

República Federativa do Brasil  
Ministério do Desenvolvimento, Indústria  
e do Comércio Exterior  
Instituto Nacional da Propriedade Industrial.

(21) **PI 0713098-8 A2**

(22) Data de Depósito: 22/05/2007  
(43) Data da Publicação: 16/10/2012  
(RPI 2180)



(51) *Int.Cl.:*  
C12Q 1/68  
G01N 33/48  
G01N 33/574  
G06F 19/00

**(54) Título:** MÉTODO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, MÉTODO DE DETECÇÃO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, SISTEMA DE DETECÇÃO, MEIO DE ARMAZENAGEM LEGÍVEL POR COMPUTADOR, ARRANJO DE ÁCIDO NUCLEICO, USO DE UM ARRANJO, MÉTODO PARA DETERMINAR O INÍCIO OU PREDISPOSIÇÃO PARA O INÍCIO DE UMA ANORMALIDADE CELULAR OU UMA CONDIÇÃO DESTINGUIDA POR UMA ANORMALIDADE CELULAR NO INTESTINO GROSSO, KIT DE DIAGNÓSTICO PARA ENSAIAR AMOSTRAS BIOLÓGICAS

**(30) Prioridade Unionista:** 22/05/2006 US 60/802312

**(73) Titular(es):** Clinical Genomics Pty. Ltd, Commonwealth Scientific And Industrial Research Organisation

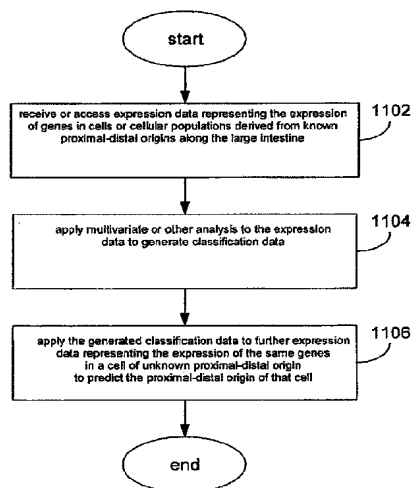
**(72) Inventor(es):** Lawrence C. Lapointe, Robert Dunne

**(74) Procurador(es):** Momsen, Leonardos & CIA.

**(86) Pedido Internacional:** PCT AU2007000703 de 22/05/2007

**(87) Publicação Internacional:** WO 2007/134395de 29/11/2007

**(57) Resumo:** MÉTODO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, MÉTODO DE DETECÇÃO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, SISTEMA DE DETECÇÃO, MEIO DE ARMAZENAGEM LEGÍVEL POR COMPUTADOR, ARRANJO DE ÁCIDO NUCLEICO, USO DE UM ARRANJO, MÉTODO PARA DETERMINAR O INÍCIO OU PREDISPOSIÇÃO PARA O INÍCIO DE UMA ANORMALIDADE CELULAR OU UMA CONDIÇÃO DESTINGUIDA POR UMA ANORMALIDADE CELULAR NO INTESTINO GROSSO, KIT DE DIAGNÓSTICO PARA ENSAIAR AMOSTRAS BIOLÓGICAS. A presente invenção diz respeito no geral a uma série de moléculas de ácido nucléico, os perfis de expressão dos quais caracterizam a origem anatômica da célula ou população de células dentro do intestino grosso. Mais particularmente, a presente invenção diz respeito a uma série de moléculas de ácido nucléico, os perfis de expressão das quais caracterizam a origem proximal ou distal da célula ou população de células dentro do intestino grosso. Os perfis de expressão da presente invenção são úteis na faixa de aplicações incluindo, mas não limitado a determinar a origem anatômica da célula ou população de células que foram derivadas do intestino grosso. Além disso, visto que a progressão da célula normal é frequentemente caracterizada pela desdiferenciação, o método da presente invenção também fornece um meio de identificar a anormalidade celular com base na expressão de um perfil de expressão incorreto em relação àquele que deve ser expressado pelas células objeto quando considerada sua localização anatômica dentro do cólon. Consequentemente, este aspecto da invenção fornece um meio valioso de identificar a existência de células colônicas do intestino grosso, estas sendo indicativas de uma anormalidade dentro do intestino grosso tal como o início ou pré disposição ao início da condição tal como um neoplasma colorretal.



“MÉTODO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, MÉTODO DE DETECÇÃO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, SISTEMA DE DETECÇÃO, MEIO DE ARMAZENAGEM LEGÍVEL POR COMPUTADOR, ARRANJO DE ÁCIDO NUCLEICO, USO DE UM ARRANJO, MÉTODO PARA DETERMINAR O INÍCIO OU PREDISPOSIÇÃO PARA O INÍCIO DE UMA ANORMALIDADE CELULAR OU UMA CONDIÇÃO DISTINGUIDA POR UMA ANORMALIDADE CELULAR NO INTESTINO GROSSO, KIT DE DIAGNÓSTICO PARA ENSAIAR AMOSTRAS BIOLÓGICAS”

#### CAMPO DA INVENÇÃO

A presente invenção diz respeito no geral ao conjunto das moléculas de ácido nucléico, aos perfis de expressão que caracterizam a origem anatômica de uma célula ou população de células dentro do intestino grosso. Mais particularmente, a presente invenção diz respeito a um conjunto de moléculas de ácido nucléico, os perfis de expressão das quais caracterizam a origem proximal ou distal de uma célula ou população de células dentro do intestino grosso. Os perfis de expressão da presente invenção são úteis dentro de uma faixa de aplicações incluindo, mas não limitadas à determinação da origem anatômica de uma célula ou população de células as quais foram derivadas do intestino grosso. Entretanto visto que a progressão de uma célula normal em direção a um estado neoplástico que é frequentemente caracterizado pela desdiferenciação fenotípica, o método da presente invenção também fornece um meio de identificar uma anormalidade celular com base na expressão de um perfil de expressão incorreto em relação àquele que deve ser expressado pelas células objeto quando consideradas em consideração com a sua localização anatômica dentro do cólon. Consequentemente, este

aspecto da invenção fornece um meio valioso de identificar a existência de células do cólon no intestino grosso, estas sendo indicativas de uma anormalidade dentro do intestino grosso tal como o início ou predisposição para o início de uma condição tal como uma neoplasia colorretal.

## 5                                   FUNDAMENTOS DA INVENÇÃO

Os detalhes bibliográficos das publicações aludidas pelo autor neste relatório descritivo são colecionados alfabeticamente no final da descrição.

10                                   Referência a qualquer técnica anterior neste relatório descritivo não é, e não deve ser interpretada como uma admissão ou qualquer forma de sugestão de que esta técnica anterior forme parte do conhecimento geral comum na Austrália.

15                                   Adenomas são tumores benignos de origem epitelial que são derivados de tecido glandular ou exibem estruturas glandulares claramente definidas. Alguns adenomas mostram elementos de tecidos reconhecíveis, tais como tecidos fibrosos (fibroadenomas), enquanto outros, como adenomas brônquicos, produzem compostos ativos que dão origem às síndromes clínicas. Tumores em certos órgãos, incluindo a glândula pituitária, são frequentemente classificados pela suas afinidades no tingimento histológico, 20 por exemplo, eosinófilos, basófilos e adenomas cromofóbicos.

25                                   Os adenomas podem se tornar carcinogênicos e são então denominados adenocarcinoma. Consequentemente, adenocarcinomas são definidos como tumores epiteliais malignos que surgem de estruturas glandulares, que são partes constituintes da maioria dos órgãos do corpo. Este termo também é aplicado aos tumores que mostram um padrão de crescimento glandular. Estes tumores podem ser subclassificados de acordo com a substância que eles produzem, por exemplo adenocarcinomas que secretam muco ou serosos, ou ao arranjo microscópico de sua células dentro de padrões, por exemplo adenocarcinomas papilares ou foliculares. Estes

carcinomas podem ser sólidos ou císticos (cistadenocarcinoma). Cada órgão pode produzir tumores mostrando uma grande variedade de tipos histológicos, por exemplo, o ovário pode produzir tanto *muconous* quanto cistadenocarcinoma. No geral, a incidência global de carcinoma dentro de um

5 adenoma é de aproximadamente 5%. Entretanto, isto está relacionado com o tamanho e, embora seja rara em adenomas de menos do que 1 centímetro, é estimado em 40 a 50% entre as lesões vilosas que são maiores do que 4 centímetros. Adenomas com graus mais altos de displasia têm uma incidência mais alta de carcinoma. Uma vez que um adenoma esporádico tenha se

10 desenvolvido, a chance de um novo adenoma ocorrer é de aproximadamente 30% dentro de 26 meses.

Os adenomas colorretais representam uma classe de adenomas que estão apresentando uma incidência crescente, particularmente nos países mais prósperos. As causas de adenoma, e sua mudança para adenocarcinoma,

15 são ainda objeto de intensa pesquisa. Até o momento, tem sido especulado que, além da predisposição genética, fatores ambientais (tais como dieta) desempenham um papel no desenvolvimento dessa condição. A maioria dos estudos indicam que os fatores ambientais relevantes dizem respeito a gordura dietética alta, baixa fibra e carboidratos refinados altos.

20 Os adenomas colônicos são proliferações localizadas de epitélio displástico que são inicialmente planos. Eles são classificados pela sua aparência grosseira quer séssil (plano) ou penduculado (tendo um talo). Enquanto adenomas pequenos (menos que 0,5 milímetros) exibem uma superfície lisa castanha, adenomas penduculados têm uma cabeça com uma

25 superfície arredondada ou lobulada marron avermelhada. Os adenomas sésseis exibem uma superfície vilosa mais delicada. Adenomas penduculados são mais prováveis de serem tubulares ou tubovilosos enquanto que as lesões sésseis são mais prováveis de serem vilosas. Os adenomas sésseis são mais comuns no ceco e no reto enquanto que os adenomas penduculados globais

são igualmente divididos entre o sigmóides-reto e o resto do intestino grosso.

Os adenomas são no geral assintomáticos, tornando difícil portanto seu diagnóstico e tratamento precoces. É tecnicamente impossível prognosticar a presença ou ausência do carcinoma com base na aparência grosseira de adenomas, embora adenomas maiores são considerados exibir uma incidência mais alta de malignidade concorrente do que os adenomas menores. Os adenomas sésseis exibem incidência mais alta de malignidade do que os adenomas penduculados do mesmo tamanho. Alguns adenomas resultam na produção de perda de sangue nas fezes microscópica. Entretanto, visto que o sangue nas fezes também pode ser indicativo de condições não adenomatosas e sintomas obstrutivos no geral não são observados na ausência de mudança maligna, o diagnóstico preciso de adenoma é tornado difícil sem a aplicação de procedimentos altamente invasivos tais como análise de biópsia. Consequentemente existe uma necessidade contínua para elucidar não apenas as causas de adenoma e sua mudança para malignidade mas para desenvolver protocolos de diagnóstico mais informativos, em particular protocolos que possibilitarão o diagnóstico rápido, de rotina e preciso de adenoma e adenocarcinoma em um estágio inicial tal como o estágio e pré maligno. Para esta finalidade estudos de adenocarcinoma colorretal têm sugerido incidência, histopatologia e prognóstico variáveis entre tumores proximais e distais.

Em termos de prosseguir nesta linha de investigação o advento do perfil de expressão de gene tem levado a um entendimento melhorado do desenvolvimento da mucosa intestinal. Por exemplo, a regulação de fatores de transcrição envolvidos na produção e manutenção do equilíbrio do eixo radial da base da cripta para o lúmen e aqueles que dão origem à diferenciação de célula epitelial são agora melhor entendidos como um resultado da análise da expressão de gene do microconjunto. [Peifer, 2002, Nature 420: 274-5, 277; Traber, 1999, Adv Exp Med Biol 470:114].

Similarmente, o entendimento tem melhorado dos eventos genéticos desenvolvimentalmente programados dentro do intestino embrionário, especialmente aqueles mecanismos de controle molecular responsáveis pelas diferenças epiteliais regionais entre o intestino delgado e o intestino grosso.

[de Santa Barbara *et al.*, 2003, *Cell Mol Life Sci* 60: 1322-1332; Park *et al.*, 2005, *Genesis* 41: 1-12] Por outro lado, pouco é conhecido a cerca da variação da expressão de gene proximal-distal ao longo do eixo longitudinal do intestino grosso. [Bates *et al.* 2002, *Gastroenterology* 122: 1467-1482] Estudos epidemiológicos de adenocarcinoma colorretal sugerem suporte para a incidência variável, histopatologia e prognóstico entre tumores proximais e distais.[Bonithon-Kopp e Benhamiche, 1999, *Eur J Cancer Prev* 8 Supl 1: S3-12; Bufill, 1990, *Ann Intern Med* 113: 779-788; Deng *et al.*, 2002, *Br J Cancer* 86: 574-579; Distler e Holt, 1997, *Dig Dis* 15: 302-311]. Assim um entendimento da variação específica de localização poderia fornecer discernimento valioso naquelas doenças que têm padrões de distribuição características ao longo do colorreto, incluindo câncer colorretal. [Birkenkamp-Demtroder *et al.*, 2005, *Gut* 54:374-384; Caldero *et al.*, 1989, *Virchows Arch A Pathol Anal- Histopathol* 415: 347-356; Garcia-Hirschfeld Garcia *et al.*, 1999, *Rev Esp Enferm Dig* 91: 481-488].

O colorreto (também chamado de intestino grosso) é frequentemente dividido por conveniência clínica em seis regiões anatômicas partindo da região terminal do íleo: o ceco; o cólon ascendente; o cólon transversal; o cólon descendente, o cólon sigmóide; e o reto. Alternativamente, estes segmentos podem ser agrupados para dividir o intestino grosso em um modelo de duas regiões compreendendo o intestino grosso proximal e distal. A região proximal (“direita”) é no geral considerada incluir o ceco, cólon ascendente, e o cólon transversal enquanto que a região distal (“esquerda”) inclui a flexura esplênica, o cólon descendente, a flexura sigmóide e o reto. Esta divisão é sustentada pela ontogênese embrionária

distinta destas regiões cuja junção é dois terços ao longo do cólon transversal e também pelo fornecimento arterial distinto a cada região. Enquanto que o intestino grosso proximal desenvolve-se a partir do intestino intermediário embrionário e é abastecido pela artéria mesentérica superior, o intestino grosso distal forma-se a partir do intestino traseiro embrionário e é abastecido pela artéria mesentérica inferior. [Yamada e Alpers, 2003, Textbook of Gastroenterology, 2 Vol. Set]. Um compreensivo de revisões de diferenças proximais-distais são fornecidas em [Iacopetta, 2002, Int J Cancer 101: 403-408].

Em trabalho que levou até a presente invenção foi determinado que um painel de genes são diferencialmente expressados entre as seções proximais e distais do intestino grosso humano. Consequentemente, este permitiu o desenvolvimento de meios para determinar se uma célula derivada do intestino grosso de interesse é de origem proximal ou origem distal. As amostras de células ou tecidos derivadas do intestino grosso normal podem ser portanto rotineiramente caracterizadas em termos de sua origem anatômica dentro do intestino grosso. Além disso ainda, visto que a maioria das condições de doença são caracterizadas por alguma mudança no perfil fenotípico ou transcrição de gene das células doentes, isto sendo particularmente verdadeiro de células que são predispostas a ou se tornaram neoplásticas, o método da presente invenção fornece um meio conveniente de identificar células anormais ou células que são predispostas a se tornarem anormais. Mais particularmente, onde uma célula de origem anatômica do intestino grosso conhecida expressa um ou mais genes ou perfis de genes que não são característicos daquela localização, a célula é classificada como anormal e pode depois passar por outra análise para elucidar a natureza desta anormalidade.

## SUMÁRIO DA INVENÇÃO

Por todo este relatório descritivo e as reivindicações que

seguem, a menos que o contexto requeira de outro modo, a palavra “compreendem”, e as variações tais como “compreende” e “compreendendo”, será entendido implicar a inclusão de um número inteiro ou etapa ou grupo de inteiros ou etapas estabelecidos mas não a exclusão de qualquer outro número

5 inteiro ou etapa ou grupo de números inteiros ou etapas.

Como aqui usado, o termo “derivado de” deve ser interpretado indicar que um número inteiro particular ou grupo de inteiro originado das espécies especificadas, mas não foi necessariamente obtido diretamente da fonte especificada. Além disso, como aqui usado as formas singulares de “a”,

10 “e” e “o” incluem referendos plurais a menos que o contexto claramente dite de outro modo.

A menos que de outro modo definido, todos os termos técnicos e científicos aqui usados têm o mesmo significado como habitualmente entendido por uma pessoa de habilidade na técnica à qual esta invenção

15 pertence.

Um aspecto da presente invenção está direcionada a um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, o dito método compreendendo medir o nível de expressão de um ou mais genes selecionados

20 de:

(i) o gene ou genes detectados pela sonda Affymetrix número:

218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

o gene detectado pela sonda Affymetrix número: 235733\_at

o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

APOBEC1, FLJ22761, KIFAP3, SLC14A2,  
 C10orf45, FTHFD, LOC375295, SLC20A1,  
 C10orf58, GCNT1, ME3, SLC23A3,  
 5 CCL8, HAS3, MEP1B, SLC38A2,  
 CLDN15, HOXB6, NPY6R, SLC9A3,  
 DEFA5, HOXD4, NR1H3, TBCC,  
 EYA2, HSD3B2, HR1H4, ZNF493,  
 OSTalfa,  
 10 PAP,

AFARP1 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
 número 202888\_s\_at,

15 CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 220753\_s\_at,

20 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206754\_s\_at,

CYP2C 18, ou o gene ou genes detectados pela sonda  
 Affymetrix número: 208126\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 214421\_x\_at ou 220017\_x\_at,

25 EPB41L3 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 222262\_s\_at ou 224453\_s\_at,

FAM45A ou o gene ou genes detectados pela sonda

Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 203639\_s\_at,

5 GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

10 HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

15 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

20 MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

25 NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219197\_s\_at,

5 SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

10 UGT1A3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221305\_s\_at ou

(ii) o gene detectado pela sonda Affymetrix número:  
230105\_at

15 o gene detectado pela sonda Affymetrix número: 230269\_at

o gene detectado pela sonda Affymetrix número: 238378\_at

o gene detectado pela sonda Affymetrix número: 239814\_at

o gene detectado pela sonda Affymetrix número: 239994\_at

o gene detectado pela sonda Affymetrix número: 240856\_at

20 o gene detectado pela sonda Affymetrix número: 242414\_at

o gene detectado pela sonda Affymetrix número: 244553\_at

ACACA, FMOD, LOC151162, S100P,

C13orf11, FRMD3, MCF2L, SCGB2A1,

C20orf56, GALNT5, MMP28, SCNN1B,

25 CAPN13, GARNL4, MUC11, SHANK2,

CLDN8, GCG, MUC12, SIAT2,

COLM, GNE, MUC17, SIAT4C,

CRIP1, HGD, MUC5B, SIAT7F,

DNAJC12, HOXB13, NEDD4L, SIDT1,

FAM3C, INSL5, PARP8, SLC13A2,  
 FBX025, IRS1, PCDH21, SLPI,  
 FLJ20366, ISL1, PI3, SPINK5,  
 FLJ20989, KIAA0703, PRAC, SST,  
 5 KIAA0830, PRAC2, TFF1,  
 KIAA1913, PTTG1IP, TNFSF11,  
 LAMA1, QPRT, TPH1,  
 LGALS2, QSCN6, WFDC2,  
 RBM24,

10 ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 213134\_x\_at ou 205548\_s\_at,

15 CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 221164\_x\_at ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
 número: 205518\_s\_at,

CRYBA2 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 220136\_s\_at

20 CTSE ou o gene ou genes detectados pela sonda Affymetrix  
 número: 205927\_s\_at,

DKFZp761N1114 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 242372\_s\_at,

25 EPB41L4A ou o gene ou genes detectados pela sonda  
 Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206070\_s\_at,

FAS ou o gene ou genes detectados pela sonda Affymetrix  
 número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

5 FLJ23548 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

10 FOXA2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix  
número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221577\_x\_at,

15 GJB3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205490\_s\_at,

HOXD13 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207397\_s\_at,

20 INSM1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda  
Affymetrix número: 212959\_s\_at,

MLPH ou o gene ou genes detectados pela sonda Affymetrix  
número: 218211\_s\_at,

25 NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda  
Affymetrix número: 203649\_s\_at,

PTPRO ou o gene ou genes detectados pela sonda Affymetrix

número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix

número: 207080\_s\_at ou 211253\_x\_at,

SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda

Affymetrix número: 207249\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda

Affymetrix número: 221024\_s\_at,

10 SPON1 ou o gene ou genes detectados pela sonda Affymetrix

número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix

número: 203769\_s\_at

TM4SF11 ou o gene ou genes detectados pela sonda

15 Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix

número: 213432\_s\_at ou 209228\_x\_at,

em uma amostra biológica do dito indivíduo em que um nível

alto de expressão dos genes do grupo (i) em relação aos níveis de controle do

20 intestino grosso distal normais é indicativo de uma origem do intestino grosso

proximal e um nível alto de expressão dos genes do grupo (ii) em relação aos

níveis de controle do intestino grosso proximal normais é indicativo de uma

origem do intestino grosso distal.

Em um outro aspecto é fornecido um método para determinar

25 a origem anatômica de uma célula ou população celular derivada do intestino

grosso de um indivíduo, o dito método compreendendo medir o nível de

expressão de um ou mais genes selecionados de:

(i) PITX2 ou o gene ou genes detectados pela sonda

Affymetrix número 207558\_s\_at,

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número 222262\_s\_at ou 224453\_s\_at,

FAM3B,

CYP2C18 ou o gene ou genes detectados pela sonda  
5 Affymetrix número 208126\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
número 219954\_s\_at,

MEP1B,

ADRA2A,

10 HSD3B2,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
número 206754\_s\_at,

SLC14A2 ou o gene ou genes detectados pela sonda  
Affymetrix número 226432\_s\_at,

15 CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
número 231576\_s\_at,

DEFA5,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número 210797\_s\_at,

20 SLC37A3,

REG1A,

MEP1B,

NR1H4; ou

(ii) DKFZp761N1114 ou o gene ou genes detectados pela  
25 sonda Affymetrix número 242374\_s\_at,

PRAC,

INSL5,

HOXB13 ou

WFDC2

em uma amostra biológica do dito indivíduo em que um nível alto de expressão dos genes do grupo (i) em relação aos níveis de controle do intestino grosso distal normais é indicativo de uma origem do intestino grosso proximal e um nível alto de expressão dos genes do grupo (ii) em relação aos níveis de controle do intestino grosso proximal normais como indicativo de uma origem do intestino grosso distal.

Em um outro aspecto, a presente invenção fornece um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

avaliar dados de treinamento, que inclui dados de treinamento de expressão que representa a expressão de genes em células ou populações celulares derivadas de origens proximais-distais conhecidas de um intestino grosso, e dados de treinamento de origem proximal-distal que representam as associações das ditas células ou populações celulares com as ditas origens proximais-distais;

processar os dados de treinamento usando análise multivariada para gerar dados de classificação para gerar dados de origem proximal-distal indicativos de uma origem proximal-distal de uma célula ou população celular adicionais derivadas de um intestino grosso, com base nos dados de expressão adicionais que representam a expressão de genes na dita célula ou população celular adicionais.

A presente invenção também fornece um método de detecção para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

avaliar os dados de expressão primários que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de pelo menos um intestino grosso;

processar os dados de expressão primários usando análise multivariada para gerar dados de modelo multivariado representativo de

associações entre os dados de expressão primários e origens proximais-distais das ditas células ou populações celulares;

5       avaliar os segundos dados de expressão que representam a expressão de genes em uma célula ou população celular derivada do intestino grosso de um indivíduo; e

      processar os segundos dados de expressão e os dados de modelo multivariado para gerar dados de origem proximal-distal representativos de uma origem proximal-distal das ditas célula ou população celular.

10       Preferivelmente, a etapa de avaliar os dados de expressão primários inclui avaliar os dados de expressão terciários dos quais os ditos os dados de expressão primários são um subconjunto, e o método inclui processar os ditos dados de expressão terciários para selecionar um subconjunto dos dados de expressão terciários que correspondam a um  
15       subconjunto de genes diferencialmente expressado sozinho ou em combinação ao longo do eixo proximal-distal do dito intestino grosso, o subconjunto selecionado sendo os ditos dados de expressão primários.

      A presente invenção também fornece um método para determinar a origem anatômica de uma célula ou população celular derivada  
20       do intestino grosso de um indivíduo, que inclui:

      avaliar os dados de expressão primários que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de um intestino grosso derivadas do intestino  
      grosso;

25       processar os dados de expressão primários usando um método de núcleo para gerar dados de classificação para processar os segundos dados de expressão que representam a expressão dos ditos genes em pelo menos uma célula ou população celular secundárias de um intestino grosso para gerar dados de origem proximal-distal que representam a origem proximal-

distal da dita pelo menos uma célula ou população celular secundárias.

A presente invenção também fornece um método de detecção para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

5                   avaliar os dados de expressão primários que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de um intestino grosso derivadas do intestino grosso;

10                   processar os dados de expressão primários usando a análise de componentes principais para gerar dados de componente principal que correspondam a pelo menos uma combinação linear da expressão dos ditos genes, os ditos dados de componente principal sendo indicativo de pelo menos uma das origens proximais-distais das ditas células ou populações celulares.

15                   A presente invenção também fornece um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

20                   avaliar dados de expressão que representem a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de pelo menos um intestino grosso; e

                  processar os dados de expressão usando análise de variação canônica para gerar dados de variação canônica indicativos de pelo menos uma das origens proximais-distais das ditas células ou populações celulares.

25                   A presente invenção também fornece um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

                  avaliar dados de treinamento, que inclui a expressão de dados de treinamento que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de pelo

menos um intestino grosso, e dados de treinamento de origem proximal-distal que representam as associações das ditas células ou populações celulares com as ditas origens proximais-distais;

5                   processar os dados de treinamento para gerar dados de  
classificação que representam uma combinação linear ou não linear de níveis  
de expressão dos ditos genes, os ditos dados de classificação sendo adaptados  
para gerar dados de origem proximal-distal adicionais indicativos de uma  
origem proximal-distal de uma célula ou subpopulação celular adicionais  
tiradas de um intestino grosso, com base nos dados de expressão adicionais  
10               que representam a expressão dos ditos genes na dita célula ou subpopulação  
celular adicionais.

A presente invenção também fornece um sistema de detecção tendo componentes para executar qualquer um dos métodos acima.

15               A presente invenção também fornece um meio de  
armazenagem legível por computador tendo armazenado nele instruções de  
programa para executar qualquer um dos métodos acima.

A presente invenção também fornece um sistema de detecção, que inclui:

20               meios para avaliar dados de treinamento, que inclui a  
expressão de dados de treinamento que representam a expressão de genes em  
células ou populações celulares derivadas de pelo menos um intestino grosso,  
e dados de treinamento de origem proximal-distal que representam as  
associações das ditas células ou populações celulares com as ditas origens  
proximais-distais;

25               meios para processar os dados de treinamento para gerar dados  
de classificação que representam uma combinação linear ou não linear de  
níveis de expressão dos ditos genes, os ditos dados de classificação sendo  
adaptados para gerar dados de origem proximal-distal indicativo de uma  
origem proximal-distal de uma célula ou população celular adicionais tiradas

de um intestino grosso, com base nos dados de expressão adicionais que representam a expressão dos ditos genes na dita célula ou população celular adicionais.

Em um outro aspecto é fornecido um método para determinar o início ou a predisposição para o início de uma anormalidade celular ou uma condição caracterizada por uma anormalidade celular no intestino grosso, o dito método compreendendo determinar, de acordo com um dos métodos mais acima descritos, o perfil de expressão de um gene proximal-distal de uma amostra biológica derivada de uma origem proximal ou distal conhecida no intestino grosso em que a detecção de um perfil de expressão de gene que é incompatível com o perfil de expressão de gene do intestino grosso proximal-distal normal é indicativo da anormalidade da célula ou população celular que expressam o dito perfil.

Um aspecto relacionado da presente invenção fornece conjunto de ácido nucléico, conjunto este que compreende uma pluralidade de:

(i) moléculas de ácido nucléico que compreendem uma sequência de nucleotídeo que corresponde a qualquer um dos genes marcadores de localização mais acima descritos ou uma sequência que exiba pelo menos 80% de identidade com esta ou um derivado, fragmento, variante ou homólogo funcionais das ditas moléculas de ácido nucléico; ou

(ii) moléculas de ácido nucléico que compreendem uma sequência de nucleotídeo capaz de hibridização a qualquer uma ou mais das sequências de (i) sob condições de severidade baixa a 42°C ou um derivado, fragmento, variante ou homólogo funcionais das ditas moléculas de ácido nucléico

(iii) sondas de ácido nucléico ou oligonucleotídeos que compreendem uma sequência de nucleotídeo capaz de hibridização a qualquer uma ou mais das sequências de (i) sob condições de severidade baixa a 42°C ou um derivado, fragmento, variante ou homólogo funcionais das ditas

moléculas de ácido nucléico

(iv) proteínas codificadas pelas moléculas de ácido nucléico de (i) ou (ii) ou um derivado, fragmento, variante ou homólogo

em que o nível de expressão do dito ácido nucléico é indicativo da origem proximal-distal de uma célula ou subpopulação celular derivada do intestino grosso.

### DESCRIÇÃO RESUMIDA DOS DESENHOS

A Figura 1 é uma representação gráfica da comparação do número de conjuntos de sonda diferenciais quando o divisor entre as regiões proximal e distal é movida.

A Figura 2 é uma representação gráfica do número relativo de transcritos elevados no intestino grosso proximal e distal.

A Figura 3 é uma representação gráfica de um exemplo típico de um modelo de dois genes.

A Figura 4 é uma representação gráfica da direção relativa de aumentar a expressão de transcritos que exibem uma mudança gradual ao longo do colorreto.

A Figura 5 é uma representação gráfica de genes que exibem comportamento de modelo de cinco segmentos.

A Figura 6a é uma representação gráfica de um exemplo típico do primeiro e segundo componentes principais gerados pela aplicação da análise de componente principal (PCA) a todos os 44.928 conjuntos de sonda do conjunto de dados de verificação, que revela pouca, se alguma, estrutura;

A Figura 6b é um gráfico do primeiro e segundo componentes principais gerados pela aplicação da PCA a um subconjunto de 115 conjuntos de sonda que são cada um diferencialmente expressados em amostras de tecido do ceco e reto (isto é, as extremidades proximal e distal extremas do intestino grosso), que revela duas classes que correspondem às porções proximal e distal do intestino grosso;

A Figura 7A é um gráfico do primeiro componente principal da Figura 6A como uma função da localização de tecido ao longo do eixo proximal-distal do intestino grosso;

5 A Figura 7B é um gráfico do primeiro componente principal da Figura 6B como uma função da localização de tecido ao longo do eixo proximal-distal do intestino grosso;

A Figura 8A é um gráfico da primeira e segunda variáveis canônicas gerado pela análise de perfil;

10 A Figura 8B é um gráfico da primeira variável canônica da Figura 8A como uma função da localização de tecido ao longo do eixo proximal-distal do intestino grosso;

A Figura 9 é um gráfico das estimativas de erro validadas cruzadas de vetores de suporte gerados a partir dos respectivos subconjuntos de genes como uma função do número de genes em cada subconjunto;

15 A Figura 10 é um diagrama de bloco de uma forma de realização preferida de um sistema de detecção; e

A Figura 11 é um diagrama de fluxo de uma forma de realização preferida de um método de detecção executado pelo sistema de detecção.

## 20 DESCRIÇÃO DETALHADA DA INVENÇÃO

A presente invenção é fundamentada, em parte, na elucidação do perfil de expressão de genes que caracterizam a origem anatômica de uma célula ou população celular do intestino grosso em termos de uma origem proximal versus uma origem distal. Esta verificação tem agora facilitado o  
25 desenvolvimento de meios de rotina de caracterizar, em termos da sua origem anatômica, uma população celular derivada do intestino grosso. Além disso ainda, visto que alguns distúrbios celulares são caracterizados por uma mudança no perfil de expressão de gene da célula doente em relação a uma célula normal correspondente, a presente invenção também fornece um meio

de triar rotineiramente células do intestino grosso, que foram derivadas de um local anatômico conhecido dentro do intestino grosso, quanto a quaisquer mudanças para o perfil de expressão de gene que esperar-se-ia ser expressado com base naquela localização particular. Onde o perfil de expressão de gene

5      correto não é observado, a célula está exibindo uma anormalidade e deve ser ainda avaliado por via de diagnosticar a condição específica da anormalidade. Em particular, seria avaliado pela pessoa de habilidade na técnica que as células neoplásticas, ou células pré dispostas a isto, algumas vezes passam pela desdiferenciação – isto sendo evidenciado por uma mudança para o

10     fenótipo de expressão de gene da célula para um fenótipo menos diferenciado. Consequentemente, qualquer mudança nas características de perfil de expressão de gene de uma célula de intestino grosso de origem proximal ou distal pode ser indicativa do início ou predisposição ao início de um neoplasma do intestino grosso, tal como um adenoma ou um adenocarcinoma.

15     Também é fornecido pela presente invenção conjuntos de ácido nucléico, tais como microarranjos, para o uso no método da invenção.

Consequentemente, um aspecto da presente invenção é direcionado a um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, o dito

20     método compreendendo medir o nível de expressão de um ou mais genes selecionados de:

(i) o gene ou genes detectados pela sonda Affymetrix número:

218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

25     o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

o gene detectado pela sonda Affymetrix número: 235733\_at

o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ABHD5,    | FAM3B,    | IGFBP2,    | POPDC3,  |
|    | ADRA2A,   | FLJ10884, | KCNG1,     | REG1A,   |
| 5  | APOBEC1,  | FLJ22761, | KIFAP3,    | SLC14A2, |
|    | C10orf45, | FTHFD,    | LOC375295, | SLC20A1, |
|    | C10orf58, | GCNT1,    | ME3,       | SLC23A3, |
|    | CCL8,     | HAS3,     | MEP1B,     | SLC38A2, |
|    | CLDN15,   | HOXB6,    | NPY6R,     | SLC9A3,  |
| 10 | DEFA5,    | HOXD4,    | NR1H3,     | TBCC,    |
|    | EYA2,     | HSD3B2,   | HR1H4,     | ZNF493,  |
|    |           |           | OSTalfa,   |          |
|    |           |           | PAP,       |          |

15 AFARP1 ou o gene ou genes detectados pela sonda Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix número 202888\_s\_at,

CCL13 ou o gene ou genes detectados pela sonda Affymetrix número: 206407\_s\_at

20 CRYL1 ou o gene ou genes detectados pela sonda Affymetrix número: 220753\_s\_at,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix número: 206754\_s\_at,

25 CYP2C18, ou o gene ou genes detectados pela sonda Affymetrix número: 208126\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix número: 214421\_x\_at ou 220017\_x\_at,

EPB41L3 ou o gene ou genes detectados pela sonda Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222262\_s\_at ou 224453\_s\_at,

FAM45A ou o gene ou genes detectados pela sonda  
Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

5 FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 203639\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219954\_s\_at,

10 GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

15 HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206601\_s\_at,

20 ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

25 MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix

número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix

número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda

Affymetrix número: 219197\_s\_at,

SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix

número: 225457\_s\_at,

10 SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

UGT1A3 ou o gene ou genes detectados pela sonda

Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda

15 Affymetrix número: 221305\_s\_at ou

(ii) o gene detectado pela sonda Affymetrix número:

230105\_at

o gene detectado pela sonda Affymetrix número: 230269\_at

o gene detectado pela sonda Affymetrix número: 238378\_at

20 o gene detectado pela sonda Affymetrix número: 239814\_at

o gene detectado pela sonda Affymetrix número: 239994\_at

o gene detectado pela sonda Affymetrix número: 240856\_at

o gene detectado pela sonda Affymetrix número: 242414\_at

o gene detectado pela sonda Affymetrix número: 244553\_at

25 ACACA, FMOD, LOC151162, S100P,  
C13orf11, FRMD3, MCF2L, SCGB2A1,  
C20orf56, GALNT5, MMP28, SCNN1B,  
CAPN13, GARNL4, MUC11, SHANK2,  
CLDN8, GCG, MUC12, SIAT2,

|    |           |           |          |          |
|----|-----------|-----------|----------|----------|
|    | COLM,     | GNE,      | MUC17,   | SIAT4C,  |
|    | CRIP1,    | HGD,      | MUC5B,   | SIAT7F,  |
|    | DNAJC12,  | HOXB13,   | NEDD4L,  | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,   | SLC13A2, |
| 5  | FBX025,   | IRS1,     | PCDH21,  | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,     | SPINK5,  |
|    | FLJ20989, | KIAA0703, | PRAC,    | SST,     |
|    |           | KIAA0830, | PRAC2,   | TFF1,    |
|    |           | KIAA1913, | PTTG1IP, | TNFSF11, |
| 10 |           | LAMA1,    | QPRT,    | TPH1,    |
|    |           | LGALS2,   | QSCN6,   | WFDC2,   |
|    |           |           | RBM24,   |          |

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

15 BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213134\_x\_at ou 205548\_s\_at,

CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164\_x\_at ou 223942\_x\_at,

20 CMAH ou o gene ou genes detectados pela sonda Affymetrix  
número: 205518\_s\_at,

CRYBA2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix  
número: 205927\_s\_at,

25 DKFZp761N1114 ou o gene ou genes detectados pela sonda  
Affymetrix número: 242372\_s\_at,

EPB41L4A ou o gene ou genes detectados pela sonda  
Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix

número: 206070\_s\_at,

FAS ou o gene ou genes detectados pela sonda Affymetrix

número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda

Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FLJ23548 ou o gene ou genes detectados pela sonda

Affymetrix número: 218187\_s\_at,

10 FN1 ou o gene ou genes detectados pela sonda Affymetrix

número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

FOXA2 ou o gene ou genes detectados pela sonda Affymetrix

número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix

15 número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix

número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix

número: 205490\_s\_at,

20 HOXD13 ou o gene ou genes detectados pela sonda

Affymetrix número: 207397\_s\_at,

INSM1 ou o gene ou genes detectados pela sonda Affymetrix

número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda

25 Affymetrix número: 212959\_s\_at,

MLPH ou o gene ou genes detectados pela sonda Affymetrix

número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix

número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda Affymetrix número: 203649\_s\_at,

PTPRO ou o gene ou genes detectados pela sonda Affymetrix número: 208121\_s\_at,

5 PYY ou o gene ou genes detectados pela sonda Affymetrix número: 207080\_s\_at ou 211253\_x\_at,

SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix número: 222258\_s\_at,

10 SLC28A2 ou o gene ou genes detectados pela sonda Affymetrix número: 207249\_s\_at

SLC2A10 ou o gene ou genes detectados pela sonda Affymetrix número: 221024\_s\_at

SPON1 ou o gene ou genes detectados pela sonda Affymetrix número: 213994\_s\_at ou 209437\_s\_at

15 STS ou o gene ou genes detectados pela sonda Affymetrix número: 203769\_s\_at

TM4SF11 ou o gene ou genes detectados pela sonda Affymetrix número: 204519\_s\_at,

20 TUSC3 ou o gene ou genes detectados pela sonda Affymetrix número: 213432\_s\_at ou 209228\_x\_at,

em uma amostra biológica do dito indivíduo em que um nível alto de expressão dos genes do grupo (i) em relação aos níveis de controle do intestino grosso distal normais é indicativo de uma origem do intestino grosso proximal e um nível alto de expressão dos genes do grupo (ii) em relação aos  
25 níveis de controle do intestino grosso proximal normais é indicativo de uma origem do intestino grosso distal.

Como detalhado mais acima, o método da presente invenção é fundamentado na determinação de que a localização distal versus proximal de uma célula dentro do intestino grosso pode agora ser averiguada em virtude

do perfil de expressão de genes que são únicos para as células de cada uma destas localizações. Consequentemente, referência para determinar a “origem anatômica” ou “localização anatômica” de uma célula ou população celular “derivadas do intestino grosso” deve ser entendido como uma referência para

5 determinar se a célula em questão origina-se da região distal do intestino grosso ou da região proximal do intestino grosso. Ainda para isto, por “origem” ou “localização” é intencionado a localização da célula ou células sob investigação exatamente antes do momento que a célula foi colhida do intestino grosso ou, onde a célula naturalmente descolou do intestino grosso

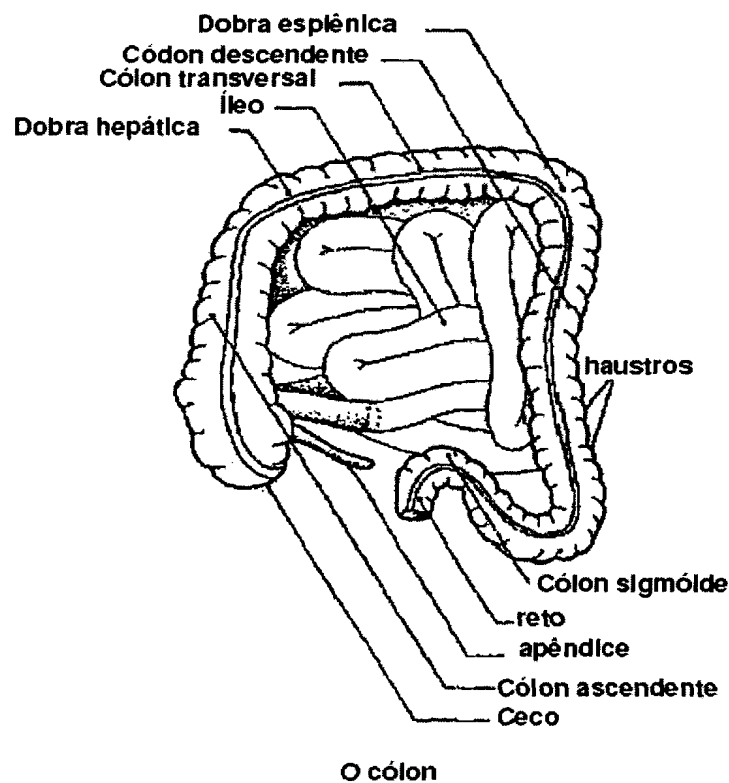
10 (por exemplo, onde ela foi descartada e é encontrada em uma amostra de fezes), no tempo imediatamente antes do descolamento da célula do intestino grosso. Sem limitar a presente invenção a qualquer teoria ou modo de ação, o intestino grosso não tem nenhuma função digestiva, como tal, mas absorve grandes quantidades de água e eletrólitos do alimento não digerido passado

15 proveniente do intestino delgado. Em intervalos regulares, os movimentos peristálticos movem os conteúdos desidratados (fezes) na direção do reto. Por conveniência clínica o intestino grosso é no geral dividido em seis regiões anatômicas começando depois da região terminal do íleo - estas sendo:

- (i) o ceco;
- 20 (ii) o cólon ascendente;
- (iii) o cólon transversal;
- (iv) o cólon descendente;
- (v) o cólon sigmóide; e
- (vi) o reto.

25 Estes segmentos também podem ser agrupados para dividir o intestino grosso em um modelo de duas regiões que compreendem o intestino grosso proximal e distal. A região proximal é no geral entendida incluir o ceco e o cólon ascendente enquanto que a região distal inclui a flexura esplênica, o cólon descendente, a flexura sigmóide e o reto. Esta divisão entre

a região proximal e distal do intestino grosso é considerada ocorrer aproximadamente dois terços ao longo ao cólon transversal. Esta divisão é sustentada pela ontogênese embrionária distinta destas regiões cuja junção é dois terços ao longo do cólon transversal e também pelo fornecimento arterial distinto a cada região. Consequentemente, os tecidos do cólon transversal podem ser proximais ou distais dependendo de qual lado desta junção corresponde ao seu ponto de origem. Seria avaliado que embora o método da presente invenção possa necessariamente não indicar de qual parte do intestino grosso proximal ou distal uma célula é originada, o mesmo fornecerá informação valiosa em relação a se o tecido é de origem proximal ou origem distal. Embora o intestino grosso proximal se desenvolva a partir do intestino intermediário embrionário e seja abastecido pela artéria mesentérica superior, o intestino grosso distal forma-se a partir do intestino posterior embrionário e é abastecido pela artéria mesentérica inferior.



