República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(21) BR 112020012280-7 A2

(22) Data do Depósito: 19/12/2018

(43) Data da Publicação Nacional: 24/11/2020

(54) Título: COMPOSIÇÕES E MÉTODOS PARA DIAGNOSTICAR CÂNCERES DE PULMÃO USANDO PERFIS DE EXPRESSÃO DE GENE

(51) Int. Cl.: C12Q 1/68; G01N 33/574; C12N 15/12.

(30) Prioridade Unionista: 19/12/2017 US 62/607,756; 29/10/2018 US 62/752,163.

(71) Depositante(es): THE WISTAR INSTITUTE OF ANATOMY AND BIOLOGY.

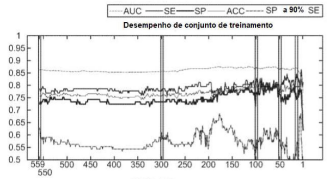
(72) Inventor(es): MICHAEL SHOWE; LOUISE C. SHOWE; ANDREW V. KOSSENKOV.

(86) Pedido PCT: PCT US2018066531 de 19/12/2018

(87) Publicação PCT: WO 2019/126343 de 27/06/2019

(85) Data da Fase Nacional: 17/06/2020

(57) Resumo: A presente invenção refere-se a métodos e composições para diagnosticar câncer de pulmão em um sujeito mamífero pelo uso de 7 ou mais genes selecionados, por exemplo, um perfil de expressão de gene do sangue do sujeito que é característico da doença. O perfil de expressão de gene inclui 7 ou mais genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX aqui.



"COMPOSIÇÕES E MÉTODOS PARA DIAGNOSTICAR CÂNCERES DE PULMÃO USANDO PERFIS DE EXPRESSÃO DE GENE".

DECLARAÇÃO RELACIONADA À PESQUISA OU DESENVOLVIMENTO COM PATROCÍNIO DO GOVERNO FEDERAL

[001] Esta invenção foi feita com o suporte do governo sob Subsídio No. CA010815 concedido pelos Institutos Nacionais de Saúde e pelo Subsídio No. 4100059200 (Marcadores de diagnóstico para câncer de pulmão em estágio inicial em amostras de sangue de gene PAX) concedido pelo Departamento de Saúde da PA. O governo tem determinados direitos sobre a invenção.

FUNDAMENTOS DA INVENÇÃO

[002] O câncer de pulmão é a causa mundial mais comum de mortalidade por câncer. Nos Estados Unidos, o câncer de pulmão é o segundo câncer mais prevalente em homens e mulheres e será responsável por mais de 174.000 novos casos por ano e mais de 162.000 mortes por câncer. De fato, o câncer de pulmão é responsável por mais mortes a cada ano do que os cânceres de mama, próstata e colorretal combinados.

[003] A alta mortalidade (80-85% em cinco anos), que mostrou pouca ou nenhuma melhora nos últimos 30 anos, enfatiza o fato de que são necessárias ferramentas novas e eficazes para facilitar o diagnóstico precoce antes da metástase nos nós regionais ou além do pulmão.

[004] Populações de alto risco incluem fumantes, ex-fumantes e indivíduos com marcadores associados a predisposições genéticas. Como a remoção cirúrgica de tumores em estágio inicial continua sendo o tratamento mais eficaz para o câncer de pulmão, houve um grande interesse na triagem de pacientes de alto risco com CT em espiral de baixa dose (LDCT). Essa estratégia identifica nódulos pulmonares não

calcificados em aproximadamente 30 a 70% dos indivíduos de alto risco, mas apenas uma pequena proporção dos nódulos detectados é diagnosticada como câncer de pulmão (0,4 a 2,7%). É provável que esse grande número de nódulos detectados resulte em tratamento excessivo e sobrecarregue os sistemas de saúde. Atualmente, a única maneira de diferenciar indivíduos com nódulos pulmonares de etiologia benigna de indivíduos com nódulos malignos é uma biópsia invasiva, cirurgia ou observação prolongada com varredura repetida. Mesmo usando os melhores algoritmos clínicos, 20-55% dos pacientes selecionados para realizar biópsia pulmonar cirúrgica por nódulos pulmonares indeterminados apresentam doença benigna e aqueles que não são submetidos à biópsia imediata ou ressecção requerem estudos sequenciais de imageamento. O uso da CT em série nesse grupo de pacientes corre o risco de atrasar a terapia curável em potencial, juntamente com os custos de varreduras repetidas, as doses de radiação não insignificantes e a ansiedade do paciente.

[005] Idealmente, um teste de diagnóstico seria facilmente acessível, barato, demonstraria alta sensibilidade e especificidade e resultaria em melhores resultados para os pacientes (médica e financeiramente). Outros mostraram que classificadores que utilizam células epiteliais têm alta precisão. No entanto, a colheita dessas células requer uma broncoscopia invasiva. Ver, Silvestri et al, N Engl J Med. 2015 July 16; 373(3): 243-251, que é aqui incorporado por referência.

[006] Esforços estão em andamento para desenvolver diagnósticos não invasivos usando escarro, sangue ou soro e analisar produtos de células tumorais, DNA tumoral metilado, RNA ou proteínas expressos no polimorfismo de nucleotídeo único (SNPs). Essa ampla faixa de testes moleculares com potencial utilidade para o diagnóstico precoce de câncer de pulmão foi discutida na literatura. Embora cada uma dessas abordagens tenha seus próprios méritos, nenhuma ainda

passou do estágio exploratório no esforço de detectar pacientes com câncer de pulmão em estágio inicial, mesmo em grupos de alto risco, ou pacientes com diagnóstico preliminar baseado em fatores radiológicos e outros fatores clínicos. Um simples exame de sangue, um evento de rotina associado a visitas regulares ao consultório clínico, seria um teste diagnóstico ideal.

SUMÁRIO DA INVENÇÃO

[007] Em um aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui pelo menos sete (7) ou mais polinucleotídeos ou oligonucleotídeos, em que cada polinucleotídeo ou oligonucleotídeo hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra de paciente. Cada gene, fragmento de gene, transcrito de gene ou produto de expressão é selecionado dos genes da Tabela I, Tabela II, Tabela III, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, pelo menos um polinucleotídeo ou oligonucleotídeo está fixado a um marcador detectável. Em uma modalidade, a composição ou kit inclui polinucleotídeos ou oligonucleotídeos que detectam o gene, fragmento de gene, transcrito de gene ou produto de expressão de cada um dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em um aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui pelo menos 8 ou mais polinucleotídeos ou oligonucleotídeos, em que cada polinucleotídeo ou oligonucleotídeo hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra de paciente. Em uma modalidade, os genes são selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em um aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui pelo menos 15 ou mais polinucleotídeos ou oligonucleotídeos, em que cada polinucleotídeo ou oligonucleotídeo

hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra de paciente. Em uma modalidade, os genes são selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em um aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui pelo menos 41 ou mais polinucleotídeos ou oligonucleotídeos, em que cada polinucleotídeo ou oligonucleotídeo hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra de paciente. Em uma modalidade, os genes são selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em um aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui pelo menos 50 ou mais polinucleotídeos ou oligonucleotídeos, em que cada polinucleotídeo ou oligonucleotídeo hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra de paciente. Em uma modalidade, os genes são selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX.

[008] Em outro aspecto, uma composição ou kit para diagnosticar ou avaliar um câncer de pulmão em um mamífero inclui 7 ou mais ligantes, em que cada ligante hibridiza com um produto de expressão de gene diferente em uma amostra de paciente. Cada produto de expressão é selecionado dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, pelo menos um ligante está fixado a um marcador detectável. Em uma modalidade, a composição ou kit inclui ligantes que detectam os produtos de expressão de cada um dos genes na Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma outra modalidade, a composição ou kit inclui ligantes que detectam os produtos de expressão de pelo menos 8 genes. Em uma outra modalidade, a composição ou kit inclui ligantes que detectam os produtos de expressão de pelo menos 15 genes. Em

uma outra modalidade, a composição ou kit inclui ligantes que detectam os produtos de expressão de pelo menos 41 genes. Em uma outra modalidade, a composição ou kit inclui ligantes que detectam os produtos de expressão de pelo menos 50 genes. Em uma modalidade, os genes são selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX.

[009] As composições descritas neste documento permitem a detecção de alterações na expressão nos genes no perfil de expressão de gene do sujeito daquele de um perfil de expressão de gene de referência. Os vários perfis de expressão de gene de referência são descritos abaixo. Em uma modalidade, a composição fornece a capacidade de distinguir um tumor cancerígeno de um nódulo não cancerígeno.

[0010] Em outro aspecto, um método para diagnosticar ou avaliar um câncer de pulmão em um mamífero envolve identificar alterações na expressão de três ou mais genes na amostra de um sujeito, os referidos genes selecionados dos genes da Tabela I, Tabela II, Tabela III , Tabela IV ou Tabela IX, e comparar os níveis de expressão de gene desse sujeito com os níveis dos mesmos genes em uma referência ou controle, em que as alterações na expressão da referida expressão de gene se correlacionam com um diagnóstico ou avaliação de um câncer de pulmão. Em uma modalidade, as alterações na expressão da referida expressão de gene fornecem a capacidade de distinguir um tumor cancerígeno de um nódulo não cancerígeno.

[0011] Em outro aspecto, um método para diagnosticar ou avaliar um câncer de pulmão em um mamífero envolve identificar um perfil de expressão de gene no sangue de um sujeito, o perfil de expressão de gene compreendendo 7 ou mais produtos de expressão de gene de 8 ou mais genes informativos, conforme descrito aqui. Os 7 ou mais genes informativos são selecionados dos genes da Tabela I, Tabela II, Tabela

III, Tabela IV ou Tabela IX. Em uma modalidade, 8 ou mais genes informativos são selecionados dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, o perfil de expressão de gene contém 15 genes selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, o perfil de expressão de gene contém 41 genes selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX. O perfil de expressão de gene do sujeito é comparado com um perfil de expressão de gene de referência a partir de uma variedade de fontes descritas abaixo. Alterações na expressão dos genes informativos se correlacionam com o diagnóstico ou avaliação de um câncer de pulmão. Em uma modalidade, as alterações na expressão da referida expressão de gene fornecem a capacidade de distinguir um tumor cancerígeno de um nódulo não cancerígeno.

[0012] Em outro aspecto, é fornecido um método para detectar câncer de pulmão em um paciente. O método inclui obter uma amostra do paciente; e detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 8 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão.

[0013] Em ainda outro aspecto, é fornecido um método para diagnosticar câncer de pulmão em um sujeito. O método inclui obter uma amostra de sangue de um sujeito; detectar uma mudança na expressão em pelo menos 8 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma

composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão; e diagnosticar o sujeito com câncer quando são detectadas alterações na expressão dos genes do sujeito em relação aos da referência.

[0014] Em outro aspecto, é fornecido um método de diagnóstico e tratamento de câncer de pulmão em um sujeito com crescimento neoplásico. O método inclui obter uma amostra de sangue de um sujeito; extrair o RNA do sangue e detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão; diagnosticar o sujeito com câncer quando são detectadas alterações na expressão dos genes do sujeito em relação aos da referência; e remover o crescimento neoplásico. Outros tratamentos adequados também podem ser fornecidos.

[0015] Outros aspectos e vantagens destas composições e métodos são descritos adicionalmente na seguinte descrição detalhada das suas modalidades preferidas.

BREVE DESCRIÇÃO DOS projetoS

[0016] As FIG. 1A e FIG. 1B são gráficos representativos que mostram o desempenho da expressão do gene Illumina em diferentes

iterações do processo SVM-RFE. FIG. 1A - Desempenho do conjunto de treinamento. FIG. 1B- Desempenho do conjunto de teste

[0017] As FIG. 2A - A FIG. 2D demonstram a precisão da classificação de nLNC. FIG. 2A- Conjunto de treinamento e validação de lesão de 8 a 20 mm configuram a curva ROC usando todos os genes para classificação de SVM. FIG.

[0018] 2B- Desempenho do conjunto de treinamento e validação usando os 100 principais genes selecionados por RFE. FIG. 2C- Desempenho dos conjuntos de treinamento e validação usando os 50 principais genes. FIG. 2D- Desempenho dos conjuntos de treinamento e validação usando os 15 principais genes.

[0019] A FIG. 3 mostra o projeto do estudo. Um total de 821 amostras únicas foram analisadas neste estudo. Os microarranjos Illumina HT12v4 e o painel NanoString PCI foram usados para selecionar sondas candidatas a biomarcadores usando 283 amostras no total. 264 amostras foram usadas para seleção de biomarcadores em microarranjos e 201 das 264 + 19 novas amostras foram usadas para selecionar os biomarcadores no painel PCI. As 51 amostras utilizadas para validação não foram utilizadas em nenhuma seleção de biomarcadores. 559 dos biomarcadores selecionados no microarranjo e nas análises do painel PCI foram projetados com sucesso para o painel personalizado NanoString. O painel personalizado foi testado com 237 das amostras usadas na seleção da sonda, para garantir que a nova plataforma reproduzisse com sucesso os resultados do microarranjo e com 346 amostras independentes adicionais não previamente testadas em nenhuma plataforma (total 583). As 583 amostras de treinamento foram usadas para criar um classificador de nódulo pulmonar NanoString (nPNC). 158 amostras adicionais que nunca foram envolvidas na seleção da sonda NanoString foram usadas para a plataforma personalizada Nanostring (346 para treinamento e 141 para

validação) para um total de 821 amostras independentes.

[0020] As FIG. 4A - A FIG. 4E mostram o desempenho da classificação do classificador de nódulos pulmonares NanoString. FIG. 4A -FIG. 4D- Comparação de ROC-AUC em conjuntos de treinamento e validação com redução progressiva do número de sondas. FIG. 4E- A probabilidade calculada de malignidade para um nódulo individual para diferentes pontuações de classificação usando o nPNC de 41 sondas.

[0021] As FIG. 5A e FIG. 5B mostras o desempenho de nPNC de 41 sondas para BN e MN na faixa de 6 a 20 mm. FIG. 5A- Comparada com os modelos clínicos de risco de câncer de pulmão da The Brock University, Mayo Clinic e Veteran's Affairs (VA). FIG. 5B- Comparada com a classificação por diâmetro máximo do nódulo

[0022] A FIG. 6 mostra o desempenho da classificação em diferentes faixas de tamanho de nódulo.

[0023] Os valores de desempenho (ROC-AUC à esquerda e especificidade com 90% de sensibilidade à direita) são fornecidos para os conjuntos de dados de Treinamento (painéis superiores), Validação (meio) e Combinado (inferior). Cada linha é rotulada no lado esquerdo da figura com a faixa de tamanho de nódulo mais baixa de Min (qualquer tamanho) a 10 mm. Os rótulos da coluna na parte inferior correspondem ao tamanho do nódulo superior, de 10 mm a Max (qualquer tamanho). Cada quadrado de um painel mostra o desempenho da classificação ao distinguir nódulos benignos de malignos que variam entre o tamanho inferior e o superior em mm, juntamente com o número de nódulos sendo comparados para as classes BN e MN. A intensidade da cor é usada para destaque visual e é proporcional aos valores de desempenho relatados com as escalas de cores mostradas na parte superior dos painéis. Por exemplo, o nPNC demonstrou o melhor desempenho ROC-AUC de 0,87 e uma especificidade de 0,64 a 90%sde sensibilidade na distinção de nódulos de 8-10 mm (o conjunto

continha 6 MN e 14 BN).

[0024] As FIG. 7A - FIG. 7D demonstram desempenho do Classificador de Nódulos Pulmonares. A. Desempenho do conjunto de treinamento de microarranjos da Illumina em diferentes iterações do processo SVM-RFE. Os destaques na caixa selecionaram o número ideal de genes. B. Desempenho do conjunto de teste. C. Comparação do desempenho ROC-AUC do classificador de expressão de gene Illumina com base na expressão de 311 sondas de genes entre conjuntos de treinamento e validação. D. Desempenho de validação cruzada de 10 vezes da expressão de gene do painel NanoString PanCancer Immune no conjunto de treinamento. AUC = área sob a curva ROC, ACC = precisão, SE = sensibilidade, SP = especificidade.

[0025] As FIG. 8A e FIG. 8B mostram uma comparação da Classificação Illumina e NanoString. FIG 8A- O gráfico mostra os escores de classificação de modelo SVM com base na validação cruzada de 10 vezes reamostrada 10 vezes usando genes selecionados por SVM na plataforma Illumina. FIG. 8B- Curvas ROC resultantes da classificação de 199 amostras são executadas no NanoString e no Illumina usando o gene selecionado do Illumina pelo SVM-RFE.

[0026] As FIG. 9A - FIG. 9C demonstram o desempenho do Classificador de Nódulos Pulmonares NanoString no conjunto de treinamento, teste e de todas as amostras usando um número diferente de genes. FIG. 9A. Conjunto de treinamento: Eliminação Recursiva de Recurso realizada em cada uma das 10 vezes no conjunto de treinamento. As vezes foram reamostradas 10 vezes. O desempenho da média dos escores de 10 reamostragens é mostrado. No eixo x, é mostrado o número de genes restantes por dobra. FIG. 9B- Conjunto de validação. Os principais genes foram selecionados de todas as 100 listas de genes em cada iteração, e seu desempenho foi avaliado no conjunto de treinamento. FIG. 9C- Conjunto combinado. Todas as 741

amostras usadas pelo SVM-RFE para classificar as sondas.

[0027] A FIG. 10 mostra gráficos do desempenho do painel personalizado nanoestruturado como comparação da curva ROC para diferentes números de genes.

DESCRIÇÃO DETALHADA DA INVENÇÃO

[0028] Os métodos e composições descritos neste documento aplicam a tecnologia de expressão de genes à triagem de sangue para a detecção e diagnóstico de câncer de pulmão. As composições e métodos aqui descritos fornecem a capacidade de distinguir um tumor cancerígeno de um nódulo não cancerígeno, determinando um perfil de expressão de RNA característico dos genes do sangue de um sujeito mamífero, preferencialmente humano. O perfil característico da expressão de gene é comparado com o perfil de um ou mais sujeitos da mesma classe (por exemplo, pacientes com câncer de pulmão ou nódulo não cancerígeno) ou um controle, para ver a qual classe o perfil de expressão de gene é mais semelhante, para fornecer um diagnóstico útil.

[0029] Esses métodos de triagem do câncer de pulmão empregam composições adequadas para a realização de um exame de sangue simples, econômico e não invasivo, utilizando o perfil de expressão de gene que poderia alertar o paciente e o médico para obter mais estudos, como uma PET ou CT adicional de radiografia de tórax, broncoscopia ou biópsia, da mesma maneira que o antígeno específico da próstata é usado para ajudar a diagnosticar e acompanhar o progresso do câncer de próstata. A aplicação desses perfis fornece diagnósticos sobrepostos e confirmatórios do tipo de doença pulmonar, começando com o teste inicial de doença maligna versus doença não maligna.

[0030] "Paciente" ou "sujeito" como utilizado neste documento significa um animal mamífero, incluindo um ser humano, um animal veterinário ou de criação, um animal doméstico ou animal de estimação,

e animais normalmente utilizados para pesquisa clínica. Em uma modalidade, o sujeito destes métodos e composições é um humano.

[0031] "Controle" ou "Sujeito de controle", como utilizado neste documento, refere-se à fonte dos perfis de expressão do gene de referência, bem como ao painel particular de sujeitos de controle aqui descrito. Em uma modalidade, o nível de controle ou referência é de um único sujeito. Em outra modalidade, o nível de controle ou referência é de uma população de indivíduos que compartilham uma característica específica.

[0032] Em ainda outra modalidade, o nível de controle ou referência é um valor atribuído que se correlaciona com o nível de um indivíduo ou população de controle específico, embora não seja necessariamente medido no momento do teste da amostra do sujeito de teste. Em uma modalidade, o sujeito ou referência de controle é de um paciente (ou população) com um nódulo não cancerígeno. Em outra modalidade, o sujeito ou referência de controle é de um paciente (ou população) com um tumor cancerígeno. Em outras modalidades, o sujeito de controle pode ser um sujeito ou população com câncer de pulmão, como um sujeito que é fumante atual ou ex-fumante com doença maligna, um sujeito com um tumor pulmonar sólido antes da cirurgia para remoção do mesmo; um sujeito com um tumor pulmonar sólido após a remoção cirúrgica do referido tumor; um sujeito com um tumor pulmonar sólido antes da terapia para o mesmo; e um sujeito com um tumor pulmonar sólido durante ou após a terapia para o mesmo.

[0033] Em outras modalidades, os controles para fins das composições e métodos aqui descritos incluem qualquer uma das seguintes classes de referência de sujeito humano sem câncer de pulmão. Esses controles não saudáveis (NHC) incluem as classes de fumantes com doença não maligna, um ex-fumante com doença não maligna (incluindo pacientes com nódulos pulmonares), um não

fumante com doença pulmonar obstrutiva crônica (COPD) e um ex-fumante com COPD. Ainda em outras modalidades, o sujeito de controle é um não fumante saudável sem doença ou um fumante saudável sem doença.

[0034] "Amostra", como utilizado neste documento, significa qualquer fluido ou tecido biológico que contém células imunes e/ou células cancerígenas. A amostra mais adequada para utilização nesta invenção inclui sangue total. Outras amostras biológicas úteis incluem, sem limitação, células mononucleares do sangue periférico, plasma, saliva, urina, líquido sinovial, medula óssea, líquido cefalorraquidiano, muco vaginal, muco cervical, secreções nasais, escarro, sêmen, líquido amniótico, amostra broncoscópica, líquido de lavagem broncoalveolar, escovações nasais e outros exsudatos celulares de um paciente com câncer. Tais amostras podem ainda ser diluídas com solução salina, tampão ou um diluente fisiologicamente aceitável. Alternativamente, essas amostras são concentradas por meios convencionais.

[0035] Como utilizado neste documento, o termo "câncer" refere-se ou descreve a condição fisiológica em mamíferos que é tipicamente caracterizada por crescimento celular não regulado. Mais especificamente, como utilizado neste documento, o termo "câncer" significa qualquer câncer de pulmão. Em uma modalidade, o câncer de pulmão é um câncer de pulmão de células não pequenas (NSCLC). Em uma modalidade mais específica, o câncer de pulmão é adenocarcinoma de pulmão (AC ou LAC). Em uma outra modalidade mais específica, o câncer de pulmão é o carcinoma de células escamosas do pulmão (SCC ou LSCC). Em outra modalidade, o câncer de pulmão é um NSCLC estágio I ou estágio II. Ainda em outra modalidade, o câncer de pulmão é uma mistura de estágios iniciais e tardios e tipos de NSCLC.

[0036] O termo "tumor", como utilizado neste documento, refere-se

a todo o crescimento e proliferação de células neoplásicas, malignas ou benignas, e a todas as células e tecidos pré-cancerígenos e cancerígenos. O termo "nódulo" refere-se a um acúmulo anormal de tecido que pode ser maligno ou benigno. O termo "tumor cancerígeno" refere-se a um tumor maligno.

[0037] Por "diagnóstico" ou "avaliação" entende-se um diagnóstico de um câncer de pulmão, um diagnóstico de um estágio de câncer de pulmão, um diagnóstico de um tipo ou classificação de um câncer de pulmão, um diagnóstico ou detecção de uma recorrência de um câncer de pulmão, um diagnóstico ou detecção de uma regressão de um câncer de pulmão, um prognóstico de um câncer de pulmão ou uma avaliação da resposta de um câncer de pulmão a uma terapia cirúrgica ou não cirúrgica. Em uma modalidade, "diagnóstico" ou "avaliação" refere-se à distinção entre um tumor cancerígeno e um nódulo pulmonar benigno.

[0038] Como utilizado neste documento, "sensibilidade" (também chamada de taxa positiva verdadeira) mede a proporção de positivos que são corretamente identificados como tais (por exemplo, a porcentagem de pessoas doentes que são corretamente identificadas como portadoras da condição).

[0039] Como utilizado neste documento, a "especificidade" (também chamada de taxa negativa verdadeira) mede a proporção de negativos que são corretamente identificados como tal (por exemplo, a porcentagem de pessoas saudáveis que são corretamente identificadas como sem a condição).

