONLLKRSLSALLPHLHSPFGMLASQ*}

>P05090JAPOD Apolipoprotein D
MYVMLLLLSALAGLFGAAEGQAFHLGKQNPVQENFDVNKYLGRWYIEKIPITTFENGRCIQANYSLMENGKIKVNLQELRADGTYNQIEGEATP
VNLTEPAKLEVKFSWFMPSAPYWLATQYENYALVYSC^{TCIQLFHVDFAWILARNPLPPE^{TVD}SLKNILTSNNIDVYKGI^{TYDQ}VNCPKLS^{YPCST}}
GRLKALHVTSA^{SLSPTPPPQKQVYRQDQGS*}

normal proteins additional peptide of proteins whose STOP codon is affected

Figura 18b

>P02649IAP0E Apolipoprotein E
MKVLAALLVTFLAGCOAKVEQAVETEPELROQTEWQSGORWELALGRFWDYLRWVQTLSEQVOEELLSSQVIOELRALMDETAKELKAY
KSELEEQLTPVAEETRAPLSKELQAAQALGADMEDVCGRLVQYRGEVGAMLGOSTEELRVRLASHLRKLRLLRDADOLQKRLAVYQAGAR
EGAERGLSAIRERLGPLVEQGRVRAATVGLAGQPLQERACAWGERLARMEEMGSRTRDLRDEVKEQVAEVRAKLEEQAQOIRLQAEAFQAR
LKSWFEPLVEDMORQWAGLVEKVQAAVGTSAAPVPSDNH¹TPKFAA²RRFAT³POLPPRSLORETLSPPQPS⁴WGGP⁵

>Q5XKQ4JAZG1 Alpha-2-glycoprotein 1, zinc-binding
MWASMSRMPLVLSLLGLPAVPOENODGRYSITYTIGLSKHVEDVPAFOALGSLNDLOFFRYNSKDRKSOPMGLWRQVEGMEDWKODSQ
LOKAREDFIMETLKDVEYYNDSNGSHVLOGRFGCEIENRSGAFWKYYVYDGKDYIEFNKEIPAWVPFDPAAQITKOKWEAEVYVQRAKAYLE
EECPATLRKYLKYSKNILDRQDPPSVVVTSHQAPGEKKLKLAYDFYPGKIDYHWTRAGEVOEPELRGDVYLHNGNGTYQSWVWVAVPPQDTAP
YSCWVQHSLSAOLVVPWEAS¹EARVGGN²VSOTL³

>P35222CHI3L1 Chitinase-3-like protein 1
MGVKASOTGFVVLVLLQCCSAYKLVCYTTSWSOYREGDGSQFPDLDRLFLCTHIYSFANISNDHIDTVEWINDVTLYGMLNTLKNRNPKLKTLLS
VGGWNFGSORFSKIASNTQSRRTFIKSVPPFLRTHGFDGLDLAWLYPGRRDQKHFTTLIKEMKAEFIKEAQPQGGKQKLLSALSAGKVTDSSYDIA
KISCHLDFISMTYDFHGAWRGTIGHSPFLFRGDEADSPDRFSNTDYAVGYMLRLGAPASKLYMGIPTFGRSFTLASSSETGVGAPISGPGIPGRFT
KEAGTLAYVEICDFLRGATVHRTLGOQVPYATKGNQWVGYDDQESVKSVQYLKDRQLAGANVWALDLDQFGSGFGGQDLRFPLTNAIKDALA
AT¹PSVLHTARGPR²RRKRLAPAKRSP³DHLPC⁴

>P10909ICLU Clusterin
MOVCSOPORGCVREQSAINTAAPP¹SAHNAASPGGARGHRVPLTEACKDSRIGGMIMKTL²LLFYGLLLTWESGOVLGQQTVDNELQEMSNQGSX
YVNKEIGNAVNGYKOKTLIEKTNEERKTL³SNLEEAKKKKEDALNETRESETKLKEPQVONETIMMALWEECKPCLKOTGMKFYARVQSRSGSL
YGRGLEEFLNOSPFTYFWNMGDRIDSLLENDROQTHMLDMQDHFSSASSIDELFODRFFFTREPQDITYHYLPFSLPHRRPHFFFPKSRIVRSIM
PFSPYEPLNFHAMFOPFLEMHQAQAMDIHFHSPAFQHPPTFEFIREGDDURTVCREIRHNSIGCLRIMKDDCKKREILSVDCSTNNP⁴SOAKLRR
ELDESLOWAERLTRKYNELLKSYQWKMLNTSSLLEQNEQFNWVSRLANLTQEDQYLYRVTTVASHTSDSDVPSGVTEVVYKLFDSDPITVTVP
VEVSRKNPKFMETVAEKALQCYRKKHREE⁵NDV⁶A⁷AP⁸TCASESSPQDELOPPRESSA⁹QVTRPQPP¹⁰QOLR¹¹QASPRSSGSC¹²TLTLD¹³SA¹⁴NI¹⁵G
K¹⁶NRAPACW¹⁷