Consequentemente, referência à região “proximal” do intestino grosso deve ser entendida como uma referência à seção do intestino grosso

que compreende o ceco e o cólon ascendente, enquanto referência à região “distal” do intestino grosso deve ser entendido como uma referência à dobra esplênica, cólon descendente, dobra sigmóide e reto. A região do cólon transversal compreende tanto a região proximal quanto a distal, as proporções relativas das quais dependerá de onde a junção dos tecidos proximal e distal ocorre. Especificamente, o tecido do cólon transversal pode ser da região proximal ou distal dependendo da distância relativa entre as dobras hepática e esplênica.

De acordo com a presente invenção, foi determinado que os genes detalhados nos parágrafos (i) e (ii), acima, são modulados, em termos de mudanças diferenciais aos seus níveis de expressão dependendo se a célula que expressa aquele gene está localizada na região proximal do intestino grosso ou na região distal do intestino grosso. Para facilidade de referência, estes genes e seus transcritos de mRNA são representados em texto em itálicos enquanto que os seus produtos de expressão de proteína são representados em texto não em itálicos. Esses genes são coletivamente aludidos como “marcadores de localização”.

Cada um dos genes detalhados nos sub-parágrafos (i) e (ii), acima, são bem conhecidos pela pessoa habilitada na técnica, como o é seus produtos de expressão de proteína codificada. A identificação destes genes como marcadores da localização de célula colorretal (intestino grosso) ocorreu em virtude da análise de expressão diferencial usando Affymetrix HG133A ou chips de gene HG133B. Para esta finalidade, cada chip de gene é caracterizada por aproximadamente 45.000 conjuntos de sonda que detectam o RNA transcrito de aproximadamente 35.000 genes. Em média, aproximadamente 11 pares de sonda detectam as regiões de sobreposição ou consecutivas do transcrito de RNA de um único gene. No geral, os genes a partir dos quais os transcritos de RNA são identificáveis pelas sondas Affymetrix são genes bem conhecidos e caracterizados. Entretanto, até o grau

em que algumas das sondas detectam transcritos de RNA que não são ainda definidos, estes genes são indicados como “o gene ou genes detectados pela sonda Affymetrix x”. Em alguns casos vários genes podem ser detectáveis por uma única sonda. Isto também é indicado onde apropriado. Deve ser entendido, entretanto, que isto não é intencionado como uma limitação de como o nível de expressão do gene objeto pode ser detectado. No primeiro caso, seria entendido que o transcrito de gene objeto também é detectável por outras sondas que estariam presentes na chip de gene da Affymetrix. A referência a uma única sonda é meramente incluída como um identificador do transcrito de gene de interesse. Em termos de triar verdadeiramente quanto ao transcrito, entretanto, pode-se utilizar uma sonda direcionada a qualquer região do transcrito e não apenas à região de transcrito de 600 pares de base terminal à qual as sondas Affymetrix são no geral direcionadas.

Referência a cada um dos genes detalhados acima e seus produtos de expressão transcritos e traduzidos devem ser portanto entendidos como uma referência a todas as formas destas moléculas e aos fragmentos, mutantes ou variantes destas. Como seria avaliado pela pessoa de habilidade na técnica, alguns genes são conhecidos exibir variação alélica entre os indivíduos. Consequentemente, a presente invenção deve ser entendida estender-se a tais variantes que, em termos das presentes aplicações em diagnóstico, alcançam o mesmo resultado a despeito do fato de que variantes genéticas menores entre as sequências de ácido nucléico reais podem existir entre os indivíduos. A presente invenção portanto deve ser entendida estender-se a todos os RNAs (por exemplo, mRNA, transcrito de RNA primário, miRNA, tRNA, rRNA etc), cDNA e isoformas peptídicas que surgem da união alternativa ou qualquer outra mutação, variação polimérica ou alélica. Também deve ser entendido incluir referência a quaisquer polipeptídeos de subunidade tais como formas precursoras que podem ser geradas, se existirem como um monômero, multímero, proteína de fusão ou

outro complexo.

Sem limitar a presente invenção a qualquer teoria ou modo de ação, embora cada um dos genes mais acima descritos seja diferencialmente expressado, isoladamente ou em combinação, como entre as células do intestino grosso distal e proximal, e é portanto diagnóstico da origem anatômica de qualquer amostra de célula dada, a expressão de alguns destes genes exibiram níveis particularmente significantes de sensibilidade, especificidade, valor preditivo positivo e/ou valor preditivo negativo. Consequentemente, em uma forma de realização preferida, poder-se-ia triar e avaliar quanto ao nível de expressão de um ou mais destes genes.

A presente invenção portanto preferivelmente fornece um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, o dito método compreendendo medir o nível de expressão de um ou mais genes selecionados de:

(i) PITX2 ou o gene ou genes detectados pela sonda Affymetrix número: 207558\_s\_at,

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix número: 222262\_s\_at ou 224453\_s\_at,

FAM3B,

CYP2C18 ou o gene ou genes detectados pela sonda Affymetrix número: 208126\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix número: 219954\_s\_at,

MEP1B,

ADRA2A,

HSD3B2,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix número: 206754\_s\_at,

SLC14A2 ou o gene ou genes detectados pela sonda Affymetrix número: 226432\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix número: 231576\_s\_at,

5 DEFA5,

OASL ou o gene ou genes detectados pela sonda Affymetrix número: 210797\_s\_at,

SLC37A3,

REG1A,

10 MEP1B,

NR1H4; ou

(ii) DKFZp761N1114 ou o gene ou genes detectados pela sonda Affymetrix número: 242374\_s\_at,

PRAC,

15 INSL5,

HOXB13 ou

WFDC2

em uma amostra biológica do dito indivíduo em que um nível alto de expressão dos genes do grupo (i) em relação aos níveis de controle do intestino grosso distal normais é indicativo de uma origem do intestino grosso proximal e um nível alto de expressão dos genes do grupo (ii) em relação aos níveis de controle do intestino grosso proximal normais como indicativo de uma origem do intestino grosso distal.

20 Preferivelmente, os ditos genes são ETNK1 e/ou GBA3 e/ou  
25 PRAC.

O método de detecção da presente invenção pode ser realizado em qualquer amostra biológica adequada. Para esta finalidade, referência a uma “amostra biológica” deve ser entendida como uma referência a qualquer amostra de material biológico derivado de um animal tal como, mas não

limitado a, material celular, biofluidos (por exemplo, sangue), fezes, espécimes de biópsia de tecido, espécimes cirúrgicos ou fluido que foi introduzido no corpo de um animal e subsequentemente removido (tal como, por exemplo, a solução recuperada de uma lavagem de enema). A amostra biológica que é testada de acordo com o método da presente invenção pode ser testada diretamente ou pode requerer alguma forma de tratamento antes de testar. Por exemplo, uma biópsia ou amostra cirúrgica podem requerer homogeneização antes de testar ou podem requerer divisão em pedaços para o teste *in situ* dos níveis de expressão qualitativos de genes individuais.

Alternativamente, uma amostra de célula pode requerer a permeabilização antes de testar. Além disso, até o grau em que a amostra biológica não esteja na forma líquida, (se tal forma é requerida para o teste) a mesma pode requerer a adição de um reagente, tal como um tampão, para mobilizar a amostra.

Até o grau em que o gene marcador de localização esteja presente em uma amostra biológica, a amostra biológica pode ser diretamente testada ou ainda todo ou algum do material de ácido nucléico presente na amostra biológica pode ser isolada antes de testar. Já em um outro exemplo, a amostra pode ser parcialmente purificada ou de outro modo enriquecida antes da análise. Por exemplo, até o grau em que uma amostra biológica compreenda uma população de célula muito diversa, pode ser desejável enriquecer uma sub-população de interesse particular. Está dentro do escopo da presente invenção para a população de célula alvo ou moléculas derivadas desta sejam pré tratadas antes de testar, por exemplo, a inativação de vírus vivo ou sendo conduzida em um gel. Também deve ser entendido que a amostra biológica pode ser recém colhida ou a mesma pode ter sido armazenada (por exemplo pelo congelamento) antes de testar ou de outro modo tratada antes de testar (tal como submentendo-se à cultura).

A escolha de qual tipo de amostra é a mais adequada para

testar de acordo com o método aqui divulgado será dependente da natureza da situação. Preferivelmente, a dita amostra é uma amostra fecal, lavagem de enema, ressecção cirúrgica ou biópsia de tecido.

5 Como detalhado mais acima, a presente invenção é planejada para caracterizar uma célula ou população celular, que sejam derivadas do intestino grosso, em termos de sua origem anatômica dentro do intestino grosso. Consequentemente, referência à “célula ou população celular” deve ser entendida como uma referência a uma célula individual ou um grupo de células. O dito grupo de células pode ser uma população difusa de células,  
10 uma suspensão de célula, uma população encapsulada de células ou uma população de células que tomam a forma de tecido.

Referência à “expressão” deve ser entendida como uma referência à transcrição e/ou tradução de uma molécula de ácido nucléico. Nesse aspecto, a presente invenção é exemplificada com respeito à triagem  
15 quanto a marcadores de localização que tomam a forma de transcritos de RNA (por exemplo, RNA primário, mRNA, miRNA, tRNA, rRNA). Referência a “RNA” deve ser entendida abranger referência a qualquer forma de RNA, tal como RNA primário, mRNA, miRNA, tRNA ou rRNA. Sem limitar a presente invenção de nenhum modo, a modulação da transcrição de  
20 gene que leva à síntese de RNA aumentada ou diminuída também correlacionar-se-á com a tradução de alguns destes transcritos de RNA (tais como mRNA) para produzir um produto de expressão. Consequentemente, a presente invenção também estende-se ao método de detecciologia que é direcionado para triar quanto aos níveis modulados ou padrões de expressão  
25 dos produtos de expressão do marcador de localização como um indicador da origem proximal ou distal de uma célula ou população celular. Embora um método seja triar quanto aos transcritos de mRNA e/ou o produto de expressão de proteína correspondente, deve ser entendido que a presente invenção não é limitada nesse aspecto e estende-se à triagem quanto a

qualquer outra forma de marcador de localização tal como, por exemplo, um transcrito de RNA primário. Está bem dentro da habilidade da pessoa de habilidade na técnica determinar o alvo de triagem mais apropriado para qualquer situação dada. Preferivelmente, os produtos de expressão de proteína são o subconjunto de análise.

Referência às “moléculas de ácido nucléico” deve ser entendida como uma referência tanto às moléculas de ácido desoxirribonucleico quanto às moléculas de ácido ribonucleico. A presente invenção portanto estende-se tanto a triar diretamente quanto aos níveis de mRNA em uma amostra biológica ou triar quanto ao cDNA complementar que foi transcrito de modo reverso a partir de uma população de mRNA de interesse. Está bem dentro da habilidade da pessoa habilitada na técnica planejar a metodologia direcionada à triagem quanto a DNA ou RNA. Como detalhado acima, o método da presente invenção também estende-se à triagem quanto ao produto de expressão da proteína traduzido a partir do mRNA objeto.

O método da presente invenção é fundamentado na correlação dos níveis de expressão dos marcadores de localização de uma amostra biológica com os níveis proximal e distal normais destes marcadores. O “nível normal” é o nível de marcador expressado por uma célula ou população celular de origem proximal no intestino grosso e o nível de marcador expressado por uma célula ou população celular de origem distal. Consequentemente, existem dois valores de nível normal que são relevantes ao método de detecção da presente invenção. Seria avaliado que estes valores de nível normal são calculados com base nos níveis de expressão de células derivadas do intestino grosso que não exibem uma anormalidade ou predisposição a uma anormalidade que alteraria os níveis ou padrões de expressão destes marcadores.

O nível normal pode ser determinado usando tecidos derivados

do mesmo indivíduo que é o objeto de teste. entretanto, seria avaliado que isto pode ser bastante invasivo para o indivíduo interessado e é portanto provável ser mais conveniente analisar os resultados de teste em relação a um resultado padrão que reflita resultados individuais ou coletivos obtidos de indivíduos saudáveis, outros que não o paciente em questão. Esta última forma de análise é de fato o método preferido de análise visto que o mesmo permite o planejamento de kits que requerem a coleta e análise de uma única amostra biológica, sendo uma amostra de teste de interesse. Os resultados padrão que fornecem os níveis de referência normais proximal e distal podem ser calculados por qualquer meio adequado que seria bem conhecido pela pessoa de habilidade na técnica. Por exemplo, uma população de tecidos normais pode ser avaliada em termos do nível de expressão dos marcadores de localização da presente invenção, fornecendo deste modo um valor padrão ou faixa de valores contra os quais todas as amostras de teste futuras são analisadas. Também deve ser entendido que os níveis de referência normais proximal e distal podem ser determinados a partir dos pacientes de um grupo específico e para o uso com respeito para testar amostras derivadas deste grupo. Consequentemente, podem ser determinados vários valores padrão ou faixas que correspondam aos grupos que diferem com respeito às características tais como idade, gênero, etnicidade ou situação de saúde. O dito “nível normal” pode ser um nível separado ou uma faixa de níveis. Os resultados de amostras biológicas que são testadas são preferivelmente avaliadas contra os níveis de referência normais tanto proximais quanto distais. Um aumento na expressão dos genes do grupo (i), mais acima definidos, em relação aos níveis distais normais é indicativo do tecido de teste sendo de origem proximal enquanto que um aumento na expressão dos genes do grupo (ii), mais acima definidos, em relação aos níveis proximais normais é indicativo do tecido sendo de origem distal. Também seria avaliado, entretanto, que também pode-se conduzir a etapa correlativa definida

analisando-se os resultados que são obtidos do ponto de vista de determinar se o resultado obtido é o mesmo como um nível normal ou distal, indicando deste modo que a amostra de teste é da mesma origem como mostra de nível de referência normal contra a qual a mesma foi avaliada.

5                   Deve ser entendido que o “indivíduo” que é o objeto de teste pode ser qualquer primata. Preferivelmente o primata é um ser humano.

                  Como detalhado mais acima, deve ser entendido que embora a presente invenção seja exemplificada com respeito à detecção de moléculas de ácido nucléico, a mesma também abrange métodos de detecção com base  
10 no teste quanto ao produto de expressão dos marcadores de localização objetos. A presente invenção também deve ser entendida significar métodos de detecção com base na identificação de produto de proteína ou material de ácido nucléico em uma ou mais amostras biológicas. entretanto, deve ser entendido que alguns dos marcadores de localização podem se correlacionar  
15 aos genes ou fragmentos de gene que não codificam um produto de expressão da proteína. Consequentemente, até o grau em que isto ocorre não seria possível testar quanto a um produto de expressão e o marcador objeto deve ser avaliado com base nos perfis de expressão de ácido nucléico.

                  O termo “proteína” deve ser entendido abranger peptídeos,  
20 polipeptídeos e proteínas. A proteína pode ser glicosilada ou não glicosilada e/ou pode conter uma faixa de outras moléculas fundidas, ligadas, unidas ou de outro modo associadas com a proteína tais como aminoácidos, lipídeos, carboidratos ou outros peptídeos, polipeptídeos ou proteínas. Referência aqui a uma “proteína” inclui uma proteína que compreende uma sequência de  
25 aminoácidos assim como uma proteína associada com outras moléculas tais como aminoácidos, lipídeos, carboidratos ou outros peptídeos, polipeptídeos ou proteínas.

                  As proteínas marcadoras de localização da presente invenção podem estar na forma multimérica significando que duas ou mais moléculas

estão associadas juntas. Onde as mesmas moléculas de proteína estão associadas juntas, o complexo é um homomultímero. Um exemplo de um homomultímero é um homodímero. Onde pelo menos uma proteína marcadora está associada com pelo menos uma proteína não marcadora, então o complexo é um heteromultímero tal como um heterodímero.

Referência a um “fragmento” deve ser entendido como uma referência a uma porção das moléculas de ácido nucléico objeto. Isto é particularmente relevante com respeito à triagem quanto a níveis de RNA modulados em amostras de fezes visto que o RNA objeto é provável de ter sido degradado ou de outro modo fragmentado devido ao ambiente do intestino. Pode-se portanto de fato detectar fragmentos da molécula de RNA objeto, fragmentos estes que são identificados em virtude do uso de uma sonda específica adequada.

Em um outro aspecto, a presente invenção fornece um método para determinar a origem anatômica de uma célula ou população celular derivadas do intestino grosso de um indivíduo, que inclui:

avaliar dados de treinamento, que incluem a expressão de dados de treinamento que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de um intestino grosso, e dados de treinamento de origem proximal-distal que representam as associações das ditas células ou populações celulares com as ditas origens proximais-distais;

processar os dados de treinamento usando a análise multivariada para gerar dados de classificação para gerar dados de origem proximal-distal indicativos de uma origem proximal-distal de uma célula ou população celular adicionais derivadas de um intestino grosso, com base nos dados de expressão adicionais que representam a expressão de genes na dita célula ou população celular adicionais.

A presente invenção também fornece um método de detecção

para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

avaliar os dados de expressão primários que representam a expressão de genes em células ou populações celulares derivadas de origens conhecidas proximal-distal de pelo menos um intestino grosso;

processar os dados de expressão selecionados usando a análise multivariada para gerar dados de modelo multivariado representativos de associações entre os dados de expressão selecionados e origens proximais-distais das ditas células ou populações celulares;

receber os segundos dados de expressão que representam a expressão de genes em uma célula ou população celular derivada do intestino grosso de um indivíduo; e

processar os segundos dados de expressão e os dados de modelo multivariado para gerar dados de origem proximal-distal representativos de uma origem proximal-distal das ditas célula ou população celular.

Preferivelmente, a etapa de avaliar os dados de expressão primários inclui avaliar os dados de expressão terciários dos quais os ditos dados de expressão primários são um subconjunto e o método inclui processar os ditos dados de expressão terciários para selecionar um subconjunto dos dados de expressão terciários que corresponde a um subconjunto de genes diferencialmente expressados sozinhos ou em combinação ao longo do eixo proximal-distal do dito intestino grosso, o subconjunto selecionado sendo os ditos dados de expressão primários.

Preferivelmente, o método inclui processar os ditos outros dados de expressão e os ditos dados de classificação multivariados para gerar os ditos dados de origem proximal-distal que representam as ditas origens proximais-distais.

Mais preferivelmente, os dados de expressão selecionados

correspondem aos genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número:

218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

5 o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

o gene detectado pela sonda Affymetrix número: 235733\_at

o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

10 o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

o gene detectado pela sonda Affymetrix número: 230105\_at

o gene detectado pela sonda Affymetrix número: 230269\_at

o gene detectado pela sonda Affymetrix número: 238378\_at

15 o gene detectado pela sonda Affymetrix número: 239814\_at

o gene detectado pela sonda Affymetrix número: 239994\_at

o gene detectado pela sonda Affymetrix número: 240856\_at

o gene detectado pela sonda Affymetrix número: 242414\_at

o gene detectado pela sonda Affymetrix número: 244553\_at

20 o gene detectado pela sonda Affymetrix número: 217320

o gene detectado pela sonda Affymetrix número: 236141

o gene detectado pela sonda Affymetrix número: 236513

o gene detectado pela sonda Affymetrix número: 238143

ABHD5, FAM3B, IGFBP2, POPDC3,

25 ADRA2A, FLJ10884, KCNG1, REG1A,

APOBEC1, FLJ22761, KIFAP3, SLC14A2,

C10orf45, FTHFD, LOC375295, SLC20A1,

C10orf58, GCNT1, ME3, SLC23A3,

CCL8, HAS3, MEP1B, SLC38A2,

CLDN15, HOXB6, NPY6R, SLC9A3,  
 DEFA5, HOXD4, NR1H3, TBCC,  
 EYA2, HSD3B2, HR1H4, ZNF493,  
 OSTalfa,  
 PAP,

5

AFARP1 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
 número 202888\_s\_at,

10

CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 220753\_s\_at,

15

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206754\_s\_at,

CYP2C18, ou o gene ou genes detectados pela sonda  
 Affymetrix número: 208126\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 214421\_x\_at ou 220017\_x\_at,

20

EPB41L3 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 222262\_s\_at ou 224453\_s\_at,

25

FAM45A ou o gene ou genes detectados pela sonda  
 Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 203639\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

5 HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
10 número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

15 MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
20 número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 207558\_s\_at,

25 PRAP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219197\_s\_at,

SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix

número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

5 UGT1A3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221305\_s\_at

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ACACA,    | FMOD,     | LOC151162, | S100P,   |
|    | C13orf11, | FRMD3,    | MCF2L,     | SCGB2A1, |
| 10 | C20orf56, | GALNT5,   | MMP28,     | SCNN1B,  |
|    | CAPN13,   | GARNL4,   | MUC11,     | SHANK2,  |
|    | CLDN8,    | GCG,      | MUC12,     | SIAT2,   |
|    | COLM,     | GNE,      | MUC17,     | SIAT4C,  |
|    | CRIP1,    | HGD,      | MUC5B,     | SIAT7F,  |
| 15 | DNAJC12,  | HOXB13,   | NEDD4L,    | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,     | SLC13A2, |
|    | FBX025,   | IRS1,     | PCDH21,    | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,       | SPINK5,  |
|    | FLJ20989, | KIAA0703, | PRAC,      | SST,     |
| 20 |           | KIAA0830, | PRAC2,     | TFF1,    |
|    |           | KIAA1913, | PTTG1IP,   | TNFSF11, |
|    |           | LAMA1,    | QPRT,      | TPH1,    |
|    |           | LGALS2,   | QSCN6,     | WFDC2,   |
|    |           | RBM24,    |            |          |

25 ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213134\_x\_at ou 205548\_s\_at,

CHSTS ou o gene ou genes detectados pela sonda Affymetrix

número: 221164\_x\_at ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix

número: 205518\_s\_at,

CRYBA2 ou o gene ou genes detectados pela sonda

5 Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix

número: 205927\_s\_at,

DKFZp761N1114 ou o gene ou genes detectados pela sonda

Affymetrix número: 242372\_s\_at

10 EPB41L4A ou o gene ou genes detectados pela sonda

Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix

número: 206070\_s\_at,

FAS ou o gene ou genes detectados pela sonda Affymetrix

15 número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix

número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda

Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

20 FLJ23548 ou o gene ou genes detectados pela sonda

Affymetrix número: 218187\_s\_at

FN1 ou o gene ou genes detectados pela sonda Affymetrix

número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

FOXA2 ou o gene ou genes detectados pela sonda Affymetrix

25 número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix

número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix

número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205490\_s\_at,

HOXD13 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207397\_s\_at,

5 INSM1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda  
Affymetrix número: 212959\_s\_at,

10 MLPH ou o gene ou genes detectados pela sonda Affymetrix  
número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda  
Affymetrix número: 203649\_s\_at,

15 PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix  
número: 207080\_s\_at ou 211253\_x\_at,

20 SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221024\_s\_at,

25 SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at,

|    |        |          |          |            |
|----|--------|----------|----------|------------|
|    | AQP8   | LGALS2   | EFNA1    | ORF51E2    |
|    | CCL11  | C6ORF105 | EMP1     | PROM1      |
| 5  | CLDN8  | CCL11    | FST      | REG3A      |
|    | MMP12  | CD69     | GHR      | SCNN1B     |
|    | P2RY14 | CLC      | HLA-DRB4 | ST3GAL4    |
|    | CCL18  | CPM      | HOXD10   | ST6GALNAC6 |
|    | ACSL1  | DEFA6    | HSD17B2  |            |
| 10 | AGR2   | DHRS9    | HSPCA    |            |
|    | ASPN   |          | IGHD     |            |
|    |        |          | MT1M     |            |

SUD ou o gene ou genes detectados pela Affymetrix sonda  
número: 200832\_s\_at,

15 ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211994\_s\_at,

BTBD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 202946\_s\_at,

20 CA1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205950\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224009\_x\_at ou 223952\_x\_at,

DKFZP564I1171 ou o gene ou genes detectados pela sonda  
Affymetrix número: 225457\_s\_at

25 EIF5A ou o gene ou genes detectados pela sonda Affymetrix  
número: 201123\_s\_at,

IGHD ou o gene ou genes detectados pela sonda Affymetrix  
número: 214973\_x\_at,

PCK1 ou o gene ou genes detectados pela sonda Affymetrix

número: 208383\_s\_at,

RBP4 ou o gene ou genes detectados pela sonda Affymetrix

número: 219140\_s\_at,

TRPM6 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 224412\_s\_at,

UGT1A6 ou o gene ou genes detectados pela sonda

Affymetrix número: 215125\_s\_at.

A presente invenção também fornece um método para  
determinar a origem anatômica de uma célula ou população celular derivada  
10 do intestino grosso de um indivíduo, que inclui:

avaliar os dados de expressão primários que representam a  
expressão de genes em células ou populações celulares derivadas de origens  
conhecidas proximal-distal de pelo menos um intestino grosso; e

processar os dados de expressão primários usando método de  
15 núcleo para gerar dados de classificação para processar os segundos dados de  
expressão que representam a expressão dos ditos genes em pelo menos uma  
célula ou população celular secundárias de um intestino grosso para gerar  
dados de origem proximal-distal que representam a origem proximal-distal da  
dita pelo menos uma célula ou população celular secundárias.

20 Preferivelmente, o método inclui processar os ditos segundos  
dados de expressão e os ditos dados de classificação para gerar dados de  
origem proximal-distal que representam as ditas localizações.

Preferivelmente, o dito método de núcleo inclui uma máquina  
vetorial de sustentação (SVM).

25 Mais preferivelmente, os ditos dados de classificação são  
representativos de genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número:  
218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

o gene detectado pela sonda Affymetrix número: 226432\_at  
 o gene detectado pela sonda Affymetrix número: 231576\_at  
 o gene detectado pela sonda Affymetrix número: 235733\_at  
 o gene detectado pela sonda Affymetrix número: 236894\_at  
 5 o gene detectado pela sonda Affymetrix número: 239656\_at  
 o gene detectado pela sonda Affymetrix número: 242059\_at  
 o gene detectado pela sonda Affymetrix número: 242683\_at  
 o gene detectado pela sonda Affymetrix número: 230105\_at  
 o gene detectado pela sonda Affymetrix número: 230269\_at  
 10 o gene detectado pela sonda Affymetrix número: 238378\_at  
 o gene detectado pela sonda Affymetrix número: 239814\_at  
 o gene detectado pela sonda Affymetrix número: 239994\_at  
 o gene detectado pela sonda Affymetrix número: 240856\_at  
 o gene detectado pela sonda Affymetrix número: 242414\_at  
 15 o gene detectado pela sonda Affymetrix número: 244553\_at  
 o gene detectado pela sonda Affymetrix número: 217320  
 o gene detectado pela sonda Affymetrix número: 236141  
 o gene detectado pela sonda Affymetrix número: 236513  
 o gene detectado pela sonda Affymetrix número: 238143  
 20 ABHD5, FAM3B, IGFBP2, POPDC3,  
 ADRA2A, FLJ10884, KCNG1, REG1A,  
 APOBEC1, FLJ22761, KIFAP3, SLC14A2,  
 ClOorf45, FTHFD, LOC375295, SLC20A1,  
 C10orf58, GCNT1, ME3, SLC23A3,  
 25 CCL8, HAS3, MEP1B, SLC38A2,  
 CLDN15, HOXB6, NPY6R, SLC9A3,  
 DEFA5, HOXD4, NR1H3, TBCC,  
 EYA2, HSD3B2, HR1H4, ZNF493,  
 OSTalfa,

## PAP,

AFARP1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202234\_s\_at,

5 ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
número: 202888\_s\_at,

CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 220753\_s\_at,

10 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206754\_s\_at,

CYP2C18, ou o gene ou genes detectados pela sonda  
Affymetrix número: 208126\_s\_at,

15 CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 214421\_x\_at ou 220017\_x\_at

EPB41L3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222262\_s\_at ou 224453\_s\_at,

20 FAM45A ou o gene ou genes detectados pela sonda  
Affymetrix número: 221804\_s\_at ou 222955\_s\_at

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 203639\_s\_at,

25 GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

5 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

10 MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

15 NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número: 210757\_s\_at,

20 PITX2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219197\_s\_at,

25 SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

UGT1A3 ou o gene ou genes detectados pela sonda

Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda

Affymetrix número: 221305\_s\_at

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ACACA,    | FMOD,     | LOC151162, | S100P,   |
| 5  | Cl3orf1,  | FRMD3,    | MCF2L,     | SCGB2A1, |
|    | C20orf56, | GALNT5,   | MMP28,     | SCNN1B,  |
|    | CAPN13,   | GARNL4,   | MUC11,     | SHANK2,  |
|    | CLDN8,    | GCG,      | MUC12,     | SIAT2,   |
|    | COLM,     | GNE,      | MUC17,     | SIAT4C,  |
| 10 | CRIP1,    | HGD,      | MUC5B,     | SIAT7F,  |
|    | DNAJC12,  | HOXB13,   | NEDD4L,    | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,     | SLC13A2, |
|    | FBX025,   | IRS1,     | PCDH21,    | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,       | SPINK5,  |
| 15 | FLJ20989, | KIAA0703, | PRAC,      | SST,     |
|    |           | KIAA0830, | PRAC2,     | TFF1,    |
|    |           | KIAA1913, | PTTG1IP,   | TNFSF11, |
|    |           | LAMA1,    | QPRT,      | TPH1,    |
|    |           | LGALS2,   | QSCN6,     | WFDC2,   |
| 20 |           |           | RBM24,     |          |

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213134\_x\_at ou 205548\_s\_at,

25 CHSTS ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164\_x\_at ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
número: 205518\_s\_at,

CRYBA2 ou o gene ou genes detectados pela sonda

Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix  
número: 205927\_s\_at,

DKFZp761N1114 ou o gene ou genes detectados pela sonda

5 Affymetrix número: 242372\_s\_at,

EPB41L4A ou o gene ou genes detectados pela sonda  
Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206070\_s\_at,

10 FAS ou o gene ou genes detectados pela sonda Affymetrix  
número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda

15 Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FLJ23548 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

20 FOXA2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix  
número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix  
25 número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205490\_s\_at,

HOXD13 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207397\_s\_at,

INSM1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda  
Affymetrix número: 212959\_s\_at,

5 MLPH ou o gene ou genes detectados pela sonda Affymetrix  
número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,

10 PLA2G2A ou o gene ou genes detectados pela sonda  
Affymetrix número: 203649\_s\_at

PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix  
número: 207080\_s\_at ou 211253\_x\_at,

15 SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at,

20 SLC2A10 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221024\_s\_at

SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at

25 TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at,

AQP8      LGALS2      EFNA1      ORF51E2

|         |          |          |            |
|---------|----------|----------|------------|
| CCL11   | C6ORF105 | EMP1     | PROM1      |
| CLDN8   | CCL11    | FST      | REG3A      |
| MMP12   | CD69     | GHR      | SCNN1B     |
| P2RY14  | CLC      | HLA-DRB4 | ST3GAL4    |
| 5 CCL18 | CPM      | HOXD10   | ST6GALNAC6 |
| ACSL1   | DEFA6    | HSD17B2  |            |
| AGR2    |          | DHRS9    | HSPCA      |
| ASPN    |          | IGHD     |            |
|         |          | MT1M     |            |

- 10 SCD ou o gene ou genes detectados pela sonda Affymetrix  
número: 200832\_s\_at,
- ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211994\_s\_at,
- BTBD3 ou o gene ou genes detectados pela sonda Affymetrix  
15 número: 202946\_s\_at,
- CA1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205950\_s\_at,
- DHRS9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224009\_x\_at ou 223952\_x\_at,
- 20 DKFZP564I1171 ou o gene ou genes detectados pela sonda  
Affymetrix número: 225457\_s\_at,
- EIF5A ou o gene ou genes detectados pela sonda Affymetrix  
número: 201123\_s\_at,
- IGHD ou o gene ou genes detectados pela sonda Affymetrix  
25 número: 214973\_x\_at,
- PCK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 208383\_s\_at,
- RBP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219140\_s\_at,

TRPM6 ou o gene ou genes detectados pela sonda Affymetrix número: 224412\_s\_at,

UGT1A6 ou o gene ou genes detectados pela sonda Affymetrix número: 215125\_s\_at,

5 Ainda mais preferivelmente, os ditos dados de classificação são representativos de um subconjunto de 13 genes.

O mais preferivelmente, os ditos 13 genes são

PRAC,

CCL11,

10 FRZB ou o gene ou genes detectados pela sonda Affymetrix número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix número: 221577\_x\_at,

CLDN8,

15 SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix número: 221577\_x\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix número: 279954\_s\_at,

DEFA5,

20 SPINK5,

OSTalfa,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix número: 202888\_s\_at, e

MUC5.

25 A presente invenção também fornece um método de detecção para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

avaliar os dados de expressão primários que representam a expressão de genes em células ou populações celulares derivadas de origens

conhecidas proximal-distal de pelo menos um intestino grosso;

processar os primeiros dados usando a análise de componentes principais para gerar dados de componente principal que correspondem a pelo menos uma combinação linear da expressão dos ditos genes, os ditos dados de componente principal sendo indicativo de pelo menos uma das ditas origens proximais-distais das ditas células ou populações celulares.

Preferivelmente, as ditas etapas de avaliar os dados de expressão primários inclui avaliar os dados de expressão terciários dos quais os ditos dados de expressão primários são um subconjunto, e o método inclui processar os ditos dados de expressão terciários para selecionar um subconjunto dos dados de expressão terciários selecionados que correspondem a um subconjunto de genes diferencialmente expressados ao longo do eixo proximal-distal do dito pelo menos um intestino grosso, o subconjunto selecionado sendo os ditos dados de expressão primários.

Preferivelmente, Os dados de expressão selecionados correspondem aos genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número: 218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

o gene detectado pela sonda Affymetrix número: 235733\_at

o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

o gene detectado pela sonda Affymetrix número: 230105\_at

o gene detectado pela sonda Affymetrix número: 230269\_at

o gene detectado pela sonda Affymetrix número: 238378\_at

o gene detectado pela sonda Affymetrix número: 239814\_at

o gene detectado pela sonda Affymetrix número: 239994\_at

o gene detectado pela sonda Affymetrix número: 240856\_at

o gene detectado pela sonda Affymetrix número: 242414\_at

5

o gene detectado pela sonda Affymetrix número: 244553\_at

o gene detectado pela sonda Affymetrix número: 217320

o gene detectado pela sonda Affymetrix número: 236141

o gene detectado pela sonda Affymetrix número: 236513

o gene detectado pela sonda Affymetrix número: 238143

10

ABHD5, FAM3B, IGFBP2, POPDC3,

ADRA2A, FLJ10884, KCNG1, REG1A,

APOBEC1, FLJ22761, KIFAP3, SLC14A2,

C10orf45, FTHFD, LOC375295, SLC20A1,

C10orf58, GCNT1, ME3, SLC23A3,

15

CCL8, HAS3, MEP1B, SLC38A2,

CLDN15, HOXB6, NPY6R, SLC9A3,

DEFA5, HOXD4, NR1H3, TBCC,

EYA2, HSD3B2, HR1H4, ZNF493,

OSTalfa,

20

PAP,

AFARP1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
número: 202888\_s\_at,

25

CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 220753\_s\_at,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix

número: 206754\_s\_at,

CYP2C18, ou o gene ou genes detectados pela sonda Affymetrix número: 208126\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
5 número: 214421\_x\_at ou 220017\_x\_at,

EPB41L3 ou o gene ou genes detectados pela sonda Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222262\_s\_at ou 224453\_s\_at,

10 FAM45A ou o gene ou genes detectados pela sonda Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 203639\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
15 número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP 1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

20 HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
25 número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

5 NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número: 210757\_s\_at,

10 PITX2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219197\_s\_at,

15 SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at

20 UGT1A3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221305\_s\_at

|    |           |         |            |          |
|----|-----------|---------|------------|----------|
|    | ACACA,    | FMOD,   | LOC151162, | S100P,   |
|    | C13orf11, | FRMD3,  | MCF2L,     | SCGB2A1, |
| 25 | C20orf56, | GALNT5, | MMP28,     | SCNN1B,  |
|    | CAPN13,   | GARNL4, | MUC11,     | SHANK2,  |
|    | CLDN8,    | GCG,    | MUC12,     | SIAT2,   |
|    | COLM,     | GNE,    | MUC17,     | SIAT4C,  |
|    | CRIP1,    | HGD,    | MUC5B,     | SIAT7F,  |

- DNAJC12, HOXB13, NEDD4L, SIDT1,  
 FAM3C, INSL5, PARP8, SLC13A2,  
 FBX025, IRS1, PCDH21, SLPI,  
 FLJ20366, ISL1, PI3, SPINK5,  
 5 FLJ20989, KIAA0703, PRAC, SST,  
 KIAA0830, PRAC2, TFF1,  
 KIAA1913, PTTG1IP, TNFSF11,  
 LAMA1, QPRT, TPH1,  
 LGALS2, QSCN6, WFDC2,  
 10 RBM24,  
 ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 201097\_s\_at,  
 BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 213134\_x\_at ou 205548\_s\_at,  
 15 CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 221164\_x\_at ou 223942\_x\_at,  
 CMAH ou o gene ou genes detectados pela sonda Affymetrix  
 número: 205518\_s\_at,  
 CRYBA2 ou o gene ou genes detectados pela sonda  
 20 Affymetrix número: 220136\_s\_at  
 CTSE ou o gene ou genes detectados pela sonda Affymetrix  
 número: 205927\_s\_at,  
 DKFZp761N1114 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 242372\_s\_at,  
 25 EPB41L4A ou o gene ou genes detectados pela sonda  
 Affymetrix número: 228256\_s\_at,  
 EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206070\_s\_at,  
 FAS ou o gene ou genes detectados pela sonda Affymetrix

número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix

número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda

5 Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FL323548 ou o gene ou genes detectados pela sonda

Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix

número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

10 FOXA2 ou o gene ou genes detectados pela sonda Affymetrix

número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix

número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix

15 número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix

número: 205490\_s\_at,

HOXD13 ou o gene ou genes detectados pela sonda

Affymetrix número: 207397\_s\_at,

20 INSM1 ou o gene ou genes detectados pela sonda Affymetrix

número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda

Affymetrix número: 212959\_s\_at,

MLPH ou o gene ou genes detectados pela sonda Affymetrix

25 número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix

número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda

Affymetrix número: 203649\_s\_at

PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix  
número: 207080\_s\_at ou 211253\_x\_at,

5 SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at,

10 SLC2A10 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221024\_s\_at,

SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at

15 TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at,

|    |        |          |          |            |
|----|--------|----------|----------|------------|
|    | AQP8   | LGALS2   | EFNA1    | ORF51E2    |
| 20 | CCL11  | C6ORF105 | EMP1     | PROM1      |
|    | CLDN8  | CCL11    | FST      | REG3A      |
|    | MMP12  | CD69     | GHR      | SCNN1B     |
|    | P2RY14 | CLC      | HLA-DRB4 | ST3GAL4    |
|    | CCL18  | CPM      | HOXD10   | ST6GALNAC6 |
| 25 | ACSL1  | DEFA6    | HSD17B2  |            |
|    | AGR2   | DHRS9    | HSPCA    |            |
|    | ASPN   |          | IGHD     |            |
|    |        |          | MT1M     |            |

SCD ou o gene ou genes detectados pela sonda Affymetrix

número: 200832\_s\_at,

ABCB1 ou o gene ou genes detectados pela sonda Affymetrix

número: 211994\_s\_at,

BTBD3 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 202946\_s\_at,

CA1 ou o gene ou genes detectados pela sonda Affymetrix

número: 205950\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix

número: 224009\_x\_at ou 223952\_x\_at,

10 DKFZP56411171 ou o gene ou genes detectados pela sonda  
Affymetrix número: 225457\_s\_at,

EIF5A ou o gene ou genes detectados pela sonda Affymetrix

número: 201123\_s\_at,

IGHD ou o gene ou genes detectados pela sonda Affymetrix

15 número: 214973\_x\_at,

PCK1 ou o gene ou genes detectados pela sonda Affymetrix

número: 208383\_s\_at,

RBP4 ou o gene ou genes detectados pela sonda Affymetrix

número: 219140\_s\_at,

20 TRPM6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224412\_s\_at,

UGT1A6 ou o gene ou genes detectados pela sonda  
Affymetrix número: 215125\_s\_at,

25 A presente invenção também fornece um método para  
determinar a origem anatômica de uma célula ou população celular derivada  
do intestino grosso de um indivíduo, que inclui:

avaliar os dados de expressão primários que representam a  
expressão de genes em uma célula ou população celular derivada de origens  
conhecidas proximal-distal de pelo menos um intestino grosso; e

processar os dados de expressão usando análise de variação canônica para gerar dados de variação canônica indicativos de pelo menos uma das origens proximais-distais das ditas células ou populações celulares.

Preferivelmente, a dita análise de variação canônica inclui a análise de perfil.