[0040] Por "mudança de expressão" entende-se uma suprarregulação de um ou mais genes selecionados em comparação com a referência ou controle; uma infrarregulação de um ou mais genes selecionados em comparação com a referência ou controle; ou uma combinação de certos genes suprarregulados e genes infrarregulados.

[0041] Por "reagente terapêutico" ou "regime" entende-se qualquer

tipo de tratamento empregado no tratamento de cânceres com ou sem tumores sólidos, incluindo, sem limitação, produtos farmacêuticos quimioterapêuticos, modificadores de resposta biológica, radiação, dieta, terapia com vitaminas, terapias hormonais, terapia de genes, ressecção cirúrgica, etc.

[0042] Por "genes informativos", como utilizado neste documento, entende-se aqueles genes cuja expressão muda (de maneira suprarregulada ou infrarregulada) caracteristicamente na presença de câncer de pulmão. Um número estatisticamente significativo de tais genes informativos forma assim perfis de expressão de gene adequados para uso nos métodos e composições. Tais genes são mostrados na Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX abaixo. Tais genes compõem o "perfil de expressão".

[0043] O termo "número estatisticamente significativo de genes" no contexto desta invenção difere dependendo do grau de alteração na expressão de gene observada. O grau de alteração na expressão de gene varia de acordo com o tipo de câncer e com o tamanho ou a disseminação do câncer ou tumor sólido. O grau de alteração também varia com a resposta imune do indivíduo e está sujeito a variações com cada indivíduo. Por exemplo, em uma modalidade desta invenção, uma grande mudança, por exemplo, 2-3 vezes aumenta ou diminui em um pequeno número de genes, por exemplo, em cerca de 5-8 genes, é estatisticamente significativa. Em outra modalidade, uma mudança relativa menor em cerca de 15 ou mais genes é estatisticamente significativa.

[0044] Assim, os métodos e composições aqui descritos contemplam o exame do perfil de expressão de um "número estatisticamente significativo de genes" variando de 5 a cerca de 50 genes em um único perfil. Em uma modalidade, os genes são selecionados da Tabela I. Em outra modalidade, os genes são

selecionados da Tabela II. Em uma outra modalidade, os genes são selecionados da Tabela III. Em uma modalidade, o perfil do gene é formado por um número estatisticamente significativo de 5 ou mais genes. Em uma modalidade, o perfil do gene é formado por um número estatisticamente significativo de 7 ou mais genes. Em uma modalidade, o perfil do gene é formado por um número estatisticamente significativo de 8 ou mais genes. Em uma modalidade, o perfil do gene é formado por um número estatisticamente significativo de 10 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 15 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 20 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 25 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 30 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 35 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 40 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 41 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 45 ou mais genes. Em uma outra modalidade, o perfil do gene é formado por um número estatisticamente significativo de 50 ou mais genes. Em outra modalidade, o perfil do gene é formado por 1, 2, 3, 4, 5, 6, 7 ou 8 genes da Tabela I. Em outra modalidade, o perfil do gene é formado por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 , 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 ou 50 genes da Tabela II. Em outra modalidade, o perfil do gene é formado por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 ou 50 genes da Tabela III. Em outra modalidade, o perfil do gene é formado por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 ou 100 da Tabela IV. Em outra modalidade, o perfil do gene é formado por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 ou 41 genes da Tabela IX.

[0045] A Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX abaixo referem-se a uma coleção de genes conhecidos úteis na discriminação entre um sujeito com câncer de pulmão, por exemplo, NSCLC, e sujeitos com nódulos pulmonares benignos (não malignos). As sequências dos genes identificados na Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX estão disponíveis ao público. Um versado na técnica pode reproduzir facilmente as composições e métodos aqui descritos, utilizando as sequências dos genes, todos disponíveis publicamente em fontes convencionais, como o GenBank. O número de acessão do GenBank para cada gene é fornecido.

[0046] O termo "microarranjo" refere-se a um arranjo ordenado de elementos de arranjo hibridizáveis, preferencialmente sondas polinucleotídicas ou oligonucleotídicas, em um substrato.

[0047] O termo "polinucleotídeo", quando utilizado na forma singular ou plural, geralmente refere-se a qualquer polirribonucleotídeo ou polideoxribonucleotídeo, que pode ser RNA ou DNA não modificado ou RNA ou DNA modificado. Assim, por exemplo, os polinucleotídeos conforme aqui definidos incluem, sem limitação, DNA de fita simples e

dupla, DNA incluindo regiões de fita simples e dupla, RNA de fita simples e dupla e RNA incluindo regiões de fita simples e dupla, moléculas híbridas compreendendo DNA e RNA que podem ser de fita simples ou, mais tipicamente, de fita dupla ou incluem regiões de fita simples e dupla. Além disso, o termo "polinucleotídeo", conforme utilizado neste documento, refere-se a regiões de cadeia tripla compreendendo RNA ou DNA ou RNA e DNA. As cadeias nessas regiões podem ser da mesma molécula ou de moléculas diferentes. As regiões podem incluir todas as uma ou mais moléculas, mas mais tipicamente envolvem apenas uma região de algumas das moléculas. Uma das moléculas de uma região tripla helicoidal é frequentemente um oligonucleotídeo. O termo "polinucleotídeo" inclui especificamente cDNAs. O termo inclui DNAs (incluindo cDNAs) e RNAs que contêm uma ou mais bases modificadas. Assim, DNAs ou RNAs com espinhas dorsais modificadas para estabilidade ou por outras razões são "polinucleotídeos", como o termo aqui é pretendido. Além disso, DNAs ou RNAs compreendendo bases incomuns, como inosina, ou bases modificadas, como bases tritiadas, estão incluídos no termo "polinucleotídeos", conforme aqui definido. Em geral, o termo "polinucleotídeo" abrange todas as formas quimicamente, enzimaticamente e/ou metabolicamente modificadas de polinucleotídeos não modificados, bem como as formas químicas de DNA e RNA características de vírus e células, incluindo células simples e complexas.

[0048] O termo "oligonucleotídeo" refere-se a um polinucleotídeo relativamente curto, incluindo, sem limitação, desoxirribonucleotídeos de fita simples, ribonucleotídeos de fita simples ou dupla, híbridos de RNA:DNA e DNAs de fita dupla. Os oligonucleotídeos, como os oligonucleotídeos de sonda de DNA de fita única, são frequentemente sintetizados por métodos químicos, por exemplo, usando sintetizadores

de oligonucleotídeos automatizados que estão disponíveis comercialmente.

[0049] No entanto, os oligonucleotídeos podem ser produzidos por uma variedade de outros métodos, incluindo técnicas mediadas por DNA recombinante in vitro e pela expressão de DNAs em células e organismos.

[0050] Os termos "gene diferencialmente expresso", "expressão diferencial de gene" e seus sinônimos, usados de forma intercambiável, referem-se a um gene cuja expressão é ativada para um nível mais alto ou mais baixo em um sujeito que sofre de uma doença, especificamente câncer, como câncer de pulmão, relativo à sua expressão em um sujeito de controle, como um sujeito com um nódulo benigno. Os termos também incluem genes cuja expressão é ativada para um nível superior ou inferior em diferentes estágios da mesma doença. Entende-se também que um gene expresso diferencialmente pode ser ativado ou inibido no nível de ácido nucleico ou nível de proteína, ou pode estar sujeito a emenda alternativa para resultar em um produto polipeptídico diferente. Tais diferenças podem ser evidenciadas por uma alteração nos níveis de mRNA, expressão da superfície, secreção ou outra partição de um polipeptídeo, por exemplo. A expressão de gene diferencial pode incluir uma comparação da expressão entre dois ou mais genes ou seus produtos de gene, ou uma comparação das razões da expressão entre dois ou mais genes ou seus produtos de gene, ou mesmo uma comparação de dois produtos processados de maneira diferente do mesmo gene, que diferem entre sujeitos normais, controles não relacionados à saúde e sujeitos que sofrem de uma doença, especificamente câncer, ou entre vários estágios da mesma doença. A expressão diferencial inclui diferenças quantitativas, bem como qualitativas, no padrão de expressão temporal ou celular em um gene ou em seus produtos de expressão entre, por exemplo, células normais

e doentes, ou entre células que sofreram diferentes eventos ou estágios da doença. Para os fins desta invenção, "expressão diferencial de gene" é considerada presente quando existe uma diferença estatisticamente significante ($p < 0,05$) na expressão de gene entre o sujeito e as amostras de controle.

[0051] O termo "superexpressão" em relação a um transcrito de RNA é usado para se referir ao nível do transcrito determinado por normalização ao nível de mRNAs de referência, que podem ser todos os transcritos medidos na amostra ou em um conjunto de referência específico de mRNAs.

[0052] A frase "amplificação de genes" refere-se a um processo pelo qual várias cópias de um gene ou fragmento de gene são formadas em uma célula ou linhagem celular específica. A região duplicada (um trecho de DNA amplificado) é frequentemente chamada de "amplicon". Geralmente, a quantidade de RNA mensageiro (mRNA) produzido, isto é, o nível de expressão de gene, também aumenta na proporção do número de cópias feitas do gene específico expresso.

[0053] No contexto das composições e métodos aqui descritos, referência a "7 ou mais", "pelo menos 7" etc. dos genes listados na Tabela I, Tabela II, Tabela III, Tabela III, Tabela IV ou Tabela IX significa qualquer um ou qualquer e todas as combinações dos genes listados. Por exemplo, perfis de expressão de gene adequados incluem perfis contendo qualquer número entre pelo menos 7 a 50 genes da Tabela II. Em outro exemplo, os perfis de expressão de gene adequados incluem perfis contendo qualquer número entre pelo menos 8 a 50 genes da Tabela III. Por exemplo, perfis de expressão de gene adequados incluem perfis contendo qualquer número entre pelo menos 7 a 100 genes da Tabela IV. Por exemplo, perfis de expressão de gene adequados incluem perfis contendo qualquer número entre pelo menos 7 a 41 genes da Tabela IX. Em uma modalidade, os perfis de genes

formados por genes selecionados em uma tabela são usados em ordem de classificação, por exemplo, os genes classificados no topo da lista demonstraram resultados discriminatórios mais significativos nos testes e, portanto, podem ser mais significativos em um perfil do que os genes de classificação inferior. No entanto, em outras modalidades, os genes que formam um perfil genético útil não precisam estar na ordem de classificação e podem ser qualquer gene da tabela. Ao referir-se à "Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX", também é contemplado que suas combinações podem ser feitas para fornecer um classificador útil aqui.

[0054] Como utilizado neste documento, "marcadores" ou "moléculas repórter" são frações químicas ou bioquímicas úteis para marcar um ácido nucleico (incluindo um único nucleotídeo), polinucleotídeo, oligonucleotídeo ou ligante de proteína, por exemplo, aminoácido ou anticorpo. "Marcadores" e "moléculas repórter" incluem agentes fluorescentes, agentes quimioluminescentes, agentes cromogênicos, agentes de extinção, radionucleotídeos, enzimas, substratos, cofatores, inibidores, partículas magnéticas e outras frações conhecidas na técnica. "Marcadores" ou "moléculas repórteres" são capazes de gerar um sinal mensurável e podem ser covalentemente ou não covalentemente unidos ou ligados a um oligonucleotídeo ou nucleotídeo (por exemplo, um nucleotídeo não natural) ou ligante.

[0055] Salvo definição em contrário neste relatório descritivo, os termos técnicos e científicos aqui utilizados têm o mesmo significado como vulgarmente entendido por um versado na técnica à qual essa invenção pertence e por referência a textos publicados, que proporcionam aos versados na técnica um guia geral para muitos dos termos utilizados no presente pedido.

I. PERFIS DE EXPRESSÃO DE GENE

[0056] Os inventores mostraram que os perfis de expressão de

gene do sangue total de pacientes com câncer de pulmão diferem significativamente daqueles observados em pacientes com nódulos pulmonares não cancerígenos. Por exemplo, alterações nos produtos de expressão de gene dos genes da Tabela I, Tabela II, Tabela III, Tabela IV e/ou Tabela IX podem ser observadas e detectadas pelos métodos desta invenção no sangue circulante normal de pacientes com tumores sólidos pulmonares em estágio inicial.

[0057] Os perfis de expressão de gene aqui descritos fornecem novos marcadores de diagnóstico para a detecção precoce do câncer de pulmão e podem impedir que os pacientes sejam submetidos a procedimentos desnecessários relacionados à cirurgia ou biópsia para um nódulo benigno. Como os riscos são muito baixos, a razão benefício/risco é muito alta. Em uma modalidade, os métodos e composições descritos neste documento podem ser usados em conjunto com fatores de risco clínicos para ajudar os médicos a tomar decisões mais precisas sobre como gerenciar pacientes com nódulos pulmonares. Outra vantagem desta invenção é que o diagnóstico pode ocorrer precocemente, uma vez que o diagnóstico não depende da detecção de células tumorais circulantes que estão presentes apenas em pequenos números que desaparecem nos cânceres de pulmão em estágio inicial.

[0058] Em um aspecto, é fornecida uma composição para classificar um nódulo como cancerígeno ou benigno em um mamífero. Em uma modalidade, a composição inclui pelo menos 7 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os 7 primeiros genes listados na Tabela IV. Em uma modalidade, a composição inclui pelo

menos 8 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os da Tabela I. Em outra modalidade, a composição inclui pelo menos 15 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os primeiros 15 genes da Tabela II. Em uma modalidade, os genes são os primeiros 15 genes da Tabela III. Em uma modalidade, os genes são os primeiros 15 genes da Tabela IV. Em uma modalidade, os genes são os primeiros 15 genes da Tabela IX. Em uma outra modalidade, a composição inclui pelo menos 41 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os genes da Tabela IX. Em uma outra modalidade, a composição inclui pelo menos 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os genes da Tabela II. Em uma outra modalidade, os genes são os genes da Tabela III. Em outra modalidade, os genes são os genes da Tabela IV. Em uma modalidade, o polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um mRNA.

Tabela I

Nome do Gene	Acessão #	Classificação T1	Classificação S1
MERTK	NM_006343.2	11	3
SLC25A20	NM_000387.5	1	6
KYNU	NM_001032998.1	15	8
P2RY5	NM_005767.5	4	14
SNORA56	NR_002984.1	44	16
IL1B	NM_000576.2	36	18
LY96	NM_015364.4	32	22
REPIN1	NM_014374.3	38	38

Tabela II

Nome do Gene	Acessão #	Classificação T1
SLC25A20	NM_000387.5	1
CD160	NM_007053.3	2
CCL3	NM_002983.2	3
P2RY5	NM_005767.5	4
CCL3L3	NM_001001437.3	5
PSMA6	NM_002791.2	6
LILRA5 b	NM_181879.2	7
CCND3	NM_001760.2	8
LDHA	NM_001165416.1	9
NME1-NME2	NM_001018136.2	10
MERTK	NM_006343.2	11
CXCR5 b	NM_001716.3	12
TAPBP	NM_003190.4	13
CAMP	NM_004345.4	14
KYNU	NM_001032998.1	15
ACAA2	NM_006111.2	16
ANXA1 b	NM_000700.1	17
CABC1	NM_020247.4	18
SOCS1	NM_003745.1	19
C4orf27	NM_017867.2	20
SPA17	NM_017425.3	21
MEN1	NM_130799.2	22
MAGEA3	NM_005362.3	23
UBA1	NM_003334.3	24
CLN8	NM_018941.3	25
ETFDH	NM_004453.3	26
RHOB	NM_004040.3	27
CD160 b	NM_007053.2	28
KIR_Activating_Subgroup_2	NM_014512.1	29
PDGFD	NM_033135.3	30
HLA-DMB	NM_002118.3	31
LY96	NM_015364.4	32
IL16	NM_004513.4	33
DPF2	NM_006268.4	34
RBX1	NM_014248.3	35
IL1B	NM_000576.2	36
LOC148137	NM_144692.1	37
REPIN1	NM_014374.3	38
PELP1	NM_014389.2	39
PRG2	NM_002728.4	40
RHOU	NM_021205.5	41
C19orf59	NM_174918.2	42
C1orf103	NM_018372.3	43
SNORA56	NR_002984.1	44
IL1R2	NM_173343.1	45
NFATC4	NM_001136022.2	46
ANP32B	NM_006401.2	47
C4B	NM_001002029.3	48
STOM	NM_004099.5	49
LPIN2	NM_014646.2	50

Tabela III

Nome do Gene	Acessão #	Classificação S1
TBCE	NM_001079515.2	1
ITGAL	NM_002209.2	2
MERTK	NM_006343.2	3
BCOR	NM_017745.5	4
GLT25D1	NM_024656.2	5
SLC25A20	NM_000387.5	6
LOC100130229	XM_001717158.1	7
KYNU	NM_001032998.1	8
BANP	NM_079837.2	9
IGFBP7	NM_001553.2	10
SFRS15	NM_020706.2	11
SH2D3C	NM_170600.2	12
DNAJB1	NM_006145.2	13
P2RY5	NM_005767.5	14
PSMB7	NM_002799.2	15
SNORA56	NR_002984.1	16
ATP5L	NM_006476.4	17
IL1B	NM_000576.2	18
CDC42EP2	NM_006779.3	19
USP34	NM_014709.3	20
AMD1	NM_001634.4	21
LY96	NM_015364.4	22
ARG1	NM_000045.3	23
DGUOK	NM_080916.2	24
TNFSF8	NM_001244.3	25
ATG5	NM_004849.2	26
SLC6A6	NM_003043.5	27
FAIM3	NM_005449.4	28
RHOG	NM_001665.3	29
CASP1	NM_033294.3	30
PHRF1	NM_020901.3	31
TMBIM6	NM_003217.2	32
FLJ10357	NM_018071.4	33
HSP90AB1	NM_007355.3	34
CDH5	NM_001795.3	35
SNX11	NM_152244.1	36
RERE	NM_001042682.1	37
REPIN1	NM_014374.3	38
REPS1	NM_001128617.2	39
RELA	NM_021975.2	40
HERC1	NM_003922.3	41
AKAP4	NM_139289.1	42
P2RY10	NM_198333.1	43
HSCB	NM_172002.3	44
TRRAP	NM_003496.3	45
SETD1B	XM_037523.11	46
ARHGAP26	NM_015071.4	47
DYNC2LI1	NM_016008.3	48
TCP1	NM_030752.2	49
DGUOK b	NM_080916.2	50

Tabela IV

Classificação	Nome do Gene	Descrição	#Acessão
1	SLC25A20	Família transportadora de soluto 25 (carnitina/acilcarnitina translocase), membro 20	NM 000387.5
2	CCL3L3	Ligando quimiocina (motif C-C) 3 – tipo 3	NM_001001437.3
3	LDHA	Lactato desidrogenase A	NM_001165416.1
4	C4orf27	Cromossomo 4 quadro leitura aberta 27	NM_017867.2
5	RHOU	Membro família homóloga ras	NM_021205.5
6	COMMD6	Domínio COMM contendo 6	NM_203497.3
7	ZNF143	Proteína dedo de zinco 143	NM_003442.5
8	CASP3	Caspase 3, cisteína peptidase relativa a apoptose	NM_032991.2
9	P2RY5	Receptor do ácido lisofosfatídico 6	NM_005767.5
10	ZNF341	Proteína dedo de zinco 341	NM_032819.4
11	CD160	Molécula CD160	NM_007053.3
12	EIF4ENIF1	Fator de iniciação de tradução eucariótica 4E fator de importação nuclear 1	NM_019843.2
13	LILRA5b	Receptor tipo imunoglobulina de leucócito A5	NM_181879.2
14	RNF114	Proteína dedo de zinco 114	NM_018683.3
15	IL16	Interleucina 16	NM_004513.4
16	REPS1	Domínio Eps associado a RALBP1 contendo 1	NM_001128617.2
17	TMEM70	Proteína transmembrana 70	NM_017866.5

18	PRG2	Proteoglicano 2, medula óssea (ativador de célula natural killer, proteína básica maior de granulo eosinófilo)	NM_002728.4
19	CCR1	Receptor de quimiocina (motif C-C)	NM_001295.5
20	LOC148137	NA	NM_144692.1
21	HOOK3	Proteína de amarração de microtúbulo de gancho	NM_032410.3
22	C1orf222	NA	NM_001003808.1
23	KYNU	quinureninase	NM_001032998.1
24	CLN8	ceroide-lipofuscinose, neuronal 8 (epilepsia, progressiva com retardo mental)	NM_018941.3
25	PDGFD	Fator de crescimento derivado de plaqueta D	NM_033135.3
26	LOC645914	NA	XM_928884.1
27	SPA17	Proteína autoantigênica de esperma 17	NM_017425.3
28	MTCH1	Transportador de mitocôndria 1	NM_014341.2
29	STOM	Estomatina	NM_004099.5
30	CCND3	Ciclina D3	NM_001760.2
31	EHD4	Domínio EH contendo 4	NM_139265.3
32	IDO1	Indolamina 2,3-dioxigenase 1	NM_002164.3
33	PPP6C	Proteína fosfatase 6, subunidade catalítica	NM_002721.4
34	IL1B	Interleucina 1, beta	NM_0000576.2

Classificação	Nome do Gene	Descrição	#Acessão
35	SETD2	Domínio SET contendo 2	NM_014159.6
36	ILIR2	Receptor da interleucina 1, tipo II	NM_173343.1
37	ATP51	ATP sintase, transporte	NM_007100.2

		de H+, complexo Fo de mitocôndria, subunidade E	
38	CTSW	Catepsina W	NM_001335.3
39	HNRNPK	Ribonucleoproteína nuclear heterogêna K	NM_031263.2
40	NFATC4	Fator nuclear de células T ativadas, citoplásmico, dependente da calcineurina 4	NM_001136022.2
41	KIAA0101	KIAA0101	NM_014736.4
42	NME1-NME2	Leitura direta NME1-NME2	NM_001018136.2
43	REPIN1	Iniciador de replicação 1	NM_014374.3
44	PELP1	Prolina, proteína rica em glutamato e leucina 1	NM_014389.2
45	FOXK2	Caixa forkhead K2	NM_004514.3
46	MAGEA1	Família de antígeno de melanoma A, 1 (dirige expressão de antígeno MZ2-E)	NM_004988.4
47	HLA-DMB	Complexo de histocompatibilidade maior, classe II, DM beta	NM_002118.3
48	C17orf51	Cromossomo 17 quadro de leitura aberto 51	XM_944416.1
49	CAMP	Peptídeo antimicrobiano catelicidina	NM_004345.4
50	SMARCC1	Regulador da cromatina relativo a SWI/SNF, associado a matriz, dependente de actina,	NM_003074.3

		subfamília c, membro 1	
51	MAGEA3	Família de antígeno de melanoma A, 3	NM_005362.3
52	TTC9	Domínio de repetição de tetratricopeptídeo 9	NM_015351.1
53	MARCKS	Substrato da proteína cinase C rica em alanina miristoilada	NM_002356.6
54	C19orf59	NA	NM_174918.2
55	MEN1	Neoplasia endócrina múltipla 1	NM_130799.2
56	PUM1	Membro da família de ligação a RNA de pumilio 1	NM_001020658.1
57	USP9Y	Peptidase específica de ubiquitina 9, ligada a Y	NM_004654.3
58	PACS1	Proteína de seleção de agrupamento acídico de fosfofurina 1	NM_018026.3
59	S100A8	Proteína de ligação a cálcio S100 A8	NM_002964.4
60	MBD1D	Proteína do domínio de ligação metil1-CpG	XM_015844.2
61	CS	Citrase sintase	NM_004077.2
62	UBE2G1	Enzima de conjugação de ubiqutina E2G 1	NM_003342.4
63	KIAA1267	Subunidade do complexo NSL regulador de KAT8 1	NM_015443.3
64	MERTK	Proto-oncogene MER, tirosina cinase	NM_006343.2
65	CTAG1B	Antígeno de câncer/testículo 1B	NM_001327.2

66	CRKL	Vírus do sarcoma aviário v-crk CT10 tipo homólogo de oncogene	NM_005207.3
67	SYNJ1	Sinaptojanina 1	NM_003895.3
68	C4B	Componente de complemento 4B (Chido grupo sanguíneo)	NM_001002029.3