>P04195HRC Histidine-rich glycoprotein
MKALIAALLITLOYSACVPTDCSAVEPEAEKALDLNKRPRRGYLFQLLRIADAHLDORVENTTVYYLVLDVQESDCSVLSRKYWNDC¹EPD²SRRP
SEIVGQCKVIATRHSHE³SD⁴LRVID⁵FNCTTSSYSSALANTKOSP⁶LID⁷FFDET⁸ERYRKOANKALEKYKEEND⁹DFASFRVORIERVARVRGCEGTGY
FYD¹⁰SVRNCPRHH¹¹PRHPNV¹²FG¹³CRADL¹⁴FYD¹⁵VEALDLESPKNLVINCEV¹⁶DP¹⁷OE¹⁸HENINGV¹⁹PP²⁰HL²¹GH²²PF²³HWGGHERS²⁴TK²⁵PP²⁶FK²⁷PHGSRD²⁸HH
PHK²⁹PH³⁰EGPP³¹PP³²DERD³³HS³⁴GPPL³⁵PG³⁶PP³⁷LP³⁸MSCSSCQ³⁹HAT⁴⁰GT⁴¹NGAQRHSHNNSSDL⁴²PH⁴³KK⁴⁴HS⁴⁵HEQ⁴⁶PH⁴⁷GH⁴⁸PH⁴⁹HA⁵⁰HP⁵¹HE⁵²HT⁵³RO
HP⁵⁴GH⁵⁵PH⁵⁶GH⁵⁷PH⁵⁸GH⁵⁹PH⁶⁰GH⁶¹PH⁶²CH⁶³GF⁶⁴QD⁶⁵Y⁶⁶GPC⁶⁷DP⁶⁸PP⁶⁹HNQ⁷⁰GH⁷¹CH⁷²GH⁷³GF⁷⁴PP⁷⁵GH⁷⁶LR⁷⁷RR⁷⁸CP⁷⁹GK⁸⁰GR⁸¹PF⁸²HO⁸³RG⁸⁴SV⁸⁵YRL⁸⁶P⁸⁷PL⁸⁸RK⁸⁹GEV
LPLPEAN⁹⁰PS⁹¹FL⁹²PH⁹³KK⁹⁴HL⁹⁵PK⁹⁶DN⁹⁷Q⁹⁸PP⁹⁹Q¹⁰⁰SV¹⁰¹SES¹⁰²CP¹⁰³GK¹⁰⁴SG¹⁰⁵FF¹⁰⁶QV¹⁰⁷SM¹⁰⁸FF¹⁰⁹TH¹¹⁰FP¹¹¹K¹¹²W¹¹³DL¹¹⁴NR¹¹⁵Q¹¹⁶AN¹¹⁷TLN¹¹⁸

Figura 18c

>P17336|IGFBP3 Insulin-like growth factor-binding protein 3
 MORARPTLWAAALILLVLRGPPVVARAGASSAGLGPVVRCEPCDARALAOCAAPPAYCAELVREPCGGCGLTCAISEGQPCGGIYTERCGSGLR
 CQSPDEARPLQALLDGRGLCVNASAVSRLRAYLLPAPPAPGNASEEEDRSAGSVESSTHRSDFPKFPLHSKIIKKKGHAKDSORVYKVQ
 YESQSTDOTNFSSESKEETEYGPORREMEDILNHLKFLNVLSPRGVHPNCKDKGFYKKKQCRPSKGRKRGRGFCWCVDKYGGPLPGYTTKGKED
 VHCYSMOSKTPPAAPLNNSSNNMPYFAOCTMKDITSSVLQFREPLEAVN*

>P05111|INH1 Inhibin alpha chain
 INVLHLLFLLLTPQGGHSCQGLEARELVIAKQRAFLDALGPPAVTREGGDPGVRRLLPRRHALGGFTHRGSEPEEEEDVSQAILFPATDASCED
 KSAARGLAOEAEEGLFRYMFPRPSOHTRSQVTSAGLVHFTGLDRQGTAAANSSEPLGLLALSPGCPVAVPMISLGHAPPNHWAYVLHATSALSLL
 THPVLVLLRCPLCTCSARPEATPFLVAHTRTPPSSGGERARRSTPLMSNWPSPSALRLLQRPPEEPAAHANCHRYVALNISFQELQWERWVWYP
 PSFIFHYCHGGGGLHIPNLSLPVPGAPPTPACPYSLPLGACAPCCAAALPGTMRPLHVRTTSDGGYSFKYTEVTPNLLTQHCACITWYGVFLNPMMA
 CGFAPTIQWELADSWEDGSHSSLLSLEQLWATPTPLESQ*