Preferivelmente, os ditos subconjuntos de gene incluem genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número: 218888\_s\_at

- 10 o gene detectado pela sonda Affymetrix número: 225290\_at
- o gene detectado pela sonda Affymetrix número: 226432\_at
- o gene detectado pela sonda Affymetrix número: 231576\_at
- o gene detectado pela sonda Affymetrix número: 235733\_at
- o gene detectado pela sonda Affymetrix número: 236894\_at
- 15 o gene detectado pela sonda Affymetrix número: 239656\_at
- o gene detectado pela sonda Affymetrix número: 242059\_at
- o gene detectado pela sonda Affymetrix número: 242683\_at
- o gene detectado pela sonda Affymetrix número: 230105\_at
- o gene detectado pela sonda Affymetrix número: 230269\_at
- 20 o gene detectado pela sonda Affymetrix número: 238378\_at
- o gene detectado pela sonda Affymetrix número: 239814\_at
- o gene detectado pela sonda Affymetrix número: 239994\_at
- o gene detectado pela sonda Affymetrix número: 240856\_at
- o gene detectado pela sonda Affymetrix número: 242414\_at
- 25 o gene detectado pela sonda Affymetrix número: 244553\_at
- o gene detectado pela sonda Affymetrix número: 217320
- o gene detectado pela sonda Affymetrix número: 236141
- o gene detectado pela sonda Affymetrix número: 236513
- o gene detectado pela sonda Affymetrix número: 238143

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ABHD5,    | FAM3B,    | IGFBP2,    | POPDC3,  |
|    | ADRA2A,   | FLJ10884, | KCNG1,     | REG1A,   |
|    | APOBEC1,  | FLJ22761, | KIFAP3,    | SLC14A2, |
|    | C10orf45, | FTHFD,    | LOC375295, | SLC20A1, |
| 5  | C10orf58, | GCNT1,    | ME3,       | SLC23A3, |
|    | CCL8,     | HAS3,     | MEP1B,     | SLC38A2, |
|    | CLDN15,   | HOXB6,    | NPY6R,     | SLC9A3,  |
|    | DEFA5,    | HOXD4,    | NR1H3,     | TBCC,    |
|    | EYA2,     | HSD3B2,   | HR1H4,     | ZNF493,  |
| 10 |           |           | OSTalfa,   |          |
|    |           |           | PAP,       |          |

AFARP1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
15 número: 202888\_s\_at,

CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 220753\_s\_at,

20 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206754\_s\_at,

CYP2C 18, ou o gene ou genes detectados pela sonda  
Affymetrix número: 208126\_s\_at,

25 CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 214421\_x\_at ou 220017\_x\_at,

EPB41L3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222262\_s\_at ou 224453\_s\_at,

FAM45A ou o gene ou genes detectados pela sonda Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix número: 203639\_s\_at,

5 GBA3 ou o gene ou genes detectados pela sonda Affymetrix número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix número: 205541\_s\_at,

10 GULP1 ou o gene ou genes detectados pela sonda Affymetrix número: 215913\_s\_at,

HOXA9 ou o gene ou genes detectados pela sonda Affymetrix número: 205366\_s\_at ou 214551\_s\_at

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix número: 206858\_s\_at,

15 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix número: 210153\_s\_at,

20 MESP1 ou o gene ou genes detectados pela sonda Affymetrix número: 224476\_s\_at,

MOCS1 ou o gene ou genes detectados pela sonda Affymetrix número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix número: 218136\_s\_at ou 221920\_s\_at,

25 NETO2 ou o gene ou genes detectados pela sonda Affymetrix número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix

número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix

número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda

5 Affymetrix número: 219197\_s\_at,

SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix

número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda

Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

10 UGT1A3 ou o gene ou genes detectados pela sonda

Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda

Affymetrix número: 221305\_s\_at

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ACACA,    | FMOD,     | LOC151162, | S100P,   |
| 15 | C13orf11, | FRMD3,    | MCF2L,     | SCGB2A1, |
|    | C20orf56, | GALNT5,   | MMP28,     | SCNN1B,  |
|    | CAPN13,   | GARNL4,   | MUC11,     | SHANK2,  |
|    | CLDN8,    | GCG,      | MUC12,     | SIAT2,   |
|    | COLM,     | GNE,      | MUC17,     | SIAT4C,  |
| 20 | CRIP1,    | HGD,      | MUC5B,     | SIAT7F,  |
|    | DNAJC12,  | HOXB13,   | NEDD4L,    | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,     | SLC13A2, |
|    | FBX025,   | IRS1,     | PCDH21,    | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,       | SPINK5,  |
| 25 | FLJ20989, | KIAA0703, | PRAC,      | SST,     |
|    |           | KIAA0830, | PRAC2,     | TFF1,    |
|    |           | KIAA1913, | PTTG1IP,   | TNFSF11, |
|    |           | LAMA1,    | QPRT,      | TPH1,    |
|    |           | LGALS2,   | QSCN6,     | WFDC2,   |

## RBM24,

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
5 número: 213134\_x\_at ou 205548\_s\_at,

CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164\_x\_at ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
número: 205518\_s\_at,

10 CRYBA2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix  
número: 205927\_s\_at,

15 DKFZp761N1114 ou o gene ou genes detectados pela sonda  
Affymetrix número: 242372\_s\_at,

EPB41L4A ou o gene ou genes detectados pela sonda  
Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206070\_s\_at,

20 FAS ou o gene ou genes detectados pela sonda Affymetrix  
número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201798\_s\_at ou 211864\_s\_at,

25 FLJ20152 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FLJ23548 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

FOXA2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix  
número: 203698\_s\_at,

5 GDF15 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205490\_s\_at,

10 HOXD13 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207397\_s\_at,

INSM1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda  
Affymetrix número: 212959\_s\_at,

15 MLPH ou o gene ou genes detectados pela sonda Affymetrix  
número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,

20 PLA2G2A ou o gene ou genes detectados pela sonda  
Affymetrix número: 203649\_s\_at,

PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix  
número: 207080\_s\_at ou 211253\_x\_at,

25 SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda

Affymetrix número: 221024\_s\_at,

SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at,

5 STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at

TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at

|    |        |          |          |            |
|----|--------|----------|----------|------------|
| 10 | AQP8   | LGALS2   | EFNA1    | ORF51E2    |
|    | CCL11  | C6ORF105 | EMP1     | PROM1      |
|    | CLDN8  | CCL11    | FST      | REG3A      |
|    | MMP12  | CD69     | GHR      | SCNN1B     |
|    | P2RY14 | CLC      | HLA-DRB4 | ST3GAL4    |
| 15 | CCL18  | CPM      | HOXD10   | ST6GALNAC6 |
|    | ACSL1  | DEFA6    | HSD17B2  |            |
|    | AGR2   | DHRS9    | HSPCA    |            |
|    | ASPN   |          | IGHD     |            |
|    |        |          | MT1M     |            |

20 SCD ou o gene ou genes detectados pela sonda Affymetrix  
número: 200832\_s\_at,

ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211994\_s\_at,

BTBD3 ou o gene ou genes detectados pela sonda Affymetrix  
25 número: 202946\_s\_at,

CA1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205950\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224009\_x\_at ou 223952\_x\_at,

DKFZP564I1171 ou o gene ou genes detectados pela sonda Affymetrix número: 225457\_s\_at,

EIFSA ou o gene ou genes detectados pela sonda Affymetrix número: 201123\_s\_at,

5 IGHD ou o gene ou genes detectados pela sonda Affymetrix número: 214973\_x\_at,

PCK1 ou o gene ou genes detectados pela sonda Affymetrix número: 208383\_s\_at,

10 RBP4 ou o gene ou genes detectados pela sonda Affymetrix número: 219140\_s\_at,

TRPM6 ou o gene ou genes detectados pela sonda Affymetrix número: 224412\_s\_at,

UGT1A6 ou o gene ou genes detectados pela sonda Affymetrix número: 215125\_s\_at,

15 A presente invenção também fornece um método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, que inclui:

avaliar dados de treinamento, que inclui a expressão de dados de treinamento que representam a expressão de genes em células ou  
20 populações celulares derivadas de origens conhecidas proximal-distal de pelo menos um intestino grosso, e dados de treinamento de origem proximal-distal que representam as associações das ditas células ou populações celulares com as ditas origens proximais-distais;

processar os dados de treinamento para gerar dados de  
25 classificação que representam uma combinação linear ou não linear de níveis de expressão dos ditos genes, os ditos dados de classificação sendo adaptados para gerar dados de origem proximal-distal adicionais indicativo de uma origem proximal-distal de uma célula ou subpopulação celular adicionais tiradas de um intestino grosso, com base nos dados de expressão adicionais

que representam a expressão dos ditos genes na dita célula ou subpopulação celular adicionais.

Vantajosamente, o dito processamento pode incluir processar os ditos dados de treinamento com GeneRave.

5 Preferivelmente, os ditos subconjuntos de gene incluem genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número:  
218888\_s\_at

10 o gene detectado pela sonda Affymetrix número: 225290\_at  
o gene detectado pela sonda Affymetrix número: 226432\_at  
o gene detectado pela sonda Affymetrix número: 231576\_at  
o gene detectado pela sonda Affymetrix número: 235733\_at  
o gene detectado pela sonda Affymetrix número: 236894\_at  
o gene detectado pela sonda Affymetrix número: 239656\_at  
15 o gene detectado pela sonda Affymetrix número: 242059\_at  
o gene detectado pela sonda Affymetrix número: 242683\_at  
o gene detectado pela sonda Affymetrix número: 230105\_at  
o gene detectado pela sonda Affymetrix número: 230269\_at  
o gene detectado pela sonda Affymetrix número: 238378\_at  
20 o gene detectado pela sonda Affymetrix número: 239814\_at  
o gene detectado pela sonda Affymetrix número: 239994\_at  
o gene detectado pela sonda Affymetrix número: 240856\_at  
o gene detectado pela sonda Affymetrix número: 242414\_at  
o gene detectado pela sonda Affymetrix número: 244553\_at  
25 o gene detectado pela sonda Affymetrix número: 217320  
o gene detectado pela sonda Affymetrix número: 236141  
o gene detectado pela sonda Affymetrix número: 236513  
o gene detectado pela sonda Affymetrix número: 238143

ABHD5, FAM3B, IGFBP2, POPDC3,

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ADRA2A,   | FLJ10884, | KCNG1,     | REG1A,   |
|    | APOBEC1,  | FLJ22761, | KIFAP3,    | SLC14A2, |
|    | C10orf45, | FTHFD,    | LOC375295, | SLC20A1, |
|    | C10orf58, | GCNT1,    | ME3,       | SLC23A3, |
| 5  | CCL8,     | HAS3,     | MEP1B,     | SLC38A2, |
|    | CLDN15,   | HOXB6,    | NPY6R,     | SLC9A3,  |
|    | DEFA5,    | HOXD4,    | NR1H3,     | TBCC,    |
|    | EYA2,     | HSD3B2,   | HR1H4,     | ZNF493,  |
|    |           |           | OSTalfa,   |          |
| 10 |           |           | PAP,       |          |

AFARP1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
número: 202888\_s\_at,

15 CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 220753\_s\_at,

20 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206754\_s\_at,

CYP2C18, ou o gene ou genes detectados pela sonda  
Affymetrix número: 208126\_s\_at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 214421\_x\_at ou 220017\_x\_at,

25 EPB41L3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222262\_s\_at ou 224453\_s\_at,

FAM45A ou o gene ou genes detectados pela sonda

Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 203639\_s\_at,

5 GBA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205541\_s\_at,

GULP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 215913\_s\_at,

10 HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206858\_s\_at,

15 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210153\_s\_at,

MESP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224476\_s\_at,

20 MOCS1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix  
número: 218136\_s\_at ou 221920\_s\_at,

25 NETO2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix  
número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 207558\_s\_at,

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219197\_s\_at,

5 SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

10 UGT1A3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221305\_s\_at

|    |           |           |            |          |
|----|-----------|-----------|------------|----------|
|    | ACACA,    | FMOD,     | LOC151162, | S100P,   |
|    | C13orf11, | FRMD3,    | MCF2L,     | SCGB2A1, |
| 15 | C20orf56, | GALNT5,   | MMP28,     | SCNN1B,  |
|    | CAPN13,   | GARNL4,   | MUC11,     | SHANK2,  |
|    | CLDN8,    | GCG,      | MUC12,     | SIAT2,   |
|    | COLM,     | GNE,      | MUC17,     | SIAT4C,  |
|    | CRIP1,    | HGD,      | MUC5B,     | SIAT7F,  |
| 20 | DNAJC12,  | HOXB13,   | NEDD4L,    | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,     | SLC13A2, |
|    | FBX025,   | IRS1,     | PCDH21,    | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,       | SPINK5,  |
|    | FLJ20989, | KIAA0703, | PRAC,      | SST,     |
| 25 |           | KIAA0830, | PRAC2,     | TFF1,    |
|    |           | KIAA1913, | PTTG1IP,   | TNFSF11, |
|    |           | LAMA1,    | QPRT,      | TPH1,    |
|    |           | LGALS2,   | QSCN6,     | WFDC2,   |
|    |           |           | RBM24,     |          |

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213134\_x\_at ou 205548\_s\_at,

5 CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164\_x\_at ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
número: 205518\_s\_at,

CRYBA2 ou o gene ou genes detectados pela sonda  
10 Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix  
número: 205927\_s\_at,

DKFZp761N1114 ou o gene ou genes detectados pela sonda  
Affymetrix número: 242372\_s\_at,

15 EPB41L4A ou o gene ou genes detectados pela sonda  
Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206070\_s\_at,

FAS ou o gene ou genes detectados pela sonda Affymetrix  
20 número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201798\_s\_at ou 211864\_s\_at,

FLJ20152 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

25 FLJ23548 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719\_s\_at ou 210495\_x\_at ou 212464\_at ou 216442\_x\_at,

FOXA2 ou o gene ou genes detectados pela sonda Affymetrix

número: 210103\_s\_at,

FRZB ou o gene ou genes detectados pela sonda Affymetrix

número: 203698\_s\_at,

GDF15 ou o gene ou genes detectados pela sonda Affymetrix

5 número: 221577\_x\_at,

GJB3 ou o gene ou genes detectados pela sonda Affymetrix

número: 205490\_s\_at,

HOXD13 ou o gene ou genes detectados pela sonda

Affymetrix número: 207397\_s\_at,

10 INSM1 ou o gene ou genes detectados pela sonda Affymetrix

número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda

Affymetrix número: 212959\_s\_at,

MLPH ou o gene ou genes detectados pela sonda Affymetrix

15 número: 218211\_s\_at,

NEBL ou o gene ou genes detectados pela sonda Affymetrix

número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda

Affymetrix número: 203649\_s\_at,

20 PTPRO ou o gene ou genes detectados pela sonda Affymetrix

número: 208121\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix

número: 207080\_s\_at ou 211253\_x\_at,

SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix

25 número: 222258\_s\_at,

SLC28A2 ou o gene ou genes detectados pela sonda

Affymetrix número: 207249\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda

Affymetrix número: 221024\_s\_at,

SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at,

5 TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at,

|    |        |          |          |            |
|----|--------|----------|----------|------------|
|    | AQP8   | LGALS2   | EFNA1    | ORF51E2    |
| 10 | CCL11  | C6ORF105 | EMP1     | PROM1      |
|    | CLDN8  | CCL11    | FST      | REG3A      |
|    | MMP12  | CD69     | GHR      | SCNN1B     |
|    | P2RY14 | CLC      | HLA-DRB4 | ST3GAL4    |
|    | CCL18  | CPM      | HOXD10   | ST6GALNAC6 |
| 15 | ACSL1  | DEFA6    | HSD17B2  |            |
|    | AGR2   | DHRS9    | HSPCA    |            |
|    | ASPN   |          | IGHD     |            |
|    |        |          | MT1M     |            |

SCD ou o gene ou genes detectados pela sonda Affymetrix  
20 número: 200832\_s\_at,

ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211994\_s\_at,

BTBD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 202946\_s\_at,

25 CA1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205950\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224009\_x\_at ou 223952\_x\_at,

DKFZP564I1171 ou o gene ou genes detectados pela sonda

Affymetrix número: 225457\_s\_at,

EIF5A ou o gene ou genes detectados pela sonda Affymetrix  
número: 201123\_s\_at,

IGHD ou o gene ou genes detectados pela sonda Affymetrix  
5 número: 214973\_x\_at,

PCK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 208383\_s\_at,

RBP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219140\_s\_at,

10 TRPM6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224412\_s\_at,

UGT1A6 ou o gene ou genes detectados pela sonda  
Affymetrix número: 215125\_s\_at.

Vantajosamente, os ditos subconjuntos de gene podem incluir  
15 7 genes.

Preferivelmente, os ditos 7 genes são SEC6L1, PRAC,  
SPINK5, SEC6L1, ANPEP, DEFA5, e CLDN8.

Em uma outra forma de realização preferida, os ditos  
subconjuntos de gene são um ou mais dos seguintes subconjuntos:

20 (i) SCD ou o gene ou genes detectados pela sonda Affymetrix  
número: 200832\_s\_at,

MMP12

P2RY14

CLDN8

25 ETNK 1

(ii) PCP4

SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at,

CCL18

RBP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 219140\_s\_at,

DKFZP564I1171

PRAC

5 (iii) EIF5A ou o gene ou genes detectados pela sonda  
Affymetrix número: 201123\_s\_at,

IGFBP2

GDF15 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221577\_s\_at,

10 DKFZP564I1171 ou o gene ou genes detectados pela sonda  
Affymetrix número: 225457\_s\_at,

MUC12

(iv) HLA-DRB4

HOXB13

15 INSL5

(v) ETNK1 ou o gene ou genes detectados pela sonda  
Affymetrix número: 222262\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix  
número: 202888\_s\_at,

20 DEFA5

CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164\_x\_at,

o gene detectado pela Sonda Affymetrix Nº 226432\_at

COLM

25 (vi) SCNN1B

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719\_x\_at,

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224453\_s\_at,

- O gene detectado pela Sonda Affymetrix N<sup>o</sup> 225290\_at  
OSTalfa  
HOXD10  
Sonda N<sup>o</sup> 230269
- 5 (vii) SLC20A1  
HSPCA
- O gene detectado pela Sonda Affymetrix No. 217320\_at  
CCL18  
HOXB13
- 10 (viii) CD69  
OLFM4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 212768\_s\_at,  
UGT1A6 ou o gene ou genes detectados pela sonda  
Affymetrix número: 215125\_s\_at,
- 15 CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 223942\_x\_at,  
O gene detectado pela Sonda Affymetrix No. 231576\_at  
MUC11
- (ix) PLA2G2A ou o gene ou genes detectados pela sonda
- 20 Affymetrix número: 203649\_s\_at,  
REG3A  
CCL13 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206407\_s\_at,  
GCG
- 25 UGT1A3 ou o gene ou genes detectados pela sonda  
Affymetrix número: 208596\_s\_at,  
FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210485\_x\_at,  
MT1M

OR51E2

(x) SLC16A1 ou o gene ou genes detectados pela sonda Affymetrix número: 202236\_s\_at,

WFDC2

5

S10013

PTPRO ou o gene ou genes detectados pela sonda Affymetrix número: 208121\_s\_at,

CCL11

ASPN

10

FAM3B

(xi) EMP1

NEBL ou o gene ou genes detectados pela sonda Affymetrix número: 203962\_s\_at,

TFF1

15

CMAH ou o gene ou genes detectados pela sonda Affymetrix número: 205518\_s\_at,

PYY ou o gene ou genes detectados pela sonda Affymetrix número: 207080\_s\_at,

ECAT11

20

NETO2 ou o gene ou genes detectados pela sonda Affymetrix número: 222774\_s\_at,

(xii) HSD17B2

HGD

CA1 ou o gene ou genes detectados pela sonda Affymetrix número: 205950\_s\_at,

25

CPM

LGALS2

IGHD ou o gene ou genes detectados pela sonda Affymetrix número: 214973\_x\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 216442\_xs\_at,

(xiii) CLC

DEFA6

5 FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 212464\_s\_at,

FST

O gene detectado pela Sonda Affymetrix No. 236513\_at

O gene detectado pela Sonda Affymetrix No. 240856\_at

10 ETNK1

(xiv) PITX2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207558\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224009\_x\_at,

15 DKFZp761N1114

KIAA1913

(xv) GHR

HSD3B2

MEP1B

20 HOXA9 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213651\_s\_at,

TRPM6 ou o gene ou genes detectados pela sonda Affymetrix  
número: 224412\_s\_at,

O gene detectado pela Sonda Affymetrix Nº 239994\_at

25 (xvi) SPINK5

PCK1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 208383\_s\_at,

ADRA2A

NQ01 ou o gene ou genes detectados pela sonda Affymetrix

número: 210519\_s\_at, GBA3

O gene detectado pela Sonda Affymetrix No. 228004\_at

(xvii) SCGB2A1

NR1H4

5 NETO2 ou o gene ou genes detectados pela sonda Affymetrix

número: 218888\_s\_at,

ST6GALNAC6

(xviii) NEBL

PROM1 ou o gene ou genes detectados pela sonda Affymetrix

10 número: 204304\_s\_at,

AGR2

REG1A

UGT1A8 ou o gene ou genes detectados pela sonda

Affymetrix número: 221305\_s\_at,

15 DKFZp761N1114 ou o gene ou genes detectados pela sonda

Affymetrix número: 242372\_s\_at,

(xix) ACSL1

ST3GAL4

GBA3 ou o gene ou genes detectados pela sonda Affymetrix

20 número: 219954\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda

Affymetrix número: 221024\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix

número: 223952\_s\_at,

25 LAMA 1

(xx) EFNA1

BTBD3 ou o gene ou genes detectados pela sonda Affymetrix

número: 202946\_s\_at,

PI3

ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 209994\_s\_at,

C10orf45

BCMP11

5 C6orf105

CAPN13

CPM

O gene detectado pela Sonda Affymetrix No. 236141\_at

O gene detectado pela Sonda Affymetrix No. 238143\_at

10 Referência à “origem proximal-distal” deve ser entendida  
 como uma referência às células ou dados de expressão de uma origem  
 proximal ou uma origem distal. Referência às “células ou subpopulações  
 celulares”, “intestino grosso”, “proximal”, “distal”, “origem”, “localizações”,  
 “gene” e “expressão” deve ser entendida ter o mesmo significado como mais  
 15 acima fornecido.

A presente invenção também fornece um sistema de detecção  
 tendo componentes para executar qualquer um dos métodos acima.

A presente invenção também fornece um meio de  
 armazenagem legível por computador tendo armazenado nele instruções de  
 20 programa para executar qualquer um dos métodos acima.

A presente invenção também fornece um sistema de detecção,  
 que inclui:

meios para avaliar dados de treinamento, que inclui a  
 expressão de dados de treinamento que representam a expressão de genes em  
 25 células ou populações celulares derivadas de pelo menos um intestino grosso,  
 e dados de treinamento de origem proximal-distal que representam as  
 associações das ditas células ou célula populações com as ditas origens  
 proximais-distais;

meios para processar os dados de treinamento usando a análise

multivariada para gerar dados de classificação que representam uma combinação linear ou não linear de níveis de expressão dos ditos genes, os ditos dados de classificação sendo adaptados para gerar dados de origem proximal-distal indicativos de uma origem proximal-distal de uma célula ou população celular adicionais tiradas de um intestino grosso, com base nos dados de expressão adicionais que representam a expressão dos ditos genes na dita célula ou população celular adicionais.

Como detalhado mais acima, o método da presente invenção é útil para identificar células anormais com base em que uma célula de origem distal ou proximal que não esteja expressando as características de perfil de expressão de gene daquela origem anatômica está exibindo um perfil de expressão anormal e portanto deve passar por outra análise para determinar o grau e natureza completos da anormalidade objeto. Por exemplo, algumas células de adenoma ou adenocarcinoma colorretais podem exibir um perfil de expressão no intestino grosso proximal-distal incorreto devido aos eventos de desdiferenciação que são característicos da transformação neoplástica destas células.

Consequentemente, em um outro aspecto é fornecido um método para determinar o início ou predisposição ao início de uma anormalidade celular ou uma condição caracterizada por uma anormalidade celular no intestino grosso, o dito método compreendendo determinar, de acordo com um dos métodos mais acima descritos, o perfil de expressão de um gene proximal-distal de uma amostra biológica derivada de uma origem proximal ou distal conhecida no intestino grosso em que a detecção de um perfil de expressão de gene que seja incompatível com o perfil de expressão de gene do intestino grosso proximal-distal normal é indicativo da anormalidade da célula ou população celular que expressam o dito perfil.

Referência a “perfil de expressão de gene” deve ser entendida como uma referência aos resultados da expressão de gene univariados ou

multivariados mais acima descritos. Por exemplo, o “perfil” pode correlacionar-se com o nível de expressão de um ou mais genes marcadores como mais acima debatido ou o resultado da análise multivariada dos genes e/ou conjuntos de gene mais acima descritos. Consequentemente, referência a

5 “perfil de expressão de gene proximal-distal” é uma referência às características de perfil de expressão de gene de células de origem do intestino grosso proximal e aquelas de células de origem do intestino grosso distal.

Seria avaliado que as células que são o objeto de análise no

10 contexto da presente invenção são de origem proximal ou distal conhecida. Esta informação pode ser determinada por qualquer método adequado mas é mais convenientemente satisfeita isolando-se a amostra biológica de uma localização definida no intestino grosso por intermédio de uma biópsia. Entretanto, outros métodos adequados de colher ou de outro modo determinar

15 a origem anatômica da amostra biológica não são excluídos.

A anormalidade de uma célula ou população celular da amostra biológica está fundamentada na detecção de um perfil de expressão de gene que é incompatível com aquele do perfil que normalmente caracterizaria uma célula de sua origem proximal ou distal particular. Por

20 “incompatível” é intencionado que o nível de expressão de um ou mais dos genes que são analisados não é compatível com aquele que é tipicamente observado em um controle normal.

O método da presente invenção é útil como um teste autônomo ou como um monitor corrente destes indivíduos considerados como estando

25 em risco do desenvolvimento de doença ou como um monitor da eficácia de regimes de tratamento terapêuticos ou profiláticos tais como a ablação de células doentes que são caracterizadas por um perfil de expressão de gene anormal. Nestas situações, o mapeamento da modulação de níveis de expressão ou perfis de expressão de marcador de localização em qualquer

uma ou mais classes de amostras biológicas é um indicador valioso da situação de um indivíduo ou da eficácia de um regime terapêutico ou profilático que esteja correntemente em uso. Consequentemente, o método da presente invenção deve ser entendido estender-se ao monitoramento para a modulação de níveis de ou perfis de expressão de marcador de localização em um indivíduo em relação a um nível normal (como mais acima definido) ou em relação a um ou mais níveis ou perfis de expressão de gene marcador mais no início determinados a partir de uma amostra biológica do dito indivíduo.

Meios de testar quanto aos marcadores de localização expressados objetos em uma amostra biológica podem ser obtidos por qualquer método adequado, que seria bem conhecido pela pessoa de habilidade na técnica, tais como mas não limitados a:

(i) Detecção *in vivo*.

A formação de imagem molecular pode ser usada a seguir da administração das sondas ou reagentes de formação de imagem capazes de divulgar a expressão alterada dos marcadores nos tecidos intestinais.

A formação de imagem molecular (Moore *et al.*, BBA, 1402: 239-249, 1988; Weissleder *et al.*, Nature Medicine 6: 351-355, 2000) é a formação de imagem *in vivo* da expressão molecular que correlaciona-se com as macro-características correntemente visualizadas usando as técnicas de formação de imagem de diagnóstico “clássicas” tais como Raio X, tomografia computadorizada (CT), MRI, tomografia de Emissão de Pósitron (PET) ou endoscopia.

(ii) A detecção de supra-regulagem da expressão de RNA nas células pela Hibridização *In Situ* Fluorescente (FISH), ou em extratos das células pelas tecnologias tais como a Transcriptase Reversa Quantitativa-Reação da Cadeia da Polimerase (QRT-PCR) ou qualificação citométrica de fluxo dos produtos da RT-PCR (Wedemeyer *et al.*, Clinical Chemistry 48: 9 1398-1405, 2002).

(iii) A avaliação de perfis de expressão de RNA a partir de extratos celulares, por exemplo pelas tecnologias de conjunto (Alon *et al.*, Proc. Natl. Acad. Sci. USA: 96, 6745-6750, Junho de 1999).

Um “microarranjo” é um conjunto linear ou multidimensional de regiões preferivelmente separadas, cada uma tendo uma área definida, formada na superfície de um suporte sólido. A densidade das regiões separadas em um microarranjo é determinada pelo número total de polinucleotídeos alvo a serem detectados na superfície de um único suporte de fase sólida, preferivelmente pelo menos cerca de  $50/\text{cm}^2$ , mais preferivelmente pelo menos cerca de  $100/\text{cm}^2$ , ainda mais preferivelmente pelo menos cerca de  $500/\text{cm}^2$ , e ainda mais preferivelmente pelo menos cerca de  $1.000/\text{cm}^2$ . Como aqui usado, um microarranjo de DNA é um conjunto de sondas de oligonucleotídeo colocadas em um chip ou outras superfícies usadas para amplificar ou clonar polinucleotídeos alvo. Visto que a posição de cada grupo particular de sondas no conjunto é conhecida, as identidades dos polinucleotídeos alvo podem ser determinadas com base na sua ligação a uma posição particular no microarranjo.

Desenvolvimentos recentes na tecnologia de microarranjo de DNA torna possível conduzir um ensaio em larga escala de uma pluralidade de moléculas de ácido nucléico alvo em um único suporte de fase sólida. A Pat. U.S. Nº 5.837.832 (Chee *et al.*) e os pedidos de patente relacionados descrevem a imobilização de um conjunto de sondas de oligonucleotídeo para a hibridização e detecção de sequências de ácido nucléico específicas em uma amostra. Os polinucleotídeos alvo de interesse isolados de um tecido de interesse são hibridizados à chip de DNA e as sequências específicas detectadas com base na preferência dos polinucleotídeos alvos e no grau de hibridização nas localizações de sonda separadas. Um uso importante dos conjuntos é na análise da expressão de gene diferencial, onde o perfil de expressão de genes em células ou tecidos diferentes, frequentemente um

tecido de interesse e um tecido de controle, é comparado e quaisquer diferenças na expressão de gene entre os respectivos tecidos são identificados. Tal informação é útil para a identificação dos tipos de genes expressados em um tipo de tecido particular e diagnóstico de condições com base no perfil de expressão.

Em um exemplo, o RNA da amostra de interesse é submetido à transcrição reversa para se obter cDNA rotulado. Ver a Pat. U.S. Nº 6.410.229 (Lockhart *et al.*). O cDNA é depois hibridizado aos oligonucleotídeos ou cDNAs de sequência conhecida ordenada em uma chip ou outra superfície em uma ordem conhecida. Em um outro exemplo, o RNA é isolado de uma amostra biológica e hibridizado a um chip no qual são ancoradas sondas de cDNA. As localizações do oligonucleotídeo ao qual o cDNA rotulado hibridiza fornece informação de sequência no cDNA, enquanto que a quantidade de RNA ou cDNA hibridizados rotulados fornece uma estimativa da representação relativa do RNA ou cDNA de interesse. Ver Schena, *et al.* Science 270: 467-470 (1995). Por exemplo, o uso de um microarranjo de cDNA para analisar os padrões de expressão de gene em câncer humano é descrito por DeRisi, *et al.* (Nature Genetics 14: 457-460 (1996)).

Em uma forma de realização preferida, as sondas de ácido nucléico que correspondem aos ácidos nucléicos objetos são fabricadas. As sondas de ácido nucléico ligadas ao biochip são planejadas como sendo substancialmente complementar aos ácidos nucléicos da amostra biológica tal que a hibridização específica da sequência alvo e das sondas da presente invenção ocorra. Esta complementaridade não precisa ser perfeita, em que pode haver qualquer número de desemparelhamentos de par de base que interferirão com a hibridização entre a sequência alvo e os ácidos nucléicos de filamento único da presente invenção. É esperado que a homologia global dos genes ao nível de nucleotídeo provavelmente serão de cerca de 40% ou maior,

provavelmente de cerca de 60% ou maior, e ainda mais provavelmente de cerca de 80% ou maior; e além disso que haverá sequências contíguas correspondentes de cerca de 8 a 12 nucleotídeos ou mais longas. Entretanto, se o número de mutações é tão grande que nenhuma hibridização possa ocorrer mesmo sob o mínimo de condições de severidade de hibridização, a sequência não é uma sequência alvo complementar. Assim, por “substancialmente complementar” aqui é intencionado que as sondas sejam suficientemente complementar às sequências alvo para hibridizar sob condições de reação normal, particularmente condições de alta severidade.

Uma sonda de ácido nucléico é no geral de filamento único mas pode ser até certo grau de filamento único e até certo grau de filamento duplo. A filamentabilidade da sonda é ditada pela estrutura, composição e propriedades da sequência alvo. No geral, as sondas de oligonucleotídeo variam de cerca de 6, 8, 10, 12, 15, 20, 30 a cerca de 100 bases de comprimento, com cerca de 10 a cerca de 80 bases sendo preferidas, e de cerca de 15 a cerca de 40 bases sendo particularmente preferida. Isto é, no geral genes inteiros são raramente usados como sondas. Em algumas formas de realização, ácidos nucléicos muito longos podem ser usados, até centenas de bases. As sondas são suficientemente específicas para hibridizar a uma sequência padrão complementar sob condições conhecidas por aqueles de habilidade na técnica. O número de desemparelhamentos entre as sequências da sonda e suas sequências padrão (alvo) complementares às quais elas hibridizam durante a hibridização no geral não excede 15%, usualmente não excede 10% e preferivelmente não excede 5%, como determinado por BLAST (ajustes de default).

As sondas de oligonucleotídeo podem incluir as bases heterocíclicas que ocorrem de modo natural normalmente encontradas em ácidos nucléicos (uracila, citosina, timina, adenina e guanina), assim como bases modificadas e análogos de base. Qualquer base modificada ou análogo

de base compatível com a hibridização da sonda a uma sequência alvo é útil na prática da invenção. A porção de açúcar ou glicosídeo da sonda pode compreender desoxirribose, ribose, e/ou formas modificadas destes açúcares, tais como, por exemplo, 2'-O-alkil ribose. Em uma forma de realização preferida, a porção de açúcar é 2'-desoxirribose; entretanto, qualquer porção de açúcar que seja compatível com a capacidade da sonda para hibridizar a um sequência alvo pode ser usada.

Em uma forma de realização, as unidades de nucleosídeo da sonda são ligadas por uma cadeia principal de fosfodiéster, como é bem conhecido na técnica. Em formas de realização adicionais, as ligações internucleotídeo podem incluir qualquer ligação conhecida por uma pessoa de habilidade na técnica que seja compatível com hibridização específica da sonda que inclui, mas não limitada a fosforotioato, metilfosfonato, sulfamato (por exemplo, Pat. U.S. Nº 5.470.967) e poliamida (isto é, ácidos nucleicos peptídicos). Os ácidos nucleicos peptídicos são descritos em Nielsen *et al.* (1991) Science 254: 1497-1500, Pat. U.S. Nº 5.714.331, e Nielsen (1999) Curr. Opin. Biotechnol. 10: 71-75.

Em certas formas de realização, a sonda pode ser uma molécula quimérica; isto é, pode compreender mais do que um tipo de base ou subunidade de açúcar, e/ou as ligações pode ser de mais do que um tipo dentro do mesmo iniciador. A sonda pode compreender uma porção para facilitar a hibridização à sua sequência alvo, como são conhecidos na técnica, por exemplo, intercaladores e/ou aglutinantes de ranhura menor. Variações das bases, açúcares, e cadeia principal internucleosídica, assim como a presença de qualquer grupo pendente na sonda, serão compatíveis com a capacidade da sonda para ligar, em uma maneira específica de semente, com a sua sequência alvo. Um número grande de modificações estruturais, são possíveis dentro destas ligações. Vantajosamente, as sondas de acordo com a presente invenção podem ter características estruturais tal que elas permitam a

amplificação de sinal, tais características estruturais sendo, por exemplo, sondas de DNA ramificadas como aquelas descritas por Urdea *et al.* (Nucleic Acids Symp. Ser., 24: 197-200 (1991)) ou na Patente Européia Nº EP-0225.807. Além disso, métodos sintéticos para preparar as várias bases heterocíclicas, açúcares, nucleosídeos e nucleotídeos que formam a sonda, e a  
5 preparação de oligonucleotídeos de sequência específica pré determinada, são bem desenvolvidos e conhecidos na técnica. Um método preferido para a síntese de oligonucleotídeo incorpora a divulgação da Pat. U.S. Nº 5.419.966.

Sondas múltiplas podem ser planejadas para um ácido nucléico alvo particular para se responsabilizarem pelo polimorfismo e/ou estrutura secundária no ácido nucléico alvo, redundância de dados e outros. Em algumas formas de realização, onde mais do que uma sonda por sequência é usada, sondas de sobreposição ou sondas para seções diferentes de um único gene alvo são usadas. Isto é, duas, três, quatro ou mais sondas, são usadas  
10 para desenvolver em uma redundância para um alvo particular. As sondas podem ser de sobreposição (isto é, ter alguma sequência em comum), ou são específicas para sequências distintas de um gene. Quando polinucleotídeos de alvo múltiplo devam ser detectados de acordo com a presente invenção, cada sonda ou grupo de sonda que corresponde a um polinucleotídeo alvo  
15 particular é situada em uma área separada do microarranjo.

As sondas podem estar em solução, tal como em reservatórios ou na superfície de um microarranjo, ou ligadas a um suporte sólido. Exemplos de materiais de suporte sólido que podem ser usados incluem um plástico, uma cerâmica, um metal, uma resina, um gel e uma membrana. Os  
20 tipos úteis de suportes sólidos incluem placas, pérolas, material magnético, micropérolas, chips de hibridização, membranas, cristais, cerâmicas e monocamadas auto-montantes. Um exemplo compreende uma matriz bidimensional ou tridimensional, tal como um gel ou chip de hibridização com sítios de ligação de sonda múltipla (Pevzner *et al.*, J. Biomol. Struc. &

Dyn. 9: 399-410, 1991; Maskos e Southern, Nuc. Acids Res. 20: 1679-84, 1992).

As lascas de hibridização podem ser usadas para construir conjuntos de sonda muito grandes que são subsequentemente hibridizadas com um ácido nucléico alvo. A análise do padrão de hibridização do chip pode ajudar na identificação da sequência de nucleotídeo alvo. Os padrões podem ser manualmente analisados ou analisados por computador, mas está claro que o sequenciamento posicional pela hibridização leva por si só à análise de computador e automação. Em um outro exemplo, pode-se usar um chip Affymetrix em um suporte estrutural de fase sólida em combinação com um método com base em pérola fluorescente. Já em um outro exemplo, pode-se utilizar um microarranjo de cDNA. Nesse aspecto, os oligonucleotídeos descritos por Lockhart *et al* (isto é, sondas de síntese Affymetrix *in situ* na fase sólida) são particularmente preferidos, isto é, fotolitografia.

Como será avaliado por aqueles na técnica, os ácidos nucléicos podem ser ligados ou imobilizados a um suporte sólido em uma ampla variedade de modos. Por “imobilizado” aqui é intencionada a associação ou ligação entre a sonda de ácido nucléico e o suporte sólido é suficiente para ser estável sob as condições de ligação, lavagem, análise e remoção. A ligação pode ser covalente ou não covalente. por “ligação não covalente” e equivalentes gramaticais aqui é intencionado uma ou mais das interações eletrostáticas, hidrofílicas e hidrofóbicas. Incluídas na ligação não covalente está a ligação covalente de uma molécula, tal como estreptavidina, ao suporte e a ligação não covalente da sonda biotinilada à estreptavidina. Por “ligação covalente” e equivalentes gramaticais aqui é intencionado que as duas porções, o suporte sólido e a sonda, sejam ligados por pelo menos uma ligação, que inclui ligações sigma, ligações pi e ligações de coordenação. As ligações covalentes podem ser formadas diretamente entre a sonda e o suporte sólido ou podem ser formadas por um reticulador ou pela inclusão de um

grupo reativo específico no suporte sólido ou na sonda ou ambas as moléculas. A imobilização também pode envolver uma combinação de interações covalentes e não covalentes.

As sondas de ácido nucléico podem ser ligadas ao suporte sólido pela ligação covalente tal como pela conjugação com um agente de ligação ou pela ligação covalente ou não covalente tais como interações eletrostáticas, ligações de hidrogênio ou ligação de anticorpo-antígeno, ou pelas suas combinações. Os agentes de ligação típicos incluem biotina/avidina, biotina/estreptavidina, proteína A de *Staphylococcus aureus*/fragmento F<sub>c</sub> de anticorpo IgG e quimeras de estreptavidina/proteína A (T. Sano e C. R. Cantor, *Bio/Technology* 9: 1378-81 (1991)), ou derivados ou combinações destes agentes. Os ácidos nucléicos podem ser ligados ao suporte sólido por uma ligação fotoclivável, uma ligação eletrostática, uma ligação de dissulfeto, uma ligação de peptídeo, uma ligação de diéster ou uma combinação destes tipos de ligações. A série também pode ser ligada ao suporte sólido por uma ligação seletivamente liberável tal como 4,4'-dimetoxitritila ou seus derivados. Verificou-se que derivados úteis incluem o ácido 3 ou 4 [bis-(4-metoxifenil)]-metil-benzóico, ácido N-succinimidil-3 ou 4 [bis-(4-metoxifenil)]-metil-benzóico, ácido N-succinimidil-3 ou 4 [bis-(4-metoxifenil)]-hidroximetil-benzóico, ácido N-succinimidil-3 ou 4 [bis-(4-metoxifenil)]-clorometil-benzóico e sais destes ácidos.

No geral, as sondas são ligadas ao biochip em uma ampla variedade de modos, como será avaliado por aqueles na técnica. Como aqui descrito, os ácidos nucléicos podem ser sintetizados primeiro, com a ligação subsequente ao biochip, ou podem ser diretamente sintetizados no biochip.

O biochip compreende um substrato sólido adequado. Por “substrato” ou “suporte sólido” ou outros equivalentes gramaticais aqui é intencionado qualquer material que possa ser modificada para conter sítios individuais separados apropriados para a ligação ou associação das sondas de

ácido nucléico e é avaliável a pelo menos um método de detecção. O suporte de fase sólida da presente invenção pode ser de qualquer material sólido e estruturas adequadas para sustentar a hibridização e síntese de nucleotídeo. Preferivelmente, o suporte de fase sólida compreende pelo menos uma

5 superfície substancialmente rígida em que os iniciadores possam ser imobilizados e a reação de transcriptase reversa realizada. Os substratos com os quais os elementos de microarranjo de polinucleotídeo são estavelmente associados e podem ser fabricados a partir de uma variedade de materiais, que incluem plásticos, cerâmicas, metais, acrilamida, celulose, nitrocelulose,

10 vidro, poliestireno, vinil acetato de polietileno, polipropileno, polimetacrilato, polietileno, óxido de polietileno, polissilicatos, policarbonatos, Teflon.RTM., fluorocarbonetos, náilon, borracha de sílica, polianidridos, ácido poliglicólico, ácido poliláctico, poliortoésteres, polipropilfumerato, colágeno, glicosaminoglicanos e poliaminoácidos. Os substratos podem ser

15 bidimensionais ou tridimensionais na forma, tais como géis, membranas, películas finas, vidros, placas, cilindros, pérolas, pérolas magnéticas, fibras ópticas, fibras tecidas, etc. Uma forma preferida de conjunto é um conjunto tridimensional. Um conjunto tridimensional preferido é uma coleção de pérolas rotuladas. Cada pérola rotulada tem iniciadores diferentes ligados a

20 ela. Os rótulos são detectáveis por meios de sinalização tais como cor (Luminex, Illumina) e campo eletromagnético (Pharmaseq) e os sinais nas pérolas rotuladas podem ser ainda remotamente detectados (por exemplo, usando fibras ópticas). O tamanho do suporte sólido pode ser de qualquer um dos tamanhos de microarranjo padrão, úteis para a tecnologia de microarranjo

25 de DNA e o tamanho pode ser feito de encomenda para se ajustar à máquina particular que é usada para conduzir uma reação da invenção. No geral, os substratos permitem a detecção óptica e não fluorescem apreciavelmente.

Em uma forma de realização, a superfície do biochip e da sonda pode ser derivada com grupos funcionais químicos para a ligação

subsequente dos dois. Assim, por exemplo, o biochip é derivado com um grupo funcional químico que inclui, mas não limitado a, grupos amino, grupos carbóxi, grupos oxo e grupos tiol, com grupos amino sendo particularmente preferidos. Usando estes grupos funcionais, as sondas podem ser ligadas usando grupos funcionais nas sondas. Por exemplo, ácidos nucleicos contendo grupos amino podem ser ligados às superfícies que compreendem grupos amino, por exemplo usando ligadores como são conhecidos na técnica; por exemplo, ligadores homo- ou heterobifuncionais como são bem conhecidos (ver o catálogo de 1994 da Pierce Chemical Company, seção técnica sobre reticuladores, páginas 155 a 200, incorporados aqui por referência). Além disso, em alguns casos, ligadores adicionais, tais como grupos alquila (que incluem grupos substituídos e heteroalquila) podem ser usados.

Nesta forma de realização, os oligonucleotídeos são sintetizados como é conhecido na técnica e depois ligados à superfície do suporte sólido. Como será avaliado por aqueles habilitados na técnica, o terminal 5' ou 3' pode ser ligado ao suporte sólido, ou a ligação pode ser por intermédio de um nucleosídeo interno. Em uma forma de realização adicional, a imobilização ao suporte sólido pode ser muito forte, embora não covalente. Por exemplo, oligonucleotídeos biotinilados podem ser fabricados, que se ligam às superfícies covalentemente revestidas com estreptavidina, resultando em ligação.

As séries podem ser produzidas de acordo com qualquer metodologia conveniente, tal como pré formando os elementos do microarranjo de polinucleotídeo e depois estavelmente associando-os com a superfície. Alternativamente, os oligonucleotídeos podem ser sintetizados na superfície, como é conhecido na técnica. Várias configurações de arranjo diferentes e métodos para a sua produção são conhecidos por aqueles de habilidade na técnica e divulgados na WO 95/25116 e WO 95/35505 (técnicas fotolitográficas), Pat. U.S. Nº 5.445.934 (síntese *in situ* pela fotolitografia),

Pat. U.S. Nº 5.384.261 (síntese *in situ* pelos caminhos de fluxo mecanicamente direcionados); e Pat. U.S. Nº 5.700.637 (síntese pelo manchamento, impressão ou ligação); a divulgação das quais são aqui incorporadas em sua totalidade por referência. Um outro método para ligar

5 DNA às pérolas usa ligandos específicos ligados à extremidade do DNA para ligar às moléculas de ligação de ligando ligadas a uma pérola. Os pares parceiros de ligação de ligando possíveis incluem biotina-avidina/estreptavidina, ou vários pares de anticorpo/antígeno tais como anticorpo de digoxigenina-antidigoxigenina (Smith *et al.*, Science 258: 1122-

10 1126 (1992)). A ligação química covalente de DNA ao suporte pode ser realizada pelo uso de agentes de ligação padrão para ligar o 5'-fosfato no DNA às microesferas revestidas através de uma ligação de fosfoamidato. Métodos para a imobilização de oligonucleotídeos aos substratos de estado sólido são bem estabelecidos. Ver Pease *et al.*, Proc. Natl. Acad. Sci. USA

15 91(11): 5022-5026 (1994). Um método preferido de ligar oligonucleotídeos aos substratos de estado sólido é descrito por Guo *et al.*, Nucleic Acids Res. 22: 5456-5465 (1994). A imobilização pode ser realizada pela síntese de DNA *in situ* (Maskos e Southern, *supra*) ou pela ligação covalente de oligonucleotídeos quimicamente sintetizados (Guo *et al.*, *supra*) em

20 combinação com tecnologias de arranjo robótico.