Classificação	Nome do Gene	Descrição	#Acessão
69	SOCS1	Supressor de sinalização de citocina 1	NM_003745.1
70	NUP153	Nucleoproteína 153kDa	NM_005124.3
71	COLEC12	Membro da subfamília da colectina 12	NM_130386.2
72	TAPBP	Proteína de ligação a TAP (tapasina)	NM_003191.4
73	IFI27L2	Interferon, proteína indutiva alfa 27 tipo 2	NM_032036.2
74	RBX1	Ring-box, ligase de proteína de ubiqutina E3	NM_014248.3
75	CR2 b	Receptor 2 de complemento C3d	NM_001006658.1
76	C1orf103	NA	NM_018372.3
77	TBCE	Cofator de dobramento de tubulina E	NM_001079515.2
78	CCL3	Ligando da quimicina (motif C-C)	NM_002983.2
79	LOC100129022	NA	XM_001716591.1
80	NCAPG	Complexo da condensina não SMC I, subunidade G	NM_022346.4
81	FLNB	Filamina B, beta	NM_001457.3
82	C3	Componente de complemento 3	NM_000064.2

83	SAP130 b	Proteína associada a Sin3A 130	NM_024545.3
84	CD160 b	Molécula CD160	NM_007053.2
85	STAG3	Antígeno estromal 3	NM_012447.3
86	SFPQ	Fator de splicing rico em prolina/glutamina	NM_005066.2
87	ITCH	Proteína ligase da ubiquitina E3 itchy	NM_001257138.1
88	HSCB	Cochaperona de agrupamento ferro-enxofre mitocondrial HscB	NM_172002.3
89	TFCP2	Fator de transcrição CP2	NM_005653.4
90	LIF	Fator inibidor de leucemia	NM_002309.3
91	BATF	Fator de transcrição de zíper de leucina básico, tipo ATF	NM_006399.3
92	SNORA56	RNA nucleolar pequeno, caixa H/ACA 56	NM_002984.1
93	ETFDH	Desidrogenase da flavoproteína de transferência de elétrons	NM_004453.3
94	BCL10	CLL de célula B/linfoma 10	NM_003921.2
95	TIAM1	Invasão e metástase de linfoma de célula T 1	NM_003253.2
96	MPDU1	Defeito de utilização de manose-P-dolicol 1	NM_004870.3
97	TRIM39	Motif tripartite contendo 39	NM_021253.3
98	RNF34	Proteína dedo de zinco 34, ligase da proteína ubiquitina E3	NM_025126.3
99	AMD1	Descarboxilase da adenosilmetionina 1	NM_001634.4

100	PSMA6	Subunidade de proteassoma (prosoma, macrodor), alfa tipo 6	NM_002791.2

TABELA IX

41 Classificador de Gene

1	SLC25A20	21	MED16
2	P2RY5	22	EMR4
3	DNAJB1	23	REPIN1
4	CCND3	24	DNAJB6
5	CD160	25	IDO1
6	MERTK	26	PSMB7
7	BCOR	27	HSP90AB1
8	ABCA5	28	CABC1
9	RNASE2	29	PRPF3
10	IGFBP7	30	PSMB8
11	ITGAL	31	TRIM39
12	DYNC2LI1	32	CD48
13	EEF1B2	33	CDH5
14	RAG1	34	KLRC1
15	DDIT4	35	TUG1
16	ARG1	36	PIM2
17	TBC1D12	37	CLPTM1
18	AZI2	38	REPS1
19	LOC100130229	39	USP9Y
20	STOM	40	AFTPH
		41	SLC6A12

[0059] Em uma modalidade, um novo perfil ou assinatura de expressão de gene pode identificar e distinguir pacientes com tumores cancerígenos de pacientes com nódulos benignos. Ver, por exemplo, os genes identificados na Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX, que podem formar um perfil de expressão de gene adequado. Em uma outra modalidade, uma porção dos genes da Tabela I forma um perfil adequado. Em ainda outra modalidade, uma porção dos genes da

Tabela II forma um perfil adequado. Em ainda uma outra modalidade, uma porção dos genes da Tabela I forma um perfil adequado. Em ainda uma outra modalidade, uma porção dos genes da Tabela IV forma um perfil adequado. Em ainda uma outra modalidade, uma porção dos genes da Tabela IX forma um perfil adequado. Como discutido aqui, esses perfis são usados para distinguir entre tumores cancerígenos e não cancerígenos, gerando uma pontuação discriminante com base nas diferenças nos perfis de expressão de gene, como exemplificado abaixo. A validade dessas assinaturas foi estabelecida em amostras coletadas em diferentes locais por diferentes grupos em uma coorte de pacientes com nódulos pulmonares não diagnosticados. Ver os Exemplos e as FIGURAS 1 e 2. As assinaturas de câncer de pulmão ou os perfis de expressão de gene aqui identificados (isto é, Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX) podem ser ainda mais otimizados para reduzir o número de produtos de expressão de gene necessários e aumentar a precisão do diagnóstico.

[0060] Em uma modalidade, a composição inclui cerca de 7 ou polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma outra modalidade, a composição inclui cerca de 5 a cerca de 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II. Em uma outra modalidade, a composição inclui pelo menos 5 a cerca de 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma

amostra selecionada dos genes da Tabela III. Em uma outra modalidade, a composição inclui cerca de 5 a cerca de 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em outra modalidade, a composição inclui 1, 2, 3, 4, 5, 6, 7 ou 8 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela I. Em outra modalidade, a composição inclui 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 ou 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela II. Em outra modalidade, a composição inclui 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47 , 48, 49 ou 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela III. Em outra modalidade, a composição inclui 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 , 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47 , 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72 , 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93,

94, 95, 96, 97 , 98, 99 ou 100 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrição de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela IV. Em outra modalidade, a composição inclui 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 ou 41 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela IX. Em uma modalidade, a composição inclui pelo menos 3 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 5 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 7 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 8 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma

amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 10 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 15 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 20 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 25 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão diferente em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 30 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 35 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de

expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 40 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 45 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 50 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, a composição inclui pelo menos 55 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 60 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 65 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da

Tabela IV. Em uma modalidade, a composição inclui pelo menos 70 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 75 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 80 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 85 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 90 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 95 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV. Em uma modalidade, a composição inclui pelo menos 100 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene,

fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela IV.

[0061] Em ainda outra modalidade, o perfil de expressão é formado pelos 3 primeiros genes na ordem de classificação da Tabela I, Tabela II, Tabela III, , Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 5 primeiros genes em ordem de classificação da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o perfil de expressão é formado pelos 8 primeiros genes na ordem de classificação da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 15 primeiros genes em ordem de classificação da Tabela II, Tabela III, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o perfil de expressão é formado pelos 20 primeiros genes na ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 25 primeiros genes em ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o perfil de expressão é formado pelos 30 primeiros genes na ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 35 primeiros genes em ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 40 primeiros genes em ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 41 primeiros genes em ordem de classificação da Tabela II, Tabela III, Tabela IV ou Tabela IX. Ainda em outra modalidade, o perfil de expressão é formado pelos 45 primeiros genes em ordem de classificação da Tabela II, Tabela III, ou Tabela IV. Em ainda outra modalidade, o perfil de expressão é formado pelos 50 primeiros genes na ordem de

classificação da Tabela II, Tabela III ou Tabela IX. Em ainda outra modalidade, o perfil de expressão é formado pelos primeiros 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69 , 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94 , 95, 96, 97, 98, 99, 100 genes em ordem de classificação da Tabela IV.

[0062] Como discutido abaixo, as composições aqui descritas podem ser usadas com os métodos de perfil de expressão de genes que são conhecidos na técnica. Assim, as composições podem ser adaptadas de acordo com o método para o qual se destinam a ser utilizadas. Em uma modalidade, pelo menos um polinucleotídeo ou oligonucleotídeo ou ligante está fixado a um marcador detectável. Em certas modalidades, cada polinucleotídeo ou oligonucleotídeo está fixado a um marcador detectável diferente, cada um capaz de ser detectado independentemente. Tais reagentes são úteis em ensaios como o nCounter, como descrito abaixo, e com os métodos de diagnóstico aqui descritos.

[0063] Em outra modalidade, a composição compreende um oligonucleotídeo ou ligante de captura, que hibridiza com pelo menos um polinucleotídeo ou oligonucleotídeo ou ligante. Em uma modalidade, esse oligonucleotídeo ou ligante de captura pode incluir uma sequência de ácido nucleico que é específica para uma porção do oligonucleotídeo ou polinucleotídeo ou ligante que é específico para o gene de interesse. O ligante de captura pode ser um peptídeo ou polipeptídeo que é específico para o ligante do gene de interesse. Em uma modalidade, o ligante de captura é um anticorpo, como em um ELISA sanduíche.

[0064] O oligonucleotídeo de captura também inclui uma fração que permite a ligação com um substrato. Esse substrato inclui, sem limitação, uma placa, cordão, lâmina, poço, cavaco ou câmara. Em uma modalidade, a composição inclui um oligonucleotídeo de captura para

cada polinucleotídeo ou oligonucleotídeo diferente que é específico para um gene de interesse. Cada oligonucleotídeo de captura pode conter a mesma fração que permite a ligação com o mesmo substrato. Em uma modalidade, a fração de ligação é biotina.

[0065] Assim, uma composição para tal diagnóstico ou avaliação em um mamífero como descrito aqui pode ser um kit ou um reagente. Por exemplo, uma modalidade de uma composição inclui um substrato sobre o qual os ligantes usados para detectar e quantificar o mRNA são imobilizados. O reagente, em uma modalidade, é um iniciador de ácido nucleico de amplificação (como um iniciador de RNA) ou par de iniciadores que amplifica e detecta uma sequência de ácido nucleico do mRNA. Em uma outra modalidade, o reagente é uma sonda polinucleotídica que hibridiza com a sequência alvo. Em uma outra modalidade, as sequências alvo estão ilustradas na Tabela III. Em uma outra modalidade, o reagente é um anticorpo ou fragmento de um anticorpo. O reagente pode incluir múltiplos referidos iniciadores, sondas ou anticorpos, cada um específico para pelo menos um gene, fragmento de gene ou produto de expressão da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Opcionalmente, o reagente pode ser associado a um marcador detectável convencional.

[0066] Em outra modalidade, a composição é um kit contendo os múltiplos polinucleotídeos ou sondas ou ligantes oligonucleotídicos relevantes, marcadores detectáveis opcionais para o mesmo, substratos de imobilização, substratos opcionais para marcadores enzimáticos, bem como outros itens de laboratório. Ainda em outra modalidade, pelo menos um polinucleotídeo ou oligonucleotídeo ou ligante está associado a um marcador detectável. Em certas modalidades, o reagente é imobilizado em um substrato. Substratos exemplificativos incluem um microarranjo, chip, placa microfluídica ou câmara.

[0067] Em uma modalidade, a composição é um kit projetado para uso com o sistema nCounter Nanostring, como discutido mais adiante.

II. MÉTODOS DE PERFIL DE EXPRESSÃO DE GENES

[0068] Os métodos de análise de expressão de gene que foram utilizados na geração dos perfis úteis nas composições e métodos aqui descritos ou na execução das etapas de diagnóstico usando as composições aqui descritas são conhecidos e bem resumidos na Patente U.S. 7.081.340. Tais métodos de criação de perfil de expressão de genes incluem métodos baseados na análise de hibridização de polinucleotídeos, métodos baseados no sequenciamento de polinucleotídeos e métodos baseados em proteômica. Os métodos mais comumente conhecidos na técnica para quantificação da expressão de mRNA em uma amostra incluem Northern blotting e hibridização in situ; Ensaios de proteção de RNAse; Análise nCounter®; e métodos baseados em PCR, como RT-PCR. Alternativamente, podem ser empregados anticorpos que possam reconhecer duplexes específicos, incluindo duplex de DNA, duplex de RNA e duplex híbrido de DNA-RNA ou duplex de proteína de DNA. Os métodos representativos para análise de expressão de gene baseada em sequenciamento incluem Análise Serial de Expressão de Genes (SAGE) e análise de expressão de gene por sequenciamento de assinatura massivamente paralela (MPSS).

[0069] Em certas modalidades, as composições aqui descritas são adaptadas para uso nos métodos de criação de perfil de expressão de gene e/ou diagnóstico aqui descritos, e aqueles conhecidos na técnica.

A. Amostra do paciente

[0070] A "amostra" ou "amostra biológica", como utilizado neste documento, significa qualquer fluido ou tecido biológico que contenha células imunes e/ou células cancerígenas. Em uma modalidade, uma amostra adequada é sangue total. Em outra modalidade, a amostra pode ser sangue venoso. Em outra modalidade, a amostra pode ser

sangue arterial. Em outra modalidade, uma amostra adequada para uso nos métodos aqui descritos inclui sangue periférico, mais especificamente células mononucleares do sangue periférico. Outras amostras biológicas úteis incluem, sem limitação, plasma ou soro. Ainda em outra modalidade, a amostra é saliva, urina, líquido sinovial, medula óssea, líquido cefalorraquidiano, muco vaginal, muco cervical, secreções nasais, escovações nasais, escarro, sêmen, líquido amniótico, líquido de lavagem broncoalveolar e outros exsudatos celulares de um sujeito com suspeita de doença pulmonar. Tais amostras podem ainda ser diluídas com solução salina, tampão ou um diluente fisiologicamente aceitável. Alternativamente, essas amostras são concentradas por meios convencionais. Deve-se entender que o uso ou referência ao longo deste relatório descritivo para qualquer amostra biológica é apenas exemplificativo. Por exemplo, onde no relatório descritivo a amostra é referida como sangue total, entende-se que outras amostras, por exemplo, soro, plasma, etc., também podem ser empregadas em outra modalidade.

[0071] Em uma modalidade, a amostra biológica é sangue total e o método emprega o sistema PaxGene Blood RNA Workflow (Qiagen). Esse sistema envolve a coleta de sangue (por exemplo, coleta de sangue único) e estabilização do RNA, seguido pelo transporte e armazenamento, seguido pela purificação dos testes de RNA total e RNA molecular. Este sistema fornece estabilização imediata do RNA e volumes consistentes de coleta de sangue. O sangue pode ser coletado no consultório ou clínica do médico e a amostra transportada e armazenada no mesmo tubo. A estabilidade do RNA a curto prazo é de 3 dias entre 18-25°C ou 5 dias entre 2-8°C. A estabilidade do RNA a longo prazo é de 4 anos a -20 a -70°C. Esse sistema de coleta de amostras permite ao usuário obter dados confiáveis sobre a expressão de gene no sangue total. Em uma modalidade, a amostra biológica é

sangue total. Embora o sistema PAXgene tenha mais ruído do que o uso de PBMC como fonte de amostra biológica, os benefícios da coleta de amostras PAXgene superam os problemas. O ruído pode ser subtraído bioinformaticamente pelo versado na técnica.

[0072] Em uma modalidade, as amostras biológicas podem ser coletadas usando o sistema de RNA sanguíneo PaxGene (PreAnalytiX, uma empresa Qiagen, BD). O sistema de RNA do sangue PAXgene compreende dois componentes integrados: o tubo de RNA do sangue PAXgene e o kit de RNA do sangue PAXgene. As amostras de sangue são coletadas diretamente nos tubos de RNA do sangue PAXgene através da técnica padrão de flebotomia. Esses tubos contêm um reagente proprietário que estabiliza imediatamente o RNA intracelular, minimizando a degradação ex-vivo ou a suprarregulação dos transcritos de RNA. A capacidade de eliminar o congelamento de amostras em lotes e de minimizar a urgência de processar amostras após a coleta aumenta significativamente a eficiência do laboratório e reduz os custos.

[0073] Depois disso, o miRNA é detectado e/ou medido usando uma variedade de ensaios.

B. Análise Nanostring

[0074] Um método quantitativo sensível e flexível que é adequado para uso com as composições e métodos aqui descritos é o sistema nCounter® Analysis (NanoString Technologies, Inc., Seattle WA). O nCounter Analysis System utiliza uma tecnologia digital de código de barras com código de cores que se baseia na medição direta multiplexada da expressão de gene e oferece altos níveis de precisão e sensibilidade (<1 cópia por célula). A tecnologia usa "códigos de barras" moleculares e imageamento de molécula única para detectar e contar centenas de transcritos únicos em uma única reação. Cada código de barras codificado por cores é fixado a uma única sonda específica do alvo (isto é, polinucleotídeo, oligonucleotídeo ou ligante)

correspondente a um gene de interesse, isto é, um gene da Tabela I, Tabela II, Tabela III, Tabela IV, Tabela IV ou Tabela IX. Misturados com os controles, eles formam um CodeSet multiplexado. Em uma modalidade, o CodeSet inclui todos os 8 genes da Tabela I. Em uma modalidade, o CodeSet inclui os 7 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui todos os 50 genes da Tabela II. Em outra modalidade, o CodeSet inclui todos os 50 genes da Tabela III. Em outra modalidade, o CodeSet inclui pelo menos 3 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 5 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 8 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 10 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 15 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 20 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 25 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 30 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o CodeSet inclui pelo menos 40 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o CodeSet inclui pelo menos 41 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o CodeSet inclui pelo menos 50 genes da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 60 genes da Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui pelo menos 70 genes da Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o CodeSet inclui pelo menos 80 genes da Tabela III, Tabela IV ou Tabela IX. Em ainda outra

modalidade, o CodeSet inclui pelo menos 90 genes da Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o CodeSet inclui todos os 100 genes da Tabela III, Tabela IV ou Tabela IX. Em ainda outra modalidade, o CodeSet inclui qualquer subconjunto de genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX, incluindo combinações dos mesmos, como aqui descrito.

[0075] A plataforma NanoString emprega duas sondas de ~ 50 bases por mRNA que hibridizam em solução. O Reporter Probe transmite o sinal; o Capture Probe permite que o complexo seja imobilizado para coleta de dados. As sondas são misturadas com a amostra do paciente. Após a hibridação, as sondas em excesso são removidas e a sonda/alvo é alinhada e imobilizada em um substrato, por exemplo, no Cartucho nCounter.

[0076] As sequências alvo utilizadas nos Exemplos abaixo para cada um dos genes da Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX são mostradas na Tabela V abaixo e são reproduzidas na listagem de sequências. Essas sequências são partes das sequências publicadas desses genes. Alternativas adequadas podem ser prontamente projetadas por um versado na técnica.

[0077] Os Cartuchos de Amostra são colocados no Digital Analyzer para coleta de dados. Os códigos de cores na superfície do cartucho são contados e tabulados para cada molécula alvo. Um benefício do uso do sistema NanoString nCounter é que nenhuma amplificação de mRNA é necessária para realizar a detecção e quantificação. No entanto, em modalidades alternativas, são utilizados outros métodos quantitativos adequados. Ver, por exemplo, Geiss et al, Direct multiplexed measurement of gene expression with color-coded probe pairs, Nat Biotechnol. 2008 Mar;26(3):317-25. doi: 10.1038/nbt1385. Epub 2008 Feb 17, que é incorporado aqui por referência na sua totalidade.

C. Técnicas de reação em cadeia da polimerase (PCR)

[0078] Outro método quantitativo adequado é o RT-PCR, que pode ser usado para comparar os níveis de mRNA em diferentes populações de amostras, em tecidos normais e tumorais, para caracterizar padrões de expressão de gene, discriminar entre mRNAs intimamente relacionados e analisar a estrutura de RNA. A primeira etapa é o isolamento do mRNA a partir de uma amostra alvo (por exemplo, normalmente o RNA total isolado de PBMC humana). O mRNA pode ser extraído, por exemplo, de amostras de tecido congeladas ou arquivadas e embebidas em parafina e fixadas (por exemplo, fixadas em formalina).

[0079] Os métodos gerais para extração de mRNA são bem conhecidos na técnica, como livros didáticos padrão de biologia molecular. Em particular, o isolamento do RNA pode ser realizado usando um kit de purificação, conjunto de tampões e protease de fabricantes comerciais, de acordo com as instruções do fabricante. Exemplos de produtos comerciais incluem mini-colunas TRI-REAGENT, Qiagen RNeasy, Kit Completo de Purificação de DNA e RNA MASTERPURE (EPICENTRE®, Madison, Wis.), Kit de Isolamento de RNA de Bloco de Parafina (Ambion, Inc.) e RNA Stat-60 (Tel-Teste). Também podem ser empregues técnicas convencionais, como centrifugação em gradiente de densidade de cloreto de césio.

[0080] A primeira etapa no perfil de expressão de gene por RT-PCR é a transcrição reversa do modelo de RNA em cDNA, seguida por sua amplificação exponencial em uma reação de PCR. As duas transcrições reversas mais comumente usadas são a transcriptase reversa do vírus avilo mieloblastose (AMV-RT) e a transcriptase reversa do vírus da leucemia murina Moloney (MMLV-RT). A etapa de transcrição reversa é tipicamente iniciada usando iniciadores específicos, hexâmeros aleatórios ou oligo-dT, dependendo das circunstâncias e do objetivo do perfil de expressão. Ver, por exemplo, as instruções do fabricante que acompanham o kit GENEAMP RNA PCR (Perkin Elmer, Califórnia,

EUA). O cDNA derivado pode então ser usado como modelo na reação subsequente de RT-PCR.

[0081] A etapa de PCR geralmente usa uma polimerase de DNA dependente de DNA termoestável, como a polimerase de Taq DNA, que possui uma atividade de nuclease 5'-3', mas não possui uma atividade de endonuclease de revisão de 3'-5'. Assim, o TAQMAN® PCR utiliza tipicamente a atividade 5'-nuclease da polimerase Taq ou Tth para hidrolisar uma sonda de hibridização ligada ao seu amplicon alvo, mas qualquer enzima com atividade equivalente à nuclease 5' pode ser usada. São utilizados dois iniciadores oligonucleotídicos para gerar um amplicon típico de uma reação de PCR. Em uma modalidade, a sequência alvo é mostrada na Tabela V. Um terceiro oligonucleotídeo, ou sonda, é projetado para detectar a sequência nucleotídica localizada entre os dois iniciadores de PCR. A sonda não é extensível pela enzima Taq DNA polimerase e é marcada com um corante fluorescente repórter e um corante fluorescente extintor. Qualquer emissão induzida por laser do corante repórter é extinta pelo corante de extinção quando os dois corantes estão localizados próximos uns dos outros, como estão na sonda. Durante a reação de amplificação, a enzima Taq DNA polimerase cliva a sonda de maneira dependente do modelo. Os fragmentos de sonda resultantes se desassociam em solução e o sinal do corante repórter liberado é livre do efeito de extinção do segundo fluoróforo. Uma molécula de corante repórter é liberada para cada nova molécula sintetizada, e a detecção do corante repórter não extinto fornece a base para a interpretação quantitativa dos dados.

[0082] O TaqMan® RT-PCR pode ser realizado usando equipamento comercialmente disponível. Em uma modalidade preferida, o procedimento de nuclease 5' é executado em um dispositivo de PCR quantitativo em tempo real, como o ABI PRISM 7900® Sequence Detection System®. O sistema amplifica amostras em um

formato de 96 poços em um termociclador. Durante a amplificação, o sinal fluorescente induzido por laser é coletado em tempo real através de cabos de fibra óptica para todos os 96 poços e detectado no CCD. O sistema inclui software para executar o instrumento e para analisar os dados. Os dados do ensaio 5'-Nuclease são inicialmente expressos como Ct, ou o ciclo do limiar. Como discutido acima, os valores de fluorescência são registrados durante cada ciclo e representam a quantidade de produto amplificado até esse ponto na reação de amplificação. O ponto em que o sinal fluorescente é registrado pela primeira vez como estatisticamente significativo é o ciclo do limiar (C_t).

[0083] Para minimizar os erros e o efeito da variação amostra a amostra, o RT-PCR geralmente é realizado usando um padrão interno. O padrão interno ideal é expresso em um nível constante entre os diferentes tecidos e não é afetado pelo tratamento experimental. Os RNAs mais frequentemente utilizados para normalizar os padrões de expressão de gene são os mRNAs para os genes de manutenção gliceraldeído-3-fosfato-desidrogenase (GAPDH) e β-actina.

[0084] A PCR em tempo real é comparável tanto à PCR competitiva quantitativa, onde o concorrente interno de cada sequência-alvo é utilizado para normalização, e com

PCR comparativa quantitativa usando um gene de normalização contido na amostra ou um gene de manutenção para RT-PCR.