>Q9UBX7|KLK11 Kallikrein-11
 MORLRWLQWKSGRGLTAAKEPGARSSPLQAMRIQLULLALATGLVGGETRIKGFCKPHSOPWQMAALFEKTRILLOGATLIAPRWILLTAAHCL
 KPRVYHLGQHNLQKEEGCEQTRTATESFPHPGFNNSLPNKDRNDIMLVKNASPVSTWAVRPLTLSSRCVTAGTSLISGWSGSTSPQLRLPH
 TLROANTIEHOKCENAYPGNIHDTNVCASVQEGGKDSQGGDGGPLVNGSLQGIISWGGDPCATRIKPGVYTKVCKYVDWIDETMKNNITOP
 [146]SPSPSISDYQAVPVSNNKKT*

>P14618|PKM2 Pyruvate kinase M1/M2
 MSKPHSEAGTAFIQTOQLHAAMADTFLEIMCRLDIDSPFITARNITGCTIGPASRVSVELKEMIKSGMNVARLNFSGHGTHEYHAETIKNVRTATES
 FASCPILYRPVAVALDTKGPEIRTGLIKSGGTAEEVLKKGATLKILONAYMEKCDENILWLDYKNICKVVEVGSKYVDDGLISLOYKKGADFLVTE
 VENGSLGSKKGVNLPGAADVLPVASEKIDDLKFGVEQDVDMFAFIRKASDVHEVRKVLCEKGNIKISKIENHEGVRRFEDELEASDQIMVA
 RGDLGIEIPAEEKVFLAOKMHIGRCNRAGKPVICATOMLESMIKPRPTRAEGSDVANAVLDGADCMILSGETAKGDYPLEAVRMOHLIAREAEA
 MFHRKLFEEELVRASSHSTDLINEAMMGVSVEASYKCLAAALVLTESGRSAHGVARYRPRAPIAVTRNPOTARQAHLRYRGIFPVLCKDPVDEAWA
 EDVDLRVNFAMNVGKARGFFKGDVWVVLTWGRPGSGFTNTIRVYVPVYVYTPPEPLIQPLSHPLPPAPHLQGLRL*

>P08747|PLG Plasminogen
 MEHKEVVLILLFLKSGGCEPLDDYVNTQGASLFSVTKKQLGAGSIEECAKCEDEEFTCRAPQYHSKEQQCVIMAEENRKSSIIIRMRDWLFEK
 KYVLSCEKGTGNKNYRGTMSTKRNIGTCQKWSSTSPHPRFSPATHPSEGLEENYCRNPNDPQGPWCYTIDPEKRYDYCDILECEEEQMHIC
 SGENYDGKISKTMISGLECQAWDSQSPHAGYIPSKFPNKMLKKNYCRNPDEIRLPWCFTTPNKRWELCDIPRCTTPPSSSGPTYOCLKGTGE
 NYRGNVAVTVSGHTCQAWSAQIPHTHNRTPENFPCKNLDENYCRNPQDKRAPWCHITNSQVRWEYCKIPSCDSSSPVSTEQLAPTAPPELTPV
 VDDCYHGDGQSYRGTSSTTTTGKKCSWSMTPHRRHOKTPENYPNAGLTIMNYCRNPQADKGPWCFTTIDPSVRWEYCNLKKCSGIEASVWAP
 PPVLLPQVETPSEEDCNFGNGKGYRGKRAITVTGTPCDDWAACEPHRHSIFTETNPRAGLEKNYCRNPQDVGGPWCYTINPRKLYDYCD
 VPQCAAPSFDCGKQVEPKKCPGRVVGCCVAHPSWPQVSLRTFGIMHFCGGTILISPEWVLTAAHCKLSPRSPSSVKVILGAHQEVLNLEPHV
 CEIEYSRLFLEPIRKDIALKLSSPAVITDKVIPACLPSPNYVADRTQCFITGWGETQGTFCAGLLKEAQLPVENKVCNRYEFLNGRVSSTELCAG
 HLAGGTDSCQSDSGGPLVCFEKDKYILQGVTSWGLGOARPNIKPGVYVRSRFTWIEGVMRNNYLDGRQSDALTTWVGG*