Além da tecnologia de fase sólida representada pelos arranjos de biochip, a expressão de gene também pode ser quantificada usando arranjos de fase líquida. Um tal sistema é a reação da cadeia da polimerase cinética (PCR). A PCR cinética permite a amplificação e quantificação

25 simultâneas de sequências de ácido nucléico específicas. A especificidade é derivada de iniciadores de oligonucleotídeo sintéticos planejados para aderir preferencialmente as sequências de ácido nucléico de filamento único que se agrupam no sítio alvo. Este par de iniciadores de oligonucleotídeo formam complexos específicos, não covalentemente ligados em cada filamento da

sequência alvo. Estes complexos facilitam a transcrição *in vitro* de DNA de filamento duplo em orientações opostas. A ciclagem de temperatura da mistura de reação cria um ciclo contínuo de ligação, transcrição e re-fusão de iniciador do ácido nucléico aos filamentos individuais. O resultado é um

5 aumento exponencial do produto de dsDNA alvo. Este produto pode ser quantificado em tempo real através do uso de um corante de intercalação ou uma sonda específica de sequência. SYBR(r) Green 1, é um exemplo de um corante de intercalação, que preferencialmente se liga ao dsDNA resultando em um aumento concomitante no sinal fluorescente. As sondas específicas de

10 sequência, tais como usadas com a tecnologia de TaqMan.RTM., consiste de um fluorocromo e uma molécula de extinção covalentemente ligada às extremidades opostas de um oligonucleotídeo. A sonda é planejada para ligar de modo selecionável a sequência de DNA alvo entre os dois iniciadores. Quando os filamentos de DNA são sintetizados durante a reação de PCR, o

15 fluorocromo é clivado da sonda pela atividade de exonuclease da polimerase que resulta no sinal de desextinção. O método de sinalização de sonda pode ser mais específica do que o método do corante de intercalação, mas em cada caso, a força do sinal é proporcional ao produto de dsDNA produzido. Cada tipo de método de quantificação pode ser usado em arranjos de fase líquida de

20 multirreservatórios com cada reservatório representando iniciadores e/ou sondas específicas para as sequências de ácido nucléico de interesse. Quando usado com preparações de RNA mensageiro de tecidos ou linhagens de célula, um arranjo de reações de sonda/iniciador pode simultaneamente quantificar a expressão de produtos de gene múltiplos de interesse. Ver

25 Germer *et al.*, Genoma Res. 10: 258-266 (2000); Heid *et al.*, Genoma Res. 6: 986-994 (1996).

(iv) Medição de níveis de proteína de marcador de localização alterados em extratos de célula, por exemplo pelo imunoensaio.

O teste quanto ao produto de expressão de marcador de

localização proteinácea em uma amostra biológica pode ser realizado por qualquer um de vários métodos adequados que são bem conhecidos por aqueles habilitados na técnica. Os exemplos de métodos adequados incluem, mas não são limitados à triagem de anticorpo de seções de tecido, espécimes de biópsia ou amostras de fluido corporal.

Até o grau em que os métodos com base em anticorpo de diagnóstico são usados, a presença da proteína marcadora pode ser determinada de vários modos tais como pela Western blotting, ELISA ou procedimentos de citometria de fluxo. Estes, naturalmente, incluem ensaios tanto de sítio único quanto de sítio duplo ou “intercalado” dos tipos não competitivos, assim como nos ensaios de ligação competitivos tradicionais. Estes ensaios também incluem a ligação direta de um anticorpo rotulado a um alvo.

Ensaio intercalados estão entre os ensaios mais úteis e habitualmente usados e são favorecidos para o uso na presente invenção. Diversas variações da técnica de ensaio de intercalação existem e todas são intencionadas a serem abrangidas pela presente invenção. Em resumo, em um ensaio avançado típico, um anticorpo não rotulado é imobilizado em um substrato sólido e a amostra a ser testada levada em contato com a molécula ligada. Depois de um período adequado de incubação, por um período de tempo suficiente para permitir a formação de um complexo de anticorpo-antígeno, um segundo anticorpo específico para o antígeno, rotulado com uma molécula repórter capaz de produzir um sinal detectável é depois adicionado e incubado, permitindo o tempo suficiente para a formação de um outro complexo de anticorpo rotulado com anticorpo-antígeno. Qualquer material não reagido é lavado e a presença do antígeno é determinada pela observação de um sinal produzido pela molécula repórter. Os resultados podem ser qualitativos, pela simples observação do sinal visível, ou pode ser quantificada comparando-se com uma amostra de controle. Variações no

ensaio avançado incluem um ensaio simultâneo, em que tanto a amostra quanto o anticorpo rotulado são adicionados simultaneamente ao anticorpo ligado. Estas técnicas são bem conhecidas por aqueles habilitados na técnica, que incluem quaisquer variações menores como estará facilmente evidente.

5                   No ensaio intercalado avançado típico, um primeiro anticorpo tendo especificidade quanto ao marcador ou partes antigênicas deste, é covalente ou passivamente ligado a uma superfície sólida. A superfície sólida é tipicamente vidro ou um polímero, os polímeros mais habitualmente usados sendo celulose, poliacrilamida, náilon, poliestireno, cloreto de polivinila ou  
10 polipropileno. Os suportes sólidos podem estar na forma de tubos, pérolas, discos de microplacas ou qualquer outra superfície adequada para conduzir um imunoensaio. Os processos de ligação são bem conhecidos na técnica e no geral consistem de reticular, ligar covalentemente ou absorver fisicamente, o complexo de polímero-anticorpo é lavado na preparação para a amostra de  
15 teste. Uma alíquota da amostra a ser testada é depois adicionada ao complexo de fase sólida e incubada por um período de tempo suficiente (por exemplo, 2 a 40 minutos) e sob condições adequadas (por exemplo, 25°C) para permitir a ligação de qualquer subunidade presente no anticorpo. A seguir do período de incubação, a fase sólida da subunidade de anticorpo é lavada e secada e  
20 incubada com um segundo anticorpo específico para uma porção do antígeno. O segundo anticorpo é ligado a uma molécula repórter que é usado para indicar a ligação do segundo anticorpo ao antígeno.

Um método alternativo envolve imobilizar as moléculas alvo na amostra biológica e depois expor o alvo imobilizado ao anticorpo  
25 específico que pode ser rotulado ou não com uma molécula repórter. Dependendo da quantidade de alvo e da força do sinal da molécula repórter, um alvo ligado pode ser detectável pela rotulação direta com o anticorpo. Alternativamente, um segundo anticorpo rotulado, específico para o primeiro anticorpo é exposto ao complexo alvo-primeiro anticorpo para formar um

complexo terciário do alvo-primeiro anticorpo-segundo anticorpo. O complexo é detectado pelo sinal emitido pela molécula repórter.

Por “molécula repórter” como usado no presente relatório descritivo, é intencionada uma molécula que, pela sua natureza química, fornece um sinal analiticamente identificável que permite a detecção de anticorpo ligado a antígeno. A detecção pode ser qualitativa ou quantitativa. As moléculas repórter mais habitualmente usadas neste tipo de ensaio são enzimas, fluoróforos ou radionuclídeo contendo moléculas (isto é, radioisótopos) e moléculas quimioluminescentes.

No caso de um imunoensaio de enzima, uma enzima é conjugada ao segundo anticorpo, no geral por meio de glutaraldeído ou periodato. Como será facilmente reconhecido, entretanto, uma ampla variedade de técnicas de conjugação diferentes existe, que são facilmente disponíveis ao técnico habilitado. As enzimas habitualmente usadas incluem peroxidase de rábano, glicose oxidase, beta-galactosidase e fosfatase alcalina, entre outras. Os substratos a serem usados com as enzimas específicas são no geral escolhidos para a produção, na hidrólise pela enzima correspondente, de uma mudança de cor detectável. Os exemplos de enzimas adequadas incluem fosfatase alcalina e peroxidase. Também é possível utilizar substratos fluorogênicos, que produzem um produto fluorescente ao invés dos substratos cromogênicos mencionados acima. Em todos os casos, o anticorpo rotulado com a enzima é adicionado ao complexo de primeiro anticorpo-hapteno, deixado ligar e depois o reagente em excesso é retirado por lavagem. Uma solução contendo o substrato apropriado é depois adicionado ao complexo de anticorpo-antígeno-anticorpo. O substrato reagirá com a enzima ligada ao segundo anticorpo, dando um sinal visual qualitativo, que pode ser ainda quantificado, de modo usual espectrofotometricamente, para dar uma indicação da quantidade de antígeno que estava presente na amostra. “Molécula repórter” também estende-se ao uso de aglutinação de célula ou

inibição da aglutinação tal como as células sangüíneas vermelhas em pérolas de látex e outros.

Alternativamente, compostos fluorescentes, tais como fluoresceína e rodamina, podem ser quimicamente ligados aos anticorpos sem  
5 alterar a sua capacidade de ligação. Quando ativado pela iluminação com luz de um comprimento de onda particular, o anticorpo rotulado com fluorocromo absorve a energia luminosa, induzindo um estado de excitabilidade na molécula, seguido pela emissão da luz em uma cor característica visualmente detectável com um microscópio óptico. Como no EIA, o anticorpo rotulado  
10 fluorescente é deixado ligar ao complexo primeiro anticorpo-hapteno. Depois de separar por lavagem o reagente não ligado, o complexo terciário remanescente é depois exposto à luz do comprimento de onda apropriado a fluorescência observada indica a presença do hapteno de interesse. As técnicas de imunofluorescência e EIA são ambas muito bem estabelecidas na  
15 técnica e são particularmente preferidas para o presente método. Entretanto, outras moléculas repórter, tais como radioisótopo, moléculas quimioluminescente ou bioluminescentes, também podem ser utilizadas.

(v) Determinar a expressão alterada de marcadores de localização de proteína na superfície da célula, por exemplo pela  
20 imunoistoquímica.

(vi) Determinar a expressão de proteína alterada com base em qualquer teste funcional adequado, teste enzimático ou teste imunológico além daqueles detalhados nos pontos (iv) e (vi) acima.

Uma pessoa de habilidade comum na técnica pode determinar,  
25 como um questão de procedimento de rotina, a adequabilidade de aplicar um dado método a um tipo particular de amostra biológica.

Sem limitar a presente invenção de nenhum modo e como detalhado acima, os níveis de expressão de gene podem ser medidos por uma variedade de métodos conhecidos na técnica. Por exemplo, transcrição de

gene ou produtos de tradução podem ser medidos. Os produtos de transcrição de gene, isto é, RNA, podem ser medidos, por exemplo, pelos ensaios de hibridização, ensaios de escoamento, Northern blots, ou outros métodos conhecidos na técnica.

5 Os ensaios de hibridização no geral envolvem o uso de sondas de oligonucleotídeo que hibridizam aos produtos de transcrição de RNA de filamento único. Assim, as sondas de oligonucleotídeo são complementares ao produto de expressão do RNA transcrito. Tipicamente, uma sonda específica de sequência pode ser direcionada para hibridizar ao RNA ou  
10 cDNA. Uma “sonda de ácido nucléico”, como aqui usada, pode ser uma sonda de DNA ou uma sonda de RNA que hibridiza a uma sequência complementar. Uma pessoa de habilidade na técnica saberia como planejar uma tal sonda tal que a hibridização específica de sequência ocorrerá. Uma pessoa de habilidade na técnica saberá ainda como quantificar a quantidade da  
15 hibridização específica de sequência como uma medida da quantidade de expressão de gene para o gene que foi transcrito para produzir o RNA específico.

A amostra de hibridização é mantida sob condições que são suficientes para permitir a hibridização específica da sonda de ácido nucléico  
20 a um produto de expressão de gene específico. “Hibridização específica”, como aqui usada, indica hibridização quase exata (por exemplo, com pouco se algum desemparelhamento). A hibridização específica pode ser realizada sob condições de alta severidade ou condições de severidade moderada. Em uma forma de realização, as condições de hibridização para a hibridização  
25 específica são de severidade alta. Por exemplo, certas condições de severidade alta podem ser usadas para distinguir ácidos nucléicos perfeitamente complementares daqueles de menos complementaridade. “Condições de severidade alta”, “condições de severidade moderada” e “condições de severidade baixa” para as hibridizações de ácido nucléico são explicadas nas

páginas 2.10.1 a 2.10.16 e páginas 6.3.1 a 6.3.6 em *Current Protocols in Molecular Biology* (Ausubel, F. *et al.*, “*Current Protocols in Molecular Biology*”, John Wiley & Sons, (1998), as divulgações inteiras das quais são aqui incorporadas por referência). As condições exatas que determinam a

5 severidade de hibridização dependem não apenas da força iônica (por exemplo, 0,2 x SSC, 0,1 x SSC), temperatura (por exemplo, temperatura ambiente, 42°C, 68°C) e a concentração de agentes desestabilizadores tais como formamida ou agentes desnaturantes tais como SDS, mas também sobre

10 fatores tais como o comprimento da sequência de ácido nucléico, composição base, desemparelhamento percentual entre sequências hibridizadoras e a frequência de ocorrência de subconjuntos daquela sequência dentro de outras sequências não idênticas. Assim, condições equivalentes podem ser

15 determinadas variando-se um ou mais destes parâmetros enquanto se mantém um grau similar de identidade ou similaridade entre as duas moléculas de ácido nucléico. Tipicamente, condições são usadas tais que as sequências de pelo menos cerca de 60%, de pelo menos cerca de 70%, de pelo menos cerca de 80%, de pelo menos cerca de 90% ou de pelo menos cerca de 95% ou mais idênticas entre si permaneçam hibridizadas entre si. Pela variação das

20 condições de hibridização de um nível de severidade na qual nenhuma hibridização ocorre a um nível no qual a hibridização é primeiro observada, as condições que permitirão que uma dada sequência hibridiza (por exemplo, seletivamente) com as sequências mais complementares na amostra podem ser determinadas.

As condições exemplares que descrevem a determinação das

25 condições de lavagem para condições de severidade moderada ou baixa são descritas em Kraus, M. e Aaronson, S., 1991. *Methods Enzymol.*, 200: 546-556; e em, Ausubel *et al.*, *Current Protocols in Molecular Biology*, John Wiley & Sons, (1998). A lavagem é a etapa em que as condições são usualmente ajustadas de modo a determinar um nível mínimo de

complementaridade dos híbridos. No geral, partindo da temperatura mais baixa na qual apenas a hibridização homóloga ocorre, cada°C em que a temperatura de lavagem final é reduzida (mantendo a concentração de SSC constante) permite um aumento de 1% na porcentagem de emparelhamento máximo entre as sequências que hibridizam. No geral, dobrando a concentração de SSC resulta em um aumento na  $T_m$  de cerca de 17°C. Usando estas diretrizes, a temperatura de lavagem pode ser determinada empiricamente para severidade alta, moderada ou baixa, dependendo do nível de emparelhamento procurado. Por exemplo, uma lavagem de severidade baixa pode compreender lavar em uma solução contendo 0,2 x SSC/0,1% de SDS por 10 minutos na temperatura ambiente; uma lavagem de severidade moderada pode compreender lavar em uma solução pré aquecida (42°C) contendo 0,2 x SSC/0,1% de SDS por 15 minutos a 42°C; e uma lavagem de severidade alta pode compreender lavar em solução pré aquecida (68°C) contendo 0,1 x SSC/0,1% de SDS por 15 minutos a 68°C. Além disso, as lavagens podem ser realizadas repetida ou sequencialmente para se obter um resultado desejado como conhecido na técnica. As condições equivalentes podem ser determinadas variando-se um ou mais dos parâmetros dados como um exemplo, como conhecido na técnica, enquanto se mantém um grau similar de complementaridade entre as moléculas de ácido nucléico alvo e o iniciador ou sonda usados (por exemplo, a sequência a ser hibridizada).

Um aspecto relacionado da presente invenção fornece conjunto de ácido nucléico, conjunto este que compreende uma pluralidade de:

(i) moléculas de ácido nucléico que compreendem uma sequência de nucleotídeo que corresponde a qualquer um dos genes marcadores de localização mais acima descritos ou uma sequência que exiba pelo menos 80% de identidade a esta ou um derivado, fragmento, variante ou homólogo funcionais das ditas moléculas de ácido nucléico; ou

(ii) moléculas de ácido nucléico que compreendem uma

sequência de nucleotídeo capaz de hibridização a qualquer uma ou mais das sequências de (i) sob condições de severidade baixa a 42°C ou um derivado, fragmento, variante ou homólogo funcionais das ditas moléculas de ácido nucléico

5 (iii) sondas de ácido nucléico ou oligonucleotídeos que compreendem uma sequência de nucleotídeo capaz de hibridização a qualquer uma ou mais das sequências de (i) sob condições de severidade baixa a 42°C ou um derivado, fragmento, variante ou homólogo funcionais das ditas moléculas de ácido nucléico

10 (iv) proteínas codificadas pelas moléculas de ácido nucléico de (i) ou (ii) ou um derivado, fragmento ou, homólogo destas em que o nível de expressão do dito ácido nucléico é indicativo da origem proximal-distal de uma célula ou subpopulação celular derivadas do intestino grosso.

Referência aqui a uma severidade baixa a 42°C inclui e  
 15 abrange de pelo menos cerca de 1% v/v a pelo menos cerca de 15% v/v de formamida e de pelo menos cerca de 1 M a pelo menos cerca de 2 M de sal para hibridização e pelo menos cerca de 1 M a pelo menos cerca de 2 M de sal para as condições de lavagem. As condições de severidade alternativas podem ser aplicadas onde necessário, tal como severidade média, que inclui e  
 20 abrange pelo menos cerca de 16% v/v, pelo menos cerca de 30% v/v de formamida e de pelo menos cerca de 0,5 M a pelo menos cerca de 0,9 M de sal para hibridização e pelo menos cerca de 0,5 M a pelo menos cerca de 0,9 M de sal para as condições de lavagem, ou severidade alta, que inclui e abrange de pelo menos cerca de 31% v/v a pelo menos cerca de 50% v/v de  
 25 formamida e de pelo menos cerca de 0,01 M a pelo menos cerca de 0,15 M de sal para a hibridização e pelo menos cerca de 0,01 M a pelo menos cerca de 0,15 M de sal para as condições de lavagem. No geral, a lavagem é realizada na  $T_m = 69,3 + 0,41 (G + C)\%$  [19] = ~12°C. Entretanto, a  $T_m$  de um DNA duplex diminui em 1°C com cada aumento de 1% no número de

desemparelhados com pares de base (Bonner *et al* (1973) J Mol. Biol. 81: 123).

Uma biblioteca ou arranjo de marcadores de ácido nucléico ou proteína fornecem informação rica e altamente valiosa. Além disso, dois ou  
5 mais arranjos ou perfis (informação obtida a partir do uso de um arranjo) de tais sequências são ferramentas úteis para comparar um conjunto de teste de resultados com uma referência, tal como uma outra amostra ou calibrador armazenado. No uso de um arranjo, membros de ácido nucléico individuais tipicamente são imobilizados em locais separados e deixadas reagir quanto as  
10 reações de ligação. Iniciadores associados com conjuntos montados de marcadores são úteis para preparar bibliotecas das sequências ou diretamente detectar marcadores de outras amostras biológicas.

Uma biblioteca (ou arranjo, quando da alusão de ácidos nucléicos fisicamente separados que correspondem a pelo menos algumas  
15 sequências em uma biblioteca) de genes marcadores exhibe propriedades altamente desejáveis. Estas propriedades são associadas com condições específicas e podem ser caracterizadas como perfis regulatórios. Um perfil, como aqui chamado refere-se a um conjunto de membros que fornecem informação de diagnóstico do tecido a partir do qual marcadores foram  
20 originalmente derivados. Um perfil em muitos casos compreende uma série de manchas em um arranjo feito de sequências depositadas.

Um perfil do paciente característico é no geral preparado pelo uso de um arranjo. Um perfil de arranjo pode ser comparado com um ou mais outros perfis de arranjo ou outros perfis de referência. Os resultados  
25 comparativos podem fornecer informação rica pertinente a estados de doença, estado de desenvolvimento, receptividade à terapia e outras informação a cerca do paciente.

Um outro aspecto da presente invenção fornece um kit de diagnóstico para ensaiar amostras biológicas que compreendem um agente

para detectar um ou mais marcadores proximais-distais e reagentes úteis para facilitar a detecção pelo agente no primeiro compartimento. Outros meios também podem ser incluídos, por exemplo, para receber uma amostra biológica. O agente pode ser qualquer molécula de detecção adequada.

5                   A presente invenção é ainda descrita pelos seguintes exemplos não limitantes:

#### EXEMPLO 1

#### MAPA DA EXPRESSÃO DE TRANSCRITO DIFERENCIAL NO INTESTINO GROSSO NORMAL

#### 10                   MATERIAIS E MÉTODOS

##### Dados da expressão de gene

Para explorar a variação da expressão de gene humano ao longo do intestino grosso não neoplástico, nós usamos dados de expressão de gene coletados usando o sistema de microarranjo GeneChip®oligonucleotide  
15 da Affymetrix (Santa Clara, CA USA) descrito em [Lipshutz *et al.*, 1999, Nat Genet 21: 20-24]. Os dados são dois conjuntos de dados do Human Genome 133 GeneChip independentes da Affymetrix (Santa Clara, CA USA): uma base de dados de microarranjo comercial grande de HGU-133 A&B chip data para ‘verificação’ e um conjunto de dados de microarranjo HGU-133 Plus 2.0  
20 menor gerado por nós para ‘validação’.

O conjunto de dados maior foi analisado para identificar padrões de expressão de gene e o conjunto de expressão secundário independentemente derivado foi usado para validar estes padrões. Assim, o primeiro conjunto de dados foi minado para a geração de hipótese enquanto  
25 que o segundo conjunto foi usado para testar a hipótese.

Os dados para este estudo são microarranjos de oligonucleotídeo hibridizados ao cRNA rotulado sintetizado a partir de transcritos de poli-A mRNA isolado de espécimes de tecido colorretal. A plataforma Affymetrix que nós usamos é planejada para quantificar

transcritos de mRNA alvo usando um painel de 11 sondas de oligonucleotídeo de emparelhamento perfeito de 25 pares de base (e 11 sondas desemparelhadas), chamado de um conjunto de sonda. Para determinar a relevância biológica da intensidade de ligação do conjunto de sonda, nós

5 anotamos as listas de conjunto de sonda resultantes usando as bibliotecas Affymetrix metafiles e BioConductor mais correntes disponíveis. Nós observamos que existem conjuntos de sonda múltiplos na plataforma de microarranjo teoricamente reativos a qualquer ‘gene’ alvo dado. Como nosso

10 foco é explorar as dinâmicas da expressão de transcrito ao longo do intestino grosso e não elucidar os mecanismos genômicos subjacentes, não exploramos mais este fenômeno.

Não obstante, estes detalhes de anotação fundamental devem ser considerados quando da interpretação da relevância biológica destes dados e nós alertamos o leitor (e outros pesquisadores que usam estas técnicas) para

15 serem cautelosos do perigo de usar os termos ‘genes’ e ‘conjunto de sonda’ intercambiavelmente.

#### Conjunto de dados de ‘Verificação’

A expressão de genes e descrições clínicas para 184 espécimes de tecido colorretal foram adquiridos da GeneLogic Inc. (Gaithersburg, MD,

20 USA).

Os dados de microarranjo de tecido individuais foram selecionados com as seguintes características: mucosa colorretal não neoplástica (confirmada pela histologia) a partir de espécime de tecido de outro modo saudável (isto é, nenhuma evidência de inflamação ou outra

25 doença no sítio de espécime) com um sítio anatomicamente identificável de ressecção designado como um de: ceco, cólon ascendente, cólon descendente, cólon sigmóide ou reto.

Para cada tecido selecionado da base de dados geneLogic, nós recebemos arquivos eletrônicos de dados brutos contendo um total de 44.928

conjuntos de sonda (HGU133A e HGU133B, combinados), descritores experimentais e clínicos para cada tecido e imagens de microscópio digitalmente arquivados das preparações de histologia. Cada registro de dados foi manualmente avaliado quanto a consistência clínica e uma amostra de registros foi aleatoriamente escolhida para auditoria de histopatologia usando imagens de histologia digitalmente arquivadas. Uma análise de controle de qualidade foi realizada para identificar e remover arranjos que não atinjam as medidas de controle de qualidade essenciais como definido pelo fabricante. [Affymetrix, 2001; Wilson e Miller, 2005, Bioinformatics].

Os níveis de expressão de gene foram calculados pelas técnicas de normalização tanto da Microarray Suite (MAS) 5.0 (Affymetrix) quanto da Robust Multichip Average (RMA). [Affymetrix, 2001; Hubbell *et al.*, 2002, Bioinformatics 18: 1585-1592; Irizarry *et al.*, 2003, Nucleic Acids Res 31: e15] Os dados normalizados MAS foram usados para realizar as rotinas de controle de qualidade padrão e o conjunto de dados finais foi normalizado com RMA para todas as análises subsequentes.

#### Conjunto de dados de 'Validação'

Os espécimes colorretais no conjunto de 'validação' foram coletados a partir de um banco de tecido hospitalar de orientação terciário no metropolitan Adelaide (Repatriation General Hospital e Flinders Medical Centre). O banco de tecido e este projeto foram aprovados pelo Research and Ethics Committee of the Repatriation General Hospital e o consentimento do paciente foi recebido para cada tecido estudado. A seguir da ressecção cirúrgica, os espécimes foram colocados em um receptáculo estéril e coletado da sala de operações. O tempo de ressecção operativa para a coleta da sala de operações foi variável mas não mais do que 30 minutos. As amostras, aproximadamente 125 mm<sup>3</sup> (5 x 5 x 5 mm) no tamanho, foram tiradas do tecido macroscopicamente normal tão distante da patologia quanto possível, definidos tanto pela região colônica assim como pela distância proximal ou

distal em relação à patologia. Os tecidos foram colocados em criofrascos, depois imediatamente imersos em nitrogênio líquido e armazenados a -150°C até o processamento.

5 As amostras congeladas foram processadas pelos autores usando protocolos padrão e kits comercialmente disponíveis. Em resumo, tecidos congelados foram homogeneizados usando um moinho de pérolas de carbeto (Mixer Mill MM 300, Qiagen, Melbourne, Austrália) na presença de Tampão de Lise de RNA Promega SV gelado (Promega, Sydney, Austrália) para neutralizar a atividade de RNase. Os lisados de tecido homogeneizado  
10 para cada tecido foram aliquotados aos volumes convenientes e armazenados a -80°C. O RNA total foi extraído de lisados de tecido usando o sistema de RNA Promega SV Total de acordo com as instruções do fabricante e a integridade foi avaliada visualmente pela eletroforese em gel.

Para medir a expressão relativa de transcritos de mRNA, as  
15 amostras de RNA de tecido foram analisadas usando Affymetrix HG U133 Plus 2.0 GeneChips (Affymetrix, Santa Clara, CA USA) de acordo com os protocolos do fabricante [Affymetrix, 2004]. O cRNA rotulado com biotina foi preparado usando 5 µg (1,0 µg/µL) de RNA total (aprox. 1 µg de mRNA) com o kit de “cDNA de um ciclo” (incorporando um iniciador de T7-  
20 oligo(dT)) e o kit de rotulação geneChip IVT. O cRNA transcrito *in vitro* foi fragmentado (20 µg) e analisado quanto aos propósitos de controle de qualidade pela espectrofotometria e eletroforese em gel antes da hibridização. Finalmente, um coquetel de hibridização foi preparado com 15 µg de cRNA (0,5 µg/µL) e hibridizados aos microarranjos HG U133 Plus 2.0 por 16 horas  
25 a 45°C em uma câmara de Hibridização Affymetrix 640. Cada amostra de cRNA foi reforçada com controles de hibridização eucarióticos padrão para monitorar a qualidade.

Os microarranjos hibridizados foram tingidos com estreptavidina ficoeritirina e lavados com uma solução contendo anticorpos

anti-estreptavidina biotinilados usando o Affymetrix Fluidics Station 450. Finalmente, os microarranjos tingidos e lavados foram escaneados com o Affymetrix Scanner 3000.

O pacote de software da Affymetrix foi usado para transformar os arquivos de imagem de microarranjo bruto para o formato digitalizado. Como para o conjunto Discovery acima, os níveis de expressão de gene para o conjunto de dados de validação foram gerados usando MAS 5.0 (Affymetrix) com propósitos de controle de qualidade e com o método da normalização RMA para dados de expressão.

#### Análise Estatística

Como mostrado na Figura 10, um sistema de detecção inclui módulos de detecção 1002 a 1007, que incluem um módulo de máquina vetorial de sustentação (SVM) 1002, um analisador de perfil 1004, um analisador principal de componente 1006 e um módulo classificador 1007. O sistema de detecção executa os métodos de detecção que geram dados de localização representativos da origem ao longo do eixo proximal-distal do intestino grosso de uma célula, ou população de célula, deste intestino. Os dados de localização são gerados pelo processamento de dados de expressão de gene que representam a expressão de genes dentro desta célula ou população de célula. Na forma de realização descrita, o sistema de detecção é um sistema de computador padrão tal como um sistema de computador com base no Intel IA-32 e os módulos de detecção 1002 a 1007 são implementados como módulos de software armazenados em armazenagem não volátil (por exemplo, disco rígido) 1020 associados com o sistema de computador. Entretanto, estará evidente que pelo menos partes dos módulos de detecção 1002 a 1007 ou os métodos de detecção aqui descritos alternativamente podem ser implementados como um ou mais componentes de hardware dedicado, tal como circuitos integrados específicos de aplicação (ASICs).

Na forma de realização descrita, o sistema de detecção

também inclui módulos C++ 1008 para fornecer suporte de linguagem C++, que inclui bibliotecas C++ e um módulo R 1012 que fornece suporte para a linguagem de programação estatística R e a biblioteca MASS descrita em [Venables e Ripley, 2002] e disponível do depósito de fonte aberta CRAN em

5 <http://cran.r-project.org>. O sistema também inclui a aplicação do software BioConductor 1010 disponível da <http://www.bioconductor.org>, que, junto com o analisador de perfil 1004 e o analisador de componente principal 1006, são implementados na linguagem de programação R, como descrito em <http://www.r-project.org>. O SVM 1002 é implementado na linguagem de

10 programação C++. O módulo classificador 1007 é a aplicação geneRave, como descrito em <http://www.bioinformatics.csiro.au/products.shtml> e referências aí fornecidas. O sistema também inclui o Microarranjo Suite (MAS) 5.0 1014 e a aplicação de normalização Robust Multichip Average (RMA) 1016, ambos disponíveis da Affymetrix e descritos em

15 <http://www.affymetrix.com>. As aplicações de software são executadas sob o controle de um sistema de operação padrão 1018, tal como Linux ou MacOS 10.4 e o sistema de computador inclui componentes de hardware de computador padrão, que inclui pelo menos um processador 1022, memória de acesso aleatória 1024, um teclado 1026, um dispositivo de indicação padrão

20 tal como um mouse 1028 e uma tela de vídeo 1030, todos os quais são interconectados por intermédio de um barramento de sistema 1032, como mostrado.

Os métodos de detecção incluem métodos de classificação da forma geral da Figura 11. Primeiro, na etapa 1102, o sistema recebe ou de

25 outro modo acessa dados de expressão que representam a expressão de genes em células de origem proximal-distal conhecida. Na etapa 1104, uma multivariada ou outra forma de classificação ou método de decisão é aplicado aos dados de expressão para gerar dados de classificação, como descrito abaixo. Tipicamente, os dados de expressão representam a expressão de genes

que, sozinhos ou em combinação, já são conhecidos como sendo diferencialmente expressados ao longo do eixo proximal-distal do intestino grosso. Entretanto, o método também pode ser usado para identificar tais genes e/ou combinações de gene, como descrito abaixo. Na etapa 1106, os dados de classificação são aplicados aos dados de expressão adicionais que representam a expressão dos mesmos genes em uma célula de origem desconhecida para prognosticar a origem proximal-distal desta célula ao longo do intestino grosso.

Além disso, estará evidente àqueles habilitados na técnica que a função classificadora ou discriminante resultante representada pelos dados de classificação inicialmente gerados pode ser ajustada com base nos princípios teóricos de decisão para melhorar os resultados de classificação e a sua utilidade. Por exemplo, um ponto de vista anterior na probabilidade de resultados pode ser incorporado, e/ou uma superfície de decisão pode ser modificada com base nos custos diferentes de casos de má classificação. Estes e outros métodos relevantes de teoria de decisão, minimizando perda de funções e custo de má classificação são descritos em [Krzanowski e Marriott, 1995].

Para todas as análises estatísticas, nós usamos o software de fonte aberta disponível da BioConductor para o ambiente de estatísticas R (R sendo uma implementação de fonte aberta do ambiente de análise estatística S). (Bioconductor, [www.bioconductor.org](http://www.bioconductor.org)) [Gautier *et al.*, 2004, Bioinformatics 20: 307-315; Gentleman *et al.*, 2004, Genome Biol 5: R80].

Os métodos lineares usados para gerar e processos lineares e combinações não lineares de níveis de expressão de gene, que incluem a regressão linear, regressão linear múltipla, análise discriminante linear, regressão de logística, modelos lineares generalizados e análise de componentes principais, são todos descritos em [Hastie, 2001], por exemplo. Estes métodos são implementados em R.

Os gradientes de expressão de gene foram analisados usando três técnicas analíticas. Primeiro, nós comparamos a variação da expressão do gene de genes individuais ao longo do intestino grosso na maneira univariada usual. Em seguida, nós exploramos ainda aqueles genes particulares que exibem diferenças de expressão estatisticamente significantes com modelos lineares para comparar mudança de expressão dicótoma (proximal vs. distal) com um modelo gradual (multissegmento) de mudança. Finalmente, nós aplicamos técnicas multivariadas para entender a variação de expressão ampla de genoma sutil ao longo do eixo proximal-distal. Tais variações de expressão amplas de genoma foram interrogadas usando métodos não paramétricos como descritos em [Ripley, 1996], que incluem métodos do vizinho mais próximo.

#### Mapas de expressão de gene individuais

##### Expressão diferencial Univariada

Os transcritos de gene diferencialmente expressados entre o intestino grosso proximal e distal foram identificados usando um teste t moderado implementado na biblioteca de Biocondutor ‘limma’ para R [Smyth, 2005]. Estimativas de significância (valores p) foram corrigidas para ajustar quanto ao teste de hipótese múltipla (MET) usando a correção de Bonferroni conservativa. Os subconjuntos de tecido limitados ao ceco vs. o reto foram similarmente testados.

Os transcritos de gene identificados como sendo diferencialmente expressados também foram avaliados nos espécimes de ‘Validação’ em uma base de conjunto de sonda-ao lado de-conjunto de sonda usando testes t modificados. Para avaliar a significância do número total de conjuntos de sonda diferenciais que foram do mesmo modo diferenciais nos dados de validação, o número de conjuntos de sonda ‘validados’ foram comparados a uma distribuição nula estimada usando uma simulação de Monte Carlo.

Comparação modelo de intestino grosso multi-segmentado vs. intestino grosso bi-segmentado

Para avaliar a natureza da variação da expressão de gene inter-segmento nós analisamos conjuntos de sonda diferencialmente expressados quanto ao ajuste relativo aos modelos lineares em uma matriz de trabalho multi-segmentada vs. uma bi-segmentada. A meta desta análise é explorar se a expressão intersegmentada de conjuntos de sonda que são conhecidas como sendo diferencialmente expressadas entre as extremidades de terminal do intestino grosso são melhores modeladas por um modelo linear de cinco segmentos que aproxima uma graduação contínua ou por um gradiente dicotômico ‘proximal’ vs. ‘distal’ mais simples. Como nossos dados são apenas identificados pela designação de segmento colorretal e não por uma medição contínua ao longo do comprimento do intestino grosso, nós aproximamos o modelo contínuo usando as localizações de segmento de tecido. Nós escolhemos conjuntos de sonda que são diferencialmente expressados entre os segmentos mais terminais (ceco e reto) de modo a maximizar a probabilidade de identificar transcritos que variam ao longo do eixo proximal-distal do intestino grosso.

Nós primeiro modelamos a expressão destes conjuntos de sonda ao longo do eixo proximal-distal do intestino grosso usando um modelo linear robusto de cinco fatores de acordo com uma matriz indicadora definida pelo segmento colorretal para cada tecido. Para este modelo cada tecido foi designado pelas localizações de biópsia a um de: ceco, ascendente, descendente, sigmóide ou reto. (Pelas razões descritas abaixo, tecidos transversais não foram incluídos nesta análise). Este modelo de cinco segmentos foi depois comparado com um modelo linear robusto de dois fatores com uma matriz de planejamento que corresponde às regiões proximal e distal teóricas do intestino grosso. Os mesmos dados foram usados para ambas as comparações de modelo, entretanto para os dois modelos de

segmento, o primeiro fator (que corresponde aos tecidos proximais) incluiu todos os tecidos do ceco e cólon ascendente enquanto que o segundo fator (que corresponde ao intestino grosso distal) incluiu todos os tecidos dos segmentos descendente, sigmóide e reto.

5 Quando da comparação destes modelos distintos para cada conjunto de sonda, nós usamos um teste F para avaliar a hipótese  $H_a$  de que o ajuste melhorado (residual de regressão reduzido) fornecido pelo modelo de cinco segmentos mais complexo foi significativamente melhor do que o modelo de dois segmentos mais simples. Uma redução residual não  
10 significativa indica uma insuficiência para rejeitar a hipótese nula  $H_0$ : de que não existe valor inerente para adotar um modelo de cinco segmentos mais complexo em relação à alternativa mais simples.

Mapeamento do padrão de expressão de gene multivariado

## RESULTADOS

15 Coleta de dados da expressão de gene

Conjuntos de Dados de verificação e Validação

Um conjunto de dados de verificação foi gerado usando dados da hibridização de cRNA aos microarranjos HG U133A/B GeneChip da Affymetrix que foram adquiridos da GeneLogic Inc.

20 Os dados de 184 tecidos normais que atingem os critérios de inclusão e os critérios de garantia de qualidade para o HG U133A/B GeneChip foram analisados e usados para a geração de hipótese. Os tecidos compreenderam subconjuntos de segmento como seguem: 29 cecos, 45 ascendentes, 13 descendentes, 54 sigmóides e 43 retos. Para cada tecido,  
25 44.928 conjuntos de sonda foram corrigidos no fundo e normalizados usando pré-processamento de RMA.

Para construir o conjunto de dados de 'validação', 19 HG U133 Plus2.0 GeneChips foram hibridizados a cRNA rotulado preparado a partir de 8 espécimes de tecido proximal e 11 espécimes distais. Devido aos

parâmetros de controle de qualidade severos para a aceitabilidade de tecido e GeneChip, estes conjuntos de dados de validação não incluíram tecidos suficientes para explorar modelos de segmento múltiplo. Cada microarranjo mediu a expressão de transcrito para 54.675 conjuntos de sonda.

5                   A junção teórica entre o intestino grosso proximal e distal é aproximadamente dois terços do comprimento do cólon transversal medido a partir da dobra hepática. [Yamada e Alpers, 2003, *supra*] Visto que os dados de amostra não foram específicos para a distância ao longo do cólon transversal, estes tecidos foram excluídos da análise de verificação.

10                   Variação de gene ao longo do intestino grosso

Mudanças da expressão de gene individual

Expressão diferencial univariada

15                   Para explorar o ponto de divisão ‘natural’ entre os segmentos anatômicos do intestino grosso, nós medimos o número absoluto de mudanças de expressão do conjunto de sonda quando o ‘divisor’ hipotético foi movido gradualmente do ceco para o reto. A Figura 1 mostra o número de conjuntos de sonda que foi diferencialmente expressado para todas as combinações contínuas inter-segmento. Embora não estatisticamente significativa, o número máximo de diferenças de conjunto de sonda, 206, ocorre quando as regiões

20                   proximal e distal são divididas entre os segmentos ascendentes e descendentes. Visto que este ponto divisor é compatível tanto com o nosso entendimento de desenvolvimento embrionário quanto a separação usual dos segmentos proximal e distal, nosso trabalho assume que os tecidos proximal e distal são separados neste modelo.

25                   Um total de 206 conjuntos de sonda, que correspondem a aproximadamente 154 alvos de gene conhecidos, foram diferencialmente expressados mais altos nas amostras colorretais proximal ou distal comparadas com a região correspondente (Bonferroni corrigido  $p < 0,05$ ). Destes 206 conjuntos de sonda, 31 (16,5%) também foram diferencialmente

expressados nos dados de validação com uma diferença significativa (31/206,  $p < 0,05$  pela estimativa Monte Carlo).

Um total de 115 conjuntos de sonda foram diferencialmente expressados apenas entre tecidos selecionados do ceco ( $n = 29$ ) e do reto ( $n = 43$ ). Embora 102 (89%) destes conjuntos de sonda estejam incluídos nos 206 conjuntos de sonda que diferem entre intestino grosso proximal e distal descrito acima, a expressão de gene do ceco vs. reto é útil, em princípio, para isolar aqueles transcritos que são diferentes entre as extremidades mais terminais do intestino grosso. Neste subconjunto, 28 conjuntos de sonda (24,3%) foram do mesmo modo diferencialmente expressados no reto vs. o ceco nos dados de validação (28/115,  $p < 10^{-5}$  pela estimativa Monte Carlo).

Os conjuntos de sonda diferencialmente expressados e as estatísticas de diferença para os conjuntos de sonda com expressão elevada em tecidos proximal e distal são mostrados nas Tabelas 1 e 2, respectivamente. A Figura 2 compara o número de conjuntos de sonda expressados significativamente mais altos no intestino proximal ( $n = 94$ ) ou distal ( $n = 126$ ) (ou ceco e reto), respectivamente.

#### Modelos de expressão de gene de segmento múltiplo

Uma análise quanto a expressão diferencial também foi feita para todas as cinco transições inter-segmento na ordem do ceco para o reto (isto é, ceco vs. ascendente, ascendente vs. transversal, etc.). De modo interessante, nenhum transcrito foi diferencialmente expressado a um grau significativo entre quaisquer dois segmentos contíguos (teste t moderado;  $p < 0,05$ ).

Para explorar a natureza exata destas mudanças de expressão de transcrito de gene, nós construímos e comparamos modelos lineares robustos adaptados para os dados de expressão com base nas localizações para cada amostra de tecido. Dois modelos lineares robustos da expressão de conjunto de sonda univariado foram comparados para cada um dos 115

conjuntos de sonda diferencialmente expressado entre os dois segmentos de terminal do intestino grosso, o ceco e reto. Em particular, nós interrogamos se a expressão destes transcritos que foram diferencialmente expressado entre estes segmentos de terminal foram melhor explicados (em termos de ajuste residual) por um modelo de dois segmentos simples ou pelo modelo de cinco segmentos mais descritivo.

Dos 115 conjuntos de sonda diferencialmente expressados, a análise falhou em rejeitar a hipótese nula de que um modelo complexo não melhora significativamente o ajuste de modelo para os dados de expressão do gene observado para 65 (57%) dos casos (teste F,  $p > 0,05$ ). Assim, mais do que metade destes transcritos diferencialmente expressados ao longo do intestino grosso são satisfatoriamente modelados pelo modelo de expressão de dois segmentos por meio do qual a expressão é dicótoma e definida pelas localizações proximal vs. distal. O conjunto de sonda mais diferencialmente expressado entre o ceco e reto é o transcrito para PRAC. Uma comparação dos modelos de dois segmentos e de segmento múltiplo para este transcrito são mostrados na Figura 3.

Para os 50 (43%) conjuntos de sonda remanescentes, a hipótese nula foi rejeitada ( $p < 0,05$ ), sugerindo que um modelo de cinco fatores dependente das localizações de segmento de fato melhora a eficácia preditiva de tal expressão de transcritos ao longo do eixo proximal-distal em uma maneira significativa. A inspeção destes modelos confirma que a maioria dos modelos são crescentemente monotônicos ou decrescentemente monotônicos em tecidos progredindo ao longo do intestino grosso.