[0085] Em outro método de PCR, ou seja, o método de perfil de expressão de genes baseado em Mass ARRAY (Sequenom, Inc., San Diego, CA), após o isolamento do RNA e a transcrição reversa, o cDNA obtido é enriquecido com uma molécula de DNA sintética (concorrente), que corresponde à região de cDNA alvo em todas as posições, exceto uma única base, e serve como um padrão interno. A mistura de cDNA/concorrente é amplificada por PCR e é submetida a um tratamento com enzima fosfatase alcalina (SAP) de camarão pós-PCR,

que resulta na desfosforilação dos nucleotídeos restantes. Após a inativação da fosfatase alcalina, os produtos de PCR do concorrente e do cDNA são submetidos à extensão do iniciador, que gera sinais de massa distintos para os produtos de PCR derivados do concorrente e do cDNA. Após a purificação, esses produtos são dispensados em um arranjo de chips, que é pré-carregada com os componentes necessários para a análise com espectrometria de massa por tempo de voo por ionização por dessorção a laser assistida por arranjo (MALDI-TOF MS). O cDNA presente na reação é então quantificado através da análise das razões das áreas de pico no espectro de massa gerado.

[0086] Ainda outras modalidades de técnicas baseadas em PCR que são conhecidas na técnica e podem ser usadas para criação de perfil de expressão de gene incluem, por exemplo, exibição diferencial, polimorfismo de comprimento de fragmento amplificado (iAFLP) e tecnologia BeadArray™ (Illumina, San Diego, CA) usando o sistema Luminex100 LabMAP disponível comercialmente e múltiplas microesferas codificadas por cores (Luminex Corp., Austin, Tex.) em um teste rápido para expressão de genes; e análise de perfil de expressão de alta cobertura (HiCEP).

D. Microarranjos

[0087] A expressão diferencial de genes também pode ser identificada ou confirmada usando a técnica de microarranjo. Assim, o perfil de expressão de genes associados ao câncer de pulmão pode ser medido em tecido fresco ou embebido em parafina, usando a tecnologia de microarranjo. Neste método, as sequências polinucleotídicas de interesse (incluindo cDNAs e oligonucleotídeos) são plaqueadas ou dispostas em um substrato de microchip. As sequências dispostas são então hibridizadas com sondas de DNA específicas de células ou tecidos de interesse. Assim como nos outros métodos e composições deste documento, a fonte de mRNA é o RNA total isolado do sangue

total dos controles e dos pacientes.

[0088] Em uma modalidade da técnica de microarranjo, insertos amplificados por PCR de clones de cDNA são aplicados a um substrato em um arranjo denso. Em uma modalidade, todas as 559 sequências de nucleotídeos da Tabela III são aplicadas ao substrato. Os genes de microarranjo, imobilizados no microchip, são adequados para hibridação sob condições rigorosas. As sondas de cDNA marcadas com fluorescência podem ser geradas através da incorporação de nucleotídeos fluorescentes por transcrição reversa de RNA extraído de tecidos de interesse. As sondas de cDNA rotuladas aplicadas ao chip hibridizam com especificidade para cada ponto do DNA no arranjo. Após uma lavagem rigorosa para remover sondas não especificamente ligadas, o chip é digitalizado por microscopia confocal a laser ou por outro método de detecção, como uma câmera CCD. A quantificação da hibridização de cada elemento disposto permite avaliar a abundância de mRNA correspondente. Com fluorescência de duas cores, sondas de cDNA marcadas separadamente, geradas a partir de duas fontes de RNA, são hibridizadas em pares ao arranjo. A abundância relativa dos transcritos das duas fontes correspondentes a cada gene especificado é assim determinada simultaneamente. A escala miniaturizada da hibridização proporciona uma avaliação conveniente e rápida do padrão de expressão para um grande número de genes. Foi demonstrado que esses métodos têm a sensibilidade necessária para detectar transcritos raros, que são expressos em poucas cópias por célula, e para detectar de forma reprodutível pelo menos aproximadamente duas vezes as diferenças nos níveis de expressão. A análise de microarranjos pode ser realizada por equipamentos disponíveis comercialmente, seguindo os protocolos do fabricante.

[0089] Outros métodos úteis resumidos pela Patente U.S. 7.081.340, e incorporados por referência aqui incluem Análise Serial de

Expressão de Genes (SAGE) e Sequenciamento de Assinatura Massivamente Paralela (MPSS). Resumidamente, a análise serial da expressão de gene (SAGE) é um método que permite a análise simultânea e quantitativa de um grande número de transcritos genéticos, sem a necessidade de fornecer uma sonda de hibridização individual para cada transcrito. Primeiro, é gerada uma etiqueta de sequência curta (cerca de 10 a 14 pb) que contém informações suficientes para identificar exclusivamente um transcrito, desde que a etiqueta seja obtida de uma posição única dentro de cada transcrito. Então, muitos transcritos são ligados para formar moléculas seriais longas, que podem ser sequenciadas, revelando a identidade das múltiplas etiquetas simultaneamente. O padrão de expressão de qualquer população de transcritos pode ser avaliado quantitativamente, determinando a abundância de marcadores individuais e identificando o gene correspondente a cada marcador. Para mais detalhes, ver, por exemplo, Velculescu et al., Science 270:484487 (1995); and Velculescu et al., Cell 88:243 51 (1997), os quais são aqui incorporados por referência.

[0090] A Análise de Expressão de Genes por Sequenciamento de Assinatura Massivamente Paralela descrita por Brenner et al., Nature Biotechnology 18: 630634 (2000) (que é incorporada aqui por referência), é uma abordagem de sequenciamento que combina sequenciamento de assinatura não baseado em gel com clonagem in vitro de milhões de modelos em microesferas separadas de 5 μm de diâmetro. Primeiro, uma biblioteca de microesferas de modelos de DNA é construída por clonagem in vitro. Isto é seguido pela montagem de uma matriz planar das microesferas contendo modelo em uma célula de fluxo a uma alta densidade (geralmente maior que $3x10^6$ microesferas/cm^2). As extremidades livres dos modelos clonados em cada microesfera são analisadas simultaneamente, usando um método

de sequenciamento de assinatura baseado em fluorescência que não requer separação de fragmentos de DNA. Foi demonstrado que este método fornece simultaneamente e com precisão, em uma única operação, centenas de milhares de sequências de assinatura de genes de uma biblioteca de cDNA de levedura.

E. Imuno-histoquímica

[0091] Os métodos imuno-histoquímicos também são adequados para detectar os níveis de expressão dos produtos de expressão de gene dos genes informativos descritos para uso nos métodos e composições aqui descritos. Anticorpos ou antissoros, preferencialmente antissoros policlonais e mais preferencialmente anticorpos monoclonais ou outros ligantes de ligação a proteínas específicos para cada marcador são usados para detectar expressão. Os anticorpos podem ser detectados marcando diretamente os próprios anticorpos, por exemplo, com marcadores radioativos, marcadores fluorescentes, marcadores de hapteno, como biotina, ou uma enzima como peroxidase de rabanete ou fosfatase alcalina. Alternativamente, o anticorpo primário não marcado é utilizado em conjunto com um anticorpo secundário marcado, compreendendo antissoros, antissoros policlonais ou um anticorpo monoclonal específico para o anticorpo primário. Protocolos e kits para análises imuno-histoquímicas são bem conhecidos na técnica e estão disponíveis comercialmente.

III. COMPOSIÇÕES DA INVENÇÃO

[0092] Os métodos para diagnosticar câncer de pulmão aqui descritos que utilizam perfis de expressão de gene definidos permitem o desenvolvimento de ferramentas de diagnóstico simplificadas para diagnosticar câncer de pulmão, por exemplo, NSCLC vs. nódulo não cancerígeno. Assim, uma composição para diagnosticar câncer de pulmão em um mamífero como descrito aqui pode ser um kit ou um reagente. Por exemplo, uma modalidade de uma composição inclui um

substrato sobre o qual os referidos polinucleotídeos ou oligonucleotídeos ou ligantes ou ligantes são imobilizados. Em outra modalidade, a composição é um kit contendo os 5 ou mais polinucleotídeos ou oligonucleotídeos ou ligantes relevantes, marcadores detectáveis opcionais para o mesmo, substratos de imobilização, substratos opcionais para marcadores enzimáticos, bem como outros itens de laboratório. Ainda em outra modalidade, pelo menos um polinucleotídeo ou oligonucleotídeo ou ligante está associado a um marcador detectável.

[0093] Em uma modalidade, uma composição para diagnosticar câncer de pulmão em um mamífero inclui 7 ou mais conjuntos de sondas iniciadoras de PCR. Cada conjunto de sonda iniciadora amplifica uma sequência polinucleotídica diferente de um produto de expressão de gene de 7 ou mais genes informativos encontrados no sangue do sujeito. Estes genes informativos são selecionados para formar um perfil de expressão de genes ou assinatura que é distinguível entre um sujeito com câncer de pulmão e um sujeito com um nódulo não cancerígeno. Alterações na expressão nos genes no perfil de expressão de genes em relação ao perfil de expressão de genes de referência estão correlacionadas com um câncer de pulmão, como o câncer de pulmão de células não pequenas (NSCLC).

[0094] Em uma modalidade desta composição, os genes informativos são selecionados dentre os genes identificados na Tabela I. Em outra modalidade desta composição, os genes informativos são selecionados dentre os genes identificados na Tabela II. Em outra modalidade desta composição, os genes informativos são selecionados dentre os genes identificados na Tabela III.

[0095] Em outra modalidade desta composição, os genes informativos são selecionados dentre os genes identificados na Tabela IV. Em outra modalidade desta composição, os genes informativos são

selecionados dentre os genes identificados na Tabela IX. Essa coleção de genes é aquela para a qual a expressão do produto de gene é alterada (ou seja, aumentada ou diminuída) versus a mesma expressão do produto de gene no sangue de um controle de referência (ou seja, um paciente com um nódulo não cancerígeno). Em uma modalidade, polinucleotídeo ou oligonucleotídeo ou ligantes, isto é, sondas, são gerados para 7 ou mais genes informativos da Tabela I, Tabela II, Tabela III, Tabela IV e/ou Tabela IX para uso na composição (o CodeSet). Um exemplo de uma tal composição contém sondas para uma porção direcionada dos genes da Tabela I. Em uma outra modalidade, são geradas sondas para todos os 8 genes da Tabela I para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 15 genes da Tabela II para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 15 genes da Tabela III para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 41 genes da Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os 50 genes da Tabela II para uso na composição. Em uma outra modalidade, sondas são geradas para os 50 genes da Tabela III para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 7 genes da Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 15 genes da Tabela I para uso na composição. Em uma outra modalidade, sondas são geradas para os primeiros 50 genes da Tabela IV para uso na composição. Em uma outra modalidade, sondas são geradas para os 3 primeiros genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os 5 primeiros genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os 10 primeiros

genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os 15 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição.

Em outra modalidade, sondas são geradas para os 20 primeiros genes da Tabela II, Tabela III,

[0096] Tabela IV ou Tabela IX para uso na composição. Em uma outra modalidade, sondas são geradas para os 25 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição.

[0097] Em ainda outra modalidade, sondas são geradas para os 30 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em ainda outra modalidade, sondas são geradas para os 35 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em ainda outra modalidade, sondas são geradas para os 40 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em ainda outra modalidade, sondas são geradas para os 45 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em ainda outra modalidade, sondas são geradas para os 50 primeiros genes da Tabela II, Tabela III, Tabela IV ou Tabela IX para uso na composição. Em ainda outra modalidade, sondas são geradas para os primeiros 55, 60, 65, 70, 75, 80, 85, 90, 95 ou 100 genes da Tabela IV para uso na composição. Os genes selecionados das Tabelas não precisam estar em ordem de classificação; antes, qualquer combinação que mostre claramente uma diferença na expressão entre o controle de referência e o paciente doente é útil nessa composição.

[0098] Em uma modalidade das composições descritas acima, o controle de referência é um controle não saudável (NHC) como descrito acima. Em outras modalidades, o controle de referência pode ser qualquer classe de controles, conforme descrito acima em "Definições".

[0099] As composições baseadas nos genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX aqui descritas, opcionalmente associadas a marcadores detectáveis, podem ser apresentadas no formato de um cartão microfluídico, um chip ou câmara ou um kit adaptado para uso com as técnicas Nanostring, PCR, RT-PCR ou Q PCR descritas acima. Em um aspecto, esse formato é um ensaio de diagnóstico usando arranjos de baixa densidade TAQMAN® Quantitative PCR. Em outro aspecto, esse formato é um teste de diagnóstico usando a plataforma Nanostring nCounter.

[00100] Para utilização nas composições acima mencionadas, os iniciadores e sondas de PCR são preferencialmente concebidos com base nas sequências de íntrons presentes no(s) gene(s) a serem amplificados selecionados a partir do perfil de expressão do gene. Sequências alvo exemplificativas são mostradas na Tabela IV. A concepção das sequências de iniciador e de sonda está dentro da técnica, uma vez que o alvo genético específico é selecionado. Os métodos particulares selecionados para o projeto do iniciador e da sonda e as sequências particulares de iniciador e de sonda não são características limitativas dessas composições. Uma explicação pronta das técnicas de projeto de iniciador e de sonda disponíveis para os versados na técnica está resumida na Patente U.S. 7.081.340, com referência a ferramentas publicamente disponíveis, como o software DNA BLAST, o programa Repeat Masker (Baylor College of Medicine), Primer Express (Applied Biosystems); Ensaio de MGB por projeto (Applied Biosystems); Primer3 ((Steve Rozen and Helen J. Skaletsky (2000) Primer3 na WWW para usuários em geral e para programadores de biólogos.

[00101] Em geral, os iniciadores e sondas de PCR ideais utilizados nas composições aqui descritas têm geralmente 17-30 bases de comprimento e contêm cerca de 20-80%, como, por exemplo, cerca de

50-60% de bases de G + C. Temperaturas de fusão entre 50 e 80°C, por exemplo, cerca de 50 a 70°C são tipicamente preferidas.

[00102] Em outro aspecto, uma composição para diagnosticar câncer de pulmão em um mamífero contém uma pluralidade de polinucleotídeos imobilizados em um substrato, em que a pluralidade de sondas genômicas hibridizam com 8 ou mais produtos de expressão de gene de 8 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito, o perfil de expressão de gene compreendendo genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outro aspecto, uma composição para diagnosticar câncer de pulmão em um mamífero contém uma pluralidade de polinucleotídeos imobilizados em um substrato, em que a pluralidade de sondas genômicas hibridizam com 8 ou mais produtos de expressão de gene de 8 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito, o perfil de expressão de gene compreendendo genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, uma composição para diagnosticar câncer de pulmão em um mamífero contém uma pluralidade de polinucleotídeos imobilizados em um substrato, em que a pluralidade de sondas genômicas hibridizam com 15 ou mais produtos de expressão de gene de 15 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito, o perfil de expressão de gene compreendendo genes selecionados da Tabela II, Tabela III, Tabela IV ou Tabela IX. Este tipo de composição baseia-se no reconhecimento dos mesmos perfis de genes descritos acima para as composições Nanostring, mas emprega as técnicas de um arranjo de cDNA. A hibridização dos polinucleotídeos imobilizados na composição para os produtos de expressão de gene presentes no sangue do paciente é empregada para quantificar a expressão dos genes informativos selecionados dentre os genes

identificados na Tabela I, Tabela II, Tabela III, Tabela IV e Tabela IX para gerar um perfil de expressão de gene para o paciente, que é então comparado ao de uma amostra de referência. Conforme descrito acima, dependendo da identificação do perfil (ou seja, o dos genes da Tabela I ou subconjuntos), o dos genes da Tabela II ou subconjuntos, o dos genes da Tabela III ou subconjuntos, o dos genes da Tabela IV ou subconjuntos), esta composição permite o diagnóstico e prognóstico de câncer de pulmão NSCLC. Novamente, a seleção das sequências polinucleotídicas, seu comprimento e marcadores utilizados na composição são determinações de rotina feitas por um versado na técnica, tendo em vista os ensinamentos de quais genes podem formar os perfis de expressão de gene adequados para o diagnóstico e prognóstico de cânceres de pulmão.

[00103] Em ainda outro aspecto, uma composição ou kit útil nos métodos aqui descritos contém uma pluralidade de ligantes que ligam a 7 ou mais produtos de expressão de gene de 7 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito. Em uma outra modalidade, uma composição ou kit útil nos métodos aqui descritos contém uma pluralidade de ligantes que ligam a 8 ou mais produtos de expressão de gene de 8 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito. O perfil de expressão de gene contém os genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX, como descrito acima para as outras composições. Em uma outra modalidade, uma composição ou kit útil nos métodos aqui descritos contém uma pluralidade de ligantes que ligam a 15 ou mais produtos de expressão de gene de 15 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito. O perfil de expressão de gene contém os genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX, como descrito acima para as outras composições. Em uma outra modalidade, uma

composição ou kit útil nos métodos aqui descritos contém uma pluralidade de ligantes que ligam a 50 ou mais produtos de expressão de gene de 50 ou mais genes informativos selecionados de um perfil de expressão de gene no sangue do sujeito. O perfil de expressão de gene contém os genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX, como descrito acima para as outras composições. Esta composição permite a detecção das proteínas expressas pelos genes nas tabelas indicadas. Embora, de preferência, os ligantes sejam anticorpos para as proteínas codificadas pelos genes no perfil, seria evidente para os versados na técnica que várias formas de anticorpos, por exemplo, policlonais, monoclonais, recombinantes, quiméricos, bem como fragmentos e componentes (por exemplo, CDRs, regiões variáveis de cadeia única, etc.) podem ser usados no lugar de anticorpos. Tais ligantes podem ser imobilizados em substratos adequados para contato com o sangue do sujeito e analisados de maneira convencional. Em certas modalidades, os ligantes estão associados a marcadores detectáveis. Estas composições também permitem a detecção de alterações nas proteínas codificadas pelos genes no perfil de expressão de gene daquelas de um perfil de expressão de gene de referência.

[00104] Tais alterações estão correlacionadas com o câncer de pulmão de uma maneira semelhante à das composições contendo PCR e polinucleotídeos descritas acima.

[00105] Para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode, em uma modalidade, incluir pelo menos os primeiros 7 ou 8 dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode, em uma modalidade, incluir pelo menos os primeiros 15 dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade

para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 50 ou mais dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 20 ou mais dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 30 ou mais dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 40 ou mais dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 50 ou mais dos genes informativos da Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 60 ou mais dos genes informativos da Tabela IV. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 70 ou mais dos genes informativos da Tabela IV. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 80 ou mais dos genes informativos da Tabela IV. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir 90 ou mais dos genes informativos da Tabela IV. Em outra modalidade para todas as formas acima de composições de diagnóstico/prognóstico, o perfil de expressão de gene pode incluir todos os 100 dos genes informativos da Tabela IV.

[00106] Estas composições podem ser usadas para diagnosticar cânceres de pulmão, como estágio I ou estágio II do NSCLC. Além disso, essas composições são úteis para fornecer um diagnóstico suplementar ou original em um sujeito com nódulos pulmonares de etiologia desconhecida.

IV. *MÉTODOS DE DIAGNÓSTICO DA INVENÇÃO*

[00107] Todas as composições descritas acima fornecem uma variedade de ferramentas de diagnóstico que permitem uma avaliação não invasiva baseada no sangue do status da doença em um sujeito. O uso dessas composições em testes de diagnóstico, que podem ser acoplados a outros testes de triagem, como radiografia de tórax ou tomografia computadorizada, aumentam a precisão do diagnóstico e/ou direcionam testes adicionais.

[00108] Assim, em um aspecto, é fornecido um método para diagnosticar câncer de pulmão em um mamífero. Este método envolve identificar um perfil de expressão de gene no sangue de um mamífero, preferencialmente humano. Em uma modalidade, o perfil de expressão gênica inclui 7, 8, 15, 41, 50 ou mais produtos de expressão de gene de 7, 8, 15, 41, 50 ou mais genes informativos tendo expressão aumentada ou diminuída no câncer de pulmão. Os perfis de expressão de são formados pela seleção de 7, 8, 15, 41, 50 ou mais genes informativos dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em outra modalidade, o perfil de expressão de gene inclui 7 ou mais produtos de expressão de gene de 7 ou mais genes informativos tendo expressão aumentada ou diminuída no câncer de pulmão. Os perfis de expressão de gene são formados por seleção de 7 ou mais genes informativos dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX. Em uma modalidade, os genes são os primeiros 7 genes da Tabela IV. Em uma outra modalidade, os genes são os primeiros 15 genes da Tabela IV. Em outra modalidade, os genes são os primeiros

50 genes da Tabela IV. Em uma outra modalidade, os perfis de expressão de gene são formados por seleção de 15 ou mais genes informativos dos genes da Tabela II, Tabela III, Tabela III, Tabela IV ou Tabela IX. A comparação do perfil de expressão de gene de um sujeito com um perfil de expressão de gene de referência permite a identificação de alterações na expressão dos genes informativos que se correlacionam com um câncer de pulmão (por exemplo, NSCLC). Este método pode ser realizado usando qualquer uma das composições descritas acima. Em uma modalidade, o método permite o diagnóstico de um tumor cancerígeno a partir de um nódulo benigno.

[00109] Em outro aspecto, o uso de qualquer uma das composições aqui descritas é fornecido para diagnosticar câncer de pulmão em um sujeito.

[00110] As composições e métodos de diagnóstico descritos neste documento fornecem uma variedade de vantagens sobre os métodos de diagnóstico atuais. Entre essas vantagens estão as seguintes. Como aqui exemplificado, os sujeitos com tumores cancerígenos são diferenciados daqueles com nódulos benignos. Esses métodos e composições fornecem uma solução para o problema prático de diagnóstico de se um paciente que se apresenta em uma clínica pulmonar com um pequeno nódulo tem uma doença maligna. Os pacientes com um nódulo de risco intermediário se beneficiariam claramente de um teste não invasivo que levaria o paciente a uma categoria de risco de doença com probabilidade muito baixa ou muito alta. Uma estimativa precisa de malignidade com base em um perfil genômico (ou seja, estimar um determinado paciente com 90% de probabilidade de ter câncer versus estimar que o paciente tem apenas 5% de chance de ter câncer) resultaria em menos cirurgias para doenças benignas, tumores em estágio mais precoce removidos em um estágio curável, menos tomografias de acompanhamento e redução dos

custos psicológicos significativos da preocupação com um nódulo. O impacto econômico provavelmente também seria significativo, como a redução do custo estimado atual de assistência médica adicional associada à triagem por CT para câncer de pulmão, ou seja, US $ 116.000 por ganho de qualidade de vida ajustado. Um teste genômico não invasivo do sangue, com sensibilidade e especificidade suficientes, alteraria significativamente a probabilidade de malignidade pós-teste e, portanto, os cuidados clínicos subsequentes.

[00111] Uma vantagem desejável desses métodos sobre os métodos existentes é que eles são capazes de caracterizar o estado da doença a partir de um procedimento minimamente invasivo, ou seja, colhendo uma amostra de sangue. Em contraste, a prática atual para classificação de tumores de câncer a partir de perfis de expressão de gene depende de uma amostra de tecido, geralmente uma amostra de um tumor. No caso de tumores muito pequenos, a biópsia é problemática e, claramente, se nenhum tumor é conhecido ou visível, uma amostra é impossível. Nenhuma purificação do tumor é necessária, como é o caso quando as amostras de tumor são analisadas. Um método recentemente publicado depende da escovação das células epiteliais do pulmão durante a broncoscopia, um método que também é consideravelmente mais invasivo do que a coleta de uma amostra de sangue. As amostras de sangue têm uma vantagem adicional: o material é facilmente preparado e estabilizado para análises posteriores, o que é importante quando o RNA mensageiro deve ser analisado.

[00112] Os classificadores de genes 7, 8, 15, 41, 50 e 100 descritos aqui tiveram desempenho semelhante a um classificador de marcador 559 anteriormente relatado pelos inventores. Veja exemplos e Tabela VII abaixo. Por exemplo, um classificador consistindo nos 50 primeiros genes da Tabela IV teve desempenho semelhante a um classificador

consistindo em 559 genes. Essas composições e métodos permitem um diagnóstico e tratamento mais precisos do câncer de pulmão. Assim, em uma modalidade, os métodos descritos incluem o tratamento do câncer de pulmão. O tratamento pode remover o crescimento neoplásico, quimioterapia e/ou qualquer outro tratamento conhecido na técnica ou descrito aqui.