Figura 18d

>P01011|SERPINA3 Alpha-1-antitrypsin
 NERMPLLALGLLAAGFCPAVLCHPNSEPLDEENLTQENQDRGTHVDQLASAWDFAFSLYKGLYLKAPDKNVFSPLSISTALAFSLGAHNTLT
 EILKGLKFNLTETSEAEIHOSFQHLLRTLNOSSDELQLSMGNAMFYKEQLSLDRFTEDAKRLYGSEAFATDFQDSAAKKLINDYKINGTRGKITD
 LIKOLDSQTMMLVNYIFFKAKWEMPEPQDTHQSRFYLSKKKVVIANPIMSLHLHTIPYFRDEELSCVTWELKYTGNASALFILPDQDKNEEVEA
 MILLPETLKRWRDSLEFREIGELYLPKFSISROYNLNDILLOLIEEAFTSKADLSGITGARNLAVSQVYHKAYLDVFEEGTEASAAATAVKITLLSALVE
 TRTIVRFNRPFLLMIIVPTDTQNIFFMISKVTNPKQA*SLPSSSGALSKELCQDAAGQGLQVWAGPQCAKPSGHOMDQPYCLNLEQDSOSLQSSHHRR
 GPWYTLQSSGSHAS*

>P02787|TF Serotransferrin
 MRLAVGALLVCAYLGLCLAVPDKTVRWCAVSEHEATKQSFROHMKSVIPSDGPSVACVKKASYLDCIRAIANEADAVTLDAGLVYDAYLAPNN
 LKPYVAEFYGSKEDPQTFYYAVAVYKKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDLPEPRKPLEKAYANFFSGSCAPCADGTDFFQL
 COLCPGCGCSTLNOYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPYDEYKDCHLACVPSHTVVARSMGGKEDLW
 ELLNGAOEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGFLKVPFRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKPVKWCALSHHERLKODE
 WSVNIVGKIECVSAETTEDCIKIMNGEADAMSLDGGFYIAGKQGLVPVLAENYNKSDNCEDTPEAGYFAYAVVYKKSASDLTWDNLKGGKKSCH
 TAVGRTAGWNIPMGLLYNKINHCRFDEFFSEGCAPGSKKDSLCKLQMSGSLNLCEPNKKEGYGTGAFRCLEVKGDAVAFVAKHOTVPQNTGG
 KNPDWAKNLNEKDYELLCLDGTGTRKPVVEYANCHLAPNPHAVTRKQKEACVHKILRQDQHLFQSNVTDCSGNFOLFRRSETKDLLFRDDTVOL
 AKLHDRNTYEKYLGEEYKAVGNLRKCS*SSLEACTFRRP*ILRGRAATKYKMGITQMIHCEALVSLAQVQCANVYKLLHSSVLPQVNLKK*

>P01137|TGFB1 Transforming growth factor beta-1
 MPSPGLRLPLPLPLLLWLLVTPGRPAAGLSTCKTIDMELVKRKRIEARGOILSKLRLASPFSGQCEVPPGPLPEAVLALYNSTRDRVAGESAEPEP
 EPEADYYAKEVTRVLMVETHNEYDKFKGSTHSYIMFFNTSELREAVPEPYLLSRAELRLRLKLVKEQHYELVQKYSNNNSWRYLSNRLLAPSDSP
 EWLSDQVTVGVROWLSRGGEIEGFRLSAHCSGDSRNTLOVDINGFTTGRGCLATHIGHNRPFLLLNATPLERAQHLOSSRHRRAIDTNYCFS
 STEKNCCVRLYIDFRKDLGWKWIHEPKGYHANFOLGCPQPYWISLDTQYSKVLALYNQHNPGASAAPCCVPQALEPLPIVYVGRKPKVEQLSN
 MIVRSCKCS*CPAPPRPDAAGAPPPQAPAAALPNGAVVQKDTTRAPSPPGAPHKMERQLRNYSLGACLGSPSLTFPHSHSLSLPCLLLPVCTPLP
 GKAAQGTSGENYCS*

>P02766|TR Transferrin
 MASHRLLLQLAGLVFVSEAGPTGTGESKPLMWKVLDAVRGSPAINVAVHVRKAADDTWEPFASGKTSSEGLHGLTTEEFVEGIYKVEIDTK
 SYWKALGISPFHEAEVVFANDSGPRRYTIAALLSPYSYSTTAVVTNPKETSTSRPYQLKDEQNDFM*

Figura 18e

Previsões

Observação

	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%		70.88%

TP: Verdadeiro positivo; TN: Verdadeiro negativo

Figura 19