De maneira interessante, 41 (82%) dos 50 modelos de segmento múltiplo mostram um aumento gradual através do intestino grosso enquanto que apenas 9 modelos (18%) indicam uma diminuição gradual da expressão proximal para distal (mostrado na Figura 4). Os modelos para ambos os transportadores de soluto orgânico, alfa (OSTalfa) e homeobox

gene B13 (HOXB13) são significativamente melhorados com o modelo de cinco segmentos como ilustrado na Figura 5.

Padrões da expressão de gene ao longo do intestino grosso

Além das análises de mudanças de gene individuais ao longo do intestino grosso, nós usamos técnicas analíticas multivariadas para explorar padrões de mudanças de gene ao longo do eixo proximal-distal.

#### Análise de Componentes Principais Supervisionada

Para visualizar e explorar a variabilidade da estrutura de expressão em um nível de órgão, a análise de componente principal (PCA) e uma variante de PCA conhecida como PCA Supervisionada foram aplicadas aos dados de expressão de gene usando o analisador de componente principal (PCA) 1006 do sistema de detecção. A PCA é descrita em [Venables e Ripley, 2002] e foi implementada em R. A descrição detalhada de PCA supervisionada pode ser encontrada em [Bair *et al.*, 2004].

Inicialmente, os dados de expressão que representam a expressão de gene de todos os 44.928 conjuntos de sonda do conjunto de dados ‘Discovery’ foram processados pelo módulo PCA 1006 usando a análise de componentes principais (PCA). A PCA é um método padrão para simplificar um conjunto de dados multidimensional pela geração de transformações lineares das dimensões do conjunto de dados para reduzir o número de dimensões. Os dados transformados são fornecidos como dados de componente principal que representam um conjunto sortido de “componentes principais”, tal que o primeiro componente principal tem a maior variação, o segundo componente principal a segunda maior variação e assim por diante. O resultado de aplicar PCA ao conjunto de dados completos inclui os dados de componente multivariado ou principal mostrados na Figura 6A, que é um gráfico em que o primeiro componente principal é plotado no eixo x e o segundo componente principal no eixo y. A inspeção desta perspectiva de dimensão não produz nenhuma estrutura óbvia dentro dos dados que são

compatíveis com segmento de tecido, sugerindo que as fontes maiores da variação da expressão de gene medidos através de todos os genes é independente da localização de tecido.

Para investigar se um subconjunto de todos os genes seria  
5 usado para gerar um ou mais componentes principais indicativos da localização de tecido, os dados de expressão foram analisados pela PCA supervisionada. Como descrito em [Bair *et al*, 2004], a PCA supervisionada é similar à análise padrão de componentes principais mas usa apenas um subconjunto das características/genes (usualmente selecionados por algum  
10 meio univariado) para gerar os componentes principais. Neste caso, o conjunto de genes diferencialmente expressado entre o ceco e o reto (isto é, as extremidades extremas do intestino grosso) foram selecionados para a análise de PCA. Entretanto, outras formas de seleção de característica alternativamente podem ser usadas. Especificamente, uma matriz de dados  
15 reduzida foi gerada pela inclusão não apenas dos 115 conjuntos de sonda que são diferencialmente expressados entre as amostras de tecido tiradas do ceco e reto, mas para todos os 184 tecidos normais de todos os segmentos do intestino grosso. A PCA padrão foi depois realizada nestes dados específicos de característica. Como mostrado na Figura 6B, um gráfico dos primeiros dois  
20 componentes principais sugere a existência de duas sub-populações amplas dentro das 184 amostras de tecido, que correspondem aproximadamente ao divisor proximal vs. distal. Esta dependência da origem da célula é visualizada mais claramente se o primeiro componente principal é colocado em gráfico como uma função da origem da célula ao longo do intestino  
25 grosso, como mostrado na Figura 7B. Os símbolos na Figura 7B representam a distância interquartílica (isto é, metade dos dados) e as “barras de erro” indicam 1,5 x a distância interquartílica. Os dados fora destes limites são considerados como sendo classificados à parte e são plotados individualmente. Embora exista talvez a sugestão de uma separação fraca

entre o cólon sigmóide e o reto, os tecidos anteriores do ceco e cólon ascendente fortemente se sobrepõem com separação deficiente.

Embora os dados de componente principal possam ser usados para prognosticar a origem de células com base na expressão de genes a partir  
5 destas células, outros métodos de análise são preferidos para esta tarefa, como descritos abaixo.

#### Análise de Perfil (Análise de Variação Canônica)

Os padrões de expressão ao longo do intestino foram também analisados pelo analisador de perfil 1004 usando a Análise de Perfil para  
10 visualizar a variação de expressão inter versus intra-segmento. Como descrito em [Kiiveri, 1992], a análise de perfil é uma modificação da análise de variação canônica padrão adaptada aos casos onde o número de variáveis excede o número de observações. O método modela a matriz de covariação de classe dentro de  $p \times p \Sigma_w$  por intermédio de um modelo analítico de fator  
15 [Kiiveri, 1992] com um número relativamente baixo de fatores independentes. Os testes de permuta são usados para determinar a significância de cada termo (isto é, gene) em cada uma das variáveis canônicas. Pelo que inclui apenas termos significantes, a análise de perfil fornece uma capacidade de seleção de característica.

20 Este método é no geral útil como uma ferramenta exploratória para caracterizar a estrutura de variação de classe. A análise de variação canônica é implementada na biblioteca R MASS, como descrito em [Venables e Ripley, 2002]. A Análise de perfil foi implementada em uma biblioteca patenteada em R, como descrito em [Kiiveri 1992].

25 Dado o conhecimento *a priori* de rótulos de segmento para tecidos, a análise de perfil tenta identificar o subespaço de transcrito de gene limitado que fornece separação inter-classe máxima de cada um dos cinco segmentos do intestino grosso enquanto se minimiza a variação intraclasses (isto é, com cada segmento). Os resultados da análise de perfil do conjunto de

dados completos incluem os dados de variável canônica mostrados na Figura 8A, como um gráfico em que a primeira variável canônica é plotada ao longo do eixo x e a segunda variável canônica ao longo do eixo y. Está evidente que os segmentos de tecido correlacionam-se com a primeira variável canônica, mas as variáveis canônicas segunda e subsequentes fornecem pouca ou nenhuma informação de separação de classe. Este resultado sugere que os mesmos conjuntos de sonda estão envolvidos na separação de cada um dos segmentos colorretais, isto é, as fontes maiores de diferença de uma perspectiva de segmento de tecido são aquelas usadas para gerar a primeira dimensão de variável canônica e por este motivo todos os segmentos são melhor agrupados por este mesmo conjunto de característica de conjuntos de sonda. Como mostrado na Figura 8B, mesmo quando a primeira variável canônica é usada, nenhum dos segmentos é perfeitamente separado, embora a ordem natural dos segmentos seja claramente preservada. Como com a PCA descrita acima, os dados de variação canônica podem ser usados para classificar a origem proximal-distal de células na origem desconhecida, mas os métodos descritos abaixo são preferidos para este propósito.

### Máquinas de Vetor de Suporte

Embora os métodos multivariados descritos acima sejam úteis para investigar a variação da expressão de gene ao longo do intestino grosso, o aprendizado de máquina supervisionada foi usado para identificar genes que também são preditivos da localização de tecido em uma maneira robusta e para identificar os subconjuntos menores de conjuntos de sonda/genes que podem ser usados para prognosticar a localização de tecido com uma taxa de erro validado de cruzamento baixo.

Nas formas de realização descritas, a forma particular de aprendizado de máquina usado é uma máquina vetorial de sustentação (SVM), como fornecida pelo módulo SVM 1002; entretanto, estará evidente ao destinatário habilitado que outros métodos de núcleo podem ser

alternativamente usados. Como descrito em [Scholkopf, 2004], métodos de núcleo são extensões de métodos lineares por meio dos quais as variáveis são mapeadas em um outro espaço onde as características essenciais deste mapeamento são capturadas por um núcleo simples. Os métodos de núcleo  
 5 podem ser particularmente vantajosos em casos onde as observações são linearmente separáveis no espaço de núcleo mas não no espaço de dados originais.

O SVM 1002 determina a combinação de características (transcritos de gene) que maximamente separa as observações (isto é, tecidos)  
 10 ao longo de uma fronteira de classe-decisão, usando a metodologia de SVM padrão, como descrito em [Cristianini e Shawe-Taylor, 2000].

Especificamente, a máquina Vetorial de suporte (SVM) 1002 foi usada para gerar dados de classificação que representam o sub-conjunto menor de conjuntos de sonda do conjunto de dados completo cuja expressão  
 15 permite a separação máxima de células que se originam do ceco e reto. O SVM 1002 foi treinado usando um núcleo linear e os dados de classificação gerados em cada iteração foi avaliado usando a validação cruzada de 10 vezes. Os transcritos de gene contribuintes mais baixos de cada subconjunto de transcritos foram recursivamente eliminados para identificar o menor  
 20 conjunto de transcritos com alta precisão de prognóstico.

A taxa de erro de SVM validada cruzada como uma função do número de conjuntos de sonda incluídos no modelo (visto que eles foram sucessivamente eliminados) é mostrada na Figura 9. O conjunto de característica menor que produz uma taxa de erro validada cruzada perfeita  
 25 (0%) inclui os 13 conjuntos de sonda mostrados na Tabela 3.

Para medir a utilidade deste modelo em um conjunto de dados independente, os dados de classificação para os treze modelos de característica foram testados quanto ao desempenho de prognóstico proximal vs. distal nos dados de validação. Usando um modelo de análise discriminante

linear tradicional construído com estes 13 conjuntos de sonda, os oito tecidos proximais e os onze distais foram prognosticados com 100% de precisão.

#### Modelo Classificador

Como uma alternativa para o SVM 1002, um classificador 1007 também foi usado para processar os dados de expressão completos de amostras de tecido tiradas de localizações conhecidas ao longo do eixo proximal-distal do intestino grosso para identificar combinações de genes que podem ser usados para identificar a origem de uma célula ou população de célula de origem desconhecida ao longo do intestino grosso. Na forma de realização descrita, o classificador GeneRave linear foi usado, como descrito em <http://www.bioinformatics.csiro.au/overview.shtml>. GeneRave é preferido em casos onde o número de variáveis excede o número de observações. entretanto, estará evidente àqueles habilitados na técnica que outros classificadores poderiam ser alternativamente usados, que incluem classificadores não lineares e classificadores com base na regressão logística regularizada.

Como descrito em [Kiiveri 2002], o classificador geneRave 1007 gera dados de classificação que representam combinações lineares de níveis de expressão para identificar subconjuntos de genes que podem ser usados para identificar com precisão as localizações de uma amostra de localizações desconhecidas. GeneRave 1007 usa um modelo de rede Bayesian para selecionar genes pela eliminação de genes que em combinação linear com outros genes não têm nenhuma correlação com as localizações das quais as amostras de tecido correspondentes foram tiradas.

O resultado da análise geneRave do conjunto de dados completo em dados de classificação que correspondem a um conjunto de 7 genes cujos níveis de expressão podem ser usados para identificar acuradamente a origem de uma célula correspondente ao longo do eixo proximal-distal do intestino grosso. Os 7 genes são SEC6L1, PRAC, SPINK5,

SEC6L1, ANPEP, DEFA5 e CLDN8.

### Debate

Um mapa da expressão diferencial de gene ao longo do intestino grosso

5                   A análise de expressão univariada identificou 206 conjuntos de sonda que correspondem a 154 alvos de gene único que são diferencialmente expressados entre as regiões do intestino grosso proximais normais e distais normais em adultos humanos. Um subconjunto de 115 conjuntos de sonda (89% comuns à lista proximal vs. distal) é do mesmo modo diferencialmente  
10                   expressado entre os segmentos colorretais terminais do ceco e reto. De maneira interessante, nós não encontramos nenhum transcrito que fosse expressado de modo significativamente diferente entre quaisquer dois segmentos adjacentes.

                  Para estimar a validade destas verificações, nós também  
15                   medimos a mudança da expressão destes transcritos de gene em um conjunto independente de dados de microarranjo. Trinta e um (31) dos 206 conjuntos de sonda diferencialmente expressados em nosso conjunto de dados de verificação inicial de 184 amostras de tecidos colorretais também foram diferencialmente expressados nos dados de validação de 19 espécimes.

20                   Usando uma simulação de Monte Carlo, nós mostramos que um tal número grande de conjuntos de sonda diferenciais em ambos os conjuntos de dados é extremamente improvável.

                  Quase todos (28/31, 90%) destes transcritos ‘validados’ foram do mesmo modo diferencialmente expressados entre os dois segmentos de  
25                   terminal do ceco e reto. 57 de 154 (37%) dos alvos de gene correspondentes foram confirmados ser diferencialmente expressados entre a intestino grosso proximal e distal pelos meios independentes.

Expressão de transcrito diferencial para genes individuais

O conjunto de sonda mais significativamente diferencial que

nós observamos em nossos dados de verificação foi contra o transcrito de gene para PRAC. PRAC é altamente expressado no intestino grosso distal em relação aos tecidos proximais. Além disso, PRAC parece ser expressado em um padrão baixo-alto ao longo do intestino grosso com uma mudança de expressão acentuada ocorrendo entre os espécimes colorretais ascendentes e descendentes.

Verificou-se oito (8) conjuntos de sonda que correspondem a sete (7) genes HOX a serem diferencialmente expressados entre o intestino grosso proximal e distal. Os 39 membros da família de gene de homeobox mamífero consistem de fatores de transcrição altamente conservados que especificam a identidade de segmentos de corpo ao longo do eixo anterior-posterior do embrião em desenvolvimento [Hostikka e Capecchi, 1998, *Mech Dev* 70: 133-145; Kosaki *et al.*, 2002, *Teratology* 65: 50-62]. Os quatro grupos de parálogos de gene HOX são expressados em uma sequência de anterior para posterior, por exemplo, de *HOXA1* para *HOX13*. [Montgomery *et al.*, 1999, *Gastroenterology* 116: 702-731] Verificou-se que: os genes HOX de número mais baixo são expressados mais altos nos tecidos proximais (*HOXD3*, *HOXD4*, *HOXB6*, *HOXC6* e *HOXA9*), enquanto que os genes de número mais alto são mais expressados no intestino grosso (*HOXB13* e *HOXD13*).

De maneira interessante, houve uma ausência visível em nossas verificações de alguns transcritos de gene que foram previamente mostrados como sendo diferencialmente expressados ao longo do eixo proximal-distal. Nossos dados não demonstram um gradiente de expressão significativa para os genes de homeobox caudais *CDX1* ou *CDX2*, fatores de transcrição que foram mostrados estar envolvidos no desenvolvimento padrão do intestino através de uma faixa de vertebrados. (Chalmers *et al.*, 2000) (James *et al.*, 1994) (Silberg *et al.*, 2000). Em particular, acredita-se que *CDX2* desempenhe um papel na manutenção do fenótipo colônico no intestino

grosso adulto e foi recentemente mostrado estar presente em concentrações relativamente altas no intestino grosso proximal mas ausente no intestino grosso distal (James *et al.*, 1994) (Silberg *et al.*, 2000). Nem a análise estatística nem a inspeção visual da expressão de conjunto de sonda para este gene mostram a expressão diferencial ao longo do intestino grosso em nossos dados (dados não mostrados).

Observou-se a expressão de transcrito diferencial significativa para vários dos genes de transporte carregadores de soluto. Enquanto a expressão de conjunto de sonda para SLC2A10, SLC13A2 e SLC28A2 é mais alta no intestino grosso distal, os membros da família de carregador de soluto SLC9A3, SLC14A2, SLC16A1, SLC20A1, SCL23A3 e SLC37A2 são mais altos nos tecidos proximais.

Nossos resultados mostram que conjuntos de sonda contra todos os três dos cinco membros do grupo do cromossoma 7q22 de mucinas ligadas à membrana previamente acreditadas serem expressadas no intestino grosso, *MUC11*, *MUC12* e *MUC17*, são diferencialmente expressados em níveis altos no intestino distal [Byrd e Bresalier, 2004, Cancer Metastasis Rev 23: 77-99; Williams *et al.*, 1999, Cancer Res 59: 4083-4089; Gum *et al.*, 2002, Biochem Biophys Res Commun 291: 466-475]. Nós também confirmamos este padrão de expressão diferencial para *MUC12* e *MUC17* nos dados de validação independentes. Relatos anteriores também incitaram a questão a cerca de se as sequências genômicas para *MUC11* e *MUC12* são de genes intimamente relacionados ou talvez ainda do mesmo gene. [Byrd e Bresalier, 2004, supra]. A análise de correlação dos conjuntos de sonda *MUC11* e *MUC12* mostram uma correlação forte, positiva na extremidade mais baixa da faixa de expressão do conjunto de sonda com uma correlação mais fraca conforme a expressão aumenta (dados não mostrados). Este perfil de correlação poderia ser devido à variabilidade aumentada em níveis mais altos de expressão ou, possivelmente, porque os níveis de expressão no intestino

grosso distal (onde eles são mais altos) reflete um controle transcricional distinto.

Além disso, embora pesquisas anteriores tenham sugerido que a mucina secretada, que forma gel MUC5B fosse apenas fracamente expressada no intestino grosso [Byrd e Bresalier, 2004, *supra*], nossos resultados mostram que os conjuntos de sonda reativos para este transcrito são expressados mais altos no intestino grosso distal como para as mucinas ligadas a membrana.

Alguns dos padrões de expressão que nós relatamos aqui para seres humanos foram mostrados ser similarmente padronizados nos tratos gastrointestinais de modelos de roedor. Entretanto, vários genes específicos anteriormente mostrados serem diferencialmente expressados ao longo dos intestinos grossos de camundongos e ratos não foram encontrados por nós como sendo assim expressados. Tais transcritos de alvos de gene incluem, anidrase carbônica IV (Fleming *et al.*, 1995), membro 1 da família 4 de carregador de soluto (também conhecido como AE1) (Raj endran *et al.*, 2000), CD36/translocase de ácido graxo (Chen *et al.*, 2001) e receptor 4 equivalente a *toll* (Ortega-Cava *et al.*, 2003). Por outro lado, nossos dados estão de acordo com estudos iniciais de expressão de aquaporina-8 (AQP8), um gene cujo produto de expressão é suspeito estar envolvido na absorção de água no intestino grosso de rato normal (Calamita *et al.*, 2001). Nós observamos que AQP8 é significativamente expressado a um nível alto no intestino grosso humano proximal comparado com os tecidos distais ( $p < 0,006$ , dados não mostrados). A família das proteínas de junção firme de claudina também podem desempenhar um papel na manutenção da integridade da barreira da água no intestino grosso (Jeansonne *et al.*, 2003). Verificou-se que a expressão de claudina-8 (CLDN8) é muito mais altamente expressada nos tecidos colorretais distais. Ao contrário, claudina-15 (CLDN15), que também acredita-se esteja localizada nas fibrilas de junção

firme foi expressada em um nível alto nos tecidos colorretais proximais (Colegio *et al.*, 2002).

A natureza da mudança da expressão de gene ao longo do intestino grosso

5                   Embora uma meta deste trabalho fosse entender quais transcritos de gene são diferencialmente expressados ao longo do intestino grosso, um segundo objetivo foi explorar a natureza destas mudanças de expressão ao longo do eixo proximal-distal na região ou item específico de segmento.

10                   Nós observamos dois padrões amplos de mudança de expressão de transcrito estatisticamente significativo ao longo do colorreto. O padrão maior é descrito por aqueles 65 transcritos de gene que foram bem adaptados por um modelo de expressão de dois segmentos. Nós sugerimos que a expressão destes transcritos seja dicótoma na natureza – elevados nos  
15                   segmentos proximais e reduzidos em segmentos distais, ou vice e versa.

Tais dados são compatíveis com a visão anatômica convencional de que o divisor ‘natural’ entre o intestino grosso proximal e distal ocorre entre o cólon ascendente e descendente. Esta verificação é contrária a um relato recente por Komuro *et al.* de que um ponto de  
20                   interrupção entre o cólon descendente e o sigmóide produz a expressão diferencial máxima (Komuro *et al.*, 2005). Entretanto, nós observamos que além de analisar este padrão em espécimes de câncer colorretal, Komuro *et al.* também escolheram incluir o cólon transversal em sua análise. Nós intencionalmente excluímos tecidos deste segmento para evitar a influência de  
25                   confusão possível relacionada com o ponto de fusão do intestino intermediário-intestino traseiro prognosticado aproximadamente dois terços do comprimento do cólon transversal.

Um segundo conjunto de 50 transcritos não demonstram uma mudança de dicótomo, mas ao invés mostram uma melhora significativa no

ajuste pela aplicação dos dados de expressão a um modelo de cinco segmentos que sustentam um gradiente de expressão mais gradual movendo ao longo do intestino grosso a partir do ceco para o reto.

Estes dois padrões de expressão característicos sugerem que a expressão de gene ao longo do eixo proximal-distal seja talvez coordenada por dois sistemas subjacentes de organização.

Nós observamos que a maioria dos transcritos diferencialmente expressados nos tecidos normais do adulto aqui medidos são expressados em um padrão que é compatível com um padrão de intestino intermediário vs. intestino traseiro de desenvolvimento embrionário. Além disso, métodos multivariados que incluem PCA supervisionada e análise de variação canônica também sugerem que a fonte primária de variação entre estes dados são explicadas pelos divisores proximal vs. distal. Em um estudo recente Glebov *et al.* verificou-se que o número de genes diferencialmente expressados entre o cólon ascendente e descendente no adulto é substancialmente maior do que o número de genes igualmente identificados em intestinos grossos de fetais de 17 a 24 semanas de idade. Glebov *et al.* levantaram a hipótese de que o padrão de expressão de gene do intestino grosso do adulto é possivelmente ajustado concorrentemente com a expressão do fenótipo colônico do adulto em ~30 semanas de gestação ou talvez ainda em resposta aos conteúdos luminais pós natal do trato gastrointestinal. Embora nós não tenhamos explorado a expressão de gene no intestino grosso fetal, nós observamos padrões de expressão no adulto que sustentam uma origem embrionária compatível com a fusão do intestino intermediário-intestino traseiro.

A maioria destes transcritos que exibem uma mudança de expressão gradual entre o ceco e o reto exibem um padrão prototípico de expressão aumentada movendo do ceco para o reto. Este padrão não é observado no intestino intermediário-intestino traseiro diferencial transcrito

onde o número de transcritos elevados proximalmente é aproximadamente igual ao número elevados na região distal. Nós propusemos que o padrão crescente distalmente característico nestes transcritos seria uma função de fatores extrínsecos em comparação com o padrão de intestino intermediário-intestino traseiro intrinsecamente definido. Tais fatores incluiriam o efeito de conteúdos luminais que se movem em uma maneira unidirecional do ceco para o reto e/ou as mudanças regionais na microflora ao longo do intestino grosso. Outro trabalho será requerido para investigar se tais controles extrínsecos estão funcionando em uma maneira positiva de induzir a atividade transcricional ou através de um silenciamento transcricional reduzido.

Mudanças na expressão de gene em adaptação ao longo do intestino grosso

Para explorar a expressão de genes em adaptação ao longo do intestino grosso, nós também aplicamos a análise de componente principal e análise de perfil destes dados de expressão. Existe evidência forte quanto a um padrão de expressão de gene proximal versus distal com estas técnicas de visualização multivariada. Além disso, a análise de perfil, que simultaneamente maximiza as diferenças de expressão inter-segmento enquanto tenta diminuir a variação intra-segmento, sugere que o mesmo conjunto de genes que são responsáveis pela variabilidade entre o ceco para o reto também melhor separam os segmentos individuais. Embora estes resultados multivariados não excluam um gradiente proximal-distal sutil, a natureza bimodal evidente destas plotagens multivariadas sugere que a fonte maior de variação de expressão nestes tecidos é compatível com um padrão derivado de intestino intermediário vs. intestino traseiro.

Um conjunto menor de genes pode ser informativo

Finalmente, o método de classificação sofisticado de máquinas de vetor de suporte é usado para selecionar um subconjunto de conjuntos de sonda informativos que podem ser usados para fornecer uma classificação

estável, robusta de tecidos proximais versus distais. Conjuntos de sonda ‘selecionados’ pelo SVM 1002 são um subconjunto dos transcritos diferenciais identificados pelos métodos univariados, acima. Pela avaliação destes 13 modelos de transcrito no conjunto de validação independente, a robustez destes prognosticadores é ainda demonstrada.

Aqueles habilitados na técnica avaliarão que a invenção aqui descrita é susceptível às variações e modificações outras que não aquelas especificamente descritas. Deve ser entendido que a invenção inclui todas de tais variações e modificações. A invenção também inclui todas das etapas, características, composições e compostos aludidos ou indicados neste relatório descritivo, individual ou coletivamente e qualquer e todas as combinações de qualquer duas ou mais das ditas etapas ou características.

### Conclusões

Nosso trabalho sugere que a abundância de transcrito e talvez a regulação transcricional, segue dois padrões amplos ao longo do eixo proximal-distal do intestino grosso. O padrão dominante é um padrão de expressão dicótoma compatível com as origens embrionárias de intestino intermediário-intestino traseiro do intestino proximal e distal. Os transcritos que seguem este padrão são de maneira aproximadamente idêntica divididos naqueles que são elevados distalmente e aqueles elevados proximalmente. O segundo padrão que nós observamos é caracterizado por uma mudança gradual nos níveis de transcrito do ceco para o reto, quase todos dos quais exibem expressão aumentada contra os tecidos distais. Nós propusemos que os tecidos que exibem os padrões de intestino intermediário-intestino traseiro dicótomos são prováveis de refletir as origens embrionárias intrínsecas do intestino grosso enquanto que aqueles que exibem uma mudança gradual refletem fatores extrínsecos tais como fluxo luminal e mudanças na microflora. Quando juntos, estes padrões constituem um mapa de expressão de gene do intestino grosso. Este é o primeiro de tal mapa de um órgão humano inteiro.

TABELA 1: Lista de genes diferencialmente expressados mais alto em tecidos proximais em relação aos tecidos distais.

(p &lt; 0,05)

| Fila | ID Conjunto de sonda | Símbolo  | Descrição                                                                             | Proximal-Distal |          | Ceco-Reito |         | Validação |          |          |          |         |
|------|----------------------|----------|---------------------------------------------------------------------------------------|-----------------|----------|------------|---------|-----------|----------|----------|----------|---------|
|      |                      |          |                                                                                       | Expr. Δ         | t        | Valor P    | Expr. Δ | t         | Valor P  | t        | CI baixo | CI alto |
| 1    | 222262_s_at          | SEC6L1   | etanolamina cinase 1                                                                  | 3,3492          | -12,9258 | 5,27E-23   | 3,5741  | -9,0521   | 6,53E-09 | 1,5891   | -0,3764  | 2,4320  |
| 2    | 225458_at            | SEC6L1   | SEC6-tipo 1 (S. cerevisiae)                                                           | 5,4422          | -12,5937 | 5,10E-22   | 6,2917  | -9,2685   | 2,57E-09 | 1,4370   | -0,7340  | 3,6253  |
| 3    | 225457_s_at          | SEC6L1   | SEC6-tipo 1 (S. cerevisiae)                                                           | 4,2221          | -12,5347 | 7,62E-22   | 4,9764  | -9,7261   | 3,59E-10 | 1,2930   | -0,8902  | 3,5413  |
| 4    | 219017_at            | ETNK1    | etanolamina cinase 1                                                                  | 4,0801          | -12,3947 | 1,98E-21   | 4,1238  | -8,1023   | 3,99E-07 | 1,1704   | -1,0423  | 3,4942  |
| 5    | 207558_s_at          | PITX2    | Fator de transcrição 2 de homeodominio aos pares                                      | 1,6252          | -12,3516 | 2,66E-21   | 1,7549  | -8,5481   | 5,79E-08 | 0,6582   | -0,6362  | 1,2099  |
| 6    | 224453_s_at          | ETNK1    | etanolamina cinase 1                                                                  | 2,0637          | -11,5429 | 6,45E-19   | 2,1692  | -8,0763   | 4,47E-07 | 1,3638   | -0,1907  | 0,7586  |
| 7    | 229230_at            | OSTA1a   | transportado alfa de soluto orgânico                                                  | 2,4793          | -10,8011 | 9,47E-17   | 2,7768  | -8,6246   | 4,15E-08 | 1,95E-01 | -0,4902  | 2,2212  |
| 8    | 206340_at            | NR1H4    | subfamília 1 do receptor nuclear, grupo H, membro 4                                   | 2,0505          | -10,3266 | 2,22E-15   | 2,4066  | -9,1541   | 4,20E-09 | 3,55E-02 | 0,0394   | 0,9527  |
| 9    | 226432_at            |          | ** nenhuma descrição**                                                                | 2,3181          | -10,0408 | 1,46E-14   | 2,5744  | -7,2261   | 1,76E-05 | 1,2193   | -0,5313  | 1,8442  |
| 10   | 209869_at            | ADRA2A   | Receptor adrenérgico, alfa-2A, 1,6585                                                 | -9,8367         | 5,55E-14 | 1,7705     | 8,0507  | 4,99E-07  | 2,45E-01 | 1,2272   | -0,4738  | 1,6677  |
| 11   | 227194_at            | FAM3B    | Família com similaridade de sequência 3, membro B                                     | 2,8282          | -9,8079  | 6,70E-14   | 3,4326  | -6,9816   | 5,00E-05 | 2,04E-01 | -0,6662  | 2,7145  |
| 12   | 207251_at            | MEP1B    | Meprina A, beta                                                                       | 1,7581          | -9,7239  | 1,16E-13   | 1,8022  | -6,5673   | 2,91E-04 | 1,52E-01 | -0,2025  | 1,1482  |
| 13   | 219954_s_at          | GBA3     | Glicosidade, beta, ácido 3 (citossólico)                                              | 1,7033          | -9,6737  | 1,60E-13   | 1,9800  | -8,3619   | 1,30E-07 | 1,76E-01 | -0,2567  | 1,1929  |
| 14   | 219955_at            | F1110884 | Proteína hipotética FLJ10884                                                          | 1,8400          | -9,1831  | 3,77E-12   | 1,9031  | -5,9016   | 4,66E-03 | 2,78E-01 | -0,0917  | 0,2976  |
| 15   | 225290_at            |          | **nenhuma descrição**                                                                 | 2,2680          | -9,1191  | 5,68E-12   | 2,4516  | -6,2630   | 1,04E-03 | 3,30E-01 | -0,8929  | 2,4715  |
| 16   | 201920_at            | SLC20A1  | Família carregadora de soluto 20 (transportador de fosfato), membro 1                 | 2,1030          | -8,5555  | 1,97E-10   | 2,3428  | -7,0466   | 3,79E-05 | 3,68E-01 | -1,0459  | 2,6359  |
| 17   | 206294_at            | HSD3B2   | Hidróxi-delta-5-esteróide Desidrogenase, 3 beta-e esteróide delta-isomerase 2         | 1,8455          | -8,2334  | 1,43E-09   | 2,0613  | -6,6283   | 2,25E-04 | 3,68E-01 | -0,9742  | 2,4564  |
| 18   | 231576_at            |          | **nenhuma descrição**                                                                 | 2,1646          | -8,0045  | 5,75E-09   |         |           |          | 1,89E-01 | -0,3026  | 1,3050  |
| 19   | 222943_at            | GBA3     | Glicosidade, beta, ácido 3 (citossólico)                                              | 2,0596          | -7,9083  | 1,03E-08   | 2,5806  | -6,9404   | 5,96E-05 | 3,62E-01 | -0,7354  | 1,8413  |
| 20   | 202236_s_at          | SLC16A1  | Família carregadora de soluto 16 (transportadores do ácido monocarboxílico), membro 1 | 1,6747          | -7,6989  | 3,58E-08   | 1,8552  | -6,9860   | 4,91E-05 | 7,30E-01 | -1,4137  | 1,0142  |
| 21   | 205366_s_at          | HOXB6    | Homeo box B6                                                                          | 1,4861          | -7,6727  | 4,18E-08   | 1,6332  | -6,0387   | 2,65E-03 | 3,75E-01 | -0,3720  | 0,8890  |

|    |             |         |                                                                                                                |        |         |          |        |         |          |         |         |        |
|----|-------------|---------|----------------------------------------------------------------------------------------------------------------|--------|---------|----------|--------|---------|----------|---------|---------|--------|
| 22 | 222774_s_at | NETO2   | Neuropilina (NRP) e<br>Equivalente a tolóide (TLL) 2<br>**nenhuma descrição**                                  | 1,6919 | -7,5826 | 7,11E-08 |        |         | 6,56E-01 | 0,4551  | -0,5353 | 0,8246 |
| 23 | 235733_at   |         | Família pseudogene                                                                                             | 1,1776 | -7,4926 | 1,21E-07 | 1,2384 | -6,0872 | 2,17E-03 | 1,8733  | -0,0196 | 0,3111 |
| 24 | 202235_at   | AFARP1  | Mesoderma posterior 1                                                                                          | 1,2859 | -7,3793 | 2,33E-07 | 1,3698 | -6,6895 | 1,73E-04 | -0,6204 | -0,9183 | 0,5044 |
| 25 | 224476_s_at | MESP1   | Homeo box C6                                                                                                   | 1,2840 | -7,2589 | 4,68E-07 |        |         |          | 1,2876  | -0,0855 | 0,3497 |
| 26 | 206858_s_at | HOXC6   | citocromo P450, família 2, subfamília C, polipeptídeo 18                                                       | 1,2640 | -7,1875 | 7,05E-07 | 1,3672 | -6,2775 | 9,82E-04 | 1,5380  | -0,1110 | 0,6535 |
| 27 | 208126_s_at | CYP2C18 | defensina, alfa 5, específico de célula de Paneth                                                              | 1,5721 | -7,0842 | 1,27E-06 |        |         |          | 0,2970  | -0,8071 | 1,0692 |
| 28 | 207529_at   | DEFA5   | homólogo ausente nos olhos 2 (Drosófila)                                                                       | 2,8342 | -7,0313 | 1,71E-06 | 3,8363 | -5,9701 | 3,51E-03 | 1,5002  | -0,4189 | 1,8957 |
| 29 | 209692_at   | EYA2    | Canal de potássio ativado por voltagem, subfamília G, Membro 1                                                 | 1,3808 | -6,9744 | 2,36E-06 | 1,4435 | -5,9334 | 4,09E-03 | 2,5104  | 0,0383  | 0,4702 |
| 30 | 214595_at   | KCNNG1  | alanil (membrana) aminopeptidase (aminopeptidase N, aminopeptidase M, aminopeptidase microssômica, CD13, p150) | 1,1633 | -6,9706 | 2,41E-06 | 1,2868 | -6,4306 | 5,17E-04 | -1,7744 | -0,5220 | 0,0453 |
| 31 | 202888_s_at | ANPEP   | proteína de ligação do fator de crescimento equivalente à insulina 2, 36 kDa                                   | 2,6011 | -6,8676 | 4,30E-06 | 3,3179 | -5,7250 | 9,58E-03 | 1,1662  | -0,9121 | 3,0790 |
| 32 | 202718_at   | IGFBP2  | família com a similaridade de sequência 45, membro A                                                           | 1,8892 | -6,8559 | 4,59E-06 |        |         |          | 0,2631  | -1,0565 | 1,3500 |
| 33 | 221804_s_at | FAM45A  | Enzima de montagem de mRNA de apolipoproteína B, polipeptídeo catalítico 1                                     | 1,3071 | -6,8456 | 4,86E-06 |        |         |          | -0,4156 | -1,7005 | 1,1551 |
| 34 | 207158_at   | APOBEC1 | família carregadora de soluto 23 (transportadores de nucleobase), membro 3                                     | 1,4298 | -6,7384 | 8,81E-06 |        |         |          | 0,1857  | -0,5250 | 0,6260 |
| 35 | 230949_at   | SLC23A3 | Transição de fase 2 de G1 para S                                                                               | 1,1622 | -6,5961 | 1,92E-05 |        |         |          | 2,0879  | -0,0267 | 1,0424 |
| 36 | 205541_s_at | GSPT2   | Família carregadora de soluto 9 (trocar de sódio/hidrogênio), isoforma 3                                       | 1,3378 | -6,5339 | 2,70E-05 | 1,4485 | -5,7155 | 9,96E-03 | 1,4047  | -0,2567 | 1,1282 |
| 37 | 207212_at   | SLC9A3  | citocromo P450, família 2, subfamília C, polipeptídeo 18                                                       | 1,2571 | -6,5310 | 2,74E-05 |        |         |          | 0,0608  | -0,2994 | 0,3171 |
| 38 | 215103_at   | CYP2C18 | citocromo P450, família 2, subfamília B, polipeptídeo 6                                                        | 1,3638 | -6,5193 | 2,92E-05 | 1,4312 | -5,9261 | 4,21E-03 | 0,0248  | -0,6717 | 0,6874 |
| 39 | 206755_at   | CYP2B6  | **nenhuma descrição**                                                                                          | 1,2980 | -6,4787 | 3,64E-05 | 1,3244 | -5,5367 | 2,05E-02 | 3,3120  | 0,1017  | 0,5198 |
| 40 | 239656_at   |         | família com similaridade de sequência 45, membro A                                                             | 1,1506 | -6,4761 | 3,69E-05 |        |         |          | 0,5545  | -0,3367 | 0,5611 |
| 41 | 222955_s_at | FAM45A  |                                                                                                                | 1,2688 | -6,4573 | 4,09E-05 |        |         |          | 0,1300  | -0,2480 | 0,2802 |

|    |             |         |                                                                                                                            |        |         |          |        |         |          |          |         |         |        |
|----|-------------|---------|----------------------------------------------------------------------------------------------------------------------------|--------|---------|----------|--------|---------|----------|----------|---------|---------|--------|
| 42 | 213181_s_at | MOCS1   | Síntese de cofator de molibidênio 1                                                                                        | 1,1617 | -6,4528 | 4,19E-05 | 1,2410 | -6,4040 | 5,78E-04 | 8,98E-01 | 0,1300  | -0,2891 | 0,3268 |
| 43 | 205522_at   | HOXD4   | homeo box D4                                                                                                               | 1,2966 | -6,4496 | 4,26E-05 | 1,4206 | -5,6334 | 1,39E-02 | 1,70E-02 | 2,7802  | 0,0674  | 0,5621 |
| 44 | 221304_at   | UGT1A8  | UDP glucosiltransferase 1<br>família, polipeptídeo A8                                                                      | 1,3599 | -6,4054 | 5,40E-05 |        |         |          | 3,32E-02 | 2,4124  | 0,0157  | 0,3156 |
| 45 | 205660_at   | OASL    | equivalente à 2'-5'-oligoadenilato<br>sintetase                                                                            | 1,5483 | -6,3676 | 6,61E-05 |        |         |          | 9,13E-02 | 1,8836  | -0,1619 |        |
| 46 | 218888_s_at |         | **nenhuma descrição**<br>família carregadora de soluto 16                                                                  | 1,6234 | -6,3647 | 6,71E-05 |        |         |          | 8,65E-01 | 0,1729  | -0,7162 |        |
| 47 | 209900_s_at | SLC16A1 | (transportadores do ácido<br>monocarboxílico), membro 1                                                                    | 1,4721 | -6,3225 | 8,41E-05 | 1,6899 | -6,0457 | 2,57E-03 | 7,73E-01 | -0,2938 | -1,3553 |        |
| 48 | 242059_at   |         | **nenhuma descrição**<br>família da UDP glucosiltransferase 1,                                                             | 1,6676 | -6,3073 | 9,12E-05 |        |         |          | 1,58E-01 | 1,5283  | -0,3359 |        |
| 49 | 221305_s_at | UGT1A8  | polipeptídeo A8                                                                                                            | 1,6300 | -6,3057 | 9,20E-05 |        |         |          | 1,16E-01 | 1,7472  | -0,0934 |        |
| 50 | 219197_s_at | SCUBE2  | peptídeo de sinal, domínio CUB,<br>equivalente a EGF 2                                                                     | 1,2723 | -6,2538 | 1,21E-04 | 1,5426 | -7,2700 | 1,45E-05 | 1,51E-01 | 1,5707  | -0,0850 |        |
| 51 | 236860_at   | NPY6R   | Receptor Y6 de neuropeptídeo Y<br>(pseudogene)                                                                             | 1,1988 | -6,2070 | 1,55E-04 |        |         |          | 1,50E-01 | 1,5108  | -0,0514 |        |
| 52 | 218739_at   | ABHD5   | Contendo 5 domínios de abidrolase                                                                                          | 1,2190 | -6,2061 | 1,56E-04 |        |         |          | 8,25E-01 | 0,2256  | -0,4494 |        |
| 53 | 210797_s_at | OASL    | Equivalente à 2'-5'-oligoadenilato<br>sintetase                                                                            | 1,4082 | -6,1890 | 1,70E-04 |        |         |          | 2,62E-01 | 1,1791  | -0,1607 |        |
| 54 | 206754_s_at | CYP2B6  | citocromo P450, família<br>2, subfamília B, polipeptídeo 6                                                                 | 1,5418 | -6,1369 | 2,24E-04 |        |         |          | 2,00E-01 | 1,3404  | -0,3312 |        |
| 55 | 203333_at   | KIFAP3  | Proteína associada com cinesina 3                                                                                          | 1,2568 | -6,1317 | 2,30E-04 |        |         |          | 5,92E-01 | 0,5550  | -0,6324 |        |
| 56 | 224454_at   | ETNK1   | etanolamina cinase 1                                                                                                       | 1,1406 | -6,1181 | 2,47E-04 |        |         |          | 3,33E-01 | 0,9980  | -0,1088 |        |
| 57 | 214651_s_at | HOXA9   | homeo box A9                                                                                                               | 1,4981 | -6,0474 | 3,57E-04 | 1,6730 | -5,8388 | 6,02E-03 | 7,54E-01 | 0,3192  | -0,9026 |        |
| 58 | 242683_at   | na      | Gene hipotético<br>sustentado por AK095347                                                                                 | 1,2426 | -5,9201 | 6,86E-04 |        |         |          | 3,97E-02 | 2,3200  | 0,0201  |        |
| 59 | 236894_at   |         | **nenhuma descrição**                                                                                                      | 1,3679 | -5,8885 | 8,07E-04 |        |         |          | 6,22E-01 | 0,5028  | -0,1866 |        |
| 60 | 218136_s_at | MSCP    | Proteína carregadora de soluto<br>mitocondrial                                                                             | 1,2016 | -5,8872 | 8,12E-04 |        |         |          | 3,93E-01 | 0,8820  | -0,1419 |        |
| 61 | 210153_s_at | ME2     | Enzima málica 2, dependente de<br>NAD(+), mitocondrial                                                                     | 1,2047 | -5,8498 | 9,82E-04 |        |         |          | 6,28E-01 | 0,5001  | -0,4716 |        |
| 62 | 209752_at   | REG1A   | Proteína 1 alfa derivada da ilhota<br>regenerante (proteína da pedra<br>pancreática, proteína de filamento<br>pancreático) | 2,7216 | -5,8414 | 1,02E-03 |        |         |          | 5,62E-01 | -0,5914 | -0,3380 |        |
| 63 | 238638_at   | SLC37A2 | família carregadora de soluto 37<br>(transportadora de glicérol-3-<br>fosfato), membro 2                                   | 1,3919 | -5,8351 | 1,06E-03 |        |         |          | 5,80E-01 | 0,5732  | -0,5148 |        |
| 64 | 214421_x_at | CYP2C9  | citocromo P450, família<br>2, subfamília C, polipeptídeo 9                                                                 |        |         | 6,79E-03 | 1,3877 | -5,8095 | 6,79E-03 | 8,26E-02 | 1,8529  | -0,0292 |        |