[00113] Em uma modalidade, é fornecido um método para diagnosticar a existência ou avaliar um câncer de pulmão em um mamífero, que inclui identificar alterações na expressão de 7, 8, 15, 41, 50 ou mais genes na amostra do referido sujeito, os referidos genes selecionados dos genes da Tabela I, Tabela Ii, Tabela III, Tabela IV ou Tabela IX. Os níveis de expressão de gene do sujeito são comparados com os níveis dos mesmos genes em uma referência ou controle, em que alterações na expressão dos genes do sujeito daqueles da referência se correlacionam com um diagnóstico ou avaliação de um câncer de pulmão.

[00114] Em uma modalidade, o diagnóstico ou avaliação compreendem um ou mais de um diagnóstico de um câncer de pulmão, um diagnóstico de um nódulo benigno, um diagnóstico de um estágio de câncer de pulmão, um diagnóstico de um tipo ou uma classificação de um câncer de pulmão, um diagnóstico ou uma detecção de uma recorrência potencial de um câncer de pulmão, um diagnóstico ou uma detecção de uma regressão de um câncer de pulmão, um prognóstico de um câncer de pulmão ou uma avaliação da resposta de um câncer de pulmão a uma terapia cirúrgica ou não cirúrgica.

[00115] Em uma outra modalidade, as mudanças compreendem uma suprarregulação de um ou mais genes selecionados em comparação com a referida referência ou o referido controle ou uma infrarregulação de um ou mais genes selecionados em comparação com a referida referência ou o referido controle.

[00116] Em outra modalidade, a referência ou controle compreende três ou mais genes da amostra da Tabela I de pelo menos um sujeito de referência. O sujeito de referência pode ser selecionado do grupo que consiste em: (a) um fumante com doença maligna, (b) um fumante com doença não maligna, (c) um ex-fumante com doença não maligna, (d) um não fumante saudável sem doença, (e) um não fumante que tem doença pulmonar obstrutiva crônica (COPD), (f) ex-fumante com COPD, (g) sujeito com um tumor de pulmão sólido antes de cirurgia para remoção do mesmo; (h) um sujeito com um tumor de pulmão sólido seguindo a remoção cirúrgica do referido tumor; (i) um sujeito com um tumor de pulmão sólido antes da terapia para o mesmo; e (j) um sujeito com um tumor de pulmão sólido durante ou em seguida à terapia para o mesmo. Em uma modalidade, o sujeito de referência ou controle (a) - (j) é o mesmo sujeito de teste em um ponto de tempo temporalmente anterior.

[00117] A amostra é selecionada dentre as aqui descritas. Em uma modalidade, a amostra é sangue periférico. Os ácidos nucleicos na amostra são, em algumas modalidades, estabilizados antes da identificação de alterações nos níveis de expressão de gene. Essa estabilização pode ser realizada, por exemplo, usando o sistema Pax Gene, aqui descrito.

[00118] Em uma modalidade, o método de detecção de câncer de pulmão em um paciente inclui a. obter uma amostra do paciente; e b. detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre

o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão.

[00119] Em outra modalidade, o método de diagnóstico de câncer de pulmão em um sujeito inclui a. obter uma amostra de sangue de um sujeito; b. detectar uma mudança na expressão em pelo menos 7 genes selecionados de Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão; e diagnosticar o sujeito com câncer quando são detectadas alterações na expressão dos genes do sujeito em relação aos da referência.

[00120] Em ainda uma outra modalidade, o método inclui a. obter uma amostra de sangue de um sujeito; b. detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão; c. diagnosticar o sujeito com câncer quando são detectadas alterações na expressão dos genes do sujeito em relação aos da referência; e d. remover o crescimento neoplásico.

V. EXEMPLOS

[00121] Agora, a invenção é descrita tendo como referência os

exemplos a seguir. Estes exemplos são fornecidos apenas para fins ilustrativos e a invenção não deve, de maneira alguma, ser interpretada como limitada a esses exemplos, mas deve ser interpretada de modo a abranger toda e qualquer variação que se torne evidente como resultado do ensino aqui fornecido.

EXEMPLO 1: POPULAÇÃO DO PACIENTE - ANÁLISE A

[00122] Para o desenvolvimento do classificador de genes aqui descrito, foram coletadas amostras de sangue e informações clínicas de vários sujeitos, alguns com diagnóstico de câncer de pulmão e outros com diagnóstico de nódulo benigno, conforme identificado na tabela abaixo. As características do paciente são mostradas na Tabela VI abaixo.

[00123] Pacientes diagnosticados com pulmão benigno ou maligno recrutados em 5 hospitais: Christiana Care Health System, Centro Médico Langone da Universidade de Nova York, Hospital da Universidade da Pensilvânia, Roswell Park e Temple University Hospital. Os sujeitos estavam sendo avaliados quanto à presença de câncer de pulmão por LDCT ou radiografia de tórax ou foram diagnosticados incidentalmente com nódulos pulmonares. A população estudada de indivíduos de alto risco tinha > 50 anos de idade e mais de 20 maços/ano de tabagismo.

[00124] A coorte "controle" foi derivada de pacientes com nódulos pulmonares benignos (por exemplo, opacidades em vidro fosco, nódulos únicos, granulomas ou hamartomas). Esses pacientes foram avaliados em clínicas pulmonares ou submetidos a cirurgia torácica por nódulo pulmonar. Todas as amostras foram coletadas antes da cirurgia, biópsia, broncoscopia ou nódulos foram classificadas como não câncer após pelo menos 2 anos de acompanhamento por imageamento, sem alteração detectável no tamanho.

[00125] Como observado abaixo, T1: amostras são executadas em

microarranjo para selecionar genes para a NanoString Platform; VI: conjunto de validação de amostras da NYU usado em papel NanoString; e SI: Conjunto completo de amostras "limpas" não utilizadas em microarranjos para selecionar genes (inclui V1).

Tabela VI

	T1		V1		S1	
	MN	BN	MN	BN	MN	BN
Total	117	105	40	134	182	232
Sexo						
Feminino	65	52	19	65	93	120
Masculino	52	53	14	68	81	111
Desconhecido	0	0	7	1	8	1
Idade	67 ± 7	65 ± 7	68 ± 7	62 ± 7	68 ± 7	62 ± 7
Raça						
Branca	18	11	4	7	20	13
Negra	95	90	22	116	121	191
Outra	4	4	14	11	41	28
Estado de Fumante						
Atual	31	35	6	52	41	84
Anterior	79	65	21	73	122	137
Nunca	7	5	6	8	11	10
Desconhecido	0	0	7	1	8	1
Maços Anos	40 ± 22	39 ± 15	27 ± 27	40 ± 15	39 ± 21	41 ± 15
Local						
HFGCC	59	17	0	0	57	10
FCCC	0	0	0	0	2	1
NYU	21	57	40	134	58	195
Oncocyte	0	0	0	0	22	1
Roswell	0	0	0	0	18	19
Temple	2	9	0	0	0	1
Upenn	35	22	0	0	25	5
Tamanho (mm)	22 ± 8	9 ± 4	15 ± 3	6 ± 2	20 ± 9	6 ± 2

Estágio				
1	86 (75%)	22 (55%)		92 (50%)
2	22 (19%)	7 (18%)		16 (8%)
3	7 (6%)	11 (28%)		49 (26%)
4	0	0		1 (0%)
Desconhecido	2	0		24 (12%)

EXEMPLO 2: PROTOCOLOS E PROCESSAMENTO DE COLETA DE AMOSTRAS

[00126] As amostras de sangue foram coletadas na clínica pelo técnico de aquisição de tecidos. As amostras de sangue foram coletadas diretamente nos tubos de RNA do sangue PAXgene através da técnica padrão de flebotomia. Esses tubos contêm um reagente proprietário que estabiliza imediatamente o RNA intracelular, minimizando a degradação ex-vivo ou a suprarregulação dos transcritos de RNA. A capacidade de eliminar o congelamento de amostras em lotes e de minimizar a urgência de processar amostras após a coleta aumenta significativamente a eficiência do laboratório e reduz os custos.

EXEMPLO 3 - PURIFICAÇÃO DE RNA E AVALIAÇÃO DA QUALIDADE

[00127] O RNA do PAXgene é preparado usando um kit comercialmente disponível da Qiagen™, que permite a purificação de mRNA e miRNA. O RNA total resultante é usado para criação de perfil de mRNA. A qualidade do RNA é determinada usando um Bioanalyzer. Apenas amostras com números de integridade do RNA > 3 foram usadas no nCounter.

[00128] Resumidamente, o RNA é isolado como se segue. Ligue a agitadora-incubadora e ajuste para 55°C antes de começar. Salvo indicação em contrário, todas as etapas deste protocolo, incluindo as etapas de centrifugação, devem ser realizadas à temperatura ambiente (15-25°C). Este protocolo assume que as amostras são armazenadas a -80°C. As amostras descongeladas que tiverem deixado um TR pelo

protocolo Qiagen de no mínimo 2 horas devem ser processadas da mesma maneira.

[00129] Descongele os tubos Paxgene na vertical em um rack de plástico. Inverta os tubos pelo menos 10 vezes para misturar antes de iniciar o isolamento. Prepare todos os tubos necessários. Para cada amostra, são necessários os seguintes: 2 tubos Eppendorf numerados de 1,5 ml; 1 tubo Eppendorf com as informações da amostra (este é o tubo final); 1 coluna de centrifugação lilás Paxgene; 1 coluna Red Paxgene Spin; e 5 tubos de processamento.

[00130] Centrifugue o tubo de RNA sanguíneo PAXgene por 10 minutos a 5000 xg usando um rotor de giro na centrífuga Qiagen. (Centrífuga Sigma 4-15°C, Rotor: Sigma Nr. 11140, 7/01, 5500/min, Suporte: Sigma 13115, 286g 14/D, Suporte do tubo interno: 18010, 125g). Nota: Após o descongelamento, verifique se a amostra de sangue foi incubada no tubo de RNA sanguíneo PAXgene por no mínimo 2 buzinas à temperatura ambiente (15-25°C), para obter uma lise completa das células sanguíneas.

[00131] Sob a cobertura - remova o sobrenadante decantando em água sanitária. Quando o sobrenadante for decantado, tome cuidado para não perturbar o sedimento e seque a borda do tubo com uma toalha de papel limpa. Descarte o sobrenadante decantado colocando o sangue coagulado em um saco e depois no lixo infeccioso, descarte a porção de líquido na pia e lave com muita água. Adicione 4 ml de água sem RNase ao pelete e feche o tubo usando um novo fechamento secundário Hemogard.

[00132] Agite no vortex até o pelete estar visivelmente dissolvido. Pesar os tubos no suporte da centrífuga novamente para garantir que estejam equilibrados e centrifugar por 10 minutos a 5000 xg usando uma centrífuga Qiagen com rotor de giro. Pequenos detritos que permanecem no sobrenadante após o vórtex, mas antes da

centrifugação, não afetarão o procedimento.

[00133] Remova e descarte o sobrenadante inteiro. Deixe o tubo de cabeça para baixo por 1 min. para drenar todo o sobrenadante. A remoção incompleta do sobrenadante inibirá a lise e diluirá o lisado e, portanto, afetará as condições de ligação do RNA à membrana do PAXgene.

[00134] Adicionar 350 µL de Tampão BM1 e pipetar para cima e para baixo para lisar o pelete.

[00135] Pipete a amostra ressuspensa em um tubo de microcentrífuga marcado de 1,5 ml. Adicionar 300 µl de Tampão BM2. Em seguida, adicione 40 µl de proteinase K. Misture em vórtex por 5 segundos e incube por 10 minutos a 55°C usando uma agitadora-incubadora na velocidade mais alta possível, 800 rpm no termomisturador Eppendorf. (Se estiver usando um banho de água com agitação em vez de um termomisturador, agite rapidamente as amostras no vórtex a cada 2-3 minutos durante a incubação. Mantenha o agitador vórtex próximo à incubadora).

[00136] Pipete o lisado diretamente para uma coluna de centrifugação PAXgene Shredder (tubo lilás) colocada em um tubo de processamento de 2 ml e centrifugue por 3 minutos a 24C a 18.500 xg na centrífuga TOMY Microtwin. Pipete cuidadosamente o lisado para a coluna de centrifugação e verifique visualmente se o lisado está completamente transferido para a coluna de centrifugação. Para evitar danos a colunas e tubos, não exceda 20.000 x g.

[00137] Transfira cuidadosamente todo o sobrenadante da fração de fluxo para um tubo de microcentrífuga novo de 1,5 ml sem perturbar o pelete no tubo de processamento. Descarte o pelete no tubo de processamento.

[00138] Adicione 700 µl de isopropanol (100%) ao sobrenadante. Misture em vórtex.

[00139] Pipete 690 µl de amostra na coluna de centrifugação PAXgene RNA (vermelha) colocada em um tubo de processamento de 2 ml e centrifugue por 1 minuto a 10.000x g. Coloque a coluna de centrifugação em um novo tubo de processamento de 2 ml e descarte o tubo de processamento antigo contendo fluxo.

[00140] Pipete a amostra restante para a coluna de centrifugação PAXgene RNA (vermelha) e centrifugue por 1 minuto a 18.500 x g. Coloque a coluna de centrifugação em um novo tubo de processamento de 2 ml e descarte o tubo de processamento antigo contendo fluxo. Pipete cuidadosamente a amostra para a coluna de centrifugação e verifique visualmente se a amostra está completamente transferida para a coluna de centrifugação.

[00141] Pipete 350 µl de tampão BM3 na coluna de centrifugação do RNA do PAXgene. Centrifugue por 15s a 10.000x g. Coloque a coluna de centrifugação em um novo tubo de processamento de 2 ml e descarte o tubo de processamento antigo contendo fluxo.

[00142] Prepare a mistura de incubação DNase I para a etapa 13. Adicionar 10 mL de solução-mãe DNase I a 70 mL de tampão RDD em um tubo de microcentrífuga de 1,5 ml. Misture agitando suavemente o tubo e centrifugue brevemente para coletar o líquido residual das laterais do tubo.

[00143] Pipete a mistura de incubação de DNase I (80 µl) diretamente na membrana da coluna de rotação PAXgene RNA e coloque na bancada (20-30°C) por 15 minutos. Certifique-se de que a mistura de incubação da DNase I seja colocada diretamente na membrana. A digestão com DNase será incompleta se parte da mistura for aplicada e permanecer nas paredes ou no O-ring da coluna de centrifugação.

[00144] Pipete 350 µl de tampão BM3 na coluna de centrifugação do PAXgene RNA e centrifugue por 15s a 18.500 x g. Coloque a coluna de

centrifugação em um novo tubo de processamento de 2 ml e descarte o tubo de processamento antigo contendo fluxo.

[00145] Pipete 500 µl de tampão BM4 na coluna de centrifugação do RNA PAXgene e centrifugue por 15s a 10.000x g. Coloque a coluna de centrifugação em um novo tubo de processamento de 2 ml e descarte o tubo de processamento antigo contendo fluxo.

[00146] Adicione outros 500 µl de tampão BM4 à coluna de centrifugação do PAXgene RNA. Centrifugue por 2 minutos a 18.500 x g.

[00147] Descarte o tubo que contém a passagem e coloque a coluna de rotação do PAXgene RNA em um novo tubo de processamento de 2 ml. Centrifugue por 1 minuto a 18.500 x g.

[00148] Descarte o tubo que contém o fluxo. Coloque a coluna de rotação do PAXgene RNA em um tubo de microcentrífuga marcado de 1,5 ml (tubo final) e pipete 40 µl de tampão BR5 diretamente na membrana da coluna de rotação do PAXgene RNA. Centrifugue por 1 minuto a 10.000 xg para eluir o RNA. É importante umedecer toda a membrana com o Tampão BR5 para obter a máxima eficiência de eluição.

[00149] Repita a etapa de eluição conforme descrito, usando 40 µl de tampão BR5 e o mesmo tubo de microcentrífuga. Centrifugue por 1 minuto a 20.000 xg para eluir o RNA.

[00150] Incubar o eluato por 5 minutos a 65°C na agitadora-incubadora sem agitar. Após a incubação, esfrie imediatamente com gelo. Esta incubação a 65°C desnatura o RNA para aplicações a jusante. Não exceda o tempo ou a temperatura de incubação.

[00151] Se as amostras de RNA não forem usadas imediatamente, armazene a -20°C ou -70°C. Como o RNA permanece desnaturado após repetidos congelamentos e descongelamentos, não é necessário repetir a incubação a 65°C.

EXEMPLO 4: MEDIÇÃO DOS NÍVEIS DE RNA

[00152] Para fornecer uma assinatura de biomarcador que possa ser usada na prática clínica para diagnosticar câncer de pulmão, um perfil de expressão de gene com o menor número de genes que mantém uma precisão satisfatória é fornecido pelo uso de mais 8 genes identificados na Tabela I, bem como pelo uso de 15 ou mais dos genes identificados na Tabela II, Tabela III, Tabela IV ou Tabela IX, ou 7 ou mais genes identificados na Tabela IV. Esses perfis ou assinaturas de genes permitem testes mais simples e práticos, fáceis de usar em um laboratório clínico padrão. Como o número de genes discriminadores é pequeno o suficiente, as plataformas NanoString nCounter® são desenvolvidas usando esses perfis de expressão de gene.

A. PROTOCOLO DE ENSAIO DE EXPRESSÃO DE GENES DA PLATAFORMA NANOSTRING NCOUNTER®

[00153] O RNA total foi isolado do sangue total usando o Kit de miRNA Paxgene Blood, como descrito acima, e as amostras foram verificadas quanto à qualidade do RNA. As amostras foram analisadas com o Agilent 2100 Bioanalyzer em um chip RNA Nano, usando o escore RIN e a imagem do eletroferograma como indicadores para uma boa integridade da amostra. As amostras também foram quantificadas no Nanodrop (espectrofotômetro ND-1000), onde as leituras 260/280 e 260/230 foram registradas e avaliadas quanto à compatibilidade com Nanostring. A partir das concentrações obtidas pelo Nanodrop, as amostras totais de RNA foram normalizadas para conter 100ng em 5µL, usando água livre de Nuclease como diluente, em tiras de tubo fornecidas pelo Nanostring. Uma alíquota de 8µL de uma mistura do Nanostring nCounter Reporter CodeSet e Tampão de Hibridização (70µL de Tampão de Hibridização, 42µL Reporter CodeSet por 12 ensaios) e 2µL de Capture ProbeSet foram adicionados a cada amostra de 5µL de RNA. As amostras foram hibridizadas durante 19 horas a

65°C no Termociclador (Eppendorf). Durante a hibridização, as Sondas Repórteres, que possuem códigos de barras fluorescentes específicos para cada mRNA de interesse do usuário, e as Sondas de Captura biotiniladas ligadas ao mRNA alvo associado, para criar complexos de sonda alvo. Após a conclusão da hibridização, as amostras foram transferidas para a nCounter Prep Station para processamento usando a configuração do protocolo padrão (Tempo de Execução: 2 h e 35 min.). O robô Prep Station, durante o Protocolo Padrão, lavou amostras para remover o excesso de Sondas Repórteres e de Captura. As amostras foram movidas para um cartucho revestido de estreptavidina, onde os complexos de sonda-alvo purificadas foram imobilizados em preparação para imageamento pelo nCounter Digital Analyzer. Após a conclusão, o cartucho foi vedado e colocado no Analisador Digital usando uma configuração de Campo de Visão (FOV) em 555. Um microscópio fluorescente tabulou as contagens brutas para cada código de barras exclusivo associado a um mRNA alvo. Os dados coletados foram armazenados em arquivos .csv e depois transferidos para o Bioinformtics Facility para análise de acordo com as instruções do fabricante.

EXEMPLO 5: SELEÇÃO DE BIOMARCADORES

[00154] A Máquina de Vetor de Suporte (SVM) pode ser aplicada a conjuntos de dados de expressão de genes para descoberta e classificação de funções de genes. Verificou-se que a SVM é mais eficiente em distinguir os casos e controles mais intimamente relacionados que residem nas margens. Primeiramente, a SVM-RFE (48, 54) foi usada para desenvolver classificadores de expressão de genes que distinguem classes clinicamente definidas de pacientes de classes clinicamente definidas de controles (fumantes, não fumantes, DPOC, granuloma, etc.). SVM-RFE é um modelo baseado em SVM utilizado na técnica que remove genes, recursivamente com base em

sua contribuição para a discriminação, entre as duas classes sendo analisadas. Os genes de pontuação mais baixa por pesos coeficientes foram removidos e os genes restantes foram pontuados novamente e o procedimento foi repetido até restarem apenas alguns genes. Este método tem sido utilizado em vários estudos para realizar tarefas de classificação e seleção de genes. No entanto, a escolha de valores apropriados dos parâmetros do algoritmo (parâmetro de penalidade, função do kernel, etc.) geralmente pode influenciar o desempenho.

SVM-RCE é um modelo baseado em SVM relacionado, na medida em que, como SVM-RFE, avalia as contribuições relativas dos genes para o classificador. O SVM-RCE avalia as contribuições de grupos de genes correlatos em vez de genes individuais. Além disso, embora ambos os métodos removam os genes menos importantes a cada etapa, o SVM-RCE pontua e remove agrupamentos de genes, enquanto o SVM-RFE pontua e remove um único ou um pequeno número de genes em cada rodada do algoritmo.

[00155] O método SVM-RCE é brevemente descrito aqui. Genes de baixa expressão (expressão média menor que 2x fundo) foram removidos, realizada a normalização quantílica e, em seguida, arranjos "outlier" cujos valores de expressão mediana diferem em mais de 3 sigma da mediana do conjunto de dados foram removidos. As amostras restantes foram submetidas ao SVM-RCE usando dez repetições de validação cruzada de 10 vezes do algoritmo. Os genes foram reduzidos pelo teste t (aplicado no conjunto de treinamento) para um valor ideal determinado experimentalmente, que produz maior precisão no resultado final. Estes genes iniciais foram agrupados por meios K em grupos de genes correlatos cujo tamanho médio é de 3-5 genes. O escore da classificação de SVM foi realizada em cada agrupamento usando reamostragem de 3 vezes repetida 5 vezes, e os piores agrupamentos de escore eliminados. A precisão é determinada no

conjunto de genes sobreviventes usando os 10% restantes das amostras (conjunto de testes) e os 100 genes com maior pontuação foram registrados. O procedimento foi repetido da etapa de agrupamento para um ponto final de 2 agrupamentos. O painel genético ideal foi considerado o número mínimo de genes que fornece a precisão máxima começando com o gene mais frequentemente selecionado. A identidade dos genes individuais neste painel não é fixa, pois a ordem reflete o número de vezes que um determinado gene foi selecionado nos 100 principais genes informativos e essa ordem está sujeita a alguma variação.

A. Seleção de biomarcadores.

[00156] Os genes que pontuam mais alto (por SVM) na discriminação de tumores cancerígenos de nódulos benignos foram examinados quanto à sua utilidade para testes clínicos. Os fatores considerados incluem: maiores diferenças nos níveis de expressão entre as classes e baixa variabilidade dentro das classes. Ao selecionar biomarcadores para validação, foi feito um esforço para selecionar genes com perfis de expressão distintos para evitar a seleção de genes correlatos e identificar genes com níveis de expressão diferencial que foram robustos por técnicas alternativas, incluindo PCR e/ou imuno-histoquímica.

B. Validação.

Três métodos de validação foram considerados.

[00157] Validação cruzada: para minimizar o ajuste excessivo em um conjunto de dados, foi usada a validação cruzada de K vezes (K geralmente igual a 10), quando o conjunto de dados é dividido em partes K aleatoriamente e partes K-l foram usadas para treinamento e 1 para teste. Assim, para K = 10, o algoritmo foi treinado em uma seleção aleatória de 90% dos pacientes e 90% dos controles e testado nos 10% restantes. Isso foi repetido até que todas as amostras tenham sido

empregadas como sujeitos de teste e o classificador acumulado faça uso de todas as amostras, mas nenhuma amostra é testada usando um conjunto de treinamento do qual faz parte. Para reduzir o impacto da randomização, a separação de K vezes foi realizada M vezes produzindo diferentes combinações de pacientes e controles em cada uma das K vezes de cada vez. Portanto, para conjuntos de dados individuais, rodadas M*K de seleção permutada de conjuntos de treinamento e teste foram usadas para cada conjunto de genes.

[00158] Validação Independente: Para estimar a reprodutibilidade dos dados e a generalidade do classificador, é necessário examinar o classificador que foi construído usando um conjunto de dados e testado usando outro conjunto de dados para estimar o desempenho do classificador. Para estimar o desempenho, a validação no segundo conjunto foi realizada usando o classificador desenvolvido com o conjunto de dados original.

[00159] Validação adicional: Para testar a generalidade de um classificador desenvolvido dessa maneira, ele foi usado para classificar conjuntos independentes de amostras que não foram usadas no desenvolvimento do classificador. As precisões de validação cruzada do classificador permutado e original foram comparadas em conjuntos de testes independentes para confirmar sua validade na classificação de novas amostras.

C. Desempenho do classificador

[00160] O desempenho de cada classificador foi estimado por diferentes métodos e várias medidas de desempenho foram usadas para comparar os classificadores entre si. Essas medições incluem precisão, área sob a curva ROC, sensibilidade, especificidade, taxa positiva verdadeira e taxa negativa verdadeira. Com base nas propriedades exigidas da classificação de interesse, diferentes medições de desempenho podem ser usadas para escolher o

classificador ideal; por exemplo, o classificador a ser usado na triagem de toda a população exigiria melhor especificidade para compensar a pequena (~ 1%) prevalência da doença e, portanto, evitar um grande número de acertos falsos positivos, enquanto um classificador de diagnóstico de pacientes no hospital deve ser mais sensível.

[00161] Para diagnosticar tumores cancerígenos a partir de nódulos benignos, é mais desejável uma sensibilidade maior que uma especificidade, pois os pacientes já estão em alto risco.

EXEMPLO 6: TESTE DOS CLASSIFICADORES

[00162] Todas as amostras de sangue periférico foram coletadas em tubos de estabilização de RNA PAXgene e o RNA foi extraído de acordo com o fabricante. As amostras foram testadas em um Nanostring nCounterTM (como descrito acima) contra um painel personalizado de 559 sondas (Tabela V).

[00163] Para o Classificador 559, 432 foram selecionados com base em dados de microarranjos anteriores, 107 sondas foram selecionadas a partir de estudos Nanostring e 20 eram genes de manutenção. Para o CQ, um padrão Universal de RNA (Agilent) foi incluído em cada lote de 36 amostras testadas. Os valores de expressão da sonda foram normalizados usando os 20 genes de manutenção, bem como os controles positivos e negativos de pico fornecidos pela Nanostring (incluídos no classificador). Os escores Z foram calculados para valores de contagem de sondas e serviram como entrada para um classificador de Máquina de Vetor de Suporte (SVM) usando um núcleo polinomial. O desempenho da classificação foi avaliado por validação cruzada de 10 vezes das amostras.

A. Classificadores

[00164] Como mostrado na Tabela VII abaixo e nas FIGURAS 1A a IB, o classificador 559 desenvolvido em todas as amostras mostrou um ROC-AUC de 0,86 no conjunto de treinamento (FIG. 1A) e 0,85 no

conjunto de teste (FIG. 1B). Com o conjunto de Sensibilidade em 90%, a especificidade é de 62,9% e 55,1%, respectivamente.

[00165] Os dados das amostras do conjunto de treinamento e teste foram analisados pelo W559 na plataforma Nanostring, a fim de identificar o número mínimo de sondas necessárias para manter o desempenho obtido com todo o painel. Utilizou-se o SVM-RFE para seleção da sonda, conforme descrito anteriormente. As amostras foram selecionadas aleatoriamente para conjuntos de treinamento e teste, como mostrado na Tabela VII abaixo. A precisão obtida no conjunto de teste é mostrada na FIG. IB. As curvas ROC para os classificadores são mostradas na FIG. 2.

[00166] Para os primeiros 100 genes da Tabela IV, com uma sensibilidade de 90%, foi mostrada uma especificidade de cerca de 63% para o conjunto de treinamento e cerca de 58% para o conjunto de validação. Para os primeiros 50 genes da Tabela IV, com uma sensibilidade de 90%, foi mostrada uma especificidade de cerca de 58% para o conjunto de treinamento e cerca de 56% para o conjunto de validação. Para os primeiros 15 genes da Tabela IV, com uma sensibilidade de 90%, foi mostrada uma especificidade de cerca de 51% para o conjunto de treinamento e cerca de 46% para o conjunto de validação. Para os primeiros 7 genes da Tabela IV, com uma sensibilidade de 90%, foi mostrada uma especificidade de cerca de 67% para o conjunto de treinamento e cerca de 46% para o conjunto de validação.

Tabela VII

Métrica de Conjunto de Desempenho		Genes N					
		559	300	100	50	15	7
Treinamento	Sensibilidade (SE)	78,6%	77,8%	79,5%	77,8%	85,9%	79,5%
	Especificidade (SP)	72,4%	73,3%	77,1%	76,2%	76,2%	79,0%
	Precisão (ACC)	75,7%	75,7%	78,4%	77,0%	79,7%	79,3%
	ROC-AUC	0,860	0,856	0,873	0,866	0,871	0,851

	(AUC)						
	SP @90% SE	62,9%	61,0%	62,9%	58,1%	51,4%	67,6%
Validação	Sensibilidade (SE)	77,5%	75,0%	80,0%	82,5%	67,5%	77,5%
	Especificidade (SP)	81,9%	82,6%	79,0%	77,5%	83,3%	76,1%
	Precisão (ACC)	80,9%	80,9%	79,2%	78,7%	79,8%	76,4%
	ROC-AUC (AUC)	0,859	0,874	0,852	0,868	0,844	0,834
	SP @90% SE	55,1%	63,0%	58,0%	56,5%	46,4%	46,4%

B. Análise adicional

[00167] Análises adicionais foram realizadas e os classificadores de genes das Tabelas I, II e III foram desenvolvidos.

EXEMPLO 7: PAINEL DE MARCADOR 15

[00168] O câncer de pulmão continua sendo a principal causa de mortes relacionadas ao câncer em todo o mundo, em parte devido à falta de protocolos adequados de detecção precoce e em parte porque os sintomas iniciais comuns são facilmente ignorados. A demonstração de que a triagem de câncer de pulmão por tomografia computadorizada de baixa dose (LDCT) pode reduzir a mortalidade entre fumantes atuais e ex-fumantes (> 50 anos, > 30 maços/ano) (1,2) levou a um aumento significativo na triagem pulmonar por LDCT. Embora o LDCT seja capaz de identificar nódulos significativamente menores do que os raios X convencionais, essa capacidade vem com o desafio subsequente de distinguir a pequena porcentagem daqueles nódulos malignos da maior porcentagem de nódulos benignos (3). O NLST detectou nódulos pulmonares com diâmetro igual ou superior a 4 mm em 40% dos pacientes 56, triados com 96,4% dos exames positivos sendo falsos positivos (4). Isso é particularmente problemático para essa classe de nódulos pulmonares indeterminados (IPNs) que variam em tamanho de 4 a 20 mm (4).

[00169] Embora o IPNS detectado pelo número crescente de exames

de LDCT represente um dilema clínico, eles também oferecem a oportunidade de detectar potencialmente cânceres de pulmão nos estágios iniciais de desenvolvimento. Nos estudos anteriores, demonstrou-se que a PBMC rapidamente purificada (em 2 horas) continha dados de expressão de gene que podiam distinguir de forma exata nódulos pulmonares benignos de malignos com alta precisão (ROC-AUC 0,86) (5). Agora, relatou-se que os tubos de estabilização de RNA no sangue do PAXgene oferecem a possibilidade de coletar amostras em muitos contextos clínicos diferentes, com perfis de RNA sendo capturados no momento da coleta. O RNA do PAXgene é estável à temperatura ambiente por 5 a 7 dias, fornecendo o potencial de ser transferido para uma instalação central de testes, como muitos outros exames de sangue de rotina (6-8). A estabilidade do RNA por > 7 anos entre -20C e -80C é um benefício adicional para grandes estudos. Essa facilidade de coleta e estabilização vem com a ressalva de que o RNA do sangue total conteria informações de uma variedade maior de células sanguíneas, incluindo granulócitos e neutrófilos não incluídos nas amostras de PBMC e que, como resultado, identificar uma assinatura precisa da expressão de gene pode ser mais complexo. Os principais objetivos deste estudo foram: (i) demonstrar se o RNA do PAXgene estabilizado do sangue total pode ser extraído com sucesso de informações de expressão de gene que possam distinguir com precisão os nódulos pulmonares malignos dos benignos do pulmão detectados por LDCT; e (ii) determinar se uma assinatura de diagnóstico desenvolvida em microarranjos de transcriptoma inteiros Illumina pode ser transferida para a plataforma NanoString nCounter, mais robusta e clinicamente estabelecida, aprovada pela FDA para o teste de prognóstico Prosigna™ Breast Cancer (9).

[00170] Agora, relata-se o desenvolvimento bem-sucedido de um classificador de nódulo pulmonar (LNC) baseado em microarranjos

genômicos inteiros Illumina com uma ROC-AUC de 0,847 na validação independente. Descreve-se ainda a transição bem-sucedida desse classificador de microarranjos para a plataforma NanoString nCounter™ com um ROC-AUC demonstrado de 0,846 na validação independente. Esse nLNC tem o potencial de abordar o presente dilema do IPN e o potencial de tratamento excessivo do câncer de pulmão.

[00171] A expressão de gene do sangue total distingue nódulos pulmonares malignos e benignos do pulmão detectados por LDCT

[00172] Amostras de sangue foram coletadas prospectivamente em tubos PAXgene estabilizadores de RNA em 5 sítios clínicos. As amostras eram de pacientes com alto risco de câncer de pulmão (> 50 anos> 20 anos-embalagem), todos com nódulos pulmonares detectados por LDCT ou raio X. Os nódulos foram confirmados como malignos (MN) ou benignos (BN) por broncoscopia, biópsia e/ou ressecção pulmonar ou por pelo menos 2 anos de acompanhamento de LDCT. O RNA da amostra PAXgene purificada (ver métodos) foi testado em microarranjos Illumina H12 v4 para avaliar a viabilidade do uso do RNA PAXgene para desenvolver um classificador de nódulos pulmonares com precisões semelhantes às obtidas com o RNA da PBMC (5). Os dados demográficos dos pacientes utilizados no estudo de microarranjos são mostrados na Tabela VIII. Os pacientes do estudo são principalmente fumantes e ex-fumantes > 50 anos de idade com > 20 maços/ano de histórico de tabagismo. Os cânceres usados para desenvolver os modelos são de estágio inicial, com câncer de estágio I + II, perfazendo 84% da população de teste e 100% dos cânceres no conjunto de validação sendo de Estágio I. Os dados foram analisados usando SVM-RFE, como descrito anteriormente (5, 10).

Tabela VIII. Dados demográficos para amostras no estudo de microarranjo.

	Illumina Conjunto de Treinamento			Illumina Validação		
Categoria	MN	BN	Valor p	MN	BM	Valor p
Total N	131	133		33	18	
Gênero						
Feminino	73	68	0,454	24	11	0,5294
Masculino	58	65		9	7	
Idade	67 ± 7	65 ± 7	0,0291	72 ± 7	64 ± 7	0,0057
Raça						
Negra	19	17		3	2	
Branca	107	108	0,1205	30	15	0,3753
Outras	5	8		0	1	
Estado de Fumante						
Atual	33	47		8	6	
Anterior	90	81	0,1652	23	10	0,4663
Nunca	8	5		2	1	
Desconhecido	0	0		0	1	
Local						
HFGCC	68	22		14	1	
NYU	22	73		15	8	
Temple	2	11	8×10^{-13}	0	1	0,0062
UPenn	39	27		4	8	
Tamanho Lesão, mm	22 ± 8	8 ± 4	1×10^{-13}	17 ± 4	15 ± 4	0,3277
Estágio de Câncer						
I	87 (66%)			33 (100%)		
II	23 (18%)			0		
III	7 (5%)			0		
IV	9 (7%)			0		
Desconhecido	5 (4%)			0		

[00173] Analisa-se um conjunto de 264 amostras de treinamento (Tabela VIII) para identificar os classificadores das sondas de genes que

distinguiram com mais precisão os nódulos pulmonares malignos e benignos e para avaliar as alterações no desempenho da classificação com um número decrescente de sondas de genes usando um método SVM-RFE de validação cruzada de 10 reamostragens de 10 vezes (5, 10). Enquanto a precisão da classificação era estável em uma ampla faixa de números de sondas, a área sob a curva ROC (AUC) diminuiu ligeiramente à medida que as sondas foram eliminadas. O uso de 1000 sondas alcançou uma AUC de 0,88, enquanto 150 produziram uma AUC de 0,85. Até 15 genes foram suficientes para classificar o conjunto de treinamento com uma AUC > 0,8. Seleciona-se o menor número de sondas de genes que mantiveram uma AUC dentro de 1% daquela alcançada pelos 1000 genes e identificou-se um classificador de 311 sondas de genes que produziram uma AUC de 0,866 (Dados não mostrados). Esse desempenho foi mantido quando o classificador foi aplicado a um conjunto de validação independente do paciente (colunas da Tabela VIII à direita) atingindo AUC de 0,847 (Dados não mostrados), confirmando que um sinal robusto associado à presença de um câncer de pulmão poderia ser detectada no RNA do PAXgene com um desempenho geral semelhante aos resultados de PBMC publicados anteriormente.

[00174] Tendo verificado que o RNA das amostras de PAXgene pode distinguir nódulos pulmonares malignos de benignos detectados por imageamento com alta precisão, desenvolveu-se uma estratégia para fazer a transição do nosso LNC baseado em microarranjos para a plataforma NanoString nCounter mais clinicamente apropriada (9).

Transição para a plataforma NanoString

[00175] Para selecionar as sondas para um painel de expressão do gene de diagnóstico NanoString, várias considerações foram levadas em consideração. Sem saber quão bem as medições da expressão do gene NanoString replicariam o desempenho da plataforma de

microarranjos Illumina, projetamos nosso painel personalizado para conter redundância suficiente para poder superar as diferenças da plataforma e incluímos sondas de genes selecionadas por diferentes critérios. SVM-RFE foi o principal método utilizado para a seleção. Embora a SVM seja uma ferramenta poderosa para o desenvolvimento de um modelo de classificação, levamos em conta a possibilidade de que algumas sondas com bom desempenho nos microarranjos não demonstrassem um desempenho equivalente na plataforma NanoString. Para aumentar no lago de genes relevantes, também selecionaram-se 59 sondas com os valores mínimos de p nas comparações (p <10-4) e 76 sondas com uma alteração máxima de dobra em p0,01 na classificação. Também identificaram-se genes candidatos de manutenção que foram expressos em > 5 vezes acima do fundo do microarranjo, com coeficientes de variação menores que 20% para valores absolutos e 2,5% para valores de expressão em escala log2. Os 12 principais candidatos menos variáveis e 8 genes de manutenção NanoString conhecidos adicionais foram selecionados para uma lista final de 20 genes de manutenção.

[00176] Uma das principais diferenças entre as plataformas de microarranjo e NanoString é que nenhum processamento enzimático é necessário para avaliar a expressão de gene no NanoString. Isso significa transcrição reversa e sem amplificação por PCR dos alvos genéticos associados à análise. Esse manuseio minimizado de amostras elimina o potencial de vieses de amplificação e torna a plataforma atraente para uma aplicação clínica, mas havia a possibilidade de que alguns dos biomarcadores selecionados pudessem ser expressos em níveis muito baixos para detecção sem a amplificação por PCR. Para resolver esse problema, analisou-se a maioria das amostras do conjunto de treinamento em microarranjos Illumina no painel NanoString PanCancer Immune (cat. XT-CSO-HIP1-

12), a fim de correlacionar os níveis de expressão de gene que podem ser detectados por cada plataforma. Analisaram-se 220 das amostras de treinamento, incluindo 115 amostras MN e 105 BN. 755 dos 770 genes representados no painel NanoString PanCancer Immune também foram representados na plataforma de microarranjos, embora as sondas genéticas fossem diferentes. Embora as sondas exatas diferissem entre as 2 plataformas, essa análise forneceu uma estimativa dos níveis de expressão de gene que poderiam ser detectados de maneira robusta em ambas as plataformas. Os estudos sugeriram que sondas detectadas em 5X os níveis de fundo de microarranjos foram detectadas de maneira robusta no NanoString.

[00177] Também usamos o estudo do painel PanCancer Immune como uma plataforma de descoberta adicional, pois essas sondas já estavam bem validadas, pois funcionavam bem juntas na plataforma NanoString. Utilizou-se SVM de 10 reamostragens de 10 vezes para estimar a precisão do desempenho do painel do painel NanoString PanCancer Immune. Embora o painel NanoString tenha demonstrado desempenho inferior com uma AUC = 0,754 em comparação com 0,866 obtido com o conjunto de treinamento de microarranjos, com 90% de sensibilidade, a especificidade ainda era de 37% e selecionou-se um conjunto de 106 sondas para inclusão em nosso painel personalizado. Também adicionam-se 55 sondas adicionais identificadas em nossos estudos PBMC anteriores como associadas ao diagnóstico, resultado e/ou alterações na expressão de gene pós-ressecção (11, 13) para consideração em estudos futuros.

[00178] Para avaliar a semelhança entre a classificação usando as plataformas Illumina e NanoString, 222 amostras de conjuntos de treinamento que foram executadas em ambas as plataformas usando as 276 sondas SVM-RFE mais uma vez, levando em consideração que as diferenças das sondas e as diferenças nas 2 tecnologias da

plataforma. Considerando apenas as 276 sondas SVM-RFE, compararam-se as pontuações SVM de validação cruzada dessas amostras usando ambas as plataformas, e observa-se uma correlação de Spearman de 0,72 entre plataformas que consideraram-se suficientes para prosseguir com a avaliação de novas amostras. Testamos 414 amostras no painel NanoString personalizado da sonda 559. Utilizaram-se 142 amostras com lesões de 8 a 20 mm, que são a faixa que compõe a população de interesse do IPN. As 272 amostras restantes foram usadas para treinar o modelo. O modelo desenvolvido neste conjunto de treinamento demonstrou uma AUC de 0,813 e alcançou uma AUC de 0,820 no conjunto de validação (Dados não mostrados). Isso confirmou que as abordagens de seleção de genes foram eficazes.

[00179] Refina-se ainda mais o modelo usando a eliminação recursiva de recurso (RFE). A aplicação de SVM-RFE ao conjunto de treinamento revelou que o desempenho do classificador é estável até 15 sondas NanoString, com o nLNC retornando uma AUC de treinamento = 0,867 (Figura 2). As AUCs determinadas no conjunto de validação independente são igualmente estáveis com uma AUC = 0,843 para 15 genes, embora a variabilidade da AUC aumente e a especificidade a 90% de sensibilidade comece a cair levemente quando menos de 50 sondas forem usadas na classificação. Em uma modalidade, os 15 ou 50 genes são selecionados dentre os da Tabela II ou Tabela III.

[00180] Os estudos anteriores sobre a expressão de gene no RNA do PBMC coletados em condições padronizadas demonstraram que o sangue periférico não apenas contém informações para distinguir nódulos pulmonares malignos dos benignos do pulmão, mas também informações correlacionadas com o prognóstico. O presente estudo abordou 2 objetivos 1) a simplificação da coleta de amostras e 2) mover

o ensaio para uma plataforma clinicamente acessível. Ao abordar o primeiro objetivo, implementou-se um método simples e facilmente padronizado para a coleta de amostras usando os tubos de coleta PAXgene. Isso permitiu a coleta de amostras não apenas em centros universitários, mas também em centros comunitários que não são facilmente incluídos em estudos anteriores. Este estudo piloto estabelece a viabilidade do uso de RNA de sangue total coletado em tubos PAXgene para distinguir com precisão pacientes com nódulos pulmonares malignos do grande número de indivíduos com nódulos pulmonares benignos detectados pelo crescente número de programas de triagem. A facilidade da coleta de amostras do PAXgene nos permitiu realizar os grandes estudos necessários para abordar a diversidade do câncer de pulmão. Neste estudo, concentramo-nos principalmente em NSCLC, o câncer de pulmão mais comum, câncer de estágio inicial menor e fumantes e ex-fumantes de alto risco.

[00181] O segundo objetivo que tinha sido definido era mover o LNC da plataforma de descoberta de microarranjos para uma plataforma de detecção mais clinicamente apropriada. A plataforma NanoString nCounter foi escolhida com base na facilidade de manipulação de amostras e na eliminação de muitas das etapas enzimáticas conhecidas por resultar em vieses e efeitos de lote que encontraram-se no estudo de PBMC. Além disso, a quantidade de amostra de teste necessária é facilmente obtida. Usamos 100 ng de RNA para todos os ensaios de NanoString aqui relatados, mas também é possível o mínimo de 10 ng. Descobriu-se que, embora grandes quantidades de RNA sejam compatíveis com a plataforma, o uso da configuração de varredura de sensibilidade mais alta do FOV 550 foi mais eficaz em aumentar o sinal do que aumentar a quantidade de RNA.

[00182] Também testou-se a estabilidade do painel genético do NanoString em relação aos lotes de reagentes, alterações de pessoal e

fontes de amostras sem detectar efeitos significativos dos lotes.

[00183] Um padrão universal de RNA humano foi inicialmente incluído em cada cassete de 12 amostras. Isso foi reduzido para um por semana para avaliar o desempenho do scanner e do teste. Embora exigiram-se Números de Integridade do RNA (RINs) > 7,5 para nossos estudos de microarranjo, e embora os RINs do PAXgene sejam rotineiramente > 8,0, também foram capazes de analisar o RNA de amostras de PAXgene que foram descongeladas e recongeladas em um defeito no congelador a -80C. Foram avaliadas 5 amostras de RNA com RINs tão baixos quanto 3,2. As intensidades gerais do sinal estavam dentro dos parâmetros de controle de qualidade NanoString. Além disso, no caso de problemas de digitalização, os cassetes podem ser digitalizados novamente dentro de 24 horas, com pouca ou nenhuma perda de informações quando armazenadas, como recomenda a NanoString.

EXEMPLO 8: PAINEL DE MARCADOR 41 (TABELA IX)

[00184] Com o tabagismo como causa raiz reconhecida, o câncer de pulmão continua sendo a principal fonte de mortes relacionadas ao câncer em todo o mundo. Isso se deve em parte à falta de protocolos adequados de detecção precoce e, em parte, porque os sintomas iniciais são muito sutis. A demonstração de que a triagem de câncer de pulmão por tomografia computadorizada de baixa dose (LDCT) reduz a mortalidade entre fumantes atuais e ex-fumantes (> 55 anos, > 30 maços/ano) (1-3) levou a um aumento geral nos programas de triagem LDCT (4). Embora o LDCT identifique nódulos significativamente menores que as radiografias convencionais, essa capacidade vem com o desafio de distinguir a pequena porcentagem de nódulos pulmonares malignos da maioria dos detectados que são benignos (3). O National Lung Screening Trial (NLST) detectou nódulos pulmonares com diâmetro ≥4 mm em 40% dos pacientes examinados, sendo 96,4%

falso-positivos nas 3 rodadas de triagem (6). Para reduzir esse alto FPR, a recente classificação de Lung-RADs (7) e as novas diretrizes do grupo Fleischner (8) definem a detecção de nódulos ≥6mm como limiar positivo. No entanto, a tomografia computadorizada positiva permanece particularmente problemática para essa classe de nódulos pulmonares indeterminados (IPNs), que variam em tamanho de 6 a 20 mm, para os quais o melhor curso de ação clínica não está bem especificado (6).