|    |             |          |                                                                                            |        |         |          |        |         |          |          |         |         |        |
|----|-------------|----------|--------------------------------------------------------------------------------------------|--------|---------|----------|--------|---------|----------|----------|---------|---------|--------|
| 65 | 205815_at   | PAP      | Proteína associada com a pancreatite                                                       | 2,0272 | -5,7979 | 1,28E-03 | 2,7965 | -5,5114 | 2,27E-02 | 1,36E-01 | 1,6661  | -0,1684 |        |
| 66 | 225351_at   | FAM45A   | família com similaridade de sequência 45 membro A                                          | 1,2592 | -5,6944 | 2,14E-03 |        |         |          | 8,22E-01 | -0,2296 | -0,9944 |        |
| 67 | 243669_s_at | PRAP1    | proteína ácida rica em prolina 1                                                           | 1,4986 | -5,6740 | 2,37E-03 |        |         |          | 4,66E-01 | 0,7466  | -0,7334 | 1,5338 |
| 68 | 228564_at   | LOC37529 | gene hipotético sustentado por BC013438                                                    | 1,1976 | -5,6664 | 2,47E-03 |        |         |          | 5,38E-02 | 2,1149  | -0,0035 | 0,3785 |
| 69 | 223541_at   | HAS3     | hialuronan sintase 3                                                                       | 1,4178 | -5,6557 | 2,60E-03 |        |         |          | 3,82E-01 | -0,8990 | -1,3977 | 0,5637 |
| 70 | 202234_s_at | AFARP1   | Pseudogene da família AKR7                                                                 | 1,4304 | -5,6464 | 2,72E-03 |        |         |          | 7,49E-01 | 0,3259  | -1,0571 | 1,4355 |
| 71 | 203920_at   | NR1H3    | subfamília do receptor nuclear 1, grupo H, membro 3                                        |        |         | 1,87E-02 | 1,3409 | -5,5600 | 1,87E-02 | 4,58E-01 | 0,7617  | -0,3137 | 0,6637 |
| 72 | 231897_at   | ZNF483   | proteína de dedo de zinco 483                                                              | 1,3192 | -5,5272 | 4,90E-03 |        |         |          | 9,53E-01 | 0,0602  | -1,1123 | 1,1762 |
| 73 | 228155_at   | CIOorf58 | Matriz de leitura aberta 58 do cromossoma 10                                               | 1,4264 | -5,5143 | 5,21E-03 |        |         |          | 8,53E-01 | 0,1888  | -1,3883 | 1,6572 |
| 74 | 206601_s_at | HOXD3    | homeo box D3                                                                               | 1,1325 | -5,5056 | 5,44E-03 |        |         |          | 3,90E-01 | 0,8826  | -0,1434 | 0,3488 |
| 75 | 215913_s_at | GULP1    | GULP, adaptador de engolfamento Contendo o domínio PTB 1                                   |        |         | 2,39E-02 | 1,4578 | -5,4985 | 2,39E-02 | 2,46E-02 | 2,4689  | 0,0299  | 0,3831 |
| 76 | 208596_s_at | UGT1A3   | família de UDP glicosiltransferase 1, polipeptídeo A3                                      | 1,6580 | -5,3741 | 1,03E-02 |        |         |          | 3,94E-01 | 0,8810  | -0,5799 | 1,3851 |
| 77 | 202495_at   | TBCC     | Chaperone c específico de tubulina                                                         | 1,1465 | -5,3411 | 1,20E-02 |        |         |          | 8,85E-01 | 0,1471  | -0,3784 | 0,4337 |
| 78 | 221920_s_at | MSCP     | Proteína carregadora de soluto mitocondrial                                                | 1,1893 | -5,3370 | 1,23E-02 |        |         |          | 3,19E-01 | 1,0688  | -0,2442 | 0,6546 |
| 79 | 223058_at   | CIOorf45 | Matriz de leitura aberta 45 do cromossoma 10                                               | 1,3829 | -5,3188 | 1,34E-02 |        |         |          | 9,93E-01 | 0,0092  | -1,2206 | 1,2307 |
| 80 | 219926_at   | POPC3    | Contendo o domínio popeye 3                                                                | 1,1296 | -5,2863 | 1,56E-02 |        |         |          | 1,73E-01 | 1,4622  | -0,0737 | 0,3604 |
| 81 | 210154_at   | ME2      | Enzima málica 2, dependente de NAD(+), mitocondrial                                        | 1,3016 | -5,2581 | 1,78E-02 |        |         |          | 4,06E-01 | 0,8804  | -0,4040 | 0,8951 |
| 82 | 220753_s_at | CRYL1    | crystalina, lambda 1                                                                       | 1,2752 | -5,2392 | 1,95E-02 |        |         |          | 9,42E-01 | 0,0735  | -0,9931 | 1,0643 |
| 83 | 205505_at   | GCNT1    | Glicosaminil (N-acetil) transferase 1, núcleo 2 (beta-1,6-N-acetilglicosaminiltransferase) | 1,1227 | -5,2361 | 1,98E-02 |        |         |          | 1,91E-01 | 1,3736  | -0,0833 | 0,3805 |
| 84 | 219640_at   | CLDN15   | Claudina 15                                                                                | 1,1692 | -5,2276 | 2,06E-02 |        |         |          | 3,03E-01 | 1,0642  | -0,1625 | 0,4894 |
| 85 | 214038_at   | CCL8     | Ligando 8 de quimiocina (motivo C-C)                                                       | 1,6140 | -5,2067 | 2,27E-02 |        |         |          | 1,29E-01 | 1,7169  | -0,2431 | 1,5559 |
| 86 | 220017_x_at | CYP2C9   | citocromo P450, família 2, subfamília C, polipeptídeo 9                                    | 1,3983 | -5,1902 | 2,46E-02 | 1,5251 | -5,4185 | 3,29E-02 | 1,56E-03 | 3,8998  | 0,1592  | 0,5472 |
| 87 | 206407_s_at | CCL13    | Ligando 13 de quimiocina (motivo C-C)                                                      | 1,4448 | -5,1730 | 2,66E-02 |        |         |          | 9,06E-02 | 1,8234  | -0,0265 | 0,3189 |
| 88 | 220585_at   | FU22761  | proteína hipotética FU22761                                                                | 1,1558 | -5,1501 | 2,96E-02 |        |         |          | 7,05E-01 | 0,3868  | -0,1662 | 0,2388 |

|    |             |         |                                                                                                                                                                                                                                                          |        |         |          |          |        |         |        |
|----|-------------|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|---------|----------|----------|--------|---------|--------|
| 89 | 217085_at   | SLC14A2 | família carregadora de soluto 14<br>(transportador de uréia), membro 2                                                                                                                                                                                   | 1,2940 | -5,1161 | 3,47E-02 | 1,69E-01 | 1,5324 | -0,3248 | 1,5292 |
| 90 | 205208      | FTHFD   | formiltehidrofoliato<br>desidrogenase                                                                                                                                                                                                                    | 1,2531 | -5,1123 | 3,53E-02 | 7,99E-01 | 0,2585 | -0,3126 | 0,3997 |
| 91 | 203639_s_at | FGFR2   | Receptor 2 do fator de crescimento<br>de fibroblasto (cinase expressada por<br>bactérias, receptor do fator de<br>crescimento de queratinócito,<br>disostose craniofacial 1, síndrome de<br>Crouzon, síndrome de Pfeiffer,<br>síndrome de Jackson-Weiss) | 1,2760 | -5,0917 | 3,89E-02 | 3,02E-01 | 1,0705 | -0,1918 | 0,5747 |
| 92 | 204663_at   | ME3     | Enzima málica 3, dependente de<br>NADP(+), mitocondrial                                                                                                                                                                                                  | 1,1447 | -5,0447 | 4,83E-02 | 5,46E-01 | 0,6203 | -0,3844 | 0,6922 |
| 93 | 211776_s_at | EPB41L3 | Proteína de membrana de eritrócito<br>3, equivalente à faixa 4,1                                                                                                                                                                                         | 1,2553 | -5,0391 | 4,95E-02 | 5,81E-01 | 0,5706 | -0,4283 | 0,7236 |

TABELA 2. Lista de genes diferencialmente expressados mais altos em tecidos distais em relação aos tecidos proximais,

| Fila | ID Conjunto de sonda | Símbolo     | Descrição                                                                                                                        | Proximal-Distal |         | Ceco-Reto |         | Validação |          |          |
|------|----------------------|-------------|----------------------------------------------------------------------------------------------------------------------------------|-----------------|---------|-----------|---------|-----------|----------|----------|
|      |                      |             |                                                                                                                                  | Expr. Δ         | t       | Valor P   | t       | Expr. Δ   | t        | Valor P  |
| 1    | 230784_at            | PRAC        | proteína nuclear pequena PRAC                                                                                                    | 10,3887         | 16,6750 | 4,56E-31  | 15,5666 | 18,2177   | 2,90-24  | 1,22E-03 |
| 2    | 230105_at            |             | ** nenhuma descrição **                                                                                                          | 2,2919          | 12,3536 | 2,62E-21  | 2,9669  | 11,1548   | 8,51E-13 | 3,09E-01 |
| 3    | 209844_at            | HOXB13      | homeo box B13                                                                                                                    | 2,4103          | 12,1639 | 9,51E-21  | 3,1342  | 10,6863   | 6,07E-12 | 6,44E-02 |
| 4    | 222571_at            | SIAT7F      | sialiltransferase 7 ((alfa-N-acetilineuraminil 2,3-beta-galactosil-1,3)-N-acetil galactosaminidase alfa-2,6-sialiltransferase) F | 1,7332          | 12,0297 | 2,38E-20  | 1,9083  | 9,5206    | 8,68E-10 | 1,74E-02 |
| 5    | 203892_at            | WFDC2       | Domínio de núcleo de dissulfeto de quatro WAP 2                                                                                  | 2,0622          | 11,7522 | 1,56E-19  | 2,3090  | 9,5105    | 9,06E-10 | 7,58E-02 |
| 6    | 214598_at            | CLDN8       | Claudina 8                                                                                                                       | 4,4296          | 10,9279 | 4,05E-17  | 5,9352  | 9,2485    | 2,505,09 | 2,97E-05 |
| 7    | 230360_at            | COLM        | Colomina                                                                                                                         | 2,1190          | 10,9209 | 4,25E-17  | 2,7368  | 10,0265   | 9,94E-11 | 8,76E-03 |
| 8    | 221091_at            | INSL5       | Equivalente à insulina 5                                                                                                         | 3,3289          | 10,2037 | 5,00E-15  | 5,0245  | 9,2341    | 2,98E-09 | 2,96E-01 |
| 9    | 221164_x_at          | CHST5       | carboidrato (N-acetilglicosamina 6-0) sulfotransferase 5                                                                         | 1,5826          | 9,8032  | 6,90E-14  | 1,7349  | 8,1540    | 3,19E-07 | 7,03E-02 |
| 10   | 229254_at            | DKFZp761N11 |                                                                                                                                  | 2,3718          | 9,5776  | 2,99E-13  | 3,0443  | 9,2865    | 2,18E-09 | 1,74E-02 |
| 11   | 230269_at            |             | proteína hipotética DKFZp761N1114                                                                                                | 1,8860          | 9,5192  | 4,36E-13  | 2,1495  | 7,9351    | 8,21E-07 | 1,84E-03 |
| 12   | 223942_x_at          | CHST5       | ** nenhuma descrição **<br>carboidrato (N-acetilglicosamina 6-0) sulfotransferase 5                                              | 1,5910          | 9,3437  | 1,35E-12  | 1,7763  | 8,2351    | 2,25E-07 | 1,56E-02 |
| 13   | 230845_at            | PRAC2       | Proteína da próstata/reto e cólon no. 2                                                                                          | 1,2645          | 9,1328  | 5,20E-12  | 1,2799  | 6,5300    | 3,40E-01 | 7,34E-01 |
| 14   | 239994_at            |             | ** nenhuma descrição **                                                                                                          | 1,7691          | 8,9650  | 1,51E-11  | 2,1086  | 7,9228    | 8,69E-07 | 3,77E-01 |
| 15   | 40284_at             | FOXA2       | box cabeça de garfo A2                                                                                                           | 1,3520          | 8,5397  | 2,17E-10  | 1,4577  | 7,3722    | 9,37E-06 | 2,71E-01 |
| 16   | 207249_s_at          | SILC28A2    | família carregadora de soluto 28 (transportador de nucleosídeo ligado ao sódio), membro 2                                        | 2,0334          | 8,5384  | 2,19E-10  | 2,6495  | 6,8463    | 8,90E-05 | 2,60E-01 |
| 17   | 242372_s_at          | DKFZp761N11 |                                                                                                                                  | 1,5715          | 8,4149  | 4,70E-10  | 1,8751  | 7,5943    | 3,60E-06 | 5,96E-02 |
| 18   | 213994_s_at          | SPON1       | proteína hipotética DKFZp761N1114                                                                                                | 1,6341          | 8,3820  | 5,75E-10  | 1,8277  | 7,5849    | 3,75E-06 | 1,77E-01 |
| 19   | 205185_at            | SPINK5      | espondina 1, proteína de matriz celular                                                                                          | 2,4067          | 8,2883  | 1,02E-09  | 3,6532  | 9,5241    | 8,54E-10 | 1,77E-01 |
| 20   | 203759_at            | SIAT4C      | Inibidor da serina protease, tipo Kazal 5                                                                                        | 1,5035          | 8,2782  | 1,09E-09  |         |           |          | 6,50E-01 |
| 21   | 240856_at            |             | sialiltransferase 4C (beta-galactosídeo alfa-2,3-sialiltransferase)                                                              | 1,7989          | 8,2080  | 1,67E-09  | 2,0481  | 7,7313    | 1,995,06 | 2,82E-01 |
| 22   | 226654_at            | MUC12       | ** nenhuma descrição **<br>Mucina 12                                                                                             | 3,0988          | 8,0394  | 4,66E-09  | 4,2406  | 7,1298    | 2,65E-05 | 4,95E-03 |
| 23   | 229499_at            | CAPN13      | Calpatina 13                                                                                                                     | 1,2187          | 7,8466  | 1,49E-08  | 1,2837  | 6,4588    | 1,59E-04 | 5,49E-01 |
| 24   | 206422_at            | GCG         | Glucagon                                                                                                                         | 3,5394          | 7,8128  | 1,82E-08  | 6,0957  | 7,7872    | 1,56E-06 | 5,68E-01 |
| 25   | 236681_at            | HOXD13      | homeo box D13                                                                                                                    | 1,4419          | 7,5188  | 1,03E-07  | 1,6533  | 6,3341    | 7,75E-01 | 2,01E-01 |
| 26   | 221024_s_at          | SILC2A10    | família carregadora de soluto 2 (transportador de glicose facilitado), membro 10                                                 | 1,5552          | 7,4735  | 1,35E-07  | 1,6301  | 5,6695    | 1,20E-02 | 7,86E-01 |
| 27   | 238862_at            | DKFZp761N11 | proteína hipotética DKFZp761N113                                                                                                 | 1,3657          | 7,4657  | 1,41E-07  | 1,5027  | 7,1762    | 2,17E-05 | 2,42E-01 |

|    |             |          |                                                                                                 |        |        |          |        |        |          |          |         |         |         |
|----|-------------|----------|-------------------------------------------------------------------------------------------------|--------|--------|----------|--------|--------|----------|----------|---------|---------|---------|
| 28 | 201482_at   | QSCN6    | Quiescina Q6                                                                                    | 1,3243 | 7,4495 | 1,55E-07 | 1,4197 | 7,2690 | 1,165,05 | 2,20E-01 | -1,2733 | -0,9080 | 0,2246  |
| 29 | 210103_s_at | FOXA2    | box cabeça de garfo A2                                                                          | 1,3894 | 7,4289 | 1,75E-07 | 1,4913 | 6,2272 | 1,21E-03 | 1,13E-01 | -1,6815 | -0,9151 | 0,1081  |
| 30 | 213993_at   | SPON1    | espondina 1, proteína de matriz extracelular                                                    | 1,4348 | 7,4099 | 1,95E-07 | 1,6082 | 6,6934 | 1,71E-01 | 1,19E-01 | -1,6442 | -0,7080 | 0,0878  |
| 31 | 209436_at   | SPON1    | espondina 1, proteína de matriz extracelular                                                    | 1,5394 | 7,1992 | 6,59E-07 | 1,7567 | 6,6098 | 2,13E-04 | 1,11E-01 | -1,6837 | -1,5765 | 0,1771  |
| 32 | 234994_at   | KIAA1913 | KIAA1913                                                                                        | 2,0243 | 7,1920 | 6,87E-07 | 2,3745 | 6,1586 | 1,515,03 | 4,51E-02 | -2,1685 | -2,3949 | -0,0299 |
| 33 | 204519_s_at | TM4SF11  | membro da superfamília 4 de transmembrana 11 (plasmolipina)                                     | 1,5123 | 7,1801 | 7,35E-07 | 1,7330 | 6,4681 | 1,12E-01 | 1,52E-01 | -2,7824 | -1,6258 | -0,2071 |
| 34 | 213134_x_at | BTG3     | família BTG, membro 3                                                                           | 1,3761 | 7,1419 | 9,14E-07 | 1,4909 | 6,1257 | 1,85E-03 | 4,03E-01 | -0,8587 | -1,0225 | 0,4315  |
| 35 | 206070_s_at | EPHA3    | receptor EPH A3                                                                                 | 1,3440 | 7,0592 | 1,46E-06 |        |        |          | 7,16E-01 | 0,3698  | -0,1398 | 0,1992  |
| 36 | 201889_at   | FAM3C    | família com similaridade de sequência 3, membro C                                               | 1,5846 | 6,9954 | 2,10E-06 | 1,8871 | 7,1044 | 2,96E-05 | 1,77E-01 | -1,4134 | -2,1726 | 0,4361  |
| 37 | 239805_at   | SLC13A2  | família carregadora de soluto 13 (transportador de dicarboxilato dependente de sódio), membro 2 | 1,4052 | 6,9691 | 2,43E-06 |        |        |          | 3,14E-01 | -1,0401 | -0,731  | 0,2496  |
| 38 | 218187_s_at | FU20989  | proteína hipotética FU20989                                                                     | 1,3131 | 6,9597 | 2,57E-06 |        |        |          | 2,67E-03 | -3,5484 | -1,9436 | -0,4900 |
| 39 | 201798_s_at | FER1L3   | fer-1-equivalente 3, mioferlina (C. elegans)                                                    | 1,4386 | 6,9150 | 3,30E-06 | 1,5077 | 5,8090 | 6,80E-03 | 6,52E-02 | -1,9885 | -2,4341 | 0,0839  |
| 40 | 207397_s_at | HOXD13   | homeo box D13                                                                                   | 1,2156 | 6,8953 | 3,68E-06 | 1,3278 | 5,4274 | 3,18E-02 | 3,01E4   | -1,0705 | -0,1530 | 0,0507  |
| 41 | 205548_s_at | BTG3     | Família BTG, membro 3                                                                           | 1,3727 | 6,8644 | 4,38E-06 | 1,4636 | 5,5270 | 2,13E-02 | 5,93E-01 | -0,5445 | -0,6543 | 0,3860  |
| 42 | 207080_s_at | PYY      | Peptídeo YY                                                                                     | 2,9642 | 6,8281 | 5,36E-06 | 4,4363 | 6,1558 | 1,63E-01 | 8,57E-01 | 0,1831  | -0,5225 | 0,6204  |
| 43 | 206104_at   | ISL1     | fator de transcrição de ISL1, LIM/homeodomínio, (ilho-1)                                        | 1,2491 | 6,7817 | 6,93E-06 | 1,3294 | 5,3926 | 3,65E-02 | 2,53E-01 | -1,187  | -0,653  | 0,1853  |
| 44 | 203961_at   | NEBL     | Nebulette                                                                                       |        |        |          |        |        |          | 2,30E-01 | -1,2620 | -1,23Z  | 0,3265  |
| 45 | 208121_s_at | PTPRO    | tirosina fosfatase de proteína, tipo receptor, 0                                                | 1,5345 | 6,6278 | 1,62E-05 | 1,8643 | 7,7938 | 1,52E-06 | 2,18E-01 | 1,2917  | -0,0552 | 0,2220  |
| 46 | 236129_at   | GALNT5   | UDP-N-acetil-alfa-D-galactosamina                                                               | 1,5772 | 6,6010 | 1,87E-05 | 1,7949 | 6,6295 | 2,23E-04 | 2,44E-01 | -2,4706 | -0,5979 | -0,0471 |
| 47 | 203698_s_at | FRZB     | Proteína relacionada com frizado                                                                | 1,3923 | 6,5855 | 2,04E-05 | 1,5111 | 6,1059 | 2,00E-0  | 2,48E-01 | 1,1964  | -0,0771 | 0,2782  |
| 48 | 204351_at   | SI00P    | 5100 proteína cálcica de ligação de P                                                           | 2,5316 | 6,5625 | 2,31E-05 | 1,6958 | 7,1867 | 2,08E-05 | 4,68E-02 | -2,1574 | -3,6312 | -0,0295 |
| 49 | 205042_at   | GNE      | glicosamina (UDP-N-acetil)-2-epimerase/N-acetilmanosamina cinase                                | 1,6163 | 6,4563 | 4,11E-05 | 2,0082 | 6,7357 | 1,13E-04 | 9,31E-01 | 2,9329  | -2,2337 | -0,3643 |
| 50 | 205979_at   | SCGB2A1  | secretoglobina, família 2A, membro 1                                                            | 1,7328 | 6,4027 | 5,48E-05 | 2,0193 | 5,5811 | 1,72E-02 | 1,14E-01 | -1,693  | -0,638  | 0,0771  |
| 51 | 205927_s_at | CTSE     | cathepsina E                                                                                    | 1,4237 | 6,3675 | 6,62E-05 | 1,5846 | 6,0712 | 2,31E-01 | 5,49E-01 | -2,0671 | -1,2770 | 0,0147  |
| 52 | 229893_at   | FRMD3    | Contendo domínio FERM 3                                                                         | 1,2730 | 6,3194 | 8,55E-05 |        |        |          | 1,83E-01 | -1,3901 | -1,1336 | 0,2342  |
| 53 | 228004_at   | C20orf56 | Matriz de leitura aberta 56 do cromossoma 20                                                    | 1,7141 | 6,2459 | 1,26E-04 |        |        |          | 6,93E-01 | -0,4040 | -0,4126 | 0,2826  |
| 54 | 208450_at   | LGALS2   | lectina, que liga galactosídeo, solúvel, 2 (galectina e)                                        | 2,0310 | 6,2396 | 1,31E-04 | 2,4773 | 5,3780 | 3,87E-02 | 7,57E-01 | -1,9311 | -1,770  | 0,0999  |
| 55 | 211253_x_at | PYY      | peptídeo YY                                                                                     | 1,3778 | 6,1703 | 1,88E-04 | 1,5825 | 5,5802 | 1,725,02 | 3,08E-01 | -1,0510 | -0,7604 | 0,2555  |
| 56 | 228821_at   | SIAT2    | sialiltransferase 2 (monosialogangliosídeo sialiltransferase)                                   | 1,2800 | 6,1437 | 2,16E-04 |        |        |          | 1,80E-01 | -1,4124 | -0,1647 | 0,0341  |
| 57 | 214601_at   | TPH1     | triptofano hidroxilase 1 (triptofano 5-monooxigenase)                                           | 1,4092 | 6,0972 | 2,75E-01 | 1,6272 | 5,3527 | 1,27E-D2 | 6,10E4   | 0,5265  | -0,1518 | 0,2462  |
| 58 | 213369_at   | PCDH21   | protocaderina 21                                                                                | 1,4794 | 6,0159 | 4,20E-04 | 1,7538 | 6,0814 | 2,22E-03 | 1,50E-01 | -1,5266 | -0,9169 | 0,1555  |
| 59 | 204686_at   | IRS1     | substrato 1 do receptor de insulina                                                             | 1,4809 | 6,0115 | 4,29E-04 |        |        |          | 2,47E-01 | -1,2000 | -1,1810 | 0,3247  |

|    |             |           |                                                                                     |        |        |          |        |        |          |          |         |         |         |
|----|-------------|-----------|-------------------------------------------------------------------------------------|--------|--------|----------|--------|--------|----------|----------|---------|---------|---------|
| 60 | 202709_at   | FMOD      | Fibromodulina                                                                       | 1,2559 | 5,9660 | 5,43E-04 | 1,2837 | 5,5315 | 2,09E-02 | 9,98E-01 | 0,0024  | -0,3258 | 0,3265  |
| 61 | 234709_at   | CAPN13    | Calpaína 13                                                                         | 1,2740 | 5,9574 | 5,67E-04 |        |        |          | 2,69E-01 | -1,1440 | -0,415  | 0,1239  |
| 62 | 218692_at   | FU20366   | proteína hipotética FU20366                                                         | 1,2335 | 5,9139 | 7,08E-04 |        |        |          | 9,13E-01 | -0,1113 | -0,5029 | 0,4540  |
| 63 | 218532_s_at | FU20152   | proteína hipotética FU20152                                                         | 1,5696 | 5,8952 | 7,79E-04 |        |        |          | 8,96E-02 | -1,8034 | -2,5468 | 0,2020  |
| 64 | 242414_at   |           | ** nenhuma descrição **                                                             | 1,1722 | 5,8510 | 9,76E-04 |        |        |          | 9,69E-01 | -0,0401 | -0,2909 | 0,2801  |
| 65 | 212935_at   | MCF2L     | derivada de linhagem de célula MCF.2<br>equivalente à sequência transformadora      | 1,2007 | 5,8489 | 9,86E-04 |        |        |          | 6,83E-01 | -0,4164 | -0,5219 | 0,3506  |
| 66 | 218510_x_at | FU20152   | proteína hipotética FU20152                                                         | 1,4942 | 5,8115 | 1,19E-03 |        |        |          | 1,77E-01 | -1,4185 | -2,5309 | 0,5086  |
| 67 | 21392-Cat   | SST       | Somatostatina                                                                       | 1,7335 | 5,8030 | 1,24E-03 |        |        |          | 5,61E-01 | -0,5941 | -0,5395 | 0,3039  |
| 68 | 232321_at   | MUC17     | Mucina 17                                                                           | 1,5373 | 5,7650 | 1,51E-03 |        |        |          | 3,94E-02 | -2,2843 | -1,2222 | -0,0353 |
| 69 | 205464_at   | SCNN1B    | Canal de sódio, não ativado por voltagem 1, beta<br>(Sindrome Liddle)               | 1,5884 | 5,7391 | 1,72E-03 |        |        |          | 3,00E-04 | -2,3775 | -2,0960 | -0,1218 |
| 70 | 212098_at   | LOC151162 | proteína hipotética LOC151162                                                       | 1,2162 | 5,7307 | 1,79E-03 |        |        |          | 8,25E-02 | -1,8581 | -1,2610 | 0,0853  |
| 71 | 219973_at   | FU23548   | proteína hipotética FU23548                                                         | 1,0946 | 5,6928 | 2,16E-03 |        |        |          | 2,63E-01 | -1,1632 | 0,0225  |         |
| 72 | 203769_s_at | STS       | esteróide sulfatase (microssômica),<br>arilsulfatase<br>C, isozima S                | 1,1896 | 5,6677 | 2,45E-03 |        |        |          | 6,13E-01 | 0,5151  | 0,3673  |         |
| 73 | 230645_at   | FRMD3     | Contendo domínio FERM 3                                                             | 1,2643 | 5,6646 | 2,49E-03 |        |        |          | 3,38E-01 | 0,9874  | 0,3083  |         |
| 74 | 213432_at   | MUC5B     | mucina 5, subtipo B, traqueobrônquico                                               |        |        | 3,09E-03 | 2,3060 | 6,0011 | 3,095,03 | 1,72E-01 | -1,4427 | 0,2553  |         |
| 75 | 204781_s_at | FAS       | Fas (membro da superfamília do receptor de TNF)                                     | 1,2457 | 5,5988 | 3,44E-03 |        |        |          | 9,60E-01 | -0,0515 | 0,5646  |         |
| 76 | 203021_at   | SLP1      | Inibidor da protease de leucócito secretora<br>(antileucoproteínase)                | 1,6300 | 5,5982 | 3,46E-03 | 2,2457 | 7,0224 | 4,205,05 | 9,88E-01 | -2,9491 | -0,5152 |         |
| 77 | 204044_at   | QPRT      | Quinolinato nucleotídeo pirofosforilase<br>(carboxilante))                          | 1,2874 | 5,5770 | 3,84E-03 |        |        |          | 1,01E-01 | -1,8025 | 0,0689  |         |
| 78 | 228256_s_at | EPB41L4A  | Equivalente à faixa 4.1 de proteína de membrana de eritrócito 4A                    | 1,2835 | 5,5607 | 4,15E-03 |        |        |          | 3,17E-01 | -1,0324 | 0,2054  |         |
| 79 | 219033_at   | PARP8     | família da poli(ADP-ribose) polimerase, membro 8                                    |        |        | 4,48E,03 | 1,2434 | 5,9109 | 4,48E-03 | 8,29E-01 | 0,2199  | 0,2877  |         |
| 80 | 235004_at   | RBM24     | proteína do motivo de ligação de RNA 24                                             | 1,3389 | 5,5145 | 5,21E-03 |        |        |          | 2,33E-01 | -1,2599 | 0,1390  |         |
| 81 | 205009_at   | TFF1      | fator de trevo 1 (câncer de mama, sequência indutível por estrogênio expressada em) | 2,2026 | 5,5133 | 5,24E-03 |        |        |          | 8,36E-02 | -2,7462 | 0,1932  |         |
| 82 | 212959_s_at | MGC4170   | Proteína MGC4170                                                                    |        |        | 5,56E-03 | 1,5719 | 5,8581 | 5,565,03 | 2,94E-01 | -1,088  | 0,6093  |         |
| 83 | 213423_x_at | TUSC3     | Candidato a supressor de tumor 3                                                    | 1,4004 | 5,4510 | 7,09E-03 |        |        |          | 7,01E-01 | -0,390  | 0,2231  |         |

|     |             |          |                                                                                            |        |        |          |        |        |          |          |                    |         |
|-----|-------------|----------|--------------------------------------------------------------------------------------------|--------|--------|----------|--------|--------|----------|----------|--------------------|---------|
| 84  | 211719_x_at | FN1      | Fibronectina 1                                                                             | 1,8475 | 5,4506 | 7,11E,03 | 1,3838 | 5,6336 | 1,19E-02 | 2,60E-(- | 1,16E<br>-0,899    | 3,1093  |
| 85  | 21328-0_at  | GARNL4   | Equivalente ao domínio 4 de Rap/RanGAP<br>que ativa GTPase                                 | 1,2152 | 5,4296 | 7,86E,01 |        |        |          | 4,51E-02 | -2,1642<br>-0,8856 | -0,0110 |
| 86  | 222258_s_at | SH3BP4   | proteína de ligação 4 do domínio de SH3                                                    | 1,2523 | 5,4281 | 7,92E-03 | 1,3838 | 5,6336 | 1,19E-02 | 7,24E-01 | -0,3598<br>-0,749  | 0,5342  |
| 87  | 205221_at   | HGD      | homogentissato 1,2-dioxigenase<br>(homogentissato oxidase)                                 | 1,3595 | 5,4277 | 7,94E-03 |        |        |          | 1,74E-(- | -1,4227<br>-0,8949 | 0,1761  |
| 88  | 226050_at   | Cl3orf11 | Matriz de leitura aberta 11 cromossoma 13                                                  | 1,2961 | 5,4095 | 8,67E-03 |        |        |          | 2,65E-01 | -1,1581<br>-1,2866 | 0,3803  |
| 89  | 225591_x_at | FBX025   | proteína 25 de F-box                                                                       | 1,1734 | 5,3977 | 9,18E-03 |        |        |          | 3,52E-01 | -0,9692<br>-0,5157 | 0,1986  |
| 90  | 209228_x_at | TUSC3    | Candidato a supressor de tumor 3                                                           | 1,3320 | 5,3700 | 1,05E,02 |        |        |          | 2,11E-01 | 1,3517<br>-0,1411  | 0,5509  |
| 91  | 214798_at   | KIAA0703 | Produto do gene KIAA0703                                                                   | 1,2832 | 5,3679 | 1,06E,02 |        |        |          | 9,82E-01 | 0,0230<br>-0,5970  | 0,6096  |
| 92  | 212573_at   | KIAA0830 | Proteína KIAA0830                                                                          |        |        |          | 1,4028 | 5,6938 | 1,09E-02 | 2,89E-02 | -2,4389<br>-3,0784 | -0,1956 |
| 93  | 220136_s_at | CRYBA2   | cristalina, beta A2                                                                        | 1,1975 | 5,3523 | 1,14E,02 |        |        |          | 5,55E-01 | -0,6017<br>-0,3532 | 0,1966  |
| 94  | 41469_at    | P 13     | inibidor da protease 3, derivado de pele<br>(SKALP)                                        | 1,5984 | 5,3485 | 1,16E-02 | 2,1561 | 5,8717 | 5,26E-03 | 4,36E-02 | -2,1967<br>-3,1937 | -0,0531 |
| 95  | 210643_at   | TNFSF11  | Superfamília do fator de necrose de tumor<br>(ligando),<br>membro 11                       | 1,0847 | 5,3372 | 1,23E-02 |        |        |          | 9,40E-01 | -0,0779<br>-0,2315 | 0,2159  |
| 96  | 203697_at   | FRZB     | Proteína relacionada com frisado                                                           |        |        | 1,38E-02 | 1,6734 | 5,6350 | 1,38E-02 | 7,36E-(- | 0,3430<br>-0,3287  | 0,4560  |
| 97  | 205081_at   | CRIP1    | proteína 1 rica em cisteína (intestinal)                                                   | 1,4710 | 5,3107 | 1,39E-02 | 1,7786 | 5,5089 | 2,29E-02 | 9,96E-02 | -1,772,<br>-2,4028 | 0,2364  |
| 98  | 212448_at   | NEDD4L   | Célula precursora neural expressada,<br>Equivalente à desenvolvimento infra-<br>regulada 4 | 1,2048 | 5,3009 | 1,46E-02 |        |        |          | 1,90E-02 | -2,5972<br>-1,0089 | -0,1038 |
| 99  | 210495_x_at | FN1      | Fibronectina 1                                                                             | 1,7618 | 5,2865 | 1,56E-02 |        |        |          | 2,53E-01 | 1,1889<br>-0,774   | 2,7368  |
| 100 | 212464_s_at | FN1      | Fibronectina 1                                                                             | 1,8202 | 5,2855 | 1,57E,02 |        |        |          | 2,72E-01 | 1,1408<br>-0,8472  | 2,8050  |
| 101 | 219734_at   | SIDT1    | Família de transmembrana SID 1, membro 1                                                   | 1,2674 | 5,2552 | 1,81E-02 |        |        |          | 5,73E-01 | -0,5770<br>-0,4726 | 0,2719  |
| 102 | 227048_at   | LAMA1    | Laminina, alfa 1                                                                           |        |        | 1,94E,02 | 1,9692 | 5,5506 | 1,94E-02 | 4,30E-(- | -2,2108<br>-2,5885 | -0,0476 |
| 103 | 216442_x_at | FN1      | Fibronectina 1                                                                             | 1,7670 | 5,2217 | 2,12E-02 |        |        |          | 2,67E-01 | 1,1493<br>-0,8418  | 2,8321  |

|     |             |          |                                                                                          |        |        |          |        |        |          |          |                 |         |
|-----|-------------|----------|------------------------------------------------------------------------------------------|--------|--------|----------|--------|--------|----------|----------|-----------------|---------|
| 104 | 209437_s_at | SPON1    | Espondina 1, proteína de matriz extracelular                                             | 1,2281 | 5,2215 | 2,12E-02 | 1,4613 | 5,5757 | 1,755,02 | 4,36E-01 | 0,81<br>-0,1527 | 0,3266  |
| 105 | 206502_s_at | INSM1    | Associada a insulinoma 1                                                                 | 1,2440 | 5,2145 | 2,19E-02 | 1,4613 | 5,5757 | 1,755,02 | 5,49E-01 | 0,61<br>-0,0582 | 0,1057  |
| 106 | 201097_s_at | ARE4     | Fator 4 de ribosilação de ADP                                                            | 1,2820 | 5,2132 | 2,21E-02 |        |        |          | 1,56E-01 | -1,5017         | 0,4863  |
| 107 | 203649_s_at | PLA2G2A  | Fosfolipase A2, grupo IIA (plaquetas, fluido sinovial)                                   | 1,9975 | 5,2082 | 2,26E-02 |        |        |          | 2,57E-01 | -1,1727         | 0,9107  |
| 108 | 218976_at   | DNAJC12  | Dna (Hsp40) homólogo, subfamília C, membro 12                                            | 1,3074 | 5,2059 | 2,28E-02 |        |        |          | 4,86E-01 | -0,7120         | 0,1421  |
| 109 | 218211_s_at | MLPH     | Melanofina                                                                               | 1,3781 | 5,1857 | 2,51E-02 |        |        |          | 6,65E-01 | -0,4411         | 0,7489  |
| 110 | 203962_s_at | NEBL     | Nebulete                                                                                 | 1,4431 | 5,1725 | 2,67E-02 | 1,6869 | 5,5034 | 2,34E-02 | 2,88E-01 | -1,1152         | 0,2690  |
| 111 | 229555_at   | GALNT5   | UDP-N-acetil-alfa-D-galactosamina                                                        | 1,1612 | 5,1681 | 2,72E-02 |        |        |          | 9,63E-01 | 0,0469          | 0,4503  |
| 112 | 237183_at   | GALNT5   | UDP-N-acetil-alfa-D-galactosamina                                                        | 1,1999 | 5,1605 | 2,82E-02 |        |        |          | 5,07E-01 | -0,6779         | 0,1249  |
| 113 | 211864_s_at | FERIL3   | fer-1-equivalente 3, mioferina (C. elegans)                                              | 1,3242 | 5,1576 | 2,86E-02 |        |        |          | 2,60E-01 | -1,1717         | 0,2648  |
| 114 | 212186_at   | ACACA    | acetil-Coenzima A carboxilase alfa                                                       | 1,1447 | 5,1422 | 3,07E-02 |        |        |          | 4,68E-01 | 0,7418          | 0,4809  |
| 115 | 239814_at   | MMP28    | ** nenhuma descrição **                                                                  |        |        |          | 1,2166 | 5,4248 | 3,21E-02 | 5,39E-01 | 0,6303          | 0,2763  |
| 116 | 219909_at   |          | metaloproteinase de matriz 28                                                            | 1,2335 | 5,1262 | 3,31E-02 |        |        |          | 9,59E-02 | -1,7646         | 0,0864  |
| 117 | 213308_at   | SHANK2   | SH3 e domínios de repetição de anquirina múltiplos 2                                     | 1,2366 | 5,1150 | 3,49E-02 |        |        |          | 6,25E-01 | 0,4985          | 0,2789  |
| 118 | 200677_at   | PTTGIP   | Proteína interventora de transformação de tumor da pituitária 1                          |        |        |          | 1,2472 | 5,4015 | 3,52E-02 | 6,80E-02 | -1,9938         | 0,0799  |
| 119 | 221577_x_at | DGF15    | Fator de diferenciação de crescimento 15                                                 | 1,7442 | 5,1093 | 3,58E-02 |        |        |          | 1,17E-01 | -1,6687         | 0,0942  |
| 120 | 205490_x_at | GIB3     | Proteína de junção de intervalo, beta 3, 31 kDa                                          | 1,2239 | 5,0952 | 3,82E-02 |        |        |          | 9,12E-02 | -1,8032         | 0,0942  |
| 121 | 231814_at   | MUC11    | (conexão 31)                                                                             | 1,7000 | 5,0934 | 3,86E-02 |        |        |          | 1,48E-01 | -1,5371         | 0,0979  |
| 122 | 205518_s_at | CMAH     | Mucina 11                                                                                | 1,3496 | 5,0848 | 4,01E-02 | 2,3413 | 5,4097 | 3,41E-02 | 6,00E-01 | 0,5344          | 0,3865  |
|     |             |          | citidina monofosfatase ácido N-acetilneuramínico hidroxilase                             |        |        |          |        |        |          |          |                 |         |
| 123 | 203691_at   | P13      | (CMP-N-acetilneuraminato monooxigenase) inibidor da protease 3, derivado de pele (SKALP) | 1,7037 | 5,0784 | 4,13E-02 | 2,3708 | 5,4493 | 2,91E-02 | 7,84E-03 | -3,0135         | -0,6304 |
| 124 | 238378_at   |          | ** nenhuma descrição **                                                                  | 1,1627 | 5,0641 | 4,41E-02 |        |        |          | 9,94E-01 | 0,0083          | 0,1090  |
| 125 | 212570_at   | KIAA0830 | Proteína KIAA0830                                                                        |        |        | 4,49E-02 | 1,2827 | 5,3405 | 4,49E-02 | 3,02E-01 | -1,0690         | 0,2309  |
| 126 | 244553_at   |          | ** nenhuma descrição **                                                                  | 1,1397 | 5,0518 | 4,67E-02 | 1,2518 | 6,4213 | 5,37E-04 | 3,56E-01 | -0,9510         | 0,0756  |

TABELA 3: Modelo de prognóstico de intestino grosso de 13 genes para a descoberta de localização de gene pelo SVM.

## Máquina Vetorial de Suporte - Modelo de 13 Genes

|         |                                                                                      |
|---------|--------------------------------------------------------------------------------------|
| PRAC    | Proteína nuclear pequena PRAC                                                        |
| CCL11   | Ligando 11 de quimiocina (motivo c-c)                                                |
| FRZB    | Proteína relacionada com frisoado secretada 2                                        |
| CDF15   | Fator de diferenciação de crescimento 15                                             |
| CLDN8   | Claudina 8                                                                           |
| SEC6L1  | Equivalente a SEC6 1 (S. cerevisae)                                                  |
| SEC6L1  | Equivalente a SEC6 1 (S. cerevisae)                                                  |
| CBA3    | Glicosidase beta, ácido 3 (citossólico)                                              |
| DEFA5   | Defensina, alfa 5, específica de célula de Paneth                                    |
| SPINK5  | Inibidor da serina protease, Kasal tipo 5                                            |
| OSTalfa | Transportador alfa de soluto orgânico                                                |
|         | Alanil (membrana) aminopeptidase (aminopeptidase N, aminopeptidase M, aminopeptidase |
| ANPEP   | aminopeptidase microssômica, CD13, p150)a                                            |
| MUC5    | Mucina 5, subtipo B, traqueobrônquica                                                |

## TABELA 4

## GenrRave - Modelo de 7 Genes

|        |                                                                                      |
|--------|--------------------------------------------------------------------------------------|
| SEC6L1 | Equivalente a SEC6 1 (S. cerevisae)                                                  |
| PRAC   | Proteína nuclear pequena PRAC                                                        |
| SPINK5 | Inibidor da serina protease, Kasal tipo 5                                            |
| SEC6L1 | Equivalente a SEC6 1 (S. cerevisae)                                                  |
|        | Alanil (membrana) aminopeptidase (aminopeptidase N, aminopeptidase M, aminopeptidase |
| ANPEP  | aminopeptidase microssômica, CD13, p150)                                             |
| DEFA5  | Defensina, alfa 5, específica de célula de Paneth                                    |
| CLDN8  | Claudina 8                                                                           |

TABELA 5 - Modelos de GeneRave para Prox vs Distal

|                                                                                                                                        | Erro X-<br>validado | Modelo | SENS  | SPEC  | PPV   | NPV   | LRP    | LRN   |
|----------------------------------------------------------------------------------------------------------------------------------------|---------------------|--------|-------|-------|-------|-------|--------|-------|
| 200832_s_at,SCD 204580_at,MMP12 206637_at,P2RY14 214598_at,CLDN8 219017_at,ETNK1                                                       | 10,828              | 1      | 0,989 | 0,968 | 0,979 | 0,984 | 30,674 | 0,011 |
| 205549_at,PCP4 207249_s_at,SLC28A2 209924_at,CCL18 219140_s_at,RBP4 225458_at,DKFZP56411171 230784_at,PRAC                             | 12,10191            | 2      | 0,947 | 0,952 | 0,968 | 0,922 | 19,579 | 0,055 |
| 201123_s_at,EJF5A 202718_at,IGFBP2 221577_x_at,GDF15 225457_s_at,DKFZP56411171 226654_at,MUC12                                         | 10,82803            | 3      | 0,989 | 0,919 | 0,949 | 0,983 | 12,269 | 0,011 |
| 209728_at,HLA-DRB4 209844_at,HOXB13 221091_at,INSL5 222262_s_at,ETNK1                                                                  | 114,46497           | 4      | 0,926 | 0,919 | 0,946 | 0,891 | 11,486 | 0,080 |
| 202888_s_at,ANPEP 207529_at,DEFA5 221164_x_at,CHST5 226432_at,NA 230360_at,COLM                                                        | 14,64968            | 5      | 0,958 | 0,935 | 0,958 | 0,935 | 14,847 | 0,045 |
| 205464_at,SCNN1B 211719_x_at,FN1 224453_s_at,ETNK1 225290_at,NA 229230_at,OSTalpha 229400_at,HOXD10 230269_at,NA                       | 15,92357            | 6      | 0,958 | 0,903 | 0,938 | 0,933 | 9,898  | 0,047 |
| 201920_at,SLC20A1 211969_at,HSPCA 217320_at,NA 32128_at,CCL18 230105_at,HOXB13                                                         | 17,83439            | 7      | 0,947 | 0,919 | 0,947 | 0,919 | 11,747 | 0,057 |
| 209795_at,CD69 212768_s_at,OLFM4 215125_s_at,UGT1A6 223942_x_at,CHST5 231576_at,NA 231814_at,MUC11                                     | 18,47134            | 8      | 0,937 | 0,919 | 0,947 | 0,905 | 11,617 | 0,069 |
| 203649_s_at,PLA2G2A 205815_at,REG3A 206407_s_at,CCL13 206422_at,GGC 208596_s_at,UGT1A3 210495_x_at,FN1 217546_at,MTIM 236121_at,OR51E2 | 18,47134            | 9      | 0,947 | 0,935 | 0,957 | 0,921 | 14,684 | 0,056 |
| 202236_s_at,SLC16A1 203892_at,WFDCC2 204351_at,S100P 208121_s_at,PTPRO 210133_at,CCL11 219087_at,ASPN 227194_at,FAM3B                  | 14,01174            | 10     | 0,947 | 0,968 | 0,978 | 0,923 | 29,368 | 0,054 |
| 201324_at,EMP1 203962_s_at,NEBL 205009_at,TFPI 205518_s_at,CMAH 207080_s_at,PYY 219955_at,ECAT11 222774_s_at,NETO2                     | 0,178439            | 11     | 0,947 | 0,919 | 0,947 | 0,919 | 11,747 | 0,057 |
| 204818_at,HSD17B2 205221_at,HGD 205950_s_at,CA1 206100_at,CPM 208450_at,LGALS2 214973_x_at,IGHD 216442_x_at,FN1                        | 15,28662            | 12     | 0,958 | 0,919 | 0,948 | 0,934 | 11,878 | 0,046 |
| 206207_at,CLC 207814_at,DEFA6 212464_s_at,FN1 226847_at,FST 236513_at,NA 240856_at,NA 242059_at,ETNK1                                  | 18,47134            | 13     | 0,937 | 0,919 | 0,947 | 0,905 | 11,617 | 0,069 |
| 207558_s_at,PITX2 224009_x_at,DHRS9 229254_at,DKFZp761N1114 234994_at,KIAA1913                                                         | 17,83439            | 14     | 0,947 | 0,871 | 0,918 | 0,915 | 7,342  | 0,060 |
| 205498_at,GHR 206294_at,HSD3B2 207251_at,MEP1B 214651_s_at,HOXA9 224412_s_at,TRPM6 239994_at,NA                                        | 18,47134            | 15     | 0,968 | 0,903 | 0,939 | 0,949 | 10,007 | 0,035 |
| 205185_at,SPINK5 208383_s_at,PCK1 209869_at,ADRA2A 210519_s_at,NQO1 222943_at,GBA3 228004_at,NA                                        | 20,38217            | 16     | 0,947 | 0,887 | 0,928 | 0,917 | 8,391  | 0,059 |

| Erro X-<br>validado                                                                                                                                                                                                                                                                                       | Modelo | SENS  | SPEC  | PPV   | NPV   | LRP    | LRN   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------|-------|-------|-------|--------|-------|
| 205979_at,SCGB2A1 206340_at,NR1H4 218888_s_at,NE:TO2<br>222571_at,ST6GALNAC6                                                                                                                                                                                                                              | 17     | 0,926 | 0,887 | 0,926 | 0,887 | 8,205  | 0,083 |
| 203961_at,NEBL 204304_s_at,PROM1 209173_at,AGR2<br>209752_at,REG1A 221305_s_at,UGT1A8<br>242372_s_at,DKFZp761N1114                                                                                                                                                                                        | 18     | 0,926 | 0,919 | 0,946 | 0,891 | 11,486 | 0,080 |
| 201963_at,ACSL1 203759_at,ST3GAL4 219954_s_at,GBA3<br>221024_s_at,SLC2A10 223952_x_at,DHRS9 227048_at,LAMA1<br>202023_at,EFNA1 202946_s_at,BTBD3 203691_at,PI3<br>209994_s_at,ABCB1 223058_at,C10orf45 228241_at,BCMP11<br>229070_at,C6orf105 234709_at,CAPN13 235706_at,CPM<br>236141_at,NA 238143_at,NA | 19     | 0,958 | 0,887 | 0,929 | 0,932 | 8,484  | 0,047 |
|                                                                                                                                                                                                                                                                                                           | 20     | 0,979 | 0,984 | 0,989 | 0,968 | 60,695 | 0,021 |

## BIBLIOGRAFIA

Affymetrix. 2001a. GeneChip Expression Analysis Data Analysis Fundamentals.

Affymetrix. 2001b. Statistical Algorithms Reference Guide.

5 Affymetrix. 2004. Gene Expression Analysis: Technical Manual. 701021 Rev 5.

Alon, A., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mach, D. e

Levine, A. J. Broad patterns of gene expression revealed by  
10 clustering analysis of tumor and normal colon tissues probed by  
oligonucleotide arrays. Proc. Natl. Acad. Sci. USA: 96, 6745-6750, Junho de  
1999

Ausubel, F. *et al.*, "Current Protocols in Molecular Biology",  
John Wiley & Sons, 1998

15 Bair, E., T. Hastie, P. Debashis e R. Tibshirani. 2004.  
Prediction by supervised principal components. Stanford University

Bara, J., J. Nardelli, C. Gadenne, M. Prade e P. Burtin. 1984.  
Differences in the expression of mucus-associated antigens between proximal  
and distal human colon adenocarcinomas. Br J Cancer 49: 495-501.

20 Bates, M. D., C. R. Erwin, L. P. Sanford, D. Wiginton, J. A.  
Bezerra, L.

C. Schatzman, A. G. Jegga, C. Ley-Ebert, S. S. Williams, K.  
A. Steinbrecher, B. W. Warner, M. B. Cohen e B. J. Aronow. 2002. Novel  
genes and functional relationships in the adult mouse gastrointestinal tract  
25 identified by microarray analysis. Gastroenterology 122: 1467-1482.

Birkenkamp-Demtroder, K., S. H. Olesen, F. B. Sorensen, S.  
Laurberg, P. Laiho, L. A. Aaltonen e T. F. Orntoft 2005. Differential gene  
expression in colon cancer of the caecum versus the sigmoid and  
rectosigmoid. Gut 54: 374-384.

Bonithon-Kopp, C. e A. M. Benhamiche. 1999. Are there several colorectal cancers? Epidemiological data. *Eur J Cancer Prev* 8 Supl 1: S3-12.

5 Bonner T. I., Brenner D. J., Neufeld B. R. e Britten R. J. (1973) Reduction in the rate of DNA reassociation by sequence divergence. *J. Mol. Biol.* 81: 123-125

Bufill, J. A. 1990. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med* 113: 779-788.

10 Byrd, J. C. e R. S. Bresalier. 2004. Mucins and mucin binding proteins in colorectal cancer. *Cancer Metastasis Rev* 23: 77-99.

Calamita, G., A. Mazzone, A. Bizzoca, A. Cavalier, G. Cassano, D. Thomas e M. Svelto. 2001. Expression and immunolocalization of the aquaporin-8 water channel in rat gastrointestinal tract. *Eur J Cell Biol* 15 80: 711-719.

Caldero, J., E. Campo, C. Ascaso, J. Ramos, M. J. Panades e J. M. Rene. 1989. Regional distribution of glycoconjugates in normal, transitional and neoplastic human colonic mucosa. A histochemical study using lectins. *Virchows Arch A Pathol Anat Histopathol* 415: 347-356.

20 Chalmers, A. D., J. M. Slack e C. W. Beck. 2000. Regional gene expression in the epithelia of the *Xenopus* tadpole gut. *Mech Dev* 96: 125-128.

Chen, M., Y. Yang, E. Braunstein, K. E. Georgeson e C. M. Harmon. 2001. Gut expression and regulation of FAT/CD36: possible role in fatty acid transport in rat enterocytes. *Am J Physiol Endocrinol Metab* 25 281: E916-23.

Colegio, O. R., C. M. Van Itallie, H. J. McCrea, C. Rahner e J. M. Anderson. 2002. Claudins create charge-selective channels in the paracellular pathway between epithelial cells. *Am J Physiol Cell Physiol* 283:

C142-7.

Cristianini, N. e J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.

- 5 Cristianini, N., Shawe-Taylor, J. Support Vector Machines. 2000. Cambridge University Press. Cambridge.

Cuff, M. A., D. W. Lambert e S. P. Shirazi-Beechey. 2002. Substrate-induced regulation of the human colonic monocarboxylate transporter, MCT1. JPhysiol 539: 361-371.

- 10 de Santa Barbara, P., G. R. van den Brink e D. J. Roberts. 2003. Development and differentiation of the intestinal epithelium. Cell Mol Life Sci 60: 1322-1332.

- 15 Deng, G., E. Peng, J. Gum, J. Terdiman, M. Sleisenger e Y. S. Kim. 2002. Methylation of hMLH1 promoter correlates with the gene silencing with a region-specific manner in colorectal cancer. Br J Cancer 86: 574-579.

DeRisi, *et al.*, Nature Genetics, 14: 457-460 (1996)

Distler, P. e P. R. Holt. 1997. Are right- and left-sided colon neoplasms distinct tumors? Dig Dis 15: 302-311.

- 20 Drmanac R., Labat I. e Crkvenjakov R., An algorithm for the DNA sequence generation from k-tuple word contents of the minimal number of random fragments../. Biomol. Struc. & Dyn. 5: 1085-1102, 1991

Filipe, M. I. e A. C. Branfoot. 1976. Mucin histochemistry of the colon. Curr Top Pathol 63: 143-178.

- 25 Fleming, R. E., S. Parkkila, A. K. Parkkila, H. Rajaniemi, A. Waheed e W. S. Sly. 1995. Carbonic anhydrase IV expression in rat and human gastrointestinal tract regional, cellular, and subcellular localization. J Clin Invest 96: 2907-2913.

Garcia-Hirschfeld Garcia, J., A. Blanes Berenguel, L. Vicioso Recio, A. Marquez Moreno, J. Rubio Garrido e A. Matilla Vicente. 1999.

Colon cancer: p53 expression and DNA ploidy. Their relation to proximal or distal tumor site. *Rev Esp Enferm Dig* 91: 481-488.

Gautier, L., L. Cope, B. M. Bolstad e R. A. Irizarry. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20: 307-315.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang e J. Zhang. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genuine Biol* 5: R80.

Germer S, Holland M J, Higuchi R. 2000, High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* 10(2): 258-66.

Glebov, O. K., L. M. Rodriguez, K. Nakahara, J. Jenkins, J. Cliatt, C. J. Humbyrd, J. DeNobile, P. Soballe, R. Simon, G. Wright, P. Lynch, S. Patterson, H. Lynch, S. Gallinger, A. Buchbinder, G. Gordon, E. Hawk e I. R. Kirsch. 2003. Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiol Biomarkers Prev* 12: 755-762.

Gum, J. R. J., S. C. Crawley, J. W. Hicks, D. E. Szymkowski e Y. S. Kim. 2002. MUC17, a novel membrane-tethered mucin. *Biochem Biophys Res Commun* 291: 466-475.

Guo Z, Guilfoyle R A, Thiel A J, Wang R, Smith L M. 1994, Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res.* 22(24): 5456-65

Hastie, T, Tibshirani, R, Friedman, J, *The Elements of Statistical Learning*. Springer, 2001. Nova Iorque. Capítulo 4: Linear Methods for Classification. Hostikka, S. L. e M. R. Capecchi. 1998. The mouse Hoxcl 1 gene: genomic structure and expression pattern. *Mech Dev* 70: 133-145.

Hubbell, E., W. M. Liu e R. Mei. 2002. Robust estimators for expression analysis. *Bioinformatics* 18: 1585-1592.

Iacopetta, B. 2002. Are there two sides to colorectal cancer? *Int J Cancer* 101: 403-408.

5 Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs e T. P. Speed. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.

James, R., T. Erler e J. Kazenwadel. 1994. Structure of the murine homeobox gene *cdx2*. Expression in embryonic and adult intestinal epithelium. *J Biol Chem* 269: 15229-15237.

Jeansonne, B., Q. Lu, D. A. Goodenough e Y. H. Chen. 2003. Claudin-8 interacts with multi-PDZ domain protein 1 (MUPP1) and reduces paracellular conductance in epithelial cells. *Cell Mol Biol (Noisy-le-grand)* 49: 13-21.

15 Kiiveri, H. T. A bayesian approach to variable selection when the number of variables is very large *Science and Statistics: A Festschrift for Terry Speed*, 2003 Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Vol. 3, páginas 127-143

20 Kiiveri, H., Thomas, M., Dunne, R., Method and Apparatus for Identifying Diagnostic Components of Asystem with a characteristic response, Pedido de Patente Internacional Nº PCT/AU2002/000934

Komuro, K., M. Tada, E. Tamoto, A. Kawakami, A. Matsunaga, K. Teramoto, G. Shindoh, M. Takada, K. Murakawa, M. Kanai, N. Kobayashi, Y. Fujiwara, N. Nishimura, J. Hamada, A. Ishizu, H. Ikeda, S. 25 Kondo, H. Katoh, T. Moriuchi e T. Yoshiki. 2005. Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis. *J Surg Res* 124: 216-224.

Kondo, T., P. Dolle, J. Zakany e D. Duboule. 1996. Function

of posterior HoxD genes in the morphogenesis of the anal sphincter. *Development* 122: 2651-2659.

Kosaki, K., R. Kosaki, T. Suzuki, H. Yoshihashi, T. Takahashi, K. Sasaki, M. Tomita, W. McGinnis e N. Matsuo. 2002. Complete  
5 mutation analysis panel of the 39 human HOX genes. *Teratology* 65: 50-62.

Krzanowski, W e Marriott, F, *Multivariate Analysis Part II. Classification Covariance Structures and Repeated Measures*. 1995. Oxford Univ Press. Oxford. UK. Lipshutz, R. J., S. P. Fodor, T. R. Gingeras e D. J. Lockhart. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* 21:  
10 20-24.

Liu, X. F., P. Olsson, C. D. Wolfgang, T. K. Bera, P. Duray, B. Lee e I. Pastan. 2001. PRAC: A novel small nuclear protein that is specifically expressed in human prostate and colon. *Prostate* 47: 125-131.

Macfarlane, G. T., G. R. Gibson e J. H. Cummings. 1992.  
15 Comparison of fermentation reactions in different regions of the human colon. *J Appl Bacteriol* 72: 57-64.

Maskos e Southern, *Nuc. Acids Res.* 20: 1679-84, 1992

Miklos, G. L. e R. Maleszka. 2004. Microarray reality checks in the context of a complex disease. *Nat Biotechnol* 22: 615-621.

20 Montgomery, R. K., A. E. Mulberg e R. J. Grand. 1999. Development of the human gastrointestinal tract: twenty years of progress. *Gastroenterology* 116: 702-731.

Moore, A., Basilion, J., Chiocca, e., e Weissleder, R., Measuring Transferrin Receptor Gene Expression by NMR Imaging. *BBA*,  
25 1402: 239-249, 1988

Ortega-Cava, C. F., S. Ishihara, M. A. Rumi, K. Kawashima, N. Ishimura, H. Kazumori, J. Udagawa, Y. Kadowaki e Y. Kinoshita. 2003. Strategic compartmentalization of Toll-like receptor 4 in the mouse gut. *J Immunol* 170: 3977-3985.

Park, Y. K., J. L. Franklin, S. H. Settle, S. E. Levy, E. Chung, L. H. Jeyakumar, Y. Shyr, M. K. Washington, R. H. Whitehead, B. J. Aronow e R. J. Coffey. 2005. Gene expression profile analysis of mouse colon embryonic development. *Genesis* 41: 1-12.

5                   Pease A C, Solas D, Sullivan E J, Cronin M T, Holmes C P, Fodor S P., 1994, Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA*. 91(11): 5022-6

                    Peifer, M. 2002. Developmental biology: colon construction. *Nature* 420: 274-5, 277.

10                  Pevzner P A., 1989, 1-Tuple DNA sequencing: computer analysis., *JBiomol Struct Dyn*. 7(1): 63-73

                    Pevzner P A, Lysov YuP, Khrapko K R, Belyavsky A V, Florentiev V L, Mirzabekov A D., 1991, Improved chips for sequencing by hybridization., *J Biomol Struct Dyn*. 9(2): 399-410

15                  R: A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Viena, Áustria, 2007, ISBN 3-900051-07-0.

                    Rajendran, V. M., J. Black, T. A. Ardito, P. Sangan, S. L. Alper, C. Schweinfest, M. Kashgarian e H. J. Binder. 2000. Regulation of  
20   DRA and AE1 in rat colon by dietary Na depletion. *Am J Physiol Gastrointest Liver Physiol* 279: G931-42.

                    Ripley, B D, Cambridge Univ Press. 1996. Pattern Recognition and Neural Networks. Capítulo 6: Non-parametric methods.

                    Sano T, Cantor C R., 1991, A streptavidin-protein A chimera  
25   that allows one-step production of a variety of specific antibody conjugates., *Biotechnology (N Y)*. 9(12): 1378-81

                    Schena, *et al.* *Science* 270: 467-470, 1995

                    Scholkopf, B, Tsuda, K, e Vert, J P *Kernel Methods in Computational Biology*. 2004. MIT Press. Cambridge MA.

Silberg, D. G., G. P. Swain, E. R. Suh e P. G. Traber. 2000. Cdx1 and cdx2 expression during intestinal development. *Gastroenterology* 119: 961-971.

5 Singh, S., R. Poulson, A. M. Hanby, L. A. Rogers, N. A. Wright, M. C. Sheppard e M. J. Langman. 1998. Expression of oestrogen receptor and oestrogen-inducible genes pS2 and ERD5 in large bowel mucosa and cancer. *J Pathol* 184: 153-160.

10 Smith S B, Finzi L, Bustamante C., 1992, Direct Mechanical Measurements of the Elasticity of Single DNA Molecules by Using Magnetic Beads, *Science* 258: 1122-1126

Smyth, G. 2005. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. (eds. Gentleman, R., V. Carey, S. Dudoit, R. Irizarry e W. Huber), pp. 397-420. Springer, Nova Iorque.

15 Traber, P. G. 1999. Transcriptional regulation in intestinal development. Implications for colorectal cancer. *Adv Exp Med Biol* 470: 1-14.

Urdea *et al.*, *Nucleic Acids Symp. Ser.*, 24: 197-200, 1991

20 Venables, W. e Ripley, B. D., *Modern Applied Statistics with S*, Springer-Verlag. Nova Iorque, 2002.

Wedemeyer, N., Potter, T., Wetzlich, S. e Gohde, W. Flow Cytometric Quantification of Competitive Reverse Transcriptase-PCR products, *Clinical Chemistry* 48: 9 1398-1405, 2002

25 Weissleder, R., Moore, A., Ph. D., Mahmood-Bhorade, U., Benveniste, H., Chiocca, E. A., Basilion, J. P. High resolution in vivo imaging of transgene expression, *Nature Medicine* 6: 351-355, 2000

Williams, S. J., M. A. McGuckin, D. C. Gotley, H. J. Eyre, G. R. Sutherland e T. M. Antalis. 1999. Two novel mucin genes down-regulated in colorectal cancer identified by differential display. *Cancer Res* 59: 4083-

4089.

Wilson, C. e C. J. Miller. 2005. Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis. *Bioinformatics*

5 Yamada, T. e D. H. Alpers. 2003. Textbook of Gastroenterology, 2 Vol. Set.

## REIVINDICAÇÕES

1. Método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, caracterizado pelo fato de que inclui:

5                   avaliar os dados de treinamento, incluindo a expressão dos dados de treinamento que representam a expressão de genes nas células ou populações celulares derivadas de origens proximal-distal conhecidas de pelo menos um intestino grosso, e a origem proximal-distal dos dados de treinamento que representam associações das ditas células ou populações  
10                   celulares com as ditas origens proximal-distal;

                    processar os dados de treinamento para gerar dados de classificação que representam uma combinação linear ou não linear de níveis de expressão do dito genes, os ditos dados de classificação sendo adaptados para gerar outros dados de origem proximal-distal indicativos de uma origem  
15                   proximal-distal de uma outra célula ou subpopulação celular tirada de um intestino grosso, com base em outros dados de expressão que representam a expressão do dito genes na dita outra célula ou subpopulação celular.

                    2. Método de acordo com a reivindicação 1, caracterizado pelo fato de que inclui processar os ditos dados de classificação e os ditos outros  
20                   dados de expressão para gerar os ditos outros dados de origem proximal-distal.

                    3. Método de acordo com a reivindicação 1 ou 2, caracterizado pelo fato de que o dito processamento é fundamentado na regressão estatística, linear generalizada, e/ou regressão linear múltipla.

25                   4. Método de acordo com qualquer uma das reivindicações de 1 a 3, caracterizado pelo fato de que o dito processar inclui processar os ditos dados de treinamento com GeneRave.

                    5. Método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo,

caracterizado pelo fato de que inclui:

5 avaliar os dados de treinamento, incluindo expressão os dados de treinamento que representam a expressão de genes nas células ou populações celulares derivadas de origens proximal-distal conhecidas de um intestino grosso, e origem proximal-distal dos dados de treinamento que representam associações das ditas células ou populações celulares com as ditas origens proximal-distal;

10 processar os dados de treinamento usando análise multivariada para gerar dados de classificação para gerar dados de origem proximal-distal indicativos de uma origem proximal-distal de uma outra célula ou população celular derivada de um intestino grosso, com base em outros dados de expressão que representam a expressão de genes na dita outra célula ou população celular.

15 6. Método de acordo com a reivindicação 5, caracterizado pelo fato de que inclui processar os ditos outros dados de expressão e os ditos dados de classificação para gerar os ditos dados de origem proximal-distal.

7. Método de detecção para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, caracterizado pelo fato de que inclui:

20 avaliar os primeiros dados de expressão que representam a expressão de genes nas células ou populações celulares derivada de origens proximal-distal conhecidas de pelo menos um intestino grosso;

25 processar os primeiros dados de expressão usando análise multivariada para gerar dados modelo multivariados representativos de associações entre os primeiros dados de expressão e origem proximal-distal das ditas células ou populações celulares, os ditos dados modelo multivariados sendo adaptados para gerar dados de origem proximal-distal representativos de origem proximal-distal de uma célula ou população celular com base na segunda dado de expressão que representa a expressão de genes

nas ou população celular derivada do intestino grosso de um indivíduo.

8. Método de acordo com a reivindicação 7, caracterizado pelo fato de que inclui avaliar os ditos segunda expressão dados que representam a expressão de genes em uma célula ou população celular derivada do intestino  
5 grosso de um indivíduo; e

processar os ditos dados de expressão e os dados modelo multivariados para gerar os ditos dados de origem proximal-distal representativos de uma origem proximal-distal das ditas célula ou população celular.

9. Método de acordo com a reivindicação 7, caracterizado pelo fato de que a dita etapa de avaliar primeiros dados de expressão inclui avaliar a terceira dados de expressão dos quais os ditos primeiros dados de expressão é um subconjunto, e o método inclui processar os ditos terceiro dados de expressão para selecionar um subconjunto do terceiro dados de expressão  
10 correspondendo a um subconjunto de genes diferencialmente expressados sozinhos ou em combinação junto o eixo proximal-distal dos ditos intestino grosso, o subconjunto selecionado sendo os ditos primeiros dados de expressão.

10. Método para determinar a origem anatômica de uma célula  
20 ou população celular derivada do intestino grosso de um indivíduo, caracterizado pelo fato de que inclui:

avaliar primeiros dados de expressão que representam a expressão de genes nas células ou populações celulares derivada de origens proximal-distal conhecidas de pelo menos um intestino grosso; e

25 processar os primeiros dados de expressão usando um método kernel para gerar dados de classificação para processar os ditos dados de expressão que representam a expressão do dito genes em pelo menos um segunda célula ou população celular de um intestino grosso para gerar dados de origem proximal-distal que representam a origem proximal-distal dos ditos

pelo menos uma segunda célula ou população celular.

11. Método de acordo com a reivindicação 10, caracterizado pelo fato de que os ditos método kernel inclui um máquina de vetor de suporte (SVM).

5 12. Método de acordo com a reivindicação 10 ou 11, caracterizado pelo fato de que o método inclui processar os ditos segunda dados de expressão e os ditos dados de classificação para gerar os ditos dados de origem proximal-distal.

10 13. Método de detecção para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, caracterizado pelo fato de que inclui:

avaliar primeiros dados de expressão que representam a expressão de genes nas células ou populações celulares derivada de origens proximal-distal conhecidas de pelo menos um intestino grosso;

15 processar os primeiros dados de expressão usando principal para gerar principal dados correspondentes a pelo menos um combinação linear a expressão do dito genes, os ditos principal dados sendo indicativos de pelo menos um dos ditos origem proximal-distal das ditas células ou populações celulares.

20 14. Método de acordo com a reivindicação 13, caracterizado pelo fato de que a dita etapa de avaliar primeiros dados de expressão inclui avaliar terceiro dados de expressão de que os ditos primeiros dados de expressão é um subconjunto, e o método inclui processar os ditos terceiro dados de expressão um subconjunto terceiro dados de expressão a um subconjunto de genes expressados junto ao eixo proximal-distal dos ditos um  
25 intestino grosso, subconjunto sendo os ditos primeiros dados de expressão.

15. Método de acordo com a reivindicação 13 ou 14, caracterizado pelo fato de que inclui processar os ditos principal dados e segunda dados de expressão que representam a expressão do dito genes em

pelo menos um segunda célula ou população celular de um intestino grosso para gerar dados de origem proximal-distal que representam a origem proximal-distal.

5 16. Método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, caracterizado pelo fato de que inclui:

avaliar primeiros dados de expressão que representam a expressão de genes in um célula ou população celular derivada de origens proximal-distal conhecidas de pelo menos um intestino grosso; e

10 processar a dados de expressão usando análise de variante canônica para gerar dados indicativos de pelo menos uma das origens proximais-distais das ditas células ou populações celulares.

17. Método de acordo com a reivindicação 16, caracterizado pelo fato de que a dita análise variada canônica inclui análise de perfil.

15 18. Método de acordo com a reivindicação 16 ou 17, caracterizado pelo fato de que os ditos dados variados canônicos correspondem a um subconjunto do dito genes.

19. Método de acordo com qualquer uma das reivindicações de 25 a 27, caracterizado pelo fato de que inclui processar os ditos dados variados canônicos e segunda dados de expressão que representam a expressão do dito genes em pelo menos um segunda célula ou população celular de um intestino grosso para gerar dados de origem proximal-distal que representam a origem proximal-distal de os ditos pelo menos uma segunda célula ou população celular.

25 20. Método de acordo com qualquer uma das reivindicações de 1 a 19, caracterizado pelo fato de que inclui modificar os dados de classificação com base classificação da origem proximal-distal indicada pelos dados de origem proximal-distal.

21. Método de acordo com qualquer uma das reivindicações de

1 a 10, caracterizado pelo fato de que os ditos origem proximal-distal é determinado usando um método não paramétrico.

22. Método de acordo com a reivindicação 21, caracterizado pelo fato de que os ditos método não paramétrico inclui um método do vizinho mais próximo.

23. Método de acordo com qualquer uma das reivindicações de 1 a 22, caracterizado pelo fato de que o dito genes incluem genes selecionados de:

o gene ou genes detectados pela sonda Affymetrix número:

10 218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

o gene detectado pela sonda Affymetrix número: 235733\_at

15 o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

o gene detectado pela sonda Affymetrix número: 230105\_at

20 o gene detectado pela sonda Affymetrix número: 230269\_at

o gene detectado pela sonda Affymetrix número: 238378\_at

o gene detectado pela sonda Affymetrix número: 239814\_at

o gene detectado pela sonda Affymetrix número: 239994\_at

o gene detectado pela sonda Affymetrix número: 240856\_at

25 o gene detectado pela sonda Affymetrix número: 242414\_at

o gene detectado pela sonda Affymetrix número: 244553\_at

o gene detectado pela sonda Affymetrix número: 217320

o gene detectado pela sonda Affymetrix número: 236141

o gene detectado pela sonda Affymetrix número: 236513

o gene detectado pela sonda Affymetrix número: 238143

ABHD5, FAM3B, IGFBP2, POPDC3,

ADRA2A, FLJ10884, KCNG1, REG1A,

APOBEC1, FLJ22761, KIFAP3, SLC14A2,

5 ClOorf45, FTHFD, LOC375295, SLC20A1,

ClOorf58, GCNT1, ME3, SLC23A3,

CCL8, HAS3, MEP1B, SLC38A2,

CLDN15, HOXB6, NPY6R, SLC9A3,

DEFA5, HOXD4, NR1H3, TBCC,

10 EYA2, HSD3B2, HR1H4, ZNF493, OSTalpha,

PAP, AFARP1 ou

o gene ou genes detectados pela sonda Affymetrix número:

202234\_s\_at,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix

15 número: 202888\_s\_at,

CCL13 ou o gene ou genes detectados pela sonda Affymetrix

número: 206407\_s\_at

CRYL1 ou o gene ou genes detectados pela sonda Affymetrix

número: 220753\_s\_at,

20 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix

número: 206754\_s\_at,

CYP2C18, ou o gene ou genes detectados pela sonda

Affymetrix número: 208126 s at,

CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix

25 número: 21442 l\_x\_at ou 220017\_x\_at,

EPB41L3 ou o gene ou genes detectados pela sonda

Affymetrix número: 211776\_s\_at

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix

número: 222262\_s\_at ou 224453\_s\_at,

FAM45A ou o gene ou genes detectados pela sonda Affymetrix número: 221804\_s\_at ou 222955\_s\_at,

FGFR2 ou o gene ou genes detectados pela sonda Affymetrix número: 203639\_s\_at,

5 GBA3 ou o gene ou genes detectados pela sonda Affymetrix número: 219954\_s\_at,

GSPT2 ou o gene ou genes detectados pela sonda Affymetrix número: 205541\_s\_at,

10 GULP1 ou o gene ou genes detectados pela sonda Affymetrix número: 215913\_s\_at,

HOXA9 ou o gene ou genes detectados pela sonda Affymetrix número: 205366\_s\_at ou 214551\_s\_at,

HOXC6 ou o gene ou genes detectados pela sonda Affymetrix número: 206858\_s\_at,

15 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix número: 206601\_s\_at,

ME2 ou o gene ou genes detectados pela sonda Affymetrix número:

20 MESP1 ou o gene ou genes detectados pela sonda Affymetrix número: 224476\_s\_at

MOCS 1 ou o gene ou genes detectados pela sonda Affymetrix número: 213181\_s\_at,

MSCP ou o gene ou genes detectados pela sonda Affymetrix número: 218136\_s\_at ou 221920\_s\_at,

25 NETO2 ou o gene ou genes detectados pela sonda Affymetrix número: 222774\_s\_at,

OASL ou o gene ou genes detectados pela sonda Affymetrix número: 210757\_s\_at,

PITX2 ou o gene ou genes detectados pela sonda Affymetrix

número: 207558 s at

PRAP1 ou o gene ou genes detectados pela sonda Affymetrix

número: 243669\_s\_at,

SCUBE2 ou o gene ou genes detectados pela sonda

5 Affymetrix número: 219197\_s\_at,

SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix

número: 225457\_s\_at,

SLC16A1 ou o gene ou genes detectados pela sonda

Affymetrix número: 202236\_s\_at ou 209900\_s\_at,

10 UGT1A3 ou o gene ou genes detectados pela sonda

Affymetrix número: 208596\_s\_at,

UGT1A8 ou o gene ou genes detectados pela sonda

Affymetrix número: 221305\_s\_at

ACACA, FMOD, LOC151162, SLOOP,

15 CL3orf1, FRMD3, MCF2L, SCGB2A1,

C20orf56, GALNT5, MMP28, SCNN1B,

CAPN13, GARNL4, MUC11, SHANK2,

CLDN8, GCG, MUC12, SIAT2,

COLM, GNE, MUC17, SIAT4C,

20 CRIP1, HGD, MUC5B, SIAT7F,

DNAJC12, HOXB13, NEDD4L, SIDT1,

FAM3C, INSL5, PARP8, SLC13A2,

FBX025, IRS1, PCDH21, SLPI,

FLJ20366, ISL1, PI3, SPINK5,

25 FLJ20989, KIAA0703, PRAC, SST,

KIAA0830, PRAC2, TFF1,

KIAA1913, PTTGLIP, TNFSF11,

LAMA1, QPRT, TPH1,

LGALS2, QSCN6, WFDC2,

RBM24,

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201097\_s\_at,

BTG3 ou o gene ou genes detectados pela sonda Affymetrix  
5 número: 213134 x at ou 205548 \_s\_at

CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221164xat ou 223942\_x\_at,

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
número: 205518\_s\_at,

10 CRYBA2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 220136\_s\_at

CTSE ou o gene ou genes detectados pela sonda Affymetrix  
número: 205927\_s\_at,

DKFZp761N1114 ou o gene ou genes detectados pela sonda  
15 Affymetrix número: 242372\_s\_at,

EPB41L4A ou o gene ou genes detectados pela sonda  
Affymetrix número: 228256 s at

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206070 s at,

20 FAS ou o gene ou genes detectados pela sonda Affymetrix  
número: 204781\_s\_at,

FER1L3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 201798sat ou 211864\_s\_at,

F1120152 ou o gene ou genes detectados pela sonda  
25 Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FLJ23548 ou o gene ou genes detectados pela sonda  
Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número:

- 211719\_s\_at ou 210495\_x\_at ou 212464 at ou 216442\_x\_at,  
FOXA2 ou o gene ou genes detectados pela sonda Affymetrix  
número: 210103\_s\_at,
- FRZB ou o gene ou genes detectados pela sonda Affymetrix  
5 número: 203698\_s\_at,  
GDF15 ou o gene ou genes detectados pela sonda Affymetrix  
número: 221577\_x\_at,  
GJB3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205490\_s\_at,
- 10 HOXD13 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207397\_s\_at,  
INSM1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 206502\_s\_at,  
MGC4170 ou o gene ou genes detectados pela sonda  
15 Affymetrix número: 212959\_s\_at,  
MLPH ou o gene ou genes detectados pela sonda Affymetrix  
número: 218211\_s\_at,  
NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,
- 20 PLA2G2A ou o gene ou genes detectados pela sonda  
Affymetrix número: 203649\_s\_at,  
PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121 s at,  
PYY ou o gene ou genes detectados pela sonda Affymetrix  
25 número: 207080\_s\_at ou 211253x\_at,  
SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix  
número: 222258\_s\_at,  
SLC28A2 ou o gene ou genes detectados pela sonda  
Affymetrix número: 207249\_s\_at

SLC2A10 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221024\_s\_at,

SPON1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213994\_s\_at ou 209437\_s\_at

5 STS ou o gene ou genes detectados pela sonda Affymetrix  
número: 203769\_s\_at

TM4SF11 ou o gene ou genes detectados pela sonda  
Affymetrix número: 204519\_s\_at,

10 TUSC3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 213432\_s\_at ou 209228\_x\_at,

|           |          |          |            |
|-----------|----------|----------|------------|
| AQP8      | LGALS2   | EFNA1    | ORF51E2    |
| CCL11     | C6ORF105 | EMP1     | PROM1      |
| CLDN8     | CCL11    | FST      | REG3A      |
| MMP12     | CD69     | GHR      | SCNN1B     |
| 15 P2RY14 | CLC      | HLA-DRB4 | ST3GAL4    |
| CCL18     | CPM      | HOXD10   | ST6GALNAC6 |
| ACSL1     | DEFA6    | HSD17B2  |            |
| AGR2      | DHRS9    | HSPCA    |            |
| ASPN      |          | IGHD     |            |

20 MT1M

SCD ou o gene ou genes detectados pela sonda Affymetrix  
número 200832\_s\_at,

ABCB1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211994\_s\_at,

25 BTBD3 ou o gene ou genes detectados pela sonda Affymetrix  
número: 202946\_s\_at

CA1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 205950\_s\_at,

DHRS9 ou o gene ou genes detectados pela sonda Affymetrix

número:

224009\_x\_at ou 223952\_x\_at,

DKFZP56411171 ou o gene ou genes detectados pela sonda Affymetrix número: 225457\_s\_at,

5 EIF5A ou o gene ou genes detectados pela sonda Affymetrix número: 201123\_s\_at,

IGHD ou o gene ou genes detectados pela sonda Affymetrix número: 214973\_x\_at,

10 PCK1 ou o gene ou genes detectados pela sonda Affymetrix número: 208383\_s\_at,

RBP4 ou o gene ou genes detectados pela sonda Affymetrix número: 219140\_s\_at,

TRPM6 ou o gene ou genes detectados pela sonda Affymetrix número: 224412\_s\_at,

15 UGT1A6 ou o gene ou genes detectados pela sonda Affymetrix número: 215125 s at.

24. Método de acordo com qualquer uma das reivindicações de 1 a 23, caracterizado pelo fato de que o dito genes incluem apenas 7 genes.

20 25. Método de acordo com qualquer uma das reivindicações de 1 a 24, caracterizado pelo fato de que o dito genes incluem SEC6L1, PRAC, SP1NK5, SEC6L1, ANPEP, DEFA5, e CLDN8.

26. Método de acordo com qualquer uma das reivindicações de 1 a 23, caracterizado pelo fato de que o dito genes incluem um ou mais dos seguintes grupos de genes:

25 (i) SCD ou o gene ou genes detectados pela sonda Affymetrix número: 200832\_s\_at, MMP12 P2RY14 CLDN8 ETNK1

(ii) PCP4 SLC28A2 ou o gene ou genes detectados pela sonda Affymetrix número: 207249\_s\_at,

CCL18

- (iii) RBP4 ou o gene ou genes detectados pela sonda  
Affymetrix número: 219140\_s\_at,  
DKFZP564I1171 PRAC
- (iv) EIF5A ou o gene ou genes detectados pela sonda  
5 Affymetrix número: 201123\_s\_at,  
IGFBP2
- (v) GDF 15 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221577 s at
- (vi) DKFZP564I1171 ou o gene ou genes detectados pela  
10 sonda Affymetrix número: 225457 s at
- (vii) ANPEP ou o gene ou genes detectados pela sonda  
Affymetrix número: 202888 s at
- (viii) CHST5 ou o gene ou genes detectados pela sonda  
Affymetrix número: 221164\_x\_at,
- (ix) O gene detectado pela sonda Affymetrix No. 226432 at  
15 COLM  
SCNN1B
- (x) FN1 ou o gene ou genes detectados pela sonda Affymetrix  
número: 211719 x at
- ETNK1 ou o gene ou genes detectados pela sonda Affymetrix  
20 número: 224453\_s\_at,  
O gene detectado pela sonda Affymetrix No. 225290\_at  
SLC20A1 HSPCA  
O gene detectado pela sonda Affymetrix No. 217320\_at
- CCL18  
25 PTPRO ou o gene ou genes detectados pela sonda Affymetrix  
número: 208121\_s\_at,  
NEBL ou o gene ou genes detectados pela sonda Affymetrix  
número: 203962\_s\_at,

TFF1

CMAH ou o gene ou genes detectados pela sonda Affymetrix

número:

O gene detectado pela sonda Affymetrix No. 239994\_at

5

SPINK5

PCK1 ou o gene ou genes detectados pela sonda Affymetrix

número: 208383\_s\_at,

27. Método de acordo com qualquer uma das reivindicações de 1 a 23, caracterizado pelo fato de que os ditos dados de classificação é representativo de um subconjunto de 13 genes.

10

28. Método de acordo com qualquer uma das reivindicações de 1 a 27, caracterizado pelo fato de que o dito genes incluem: PRAC,

CCL11,

FRZB ou o gene ou genes detectados pela sonda Affymetrix

15

número: 203698\_s\_at, GDF 15 ou o gene ou genes detectados pela sonda Affymetrix número: 221577\_x\_at, CLDN8,

SEC6L1 ou o gene ou genes detectados pela sonda Affymetrix número: 221577\_x\_at, GBA3 ou o gene ou genes detectados pela sonda Affymetrix número: 279954\_s\_at, DEFA5,

20

SPINK5,

ANPEP ou o gene ou genes detectados pela sonda Affymetrix

número: 202888\_s\_at, e MUC 5.

29. Sistema de detecção, caracterizado pelo fato de que tem componentes para executar o método como definido em qualquer uma das reivindicações de 1 a 28.

25

30. Meio de armazenagem legível por computador, caracterizado pelo fato de que tem armazenado nele instruções de programa para executar o método de acordo com qualquer uma das reivindicações de 1 a 28.

31. Sistema de detecção, caracterizado pelo fato de que inclui:

meios de avaliar os dados de treinamento, que inclui expressão os dados de treinamento que representam a expressão de genes nas células ou populações celulares derivada de pelo menos um intestino grosso, e origem proximal-distal os dados de treinamento que representam associações das ditas células ou célula populações com os ditos origem proximal-distals;

meios para processar os dados de treinamento para gerar dados de classificação que representam um combinação linear ou não linear de níveis de expressão do dito genes, os ditos dados de classificação para gerar dados de origem proximal-distal indicativos de uma origem proximal-distal de um outra célula ou população celular tomada de um intestino grosso, com base em outros dados de expressão que representam a expressão do dito genes na dita outra célula ou população celular.

32. Método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, o dito método caracterizado pelo fato de que compreende medir a expressão de um ou mais genes selecionados de:

(i) PITX2 ou o gene ou genes detectados pela sonda Affymetrix número 207558\_s\_at,

ETNK1 ou o gene ou genes detectados pela sonda Affymetrix número 222262\_s\_at ou 224453\_s\_at,

FAM3B,

CYP2C 18 ou o gene ou genes detectados pela sonda Affymetrix número 208126\_s\_at,

GBA3 ou o gene ou genes detectados pela sonda Affymetrix número 219954\_s\_at,

MEP1B, ADRA2A, HSD3B2,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix número 206754\_s\_at, SLC14A2 ou o gene ou genes detectados pela sonda

Affymetrix número 226432\_s\_at, CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix número 231576\_s\_at, DEFA5,

OASL ou o gene ou genes detectados pela sonda Affymetrix número 210797\_s\_at, SLC37A3,

5 REG1A, MEP 1B, NR1H4; ou

(ii) DKFZp761N1114 ou o gene ou genes detectados pela sonda Affymetrix número 242374\_s\_at, PRAC,

INSL5,

HOXB13 ou

10 WFDC2

em uma amostra de os ditos indivíduo em que um de expressão de o (i) relativo ao normal distal intestino grosso é indicativo de uma origem intestino grosso proximal e um nível de expressão mais alto de expressão dos genes do grupo (ii) relativo a níveis de controle do intestino grosso proximal normal como indicativo de um intestino grosso distal.