[00185] Os estudos anteriores demonstraram que células mononucleares do sangue periférico (PBMC) rapidamente purificadas (dentro de 2 horas) contêm dados de expressão de gene que podem distinguir nódulos pulmonares benignos de malignos com alta precisão (9). Este trabalho estabeleceu um novo paradigma no diagnóstico de nódulos, mostrando que mesmo um câncer em estágio inicial no pulmão afeta a expressão de gene no PBMC, que é preditivo de malignidade. No entanto, essa abordagem foi limitada pela necessidade de purificar rapidamente PBMCs de amostras de sangue, a fim de manter a consistência da amostra e a integridade do RNA. Isso dificultou a coleta de amostras em ambientes onde o rápido isolamento do PBMC não era possível, incluindo a maioria das clínicas comunitárias e consultórios médicos. Além disso, os microarranjos, tão úteis para o desenvolvimento do diagnóstico, são tecnicamente complicados e propensos a variabilidades associadas a lotes de reagentes e processos enzimáticos, tornando-os menos passíveis de aplicações clínicas. A alta qualidade do RNA necessário para estudos de microarranjo também é potencialmente problemática para estudos com amostras derivadas de pacientes (10). O presente estudo retrospectivo/prospectivo procurou determinar se precisões semelhantes às obtidas em nossos estudos de PBMC (9) poderiam ser alcançadas com RNA de sangue total coletado em tubos PAXgene™ estabilizadores de RNA. O RNA do PAXgene é estabilizado no momento da coleta, fixando imediatamente os padrões

de expressão de gene. O RNA é estável a 15-25°C por 5 dias e a -20 a -70°C por 8 anos. Isso permite que as amostras sejam coletadas em qualquer ambiente clínico em que o sangue seja coletado sem a necessidade de equipamento especial para armazenamento ou purificação de células (11-13) e permite que as amostras sejam transferidas para uma instalação central para testes, tão rotineiramente quanto com outros testes de sangue. Além disso, o armazenamento a longo prazo sem perda da integridade do RNA torna o sistema adequado para análises retrospectivas. Também perguntou-se se uma assinatura PAXgene desenvolvida em microarranjos Illumina poderia ser transferida para a plataforma NanoString nCounter já aprovada pela FDA para o teste de prognóstico Prosigna™ para câncer de mama (14) e mais recentemente usada para desenvolver um teste de nível clínico que prediz resposta clínica ao bloqueio de ponto de verificação de PD-1. Atualmente, esse ensaio PD-1 está sendo avaliado em ensaios clínicos em andamento com pembrolizumabe (15). Como os ensaios de NanoString não incluem reações enzimáticas ou etapas de amplificação, o sistema evita possíveis efeitos de lote de reagentes e vieses de PCR, enquanto diminui as oportunidades de contaminação cruzada, minimizando o manuseio da amostra. Embora tenha reconhecido que um perfil de expressão de gene do sangue total seria de maior complexidade e poderia resultar potencialmente em uma redução de importantes sinais de diagnóstico, havia também a perspectiva de que tipos de células adicionais importantes pudessem contribuir para o desempenho do classificador.

[00186] Atualmente, relatou-se que a expressão de gene no sangue total, coletada usando tubos de estabilização de RNA PAXgene, pode distinguir nódulos pulmonares benignos de malignos detectados por LDCT com alta precisão na validação independente e também relatou-se a transição bem-sucedida desse classificador de nódulos

pulmonares da plataforma de desenvolvimento de microarranjos para a plataforma NanoString nCounter™.

Projeto do estudo

[00187] O processo de seleção e validação de biomarcadores em todos os estudos está resumido na Fig. 3. Um total de 821 amostras de pacientes com nódulos pulmonares malignos (MN) e benignos (BN) foram analisadas em 3 plataformas: microarranjos Illumina, painel NanoString Pan Cancer Immune (PCI) e finalmente um painel personalizado NanoString personalizado., Dados de Microarranjo de 264 amostras de pacientes (Tabela VIII, Tabela XIII) de 4 locais clínicos foram usadas para o desenvolvimento do modelo de microarranjo. As estimativas de desempenho foram baseadas em um conjunto de validação independente de 51 amostras. Além disso, 220 amostras, incluindo 201 das 264 amostras de microarranjos, foram analisadas na plataforma NanoString PCI para selecionar biomarcadores adicionais a serem incluídos no painel NanoString personalizado. Amostras de um 5º local de coleta não incluídas no processo de seleção de biomarcadores foram analisadas apenas na plataforma NanoString personalizada. O nPNC final foi desenvolvido com base nos dados gerados no painel Nanostring personalizado usando 583 amostras de treinamento (incluindo 215 amostras usadas originalmente no conjunto de treinamento de microarranjos e 368 amostras (70%) nunca usadas para a seleção de biomarcadores) e validadas usando um conjunto de 158 amostras independentes nunca envolvidas na seleção da sonda.

Tabela XIII

	Illumina Conjunto de treinamento Completo		Illumina Treinamento Subconjunto A		Illumina Treinamento Subconjunto B		Ilumina Validação	
Categoria	MN	BN	MN	BN	MN	BN	MN	BN
Total N	143	140	71	73	60	60	33	18

Gênero								
Masculino	81	70	38	38	35	30	24	11
Feminino	62	70	23	35	25	20	5	7
Desconhecido	0	0	0	0	0	0	0	0
Idade	67 ± 7	65 ± 7	66 ± 7	64 ± 7	67 ± 7	66 ± 9	72 ± 7	64 ± 7
Raça								
Negra	20	17	14	5	5	12	3	2
Branca	117	120	54	67	53	46	30	15
Outras	6	3	3	1	2	2	0	1
Fumante								
Atual	35	51	29	28	13	19	8	6
Anterior	99	34	48	42	42	39	23	10
Nunca	3	5	3	3	5	2	2	1
Desconhecido	0	0	0	0	0	0	1	1
Maços Anos	40 ± 22	38 ± 14	39 ± 17	38 ± 14	40 ± 27	38 ± 15	36 ± 19	41 ± 21
Lesão, mm	22 ± 9	8 ± 6	22 ± 8	8 ± 2	25 ± 12	14 ± 6	17 ± 4	15 ± 4
Estágio do câncer								
I	98 (69%)		55 (77%)		32 (53%)		33 (100%)	
II	24 (17%)		16 (23%)		7 (12%)		0	
III	7 (5%)		0		7 (12%)		0	
IV	9 (6%)		0		9 (13%)		0	
Desconhecido	5 (3%)		0		5 (8%)		0	
Local Clínica								
HFGCC	70	23	35	11	33	11	14	1
NYU	29	79	11	44	11	29	15	8
Temple	2	11	2	3	0	8	0	1
UPenn	42	27	23	15	16	12	4	8

Mediano +/- IQR são dados para valores contínuos, p-valores indicam significância de comparação entre grupo BN ou MN

População do estudo

[00188] As amostras foram coletadas prospectivamente de sujeitos

incidentais com um LDCT positivo em 5 locais clínicos, incluindo o Helen F. Graham Cancer Center, o Hospital da Universidade da Pensilvânia, o Roswell Park Comprehensive Cancer Center, o Temple University Hospital e os sujeitos do New York University Langone Medical Center. Os sujeitos da NYU incluíram pacientes recrutados como parte de um programa de triagem pulmonar da EDRN na NYU. O estudo foi aprovado pelo IRB em cada local participante e conduzido de acordo com os princípios expressos na Declaração de Helsinque. Todos os participantes assinaram um termo de consentimento livre e esclarecido antes de serem inscritos. A população do estudo era principalmente fumantes e ex-fumantes, com idade > 50 anos, com > 20 maços/ano de histórico de tabagismo e sem câncer prévio nos últimos 5 anos (exceto para câncer de pele não melanoma). Os nódulos foram confirmados como malignos (MN) ou benignos (BN) por imageamento repetido ou por diagnóstico patológico por broncoscopia, biópsia e/ou ressecção pulmonar. Além disso, > 97% dos nódulos benignos tiveram quatro ou mais anos de acompanhamento e o restante, dois ou mais anos no momento da análise. As amostras associadas aos MNs foram coletadas dentro de 3 meses após o diagnóstico definitivo ou antes de qualquer procedimento invasivo, incluindo cirurgia curativa. Descobriu-se que um pequeno número de participantes nunca foi fumante após o teste. O efeito no desempenho do classificador ao incluir essas amostras foi avaliado. Nos casos em que múltiplos nódulos estavam presentes, foi relatado o diâmetro do maior nódulo.

Purificação de RNA, Avaliação da Qualidade e Microarranjos

[00189] Cada local de coleta foi fornecido com um protocolo padrão para coleta e armazenamento de amostras, conforme especificado por Preanalytix (https://www.preanalytix.com/products/blood/rna para o tubo de RNA do sangue PAXgene (IVD)). As amostras foram armazenadas no local e depois transferidas a granel durante a noite em

gelo seco, ou foram transferidas para a Wistar por correio no dia da coleta e armazenadas a -70C até o processamento. O RNA total foi isolado usando o kit PAXgene miRNA (Qiagen), para capturar miRNAs e mRNAs. As amostras foram quantificadas com o espectrofotômetro NanoDrop 1000 (Thermo Fisher Scientific) e analisadas quanto à integridade do RNA no Agilent 2100 BioAnalyzer. Os rendimentos médios de RNA foram de 3 µg/2,5 ml de sangue e, em média, os números de RIN são > 8. Apenas amostras com números de integridade do RNA (RIN) > 7,5 foram usadas para os estudos de microarranjo. Uma quantidade constante (100 ng) de RNA total foi amplificada (aRNA) usando o kit de amplificação de RNA aprovado pela Illumina (Epicenter) e hibridada com os arranjos humanos de esferas genômicas inteiras Human-HT12 v4. Os microarranjos foram processados em conjuntos de 48 para minimizar os potenciais efeitos de lote.

Condições de ensaio NanoString

[00190] A hibridização NanoString foi realizada por constantes 19 (dentro das 12 a 25 horas recomendadas) horas a 65°C. O processamento pós-hibridação na Estação de Preparação nCounter usou as configurações padrão. O parâmetro de varredura do cassete foi definido como alto (555 FOV). A configuração 555 FOV aumenta significativamente o sinal geral. (Dados não mostrados). Todos os ensaios foram realizados usando esta configuração. O tamanho padrão da amostra era de 100 ng, uma quantidade que esperávamos ter em todas as amostras. Encontramos estabilidade do ensaio em várias repetições da amostra de controle de RNA humano universal (UHR). (Dados não mostrados). Variações inferiores a 5% foram observadas para a maioria das sondas de genes com 50 ou mais contagens detectadas. Verificou-se que, embora a maioria dos números de RIN da amostra estivesse acima de 8, mesmo as amostras com RINs ≤ 3 atendiam a todas as 4 medidas de qualidade da NanoString, suportando

o utilitário de plataformas com amostras de RNA degradadas. Não encontramos impacto significativo nos perfis gerais de expressão do RNA degradado. (Dados não mostrados).

Análise dos Dados

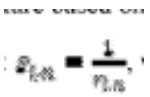
[00191] Microarranjos: Os dados de expressão bruta de microarranjo foram exportados para análise usando o software Genome Studio. Os dados brutos foram normalizados em quantis e em escala $\log_2$-. Genes com valores médios de expressão ≥2X os níveis de fundo foram usados para desenvolver o PNC usando SVM-RFE e validação cruzada de 10 reamostragens de 10 vezes (veja detalhes abaixo). As sondas mais bem classificadas (contagem de Borda) que mais distinguiram os nódulos malignos de benignos foram selecionadas como candidatas ao painel personalizado do NanoString. O conjunto de treinamento SVM também foi estratificado no subconjunto A, que contém nódulos menores e cânceres nos estágios I e II, e no subconjunto B, que continha nódulos malignos e benignos, balanceados para o tamanho da lesão (Tabela Suplementar SI). As análises adicionais dos conjuntos A e B consideraram classe de nódulos isoladamente (MN ou BN) ou classe de amostra mais local de coleta como fatores em um modelo de regressão linear para cada expressão de gene observada. Isso resultou em 6 modelos de regressão diferentes e dois conjuntos adicionais de genes foram selecionados nessas análises para inclusão no modelo NanoString com base nos seguintes parâmetros: (i) 59 genes com um valor p mínimo nas comparações usando o limiar de valor $p < 10^{-4}$ e (ii) 76 genes com um coeficiente de regressão máximo $b > \log_2(1,2)$ no valor de $p < 0,01$. Os genes de manutenção (HK) foram selecionados a partir de um pool candidato de genes bem expressos (> 5X background) com coeficientes de variação (CV) para a expressão absoluta e log2 em escala menor que 20% e 2,5%, respectivamente. Finalmente, 20 genes candidatos HK foram selecionados: os 12 genes candidatos HK com

menos CV e 8 genes candidatos que se sobrepuseram às sondas NanoString HK existentes.

[00192] NanoString: A correção de fundo foi realizada nas amostras do NanoString PanCancer Immune Panel subtraindo a média geométrica das contagens de controles negativos. As contagens da amostra foram normalizadas escalonando todos os valores pelas razões da média geométrica dos controles da amostra para a média geométrica geral das contagens dos genes de controle em todas as amostras. Isso foi feito tanto para controles positivos de pico quanto para genes de manutenção. O painel personalizado do NanoString foi quantilizado e as diferenças de lote do conjunto de códigos NanoString foram corrigidas usando as razões de expressão das amostras replicadas entre os conjuntos de códigos, conforme recomendação da NanoString. Os escores Z foram calculados a partir dos valores finais das contagens personalizadas do painel e usados como entradas para SVM-RFE.

[00193] Análise de dados SVM-RFE: A classificação supervisionada usando Máquina de Vetor de Suporte de núcleo linear (SVM) com eliminação recursiva de recursos (RFE) (16) foi usada para analisar um conjunto de dados de expressão de gene transformada com escore z para desenvolver o classificador de microarranjos com base em um conjunto de treinamento que possa distinguir as classes de pacientes MN e BN. Um conjunto equilibrado de casos e controles foi usado no desenvolvimento de classificadores, já que foi demonstrado que a SVM exige uma entrada equilibrada para o desenvolvimento dos classificadores mais precisos (17). A validação independente que testa a validade do classificador desenvolvido no treinamento em um conjunto de amostras completamente novo é cega para a identificação de amostras como casos ou controles. Como descrito anteriormente (9), empregou-se uma abordagem de validação cruzada de 10 vezes, com vezes reamostradas 10 vezes (100 modelos de divisão de treinamento

- teste). Para cada divisão dos dados de microarranjo, as 1000 sondas principais classificadas por valor-p (teste t bicaudal em 9 vezes) foram selecionadas e a SVM de núcleo linear foi treinada em 9 vezes e testada na dobra restante. Cada iteração de RFE eliminou 10% dos recursos com o menor peso absoluto do modelo, em cada rodada, conforme descrito por Guy on et al (16). Uma única eliminação de recurso por iteração SVM foi usada para análise de dados do NanoString. Os escores médios finais foram calculados da seguinte forma. A pontuação final para qualquer amostra em um conjunto de treinamento é calculada como uma média entre as pontuações geradas para essa amostra em todas as vezes de teste (10 dessas vezes entre todas as 100 divisões). A pontuação final para qualquer amostra independente em um conjunto de validação é calculada como uma média entre todos os 100 modelos de divisão. Cada amostra é então atribuída a uma classe usando os escores médios finais e um limiar de pontuação determinado a partir do conjunto de treinamento (0 para precisão imparcial, ou em um limiar fixo correspondente a 90% de sensibilidade) e a sensibilidade, especificidade e precisão foram calculadas. As sondas classificadas em todas as 100 divisões foram combinadas de acordo com o procedimento a seguir, com base no método de contagem de Borda. Em cada lista classificada n, a cada gene i foi atribuída uma pontuação: onde $r_{i,n}$ é a classificação do gene i na lista n. Um escore final FS para cada gene i foi calculado pela soma dos escores do gene i em todas as 100 listas:



Os escores finais resultantes para cada gene foram usados para atribuir sua classificação no classificador. A classificação final das sondas foi produzida usando todas as 741 amostras NanoString disponíveis.

[00194] O número ideal de sondas para dados de microarranjos foi determinado como o número mínimo de sondas que mantinham uma ROC-AUC dentro de 1% da ROC-AUC alcançada pelo SVM com as 1000 principais sondas genéticas. Para o painel personalizado NanoString, o número ideal de genes foi escolhido determinando o ponto em que a remoção de genes/sondas adicionais resultou em um declínio no desempenho da classificação. O desempenho foi avaliado pela determinação da ROC-AUC após a remoção de cada gene usando a média móvel com um tamanho de janela de suavização de 5. O número da sonda em que o ROC-AUC estava no máximo foi selecionado como o classificador ideal final.

Resultados

Teste para uma assinatura genética relacionada ao câncer de pulmão usando RNA de sangue periférico

[00195] As características demográficas dos 315 pacientes utilizados para desenvolver e validar a assinatura do câncer de pulmão por microarranjo são mostradas na Tabela VIII. As amostras utilizadas para a construção do modelo foram principalmente o NSCLC em estágio inicial, com cânceres em Estágio I + II, compreendendo 84% da população do conjunto de treinamento e 100% dos cânceres no conjunto de validação independente sendo o Estágio I.

[00196] A expressão de gene de 264 amostras (Tabela VIII, conjunto de Treinamento Illumina) foi usada para selecionar a assinatura genética do microarranjo que distinguia com mais precisão os nódulos pulmonares malignos dos benignos do pulmão usando SVM-RFE (9, 16). A precisão da classificação foi estável em uma ampla faixa de números de sondas (Fig. 7A) e um painel das 1000 sondas com classificação mais alta alcançou uma ROC-AUC de 0,878. Como o desempenho diminuiu lentamente com a eliminação de sondas de classificação mais baixa, selecionou-se o menor número de sondas que

mantinham uma ROC-AUC dentro de 1% dos 0,878 alcançados pelas 1000 sondas genéticas SVM superiores. Identificaram-se 311 sondas que retornaram uma AUC de 0,866 (IC 95%: 0,824-0,910) no conjunto de treinamento (sensibilidade 77,9%, especificidade 74,4%). O desempenho foi bem mantido na validação independente (Tabela VIII, Validação Illumina), atingindo uma AUC de 0,847 (IC 95%: 0,742-0,951) (sensibilidade 72,7%, especificidade 88,9%), semelhante ao desempenho do conjunto de treinamento (Fig. 7B-C). Essa precisão de previsão demonstrada é semelhante à do classificador de 29 genes relatado em nosso estudo purificado de PBMC (9) e indica que a presença de câncer no pulmão também pode ser detectada no RNA do sangue coletado do PAXgene com um valor igual e, em alguns casos, melhor desempenho. É importante ressaltar que o desempenho é mantido à medida que o número de sondas é reduzido (Fig. 7A-B), indicando que uma assinatura robusta é mantida em diferentes números de genes.

Transição do PNC de microarranjos para a plataforma NanoString

[00197] Tendo verificado que a expressão de mRNA de amostras de PAXgene pode distinguir nódulos pulmonares malignos de benignos, desenvolveu-se uma estratégia para fazer a transição do classificador de nódulos pulmonares (PNC) baseado em microarranjo para a plataforma NanoString nCounter (14). Como era difícil saber, a priori, como as medidas de expressão de microarranjos se replicariam na plataforma NanoString, projetamos o painel personalizado para conter redundância suficiente para reduzir as diferenças de plataforma. Incluímos os 300 biomarcadores mais bem classificados do painel genético da Illumina identificados pela análise SVM-RFE. Também foi incluído um conjunto adicional de 59 marcadores que representam as sondas expressas diferencialmente mais significativamente com valor de $p < 10^{-4}$ e 79 sondas que exibiram a maior alteração na expressão

entre os grupos MN e BN, mantendo o valor de p 0,01. Também foi adicionado um conjunto de 20 genes de manutenção (HK) com a expressão mais consistente nos microarranjos (consulte Métodos e Materiais). Como as amostras de mRNA a serem processadas na plataforma NanoString não foram submetidas à transcrição reversa e à amplificação por PCR, estávamos preocupados com o fato de que algumas das sondas de microarranjo selecionadas pudessem ser expressas em níveis muito baixos para detecção sem amplificação. Para estabelecer critérios de desempenho, analisou-se 220 das amostras de mRNA com o painel NanoString PanCancer Immune (PCI) (cat. XT-CSO-HIP1-12) (115 MN, 105 BN). Embora as sondas reais não fossem idênticas às sondas Illumina, 755 dos 770 genes representados no painel PCI também foram representados na plataforma de microarranjos. Este estudo nos permitiu correlacionar níveis detectáveis de expressão de gene entre as 2 plataformas, fornecendo uma estimativa dos níveis de expressão que poderiam ser detectados de maneira robusta em ambas as plataformas. Os resultados sugeriram que as sondas detectadas em 2X os níveis de segundo plano em microarranjos foram robustamente detectadas na plataforma NanoString.

[00198] Os dados de PCI também foram analisados usando SVM-RFE com validação cruzada de 10 vezes e 10 amostras e, embora o painel PCI tenha demonstrado apenas uma ROC-AUC = 0,754 em comparação aos 0,866 alcançados com os microarranjos (Fig. 7D), selecionaram-se 106 das sondas PCI mais discriminatórias para inclusão em nosso painel personalizado. Também foram adicionadas 55 sondas para genes identificados como associados ao resultado nos estudos de microarranjos de PBMC (18, 19), aumentando o número para 619 sondas em potencial para o painel personalizado NanoString. A Fig. 12 resume as fontes da lista final dos biomarcadores candidatos

selecionados para o painel NanoString personalizado. As sondas NanoString foram então projetadas para atingir as mesmas regiões de transcriptoma ou localizações próximas das que são direcionadas pelas sondas de microarranjo Illumina sempre que possível. As sondas que atendiam aos critérios de controle de qualidade NanoString foram projetadas com sucesso para 559 dos 619 biomarcadores selecionados (dados não mostrados).

Desenvolvimento, refino e validação do classificador de nódulos pulmonares NanoString

[00199] Primeiro, avaliamos quão bem as exatidões de classificação foram mantidas entre as plataformas Illumina e NanoString, analisando novamente 199 das amostras do conjunto de treinamento de microarranjos. Para a comparação das precisões de classificação, usamos apenas os 276 biomarcadores de microarranjos que foram projetados com sucesso como sondas NanoString. Observou-se uma correlação de Spearman rho = 0,73 (valor de $p < 10^{-12}$) para os escores de classificação da amostra entre as 2 plataformas (Fig. 8A). A ROC-AUC baseada nas 276 sondas foi de 0,881 para os microarranjos e 0,838 para NanoString (Fig. 8B), indicando que a transição da plataforma foi bem-sucedida.

[00200] Para realizar uma avaliação imparcial do desempenho do painel personalizado, analisou-se um total de 741 amostras de pacientes, incluindo amostras de um novo 5º local de coleta. O conjunto final de treinamento de nPNC de 583 amostras e a validação de 158 amostras apresentaram números balanceados de amostras de MN e BN (Tabela XI) para fornecer as melhores condições para a seleção de um classificador com boa sensibilidade e especificidade (17).

Tabela XI - Dados demográficos das amostras analisadas com o painel personalizado NanoString. A mediana ± IQR é dada para valores contínuos, os valores de p indicam significância da comparação entre

os grupos de nódulos malignos e benignos.