15

33. Método para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo, os ditos método caracterizado pelo fato de que compreende medir o nível de expressão de um ou mais genes selecionados de:

20

(i) o gene ou genes detectados pela sonda Affymetrix número: 218888\_s\_at

o gene detectado pela sonda Affymetrix número: 225290\_at

o gene detectado pela sonda Affymetrix número: 226432\_at

o gene detectado pela sonda Affymetrix número: 231576\_at

25

o gene detectado pela sonda Affymetrix número: 235733\_at

o gene detectado pela sonda Affymetrix número: 236894\_at

o gene detectado pela sonda Affymetrix número: 239656\_at

o gene detectado pela sonda Affymetrix número: 242059\_at

o gene detectado pela sonda Affymetrix número: 242683\_at

- ABHD5, FAM3B, IGFBP2, POPDC3,  
 ADRA2A, FLJ10884, KCNG1, REG1A,  
 APOBEC1, FLJ22761, KIFAP3, SLC14A2,  
 ClOorf45, FTHFD, LOC375295, SLC20A1,  
 5 ClOorf58, GCNT1, ME3, SLC23A3,  
 CCL8, HAS3, MEP1B, SLC38A2,  
 CLDN15, HOXB6, NPY6R, SLC9A3,  
 DEFA5, HOXD4, NR1H3, TBCC,  
 EYA2, HSD3B2, HR1H4, ZNF493,  
 10 OSTalpha,  
 PAP,
- AFARP1 ou o gene ou genes detectados pela sonda Affymetrix número:  
 202234\_s\_at,
- ANPEP ou o gene ou genes detectados pela sonda Affymetrix número  
 15 202888\_s\_at,
- CCL13 ou o gene ou genes detectados pela sonda Affymetrix número:  
 206407\_s\_at
- CRYL1 ou o gene ou genes detectados pela sonda Affymetrix número:  
 220753\_s\_at,
- 20 CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix número:  
 206754\_s\_at,
- CYP2C18, ou o gene ou genes detectados pela sonda Affymetrix número:  
 CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix número:  
 214421\_x\_at ou 220017\_x\_at,
- 25 EPB41L3 ou o gene ou genes detectados pela sonda Affymetrix número:  
 211776\_s\_at
- ETNK1 ou o gene ou genes detectados pela sonda Affymetrix número:  
 222262\_s\_at ou 224453\_s\_at
- FAM45A ou o gene ou genes detectados pela sonda Affymetrix número:

- 221804\_s\_at ou 222955\_s\_at  
 FGFR2 ou o gene ou genes detectados pela sonda Affymetrix número:  
 203639\_s\_at,  
 HOXC6 ou o gene ou genes detectados pela sonda Affymetrix número:  
 5 206858\_s\_at,  
 HOXD3 ou o gene ou genes detectados pela sonda Affymetrix número:  
 206601\_s\_at,  
 ME2 ou o gene ou genes detectados pela sonda Affymetrix número:  
 210153\_s\_at,  
 10 MESP1 ou o gene ou genes detectados pela sonda Affymetrix número:  
 224476\_s\_at,  
 MOCS 1 ou o gene ou genes detectados pela sonda Affymetrix número:  
 213181\_s\_at,  
 MSCP ou o gene ou genes detectados pela sonda Affymetrix número:  
 15 (ii) o gene detectado pela sonda Affymetrix número: 230105\_at  
 o gene detectado pela sonda Affymetrix número: 230269\_at  
 o gene detectado pela sonda Affymetrix número: 238378\_at  
 o gene detectado pela sonda Affymetrix número: 239814\_at  
 o gene detectado pela sonda Affymetrix número: 239994\_at  
 20 o gene detectado pela sonda Affymetrix número: 240856\_at  
 o gene detectado pela sonda Affymetrix número: 242414\_at  
 o gene detectado pela sonda Affymetrix número: 244553\_at  
 ACACA, FMOD, LOC151162, SLOOP,  
 Cl3orf11, FRMD3, MCF2L, SCGB2A1,  
 25 C20orf56, GALNT5, MMP28, SCNN1B,  
 CAPN13, GARNL4, MUC11, SHANK2,  
 CLDN8, GCG, MUC12, SIAT2,  
 COLM, GNE, MUC17, SIAT4C,  
 CRIP1, HGD, MUC5B, SIAT7F,

|    |           |           |          |          |
|----|-----------|-----------|----------|----------|
|    | DNAJC12,  | HOXB13,   | NEDD4L,  | SIDT1,   |
|    | FAM3C,    | INSL5,    | PARP8,   | SLC13A2, |
|    | FB X025,  | IRS1,     | PCDH21,  | SLPI,    |
|    | FLJ20366, | ISL1,     | PI3,     | SPINK5,  |
| 5  | F1120989, | KIAA0703, | PRAC,    | SST      |
|    |           | KIAA0830, | PRAC2    | THF1,    |
|    |           | KIAA1913, | PTTG1IP, | TNFSF11, |
|    |           | LAMA1,    | QPRT,    | TPH1,    |
|    |           | LGALS2,   | QSCN6,   | WFDC2,   |
| 10 |           |           | RBM24,   |          |

ARF4 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 201097\_s\_at, BTG3 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 213134\_x\_at ou 205548\_s\_at,

CHST5 ou o gene ou genes detectados pela sonda Affymetrix  
 15 número: 221164\_x\_at ou 223942\_x\_at

CMAH ou o gene ou genes detectados pela sonda Affymetrix  
 número: 205518\_s\_at, CRYBA2 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 220136\_s\_at CTSE ou o gene ou genes detectados pela  
 sonda Affymetrix número: 205927\_s\_at, DKFZp761N1114 ou o gene ou  
 20 genes detectados pela sonda Affymetrix número:

EPB41L4A ou o gene ou genes detectados pela sonda  
 Affymetrix número: 228256\_s\_at,

EPHA3 ou o gene ou genes detectados pela sonda Affymetrix  
 número: 206070\_s\_at, FAS ou o gene ou genes detectados pela sonda  
 25 Affymetrix número: 204781\_s\_at, FER1L3 ou o gene ou genes detectados  
 pela sonda Affymetrix número: 201798\_s\_at

ou 211864\_s\_at,

F1120152 ou o gene ou genes detectados pela sonda  
 Affymetrix número: 218532\_s\_at ou 218510\_x\_at,

FLJ23548 ou o gene ou genes detectados pela sonda Affymetrix número: 218187\_s\_at,

FN1 ou o gene ou genes detectados pela sonda Affymetrix número: 211719\_s\_at ou 210495 x at ou 212464 at ou 216442 x at

5 FOXA2 ou o gene ou genes detectados pela sonda Affymetrix número: 210103\_s\_at, FRZB ou o gene ou genes detectados pela sonda Affymetrix número: 203698\_s\_at, GDF15 ou o gene ou genes detectados pela sonda Affymetrix número: 221577\_x\_at, GJB3 ou o gene ou genes detectados pela sonda Affymetrix número: 205490\_s\_at, HOXD13 ou o gene ou genes  
10 detectados pela sonda Affymetrix número: 207397\_s\_at, INSM1 ou o gene ou genes detectados pela sonda Affymetrix número: 206502\_s\_at,

MGC4170 ou o gene ou genes detectados pela sonda Affymetrix número: 212959\_s\_at,

15 MLPH ou o gene ou genes detectados pela sonda Affymetrix número: 218211\_s\_at, NEBL ou o gene ou genes detectados pela sonda Affymetrix número: 203962\_s\_at,

PLA2G2A ou o gene ou genes detectados pela sonda Affymetrix número:

203649\_s\_at,

20 PTPRO ou o gene ou genes detectados pela sonda Affymetrix número: 208121\_s\_at, PYY ou o gene ou genes detectados pela sonda Affymetrix número: 207080\_s\_at ou 211253\_x\_at,

25 SH3BP4 ou o gene ou genes detectados pela sonda Affymetrix número: 222258\_s\_at, SLC28A2 ou o gene ou genes detectados pela sonda Affymetrix número: 207249\_s\_at,

SLC2A10 ou o gene ou genes detectados pela sonda Affymetrix número: 221024\_s\_at,

SPON1 ou o gene ou genes detectados pela sonda Affymetrix número: 213994\_s\_at ou 209437\_s\_at,

STS ou o gene ou genes detectados pela sonda Affymetrix número: 203769\_s\_at TM4SF11 ou o gene ou genes detectados pela sonda Affymetrix número: 204519\_s\_at,

5 TUSC3 ou o gene ou genes detectados pela sonda Affymetrix número: 213432\_s\_at ou 209228\_x\_at,

em uma amostra biológica de os ditos indivíduo em que um de expressão de normal distal intestino grosso é indicativos de um proximal intestino grosso e um expressão de o genes (ii) relativa ao intestino grosso proximal normal é indicativo de uma origem de intestino grosso distal.

10 34. Método de acordo com a reivindicação 32 ou 33 caracterizado pelo fato de que a dita região proximal compreende o ceco e o cólon ascendente.

15 35. Método de acordo com a reivindicação 33 caracterizado pelo fato de que os ditos região distal compreende, cólon descendente, flexão sigmóide e reto.

36. Método de acordo com a reivindicação 32 ou 33, 34 ou 35 caracterizado pelo fato de que o dito gene é ETNK1.

37. Método de acordo com a reivindicação 32 ou 33 ou 34 ou 35 caracterizado pelo fato de que o dito gene é GBA3.

20 38. Método de acordo com a reivindicação 32 ou 33 ou 34 ou 35, caracterizado pelo fato de que o dito gene é PRAC.

39. Método de acordo com qualquer uma das reivindicações de 32 a 38 caracterizado pelo fato de que a dita amostra biológica é uma amostra fecal, lavagem de enema, ressecção cirúrgica ou biópsia de tecido.

25 40. Arranjo de ácido nucleico, caracterizado pelo fato de que compreende uma pluralidade de:

(i) moléculas de ácido nucleico que compreende uma sequência de nucleotídeo que corresponde a qualquer uma das locação de genes marcadores listados na reivindicação 33 ou uma sequência exibindo

pelo menos 80% de identidade ou um derivado funcional, fragmento, variante ou homólogo das ditas moléculas de ácido nucleico; ou

(ii) moléculas de ácido nucleico que compreende um sequência de nu capaz de hibridizar a qualquer sequência de (i) sob condições de severidade baixa a 42°C, fragmento, variante ou homólogo as ditas moléculas de ácido nucleico

(iii) sondas de ácido nucleico ou oligonucleotídeos que compreendem uma sequência de nucleotídeo capaz de hibridizar a 42°C ou um derivado funcional, fragmento, variante ou homólogo das ditas moléculas de ácido nucleico

em que o nível de expressão dos ditos ácido nucleico é indicativos da origem proximal-distal de uma célula ou subpopulação celular derivada do intestino grosso.

41. Arranjo de acordo com a reivindicação 40, caracterizado pelo fato de que os ditos marcadores de localização são os marcadores listados nas reivindicações 32 ou 33.

42. Arranjo de ácido nucleico, caracterizado pelo fato de que compreende uma pluralidade de:

(i) moléculas de ácido nucleico que compreende uma sequência de nucleotídeo que corresponde a qualquer uma das locação de genes marcadores listados na reivindicação 33 ou uma sequência exibindo pelo menos 80% de identidade ou um derivado funcional, fragmento, variante ou homólogo das ditas moléculas de ácido nucleico; ou

(ii) moléculas de ácido nucleico que compreende um sequência de nu capaz de hibridizar a qualquer sequência de (i) sob condições de severidade baixa a 42°C, fragmento, variante ou homólogo as ditas moléculas de ácido nucleico

(iii) sondas de ácido nucleico ou oligonucleotídeos que compreendem uma sequência de nucleotídeo capaz de hibridizar a 42°C ou

um derivado funcional, fragmento, variante ou homólogo das ditas moléculas de ácido nucleico

em que o nível de expressão dos ditos ácido nucleico é indicativos da origem proximal-distal de uma célula ou subpopulação celular derivada do intestino grosso.

43. Arranjo de acordo com qualquer uma das reivindicações de 40 a 42 caracterizado pelo fato de que o dito arranjo é usado no método como definido em qualquer uma das reivindicações de 32 a 39.

44. Uso de um arranjo como definido em qualquer uma das reivindicações de 40 a 42, caracterizado pelo fato de ser para determinar a origem anatômica de uma célula ou população celular derivada do intestino grosso de um indivíduo.

45. Método de acordo com a reivindicação 23, caracterizado pelo fato de que os ditos genes são selecionados de:

PITX2 ou o gene ou genes detectados pela sonda Affymetrix número 207558\_s\_at, ETNK1 ou o gene ou genes detectados pela sonda Affymetrix número 222262\_s\_at ou 224453\_s\_at,

FAM3B,

CYP2C 18 ou o gene ou genes detectados pela sonda Affymetrix número 208126\_s\_at, GBA3 ou o gene ou genes detectados pela sonda Affymetrix número 219954\_s\_at,

MEP1B, ADRA2A, HSD3B2,

CYP2B6 ou o gene ou genes detectados pela sonda Affymetrix número 206754\_s\_at,

SLC 14A2 ou o gene ou genes detectados pela sonda Affymetrix número 226432\_s\_at, CYP2C9 ou o gene ou genes detectados pela sonda Affymetrix número 231576\_s\_at, DEFA5,

OASL ou o gene ou genes detectados pela sonda Affymetrix número 210797\_s\_at, SLC37A3,

REG1A, MEP1B, NR1H4; ou  
 DKFZp761N1114 ou o gene ou genes detectados pela sonda  
 Affymetrix número 242374\_s\_at,

5 PRAC,  
 INSL5,  
 HOXB13 ou  
 WFDC2

46. Método de acordo com qualquer uma das reivindicações de  
 1 a 39 ou 45, caracterizado pelo fato de que os ditos níveis de expressão é a  
 10 expressão de proteína.

47. Método de acordo com qualquer uma das reivindicações de  
 1 a 39 ou 45, caracterizado pelo fato de que os ditos níveis de expressão é  
 expressão de mRNA.

48. Método para determinar o início ou predisposição para o  
 15 início de uma anormalidade celular ou uma condição distinguida por uma  
 anormalidade celular no intestino grosso, o dito método caracterizado pelo  
 fato de que compreende determinar, com o método de acordo com qualquer  
 uma das reivindicações de 1 a 39 ou 45 a 47, o perfil de expressão de gene  
 proximal-distal de uma amostra biológica derivada de uma origem proximal  
 20 ou distal no intestino grosso em que a detecção proximal-distal intestino  
 grosso é indicativo da célula ou população celular expressando os ditos perfis.

49. Kit de diagnóstico para ensaiar amostras biológicas,  
caracterizado pelo fato de que compreende um agente para detectar um ou  
 mais marcadores proximal-distal e reagentes úteis para facilitar a detecção  
 25 pelos ditos agentes.

50. Kit de acordo com a reivindicação 49, caracterizado pelo  
 fato de que é usado o método de acordo com qualquer uma das reivindicações  
 de 1 a 39 ou 45 a 47.

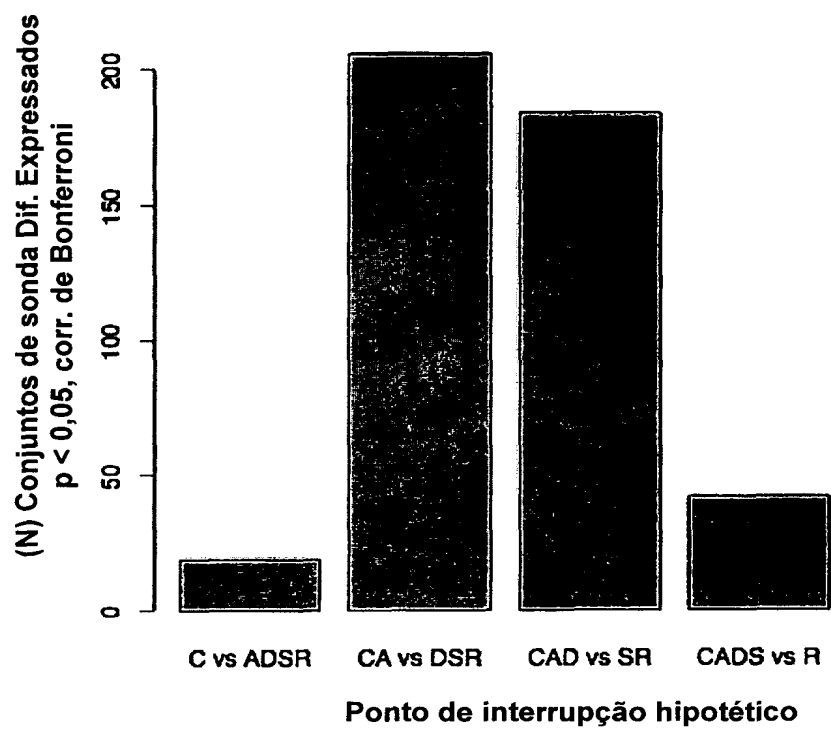
**FIGURA 1**

FIGURA 2

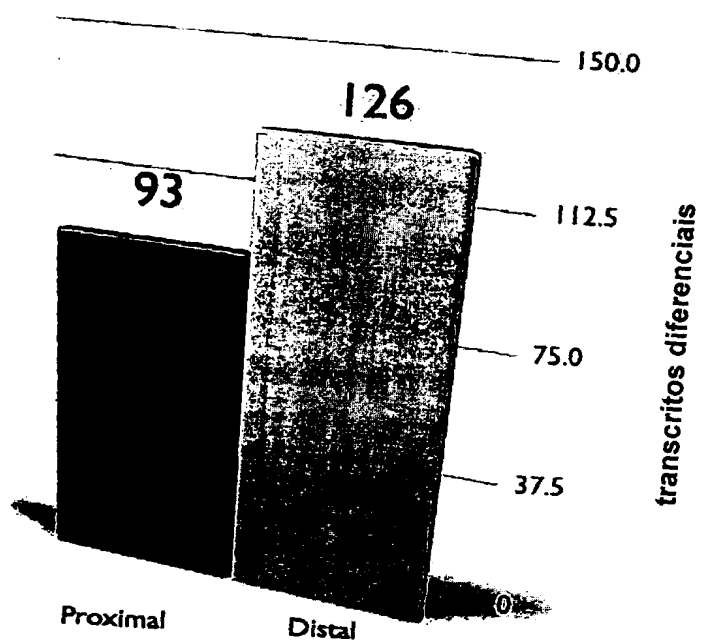


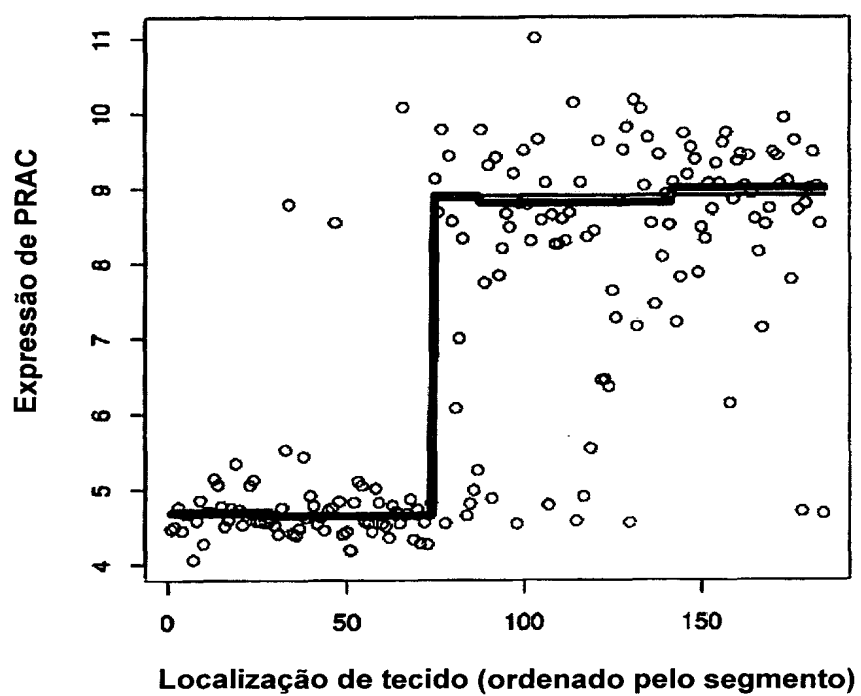
**FIGURA 3****Modelo de Comparação de PRAC de segmentos 2v5**

FIGURA 4

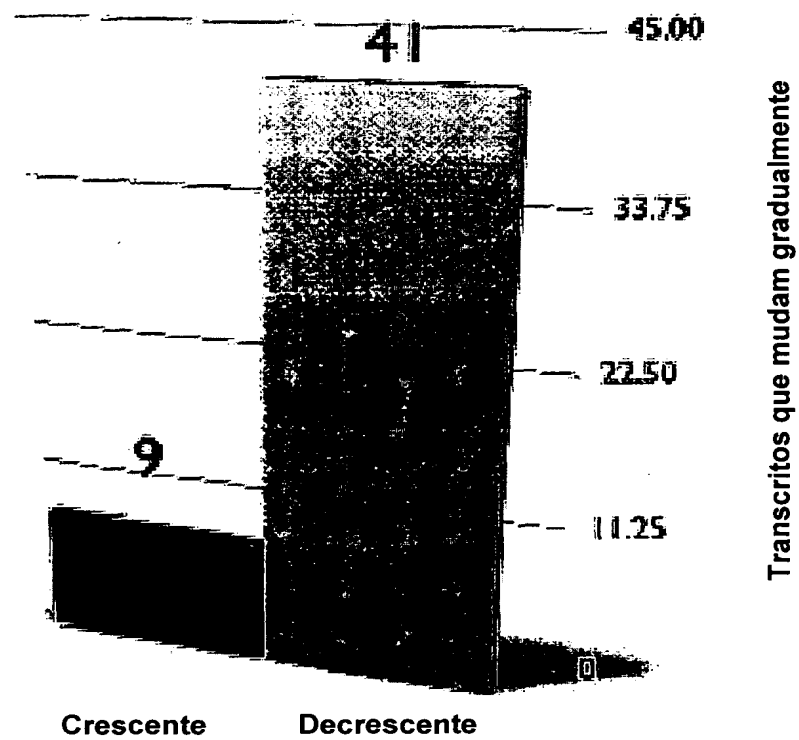


FIGURA 5

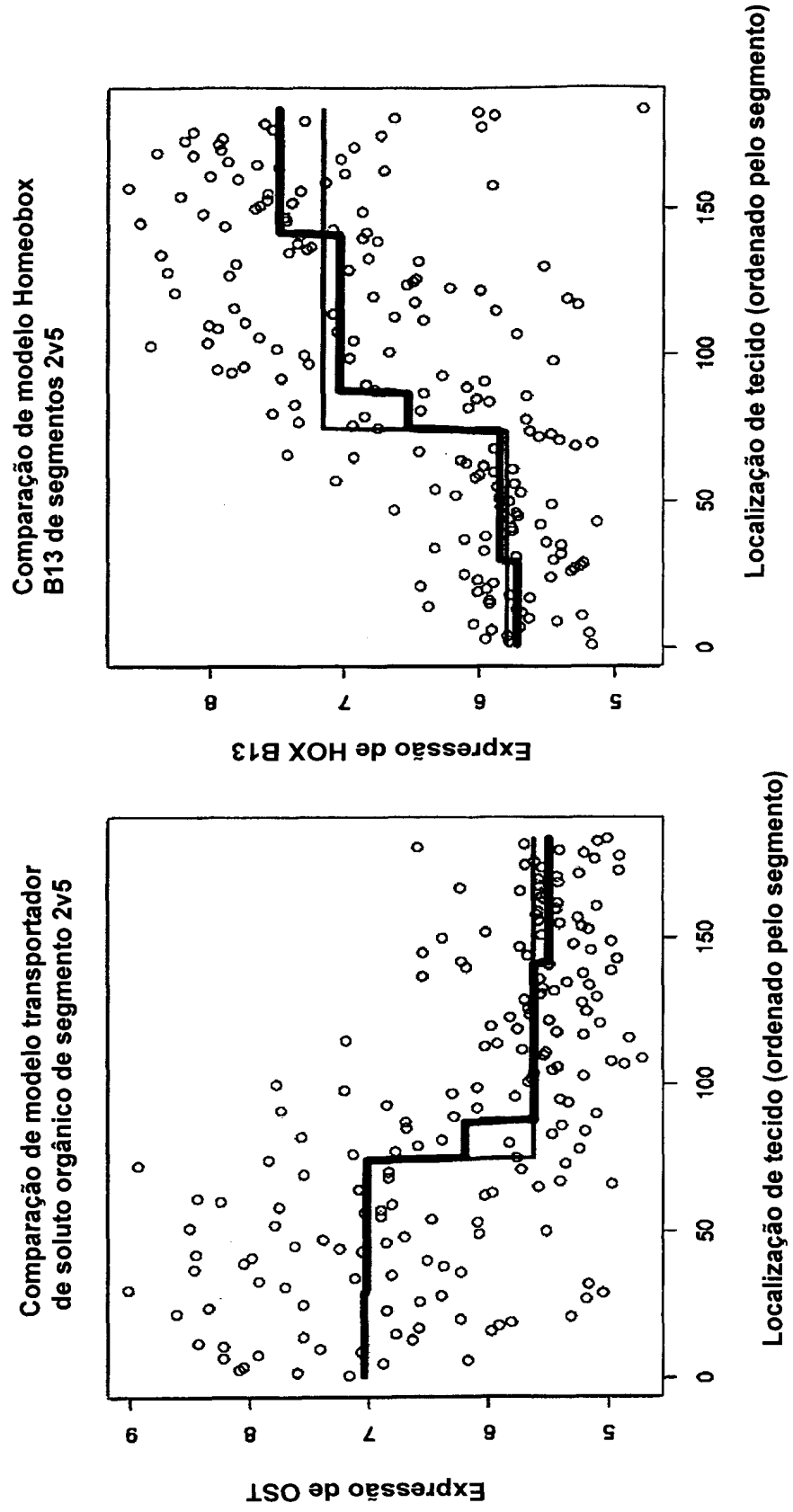


FIGURA 6A

Datos de todos os componentes principais

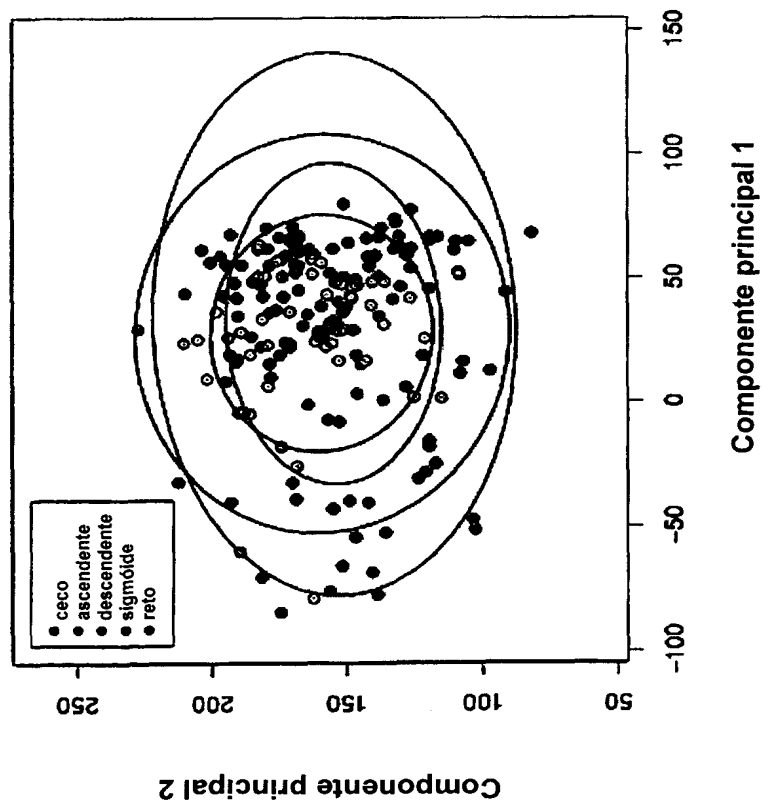


FIGURA 6B

Componentes principais supervisionados

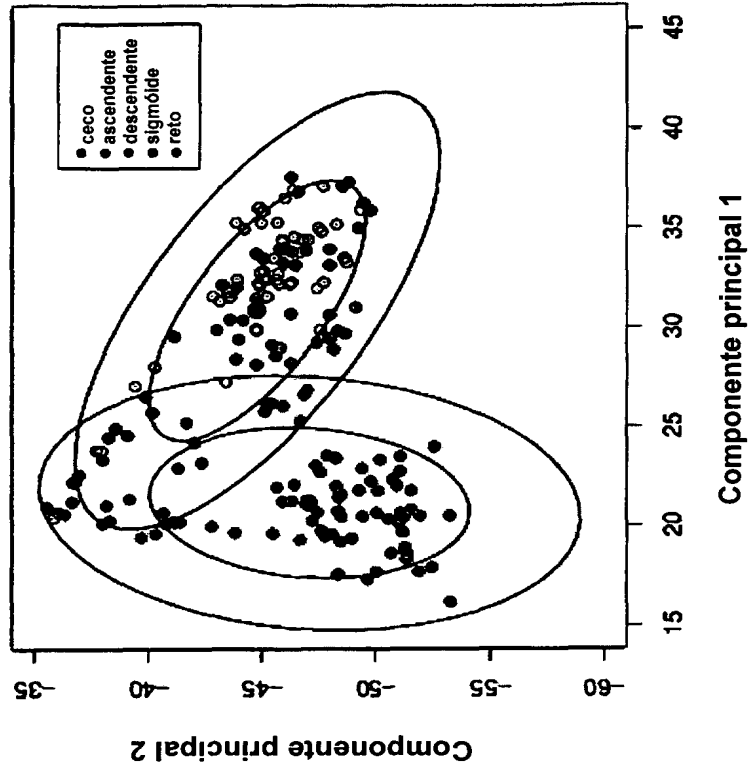


FIGURA 7B

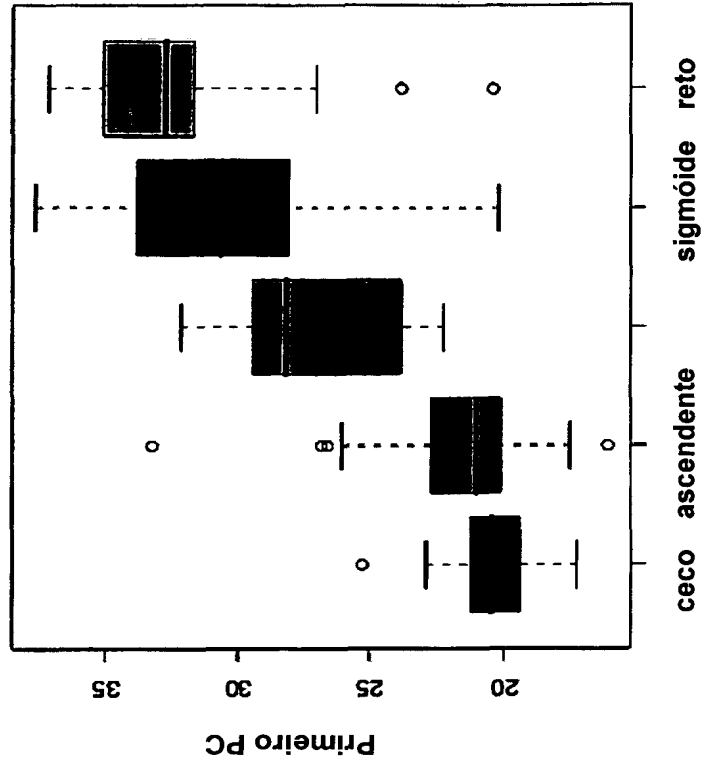


FIGURA 7A

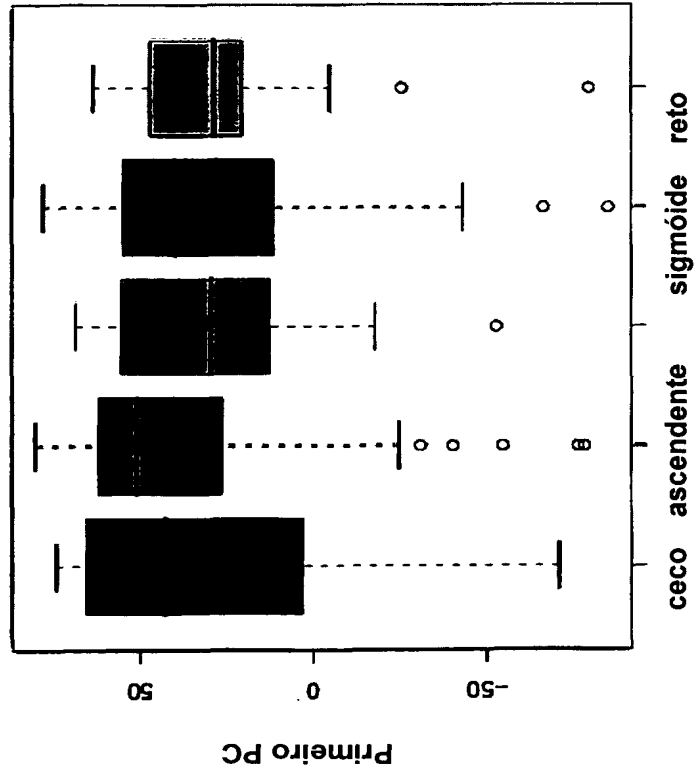


FIGURA 8B

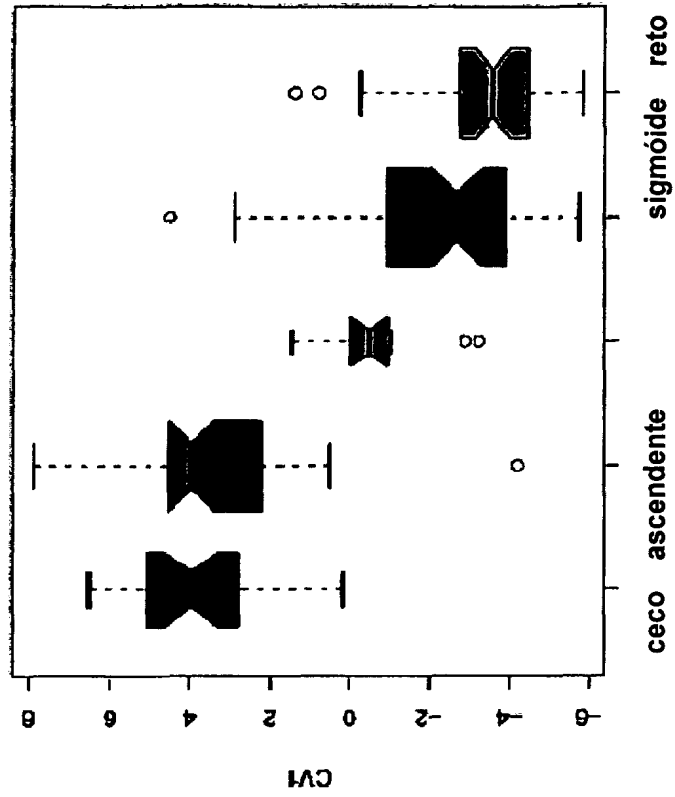


FIGURA 8A

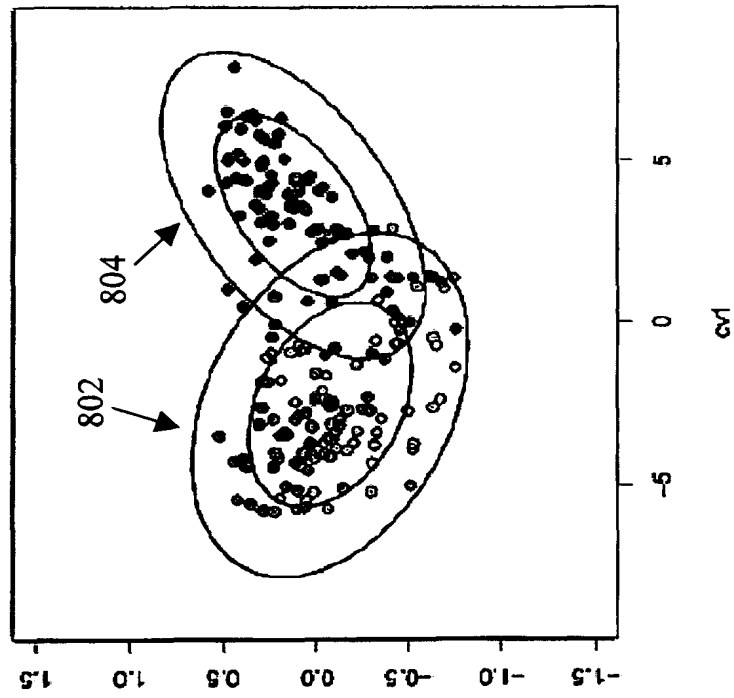


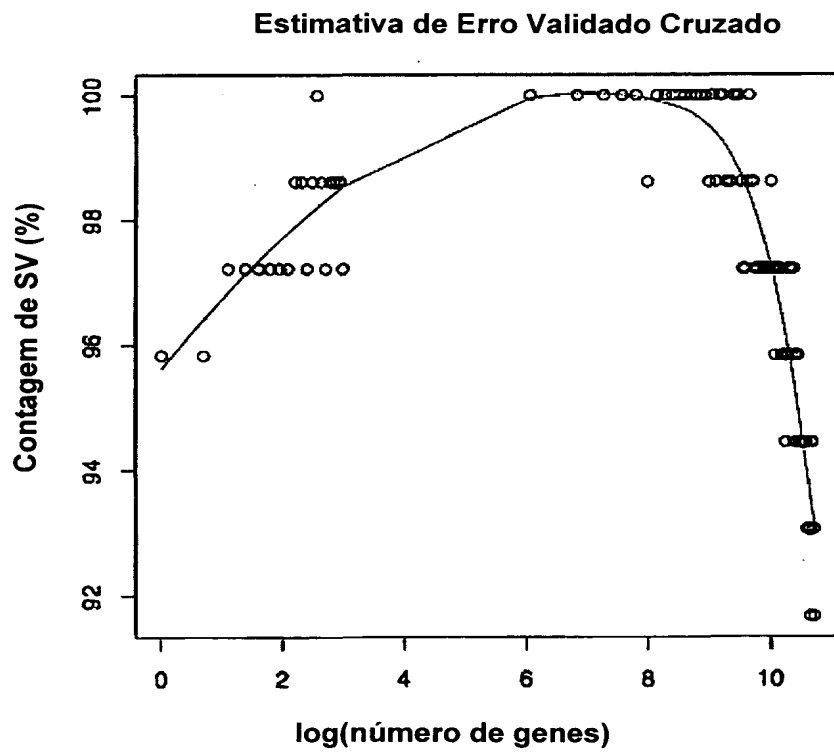
**FIGURA 9**

FIGURA 10

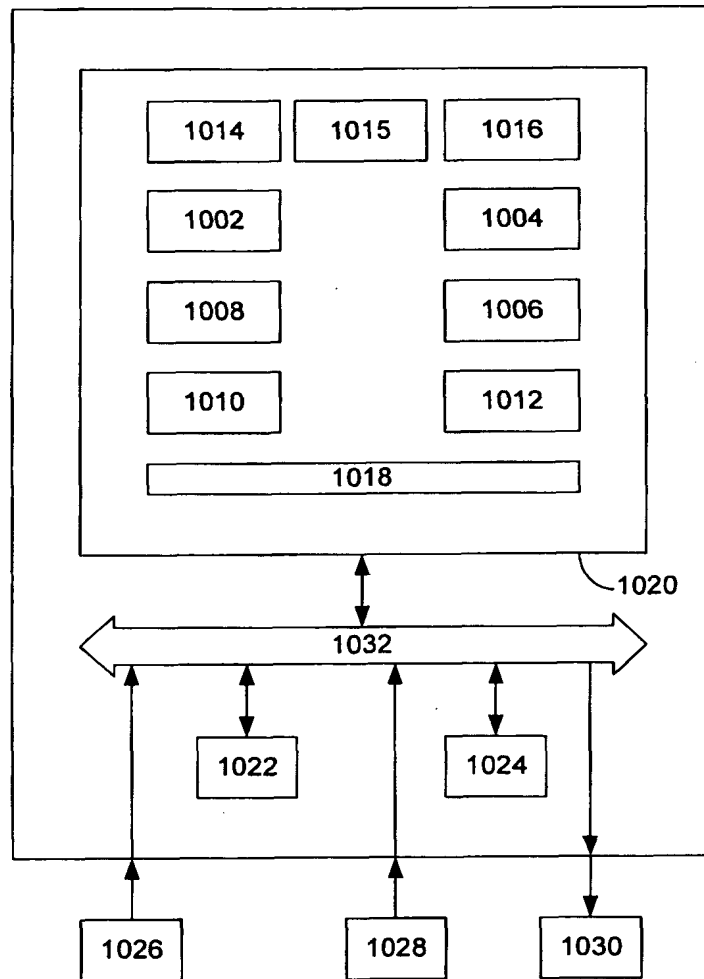
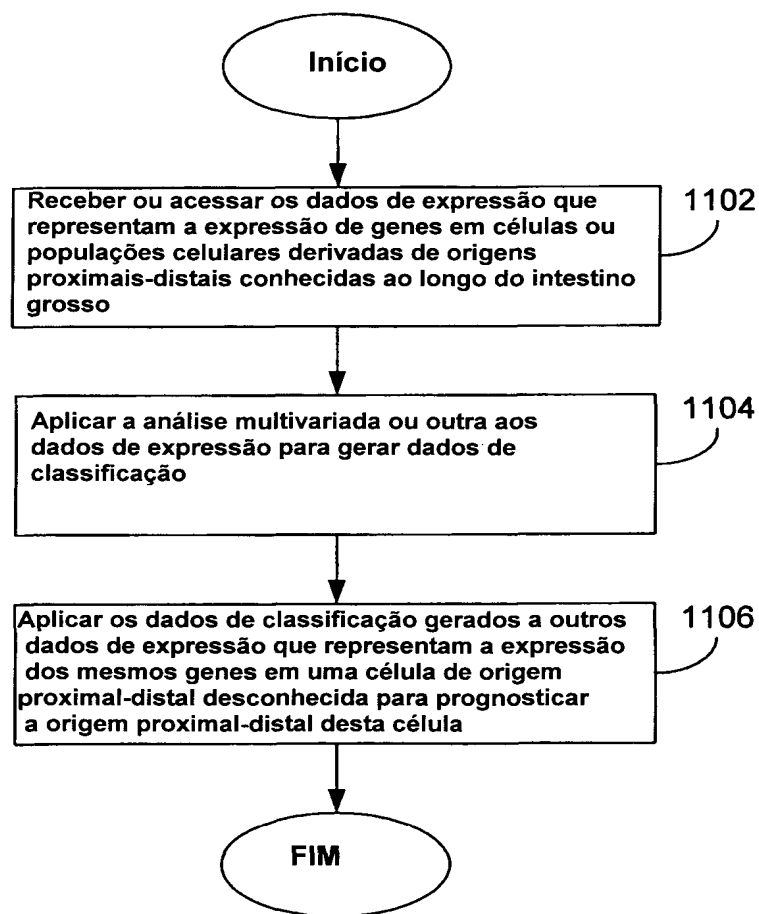


**FIGURA 11**

## RESUMO

“MÉTODO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, MÉTODO DE DETECÇÃO PARA DETERMINAR A ORIGEM ANATÔMICA DE UMA CÉLULA OU POPULAÇÃO CELULAR DERIVADA DO INTESTINO GROSSO DE UM INDIVÍDUO, SISTEMA DE DETECÇÃO, MEIO DE ARMAZENAGEM LEGÍVEL POR COMPUTADOR, ARRANJO DE ÁCIDO NUCLEICO, USO DE UM ARRANJO, MÉTODO PARA DETERMINAR O INÍCIO OU PREDISPOSIÇÃO PARA O INÍCIO DE UMA ANORMALIDADE CELULAR OU UMA CONDIÇÃO DISTINGUIDA POR UMA ANORMALIDADE CELULAR NO INTESTINO GROSSO, KIT DE DIAGNÓSTICO PARA ENSAIAR AMOSTRAS BIOLÓGICAS”

A presente invenção diz respeito no geral a uma série de moléculas de ácido nucléico, os perfis de expressão dos quais caracterizam a origem anatômica da célula ou população de células dentro do intestino grosso. Mais particularmente, a presente invenção diz respeito a uma série de moléculas de ácido nucléico, os perfis de expressão das quais caracterizam a origem proximal ou distal da célula ou população de células dentro do intestino grosso. Os perfis de expressão da presente invenção são úteis na faixa de aplicações incluindo, mas não limitado a determinar a origem anatômica da célula ou população de células que foram derivadas do intestino grosso. Além disso, visto que a progressão da célula normal é frequentemente caracterizada pela desdiferenciação, o método da presente invenção também fornece um meio de identificar a anormalidade celular com base na expressão de um perfil de expressão incorreto em relação àquele que deve ser expressado pelas células objeto quando considerada sua localização anatômica dentro do cólon. Consequentemente, este aspecto da invenção fornece um meio valioso de identificar a existência de células colônicas do

intestino grosso, estas sendo indicativas de uma anormalidade dentro do intestino grosso tal como o início ou pré disposição ao início da condição tal como um neoplasma colorretal.