	NanoString Classificador de Nódulo de Pulmão Conjunto de Treinamento			NanoString Classificador de Nódulo de Pulmão Conjunto de Validação		
Categoria	Nódulos Malignos	Nódulos Benignos	*Valor p*	Nódulos Malignos	Nódulos Benignos	*Valor p*
Total N	290	293		74	84	
Sexo						
Masculino	155 (53%)	145 (49%)		45 (61%)	44 (52%)	
Feminino	135 (47%)	146 (50%)	0,2530	29 (39%)	38 (45%)	0,2728
Desconhecido	0 (0%)	2 (1%)		0 (0%)	2 (2%)	
Idade	68 ±6	62 ±6	$3x10^{-10}$	69 ±7	65 ±7	0,0097
Raça						
Negra	34 (12%)	20 (7%)		5 (7%)	10 (12%)	
Outras	35 (12%)	17 (6%)	0,0020	8 (11%)	10 (12%)	0,5134
Branca	221 (76%)	256 (87%)		61 (82%)	64 (76%)	
Fumante						
Atual	73 (25%)	110 (36%)		23 (31%)	22 (26%)	
Anterior	196 (68%)	170 (58%)	0,0124	45 (61%)	54 (64%)	0,8895
Novo	15 (5%)	11 (4%)		5 (7%)	6 (7%)	
Desconhecido	4 (1%)	2 (1%)		1 (1%)	2 (2%)	
Maços Anos	40 ± 20	38 ± 13	0,6066	42 ± 19	42 ± 15	0,5767
Tamanho lesão, mm	22 ± 10	8 ± 4	4×10^{-17}	18 ± 6	7 ± 4	6×10^{-10}
Estágio do Câncer						
I	173 (60%)			47 (64%)		
II	41 (14%)			9 (12%)		
III	49 (17%)			10 (14%)		
IV	1 (1%)			1 (1%)		
Desconhecido	23 (19%)			7 (9%)		
Local[a]						
HFGCC	146 (50%)	49 (17%)		37 (50%)	15 (18%)	
NYU	74 (26%)	198 (68%)		24 (32%)	62 (74%)	
Roswell-Park	11 (4%)	13 (4%)	6×10^{-20}	7 (9%)	6 (7%)	2×10^{-6}
Temple	3 (1%)	11 (4%)		0 (0%)	0 (0%)	
UPenn	56 (19%)	22 (8%)		6 (8%)	1 (1%)	

[00201] O modelo de classificação usando todas as 559 sondas demonstrou uma ROC-AUC de 0,833 (95% Cl: 0,799-0,864) no conjunto de treinamento e ROC-AUC de 0,826 (95% CI: 0,760-0,891) no conjunto de validação independente (Fig. 4A). O desempenho do conjunto de treinamento permaneceu estável durante o processo de eliminação recursiva do recurso (Fig. 9A). Conjuntos de sondas de diminuição incremental alcançaram ROC-AUCs semelhantes (Fig. 2B-D, Fig. 9B). Sensibilidades, especificidades e valores preditivos positivos e negativos (VPP, VPN) também são semelhantes nos conjuntos de treinamento e validação (Tabela XII).

Tabela XII - Desempenho da classificação usando diferentes números de sondas

		Sondas N			
Conj.	Métrica de Desempenho	559	100	41	6
Treinamento	Sensibilidade	76,5%	73,4%	74,7%	73,7%
	Especificidade	76,6%	73,8%	74,8%	73,4%
	Precisão	75,8%	74,6%	74,3%	73,6%
	ROC-AUC	0,833	0,825	0,834	0,800
	ROC-AUC 95% CI	0,799-0,864	0,790-0,857	0,800-0,865	0,765-0,836
	Especificidade a 90% de Sensibilidade	53,2%	52,9%	51,9%	45,1%
	Valor Preditivo Positivo (PPV)[a]	0,056	0,049	0,052	0,048
	Valor Preditivo Positivo (NPV)[a]	0,994	0,993	0,994	0,993
Validação	Sensibilidade	67,6%	67,6%	68,9%	52,7%
	Especificidade	83,3%	83,3%	82,1%	85,7%
	Precisão	75,9%	75,9%	75,9%	70,3%
	ROC-AUC	0,826	0,817	0,825	0,782
	ROC-AUC 95% CI	0,760-0,891	0,749-0,885	0,759-0,890	0,709-0,855
	Especificidade a 90% de	46,4%	36,9%	51,2%	32,1%

	Sensibilidade				
	Valor Preditivo Positivo (PPV)[a]	0,069	0,069	0,066	0,063
	Valor Preditivo Positivo (NPV)[a]	0,993	0,993	0,993	0,990

[a] Calculado usando prevalência de 1,8% de câncer de pulmão observado no Ensaio de Triagem de Pulmão Nacional (NLST)

[00202] Enquanto o classificador da sonda 41 (Tabela IX) alcançou uma AUC de 0,834 (IC 95%: 0,800- 0,865) para treinamento e 0,825 (IC 95%: 0,759-0,890) para a validação independente (Fig. 4C), mesmo usando o mínimo de 6 sondas mantiveram a ROC-AUC acima de 0,8, embora haja uma ligeira queda no desempenho do conjunto de validação (Fig. 4D).

[00203] A assinatura ideal de 41 genes (Tabela IX) apresentou sensibilidade e especificidade imparcial de 68,1% e 82,1%, respectivamente. Atingiu uma especificidade de 51%, com uma sensibilidade de 90%, tanto para treinamento quanto para validação, com valores de VPN e VPP de 0,99 e 0,0066, respectivamente, para a validação independente. O classificador detectou cânceres com 64% de sensibilidade para o Estágio I e 70% para cânceres de estágio posterior. As probabilidades de malignidade em uma variedade de classificações de nPNC são mostradas na Fig. 4E. Cabe ressaltar que um pequeno número de indivíduos com nódulos pulmonares malignos ou benignos sem histórico de tabagismo foram incluídos na análise. A remoção desses sujeitos do estudo não alterou o desempenho da validação da AUC para o classificador de 41 sondas (diferença de AUC de 0,001) ou o classificador completo de 559 sondas (diferença de AUC de 0,010). A adição de idade, sexo, raça e histórico de tabagismo como fatores adicionais não teve impacto na classificação que produz ROC-AUC de 0,837 em comparação com 0,840 quando apenas a expressão de gene foi usada.

O classificador nPNC supera os modelos clínicos existentes

[00204] Focando no classificador de 41 painéis de biomarcadores, comparou-se o desempenho de nPNC em todas as amostras na faixa de tamanho de difícil avaliação de 6 a 20 mm com o desempenho de três algoritmos clínicos, a Universidade Brock desenvolveu em uma população de alto risco (20, 21) e os modelos da Clínica Mayo (22) e VA (23) desenvolvidos usando dados de uma população de nódulos mais incidentais. Esses algoritmos avaliam o risco de câncer de um nódulo pulmonar com base em uma variedade de parâmetros demográficos e patológicos, incluindo o tamanho e a localização do nódulo (Fig. 5A). O nPNC supera todos os 3 modelos clínicos em nódulos na faixa de 6 a 20 mm de diâmetro. Como o tamanho do nódulo é um fator de risco bem aceito incluído em cada um dos modelos clínicos, também demonstrou-se uma precisão aumentada da classificação em comparação com uma classificação que utiliza apenas tamanho para as amostras na faixa de 6 a 20 mm (Fig. 5B).

Desempenho do classificador para diferentes faixas de tamanho de nódulos

[00205] Como o tamanho dos nódulos é um fator de risco importante e a definição de NIPs não está muito bem definida (6) com as mudanças nas diretrizes, examinou-se o desempenho ao comparar nódulos pulmonares malignos e benignos que foram semelhantes em faixas de tamanho. Calculou-se o desempenho do classificador de 41 sondas nas várias faixas de tamanho de nódulos usando limiares positivos iniciais de 4 mm do estudo NLST (24) 6 mm e 8 mm, conforme discutido nos relatórios recentes da Fleischner Society (8) e do Lung Rads (7), bem como um limiar de linha de base de 10 mm. As ROC-AUCs e as especificidades quando a sensibilidade é mantida em um desempenho de 90% foram calculadas para todas as faixas possíveis dos limiares selecionados para os conjuntos de treinamento e validação

independentes (Fig. 6). No geral, o desempenho do conjunto de treinamento e validação foi altamente conservado em todas as faixas de tamanho, exceto quando apenas algumas amostras do conjunto de validação se enquadram em uma faixa de tamanho específica, como é evidente no conjunto de validação menor. O nPNC da sonda 41 tem um desempenho particularmente bom na validação independente com nódulos na faixa de difícil diagnóstico de 8 a 14 mm, atingindo uma especificidade de 64% com sensibilidade de 90%, embora a especificidade caia para 48% no conjunto maior de dados combinados. Se 4 mm ou 6 mm são usados como limiar para uma tela positiva, nosso classificador demonstra sua utilidade na classificação de IPNs, executando bem em todas as faixas com uma ROC-AUC de 0,83 e 0,81 respectivamente no conjunto de dados combinado e um limiar de 8 mm apenas reduz a AUC para 0,80. As especificidades com 90% de sensibilidade são igualmente estáveis e são calculadas como 0,50, 0,46 e 0,48 para 4, 6 e 8 mm, respectivamente.

Discussão

[00206] Os benefícios gerais dos programas de triagem de câncer de pulmão usando LDCT são evidentes no aumento relatado de 20% na sobrevida dos pacientes. No entanto, esse sucesso vem com o problema associado de como avaliar o grande número de IPNs principalmente benignos sendo detectados e a preocupação com o excesso de gerenciamento (25). A recente avaliação Lung-RADs também sugeriu que a implementação de um limiar de triagem positivo de 6-7 mm, em vez dos 4 mm utilizados no estudo NLST, pode ser mais apropriada no gerenciamento dos resultados da Triagem do Câncer de Pulmão (7) e que essa mudança reduziria a magnitude do problema do IPN com um efeito mínimo no atendimento ao paciente (8). Mesmo com as novas diretrizes, o potencial de gerenciamento excessivo dos estimados 1,6 milhões de nódulos pulmonares detectados a cada ano

nos EUA continua sendo um desafio significativo, particularmente para nódulos ≥ 6 mm e menos de ≤ 20 mm, onde o risco de malignidade pode variar de ~ 8 a 64% (26). O desenvolvimento de abordagens não invasivas alternativas para avaliar esses IPNs de maneira clinicamente significativa é um objetivo importante na medicina pulmonar.

[00207] A maioria das abordagens não invasivas de detecção precoce dependeu da identificação de ácidos nucleicos, anticorpos ou proteínas derivados de tumores presentes no sangue, plasma, soro ou escarro (27-29), com a ressalva de que esses analitos são frequentemente raros na presença de cânceres menores em estágio inicial, mais propensos à cirurgia curativa e que agora estão sendo mais facilmente detectados por LDCT. Estudos adicionais que evitaram esse problema combinaram a broncoscopia com a expressão de gene em células epiteliais das vias aéreas normais ou com a expressão de gene associada à escovação nasal. Essa abordagem é baseada no conceito de "cancerização de campo", pelo qual o tumor induz alterações na expressão de gene no trato respiratório não envolvido, que diferem com a presença de um nódulo pulmonar maligno ou benigno. Essas abordagens funcionam bem para nódulos que podem ser acessados por broncoscopia (27, 30, 31), mas são menos eficazes com IPNs menores que também representam uma grande preocupação de gerenciamento.

[00208] Mostrou-se anteriormente que uma lesão maligna no pulmão pode estender sua influência além do campo do câncer pulmonar para o sangue periférico, uma vez que a expressão de gene no RNA derivado de PBMC distingue de forma eficaz os nódulos pulmonares malignos dos benignos (9). A existência desse efeito extra-pulmonar é apoiada por relatos iniciais de modelos de camundongos para câncer de pulmão, demonstrando que fatores solúveis produzidos por lesões pré-malignas no pulmão influenciaram a expressão de marcadores de ativação específicos em macrófagos da medula óssea e que esse efeito foi

aprimorado com progressão de tumor (32-34). Embora os estudos de PBMC tenham fornecido uma prova importante do comprometimento extra pulmonar, a necessidade de uma rápida purificação das amostras de PBMC para estabilizar os perfis transcricionais foi um obstáculo à expansão para locais de coleta fora dos ambientes acadêmicos e ao desenvolvimento de uma robusta plataforma clínica. Agora demonstrou-se que o RNA do sangue total, facilmente coletado em tubos de estabilização de RNA do PAXgene, também pode ser extraído para obter informações de expressão de gene que distingam os nódulos pulmonares malignos dos benignos do pulmão. Esse sistema de coleta de sangue minimamente invasivo de 2,5 ml permite que as amostras sejam coletadas não apenas nos principais centros médicos, mas onde quer que o sangue seja rotineiramente coletado. A estabilidade do RNA em temperatura ambiente por 5 dias significa que nenhum sistema de armazenamento especial é necessário para manter a integridade da amostra, facilitando a coleta de amostras e a subsequente transferência para uma instalação central de testes, mesmo a partir de locais remotos. A qualidade do RNA o torna passível de análise em uma ampla variedade de plataformas, incluindo uma variedade de plataformas de sequenciamento que requerem RNA de alta qualidade.

[00209] Testou-se a utilidade do sistema de coleta PAXgene usando amostras coletadas em 4 centros pulmonares acadêmicos e em um hospital comunitário. As amostras foram coletadas armazenadas e transferidas a granel ou diariamente, depois transferidas pelo correio para o nosso local de teste para armazenamento e processamento final sem nenhum efeito detectável no desempenho da plataforma. Construiu-se o modelo de diagnóstico a partir da expressão de gene global testada em microarranjos Illumina com cânceres que eram principalmente de Estágio I (69%) e II (17%) e nódulos que variavam em tamanho desde a medição do limiar de 4 mm do estudo NLST até 20

mm, abrangendo a faixa de risco de malignidade de <1% a 64% (8). É importante ressaltar que o classificador de microarranjos PAXgene manteve uma ROC-AUC de 0,847 (IC 95%: 0,742-0,951) em validação independente, quase idêntica ao do conjunto de treinamento usado para o desenvolvimento do classificador. Em muitos estudos, a precisão do conjunto de validação diminuiu um pouco, sugerindo que o modelo usado para o desenvolvimento do classificador não era grande o suficiente para capturar adequadamente a diversidade potencial de sujeitos (9, 35, 36). Avançando da plataforma de desenvolvimento de microarranjos, fizemos a transição com sucesso do classificador de nódulos para a plataforma NanoString nCounter. A plataforma nCounter requer manipulação mínima de amostras, é tecnicamente simples e tem a capacidade de avaliar o RNA degradado e não degradado no mesmo ensaio. A aprovação da FDA do Ensaio Prognóstico de Câncer de Mama Prosigna™ baseado em NanoString, com base na assinatura do gene PAM50 (14), e o desenvolvimento mais recente de uma assinatura imune baseada em NanoString, que prevê a resposta clínica ao bloqueio de PD1 (37), apoia ainda mais a utilidade clínica desta plataforma.

[00210] Embora o painel genético preliminar para o classificador baseado em NanoString incluísse 559 biomarcadores, esse número poderia ser reduzido para 41 sondas, mantendo a ROC-AUC e, portanto, sugere o potencial de simplificação da plataforma de teste. Ao avaliar as contribuições das várias sondas representadas nas 41 sondas, 46% das sondas mais classificadas vieram da análise SVM, 29% do painel PCI e com o menor número de candidatos selecionado pelo valor de p. Os genes relacionados ao mieloide ligados à sobrevida nos estudos de PBMC (18) não estavam representados no classificador de 41 sondas, mas estavam bem representados nas 100 melhores sondas classificadas, enquanto as sondas relacionadas a NK estavam principalmente na metade inferior do conjunto de sondas, talvez porque

o sinal NK é significativamente diluído nas amostras PAXgene. À medida que os dados dos resultados dos pacientes são acumulados, avalia-se ainda mais a utilidade dos biomarcadores prognósticos incluídos no painel que foram selecionados devido a uma associação com recorrência/sobrevida em nossos estudos de PBMC anteriores (18, 19, 38).

[00211] Embora o desempenho técnico robusto seja importante para qualquer plataforma clínica, o benefício resultante para o paciente é primário. O desempenho do painel personalizado NanoString nas 741 amostras analisadas nessa plataforma tem implicações clínicas significativas, com potencial para impactar o uso de abordagens invasivas para avaliar algumas classes de IPN difíceis de diagnosticar. O estudo não depende da presença de células tumorais circulantes, proteínas tumorais ou RNA tumoral cuja presença é mais consistente com cânceres mais avançados. Neste estudo, aborda-se principalmente a classe de nódulos pulmonares indeterminados com diâmetro de 6 a 20 mm, de risco moderado a alto (39) e frequentemente não facilmente acessíveis por broncoscopia ou biópsia por agulha fina e cânceres em estágio inicial que são mais favoráveis às abordagens cirúrgicas. Também avalia-se o desempenho com nódulos menores na faixa de 4-6 mm, onde o risco de malignidade é pequeno, mas cuja presença pode permanecer preocupante.

[00212] É importante ressaltar que nossos nPNC superaram os algoritmos clínicos atualmente usados para estratificar candidatos com NPI para tratamento ou acompanhamento, incluindo os modelos clínicos da Universidade Brock (20, 21) Clínica Mayo (22) e VA (23) na faixa de 6 a 20 mm. Embora esses algoritmos funcionem bem quando aplicados a conjuntos de dados que incluem principalmente nódulos benignos menores e cânceres maiores, o desempenho diminui um pouco quando aplicado apenas a MN e BN na faixa de tamanho

problemática. Embora a faixa de tamanho dos nódulos pulmonares que analisou-se seja importante, ainda existe uma diferença significativa no tamanho médio entre o MN e o BN no estudo. Será importante abordar como os biomarcadores e os algoritmos clínicos funcionam quando os nódulos BN e MN são mais próximos do tamanho e onde é provável que os algoritmos clínicos tenham um desempenho ruim. Tentou-se testar esse tipo de comparação, como mostra a Figura 4. Embora as AUCs, sensibilidades e especificidades gerais sejam bem conservadas, se usar 4, 6 ou 8 mm como limiar positivo, à medida que as comparações se tornam mais granulares, algumas comparações são significativamente mais precisas que outras, e isso é particularmente evidente no estudo de validação onde o número das amostras é menor. Atingiu-se uma especificidade de 64% a 90% de sensibilidade para BN e MN na faixa de tamanho de 8-14 mm, caindo para 40% na faixa de 6-14 mm.

[00213] O tamanho dos nódulos é a principal consideração em como os IPN são tratados (40). Embora este estudo tenha interrogado um grande número de amostras de pacientes e demonstrado utilidade potencial, uma avaliação posterior com um número maior de amostras onde MN e BN estão mais intimamente relacionados por tamanho ampliará essa utilidade, pois esse é o cenário em que o tamanho não é mais informativo. Ao avançar, será importante abordar mais completamente a questão comparando BN e MN de tamanhos semelhantes em toda a faixa de tamanhos de nódulos que permanecem problemáticos. O método altamente simplificado e comprovado para a aquisição de um grande número de amostras de qualidade consistente de vários locais facilitará esse processo. Estudos expandidos também permitem abordar a base biológica para as diferenças que foram detectadas entre as classes de pacientes e avaliar se essas diferenças podem ter implicações terapêuticas.

[00214] Toda e qualquer patente, pedido de patente, publicação, incluindo o Pedido Internacional de Patente PCT/US17/38571 e a sequência de genes publicamente disponíveis citadas ao longo da divulgação são aqui expressamente incorporadas por referência em sua totalidade. São também incorporados por referência documentos prioritários, o Pedido de Patente U.S. 62/607.756, depositado em 19 de dezembro de 2017 e o Pedido de Patente U.S. 62/752.163, depositado em 29 de outubro de 2018. Embora esta invenção tenha sido descrita com referência a modalidades específicas, é evidente que outras modalidades e variações desta invenção são concebidas por outros versados na técnica sem afastamento do verdadeiro espírito e escopo da invenção. As reivindicações anexas incluem tais modalidades e variações equivalentes.

REIVINDICAÇÕES

1. Composição para diagnosticar a existência ou avaliar a progressão de um câncer de pulmão em um sujeito mamífero, caracterizada pelo fato de que compreende pelo menos 7 polinucleotídeos ou oligonucleotídeos ou ligantes, em que cada polinucleotídeo ou oligonucleotídeo ou ligante hibridiza com um gene, fragmento de gene, transcrito de gene ou produto de expressão em uma amostra selecionada dos genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX.

2. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que pelo menos um polinucleotídeo ou oligonucleotídeo ou ligante está fixado a um marcador detectável.

3. Composição, de acordo com a reivindicação 2, caracterizada pelo fato de que cada polinucleotídeo ou oligonucleotídeo ou ligante está fixado a um marcador detectável diferente.

4. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende ainda um oligonucleotídeo de captura que hibridiza com pelo menos um polinucleotídeo ou oligonucleotídeo.

5. Composição, de acordo com a reivindicação 4, caracterizada pelo fato de que o oligonucleotídeo de captura é capaz de hibridizar com cada polinucleotídeo ou oligonucleotídeo.

6. Composição, de acordo com a reivindicação 4 ou 5, caracterizada pelo fato de que o oligonucleotídeo de captura se liga a um substrato.

7. Composição, de acordo com a reivindicação 6, caracterizada pelo fato de que compreende ainda um substrato ao qual o oligonucleotídeo de captura se liga.

8. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende pelo menos 15

polinucleotídeos ou oligonucleotídeos.

9. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende pelo menos 20 polinucleotídeos ou oligonucleotídeos.

10. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende pelo menos 25 polinucleotídeos ou oligonucleotídeos.

11. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende pelo menos 41 ou pelo menos 50 polinucleotídeos ou oligonucleotídeos.

12. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada gene, fragmento de gene, transcrito de gene ou produto de expressão diferente listado na Tabela I.

13. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada gene, fragmento de gene, transcrito de gene ou produto de expressão diferente listado na Tabela II.

14. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada gene, fragmento de gene, transcrito de gene ou produto de expressão diferente listado na Tabela III.

15. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada um dos primeiros 15 genes, fragmentos de genes, transcritos de genes ou produtos de expressão listados na Tabela II.

16. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada um dos primeiros 15 genes, fragmentos de genes, transcritos de genes ou produtos de expressão listados na Tabela III.

17. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada um dos primeiros 15 genes, fragmentos de genes, transcritos de genes ou produtos de expressão listados na Tabela IV.

18. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada um dos primeiros 50 genes, fragmentos de genes, transcritos de genes ou produtos de expressão listados na Tabela III.

19. Kit, caracterizado pelo fato de que compreende a composição, como definida em qualquer uma das reivindicações 1 a 18 e um aparelho para coleta de amostra.

20. Kit, de acordo com a reivindicação 19, caracterizado pelo fato de que o referido aparelho para coleta de amostra compreende um tubo para reter sangue, que contém um reagente que estabiliza a amostra.

21. Kit, de acordo com a reivindicação 20, caracterizado pelo fato de que o reagente estabiliza o mRNA na amostra.

22. Método para diagnosticar a existência ou avaliar um câncer de pulmão em um sujeito mamífero, caracterizado pelo fato de que compreende identificar mudanças na expressão de 7 ou mais genes na amostra do referido sujeito, os referidos genes selecionados dos genes da Tabela I ou dos genes da Tabela II ou dos genes da Tabela III ou dos genes da Tabela IV; e comparar os níveis de expressão de gene

do referido sujeito com os níveis dos mesmos genes em uma referência ou um controle, em que mudanças na expressão dos genes do sujeito daqueles da referência se correlacionam com um diagnóstico ou uma avaliação de um câncer de pulmão.

23. Método, de acordo com a reivindicação 22, caracterizado pelo fato de que o referido diagnóstico ou a referida avaliação compreende um ou mais de um diagnóstico de um câncer de pulmão, um diagnóstico de um nódulo benigno, um diagnóstico de um estágio de câncer de pulmão, um diagnóstico de um tipo ou uma classificação de um câncer de pulmão, um diagnóstico ou uma detecção de uma recorrência de um câncer de pulmão, um diagnóstico ou uma detecção de uma regressão de um câncer de pulmão, um prognóstico de um câncer de pulmão ou uma avaliação da resposta de um câncer de pulmão a uma terapia cirúrgica ou não cirúrgica.

24. Método, de acordo com a reivindicação 23, caracterizado pelo fato de que as referidas mudanças compreendem uma suprarregulação de um ou mais genes selecionados em comparação com a dita referência ou o referido controle ou uma infrarregulação de um ou mais genes selecionados em comparação com a dita referência ou o referido controle.

25. Método, de acordo com a reivindicação 23, caracterizado pelo fato de que compreende ainda identificar o tamanho de um nódulo de pulmão no sujeito.

26. Método, de acordo com a reivindicação 22, caracterizado pelo fato de que compreende ainda usar a composição, como definida em qualquer uma das reivindicações 1 21 para o referido diagnóstico.

27. Método, de acordo com a reivindicação 26, caracterizado pelo fato de que a dita referência ou o referido controle compreende sete ou mais genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX da amostra de pelo menos um sujeito de referência, o referido

sujeito de referência selecionado do grupo consistindo em: (a) um fumante com doença maligna, (b) um fumante com doença não maligna, (c) um ex-fumante com doença não maligna, (d) um não fumante saudável sem doença, (e) um não fumante que tem doença pulmonar obstrutiva crônica (COPD), (f) ex-fumante com COPD, (g) sujeito com um tumor de pulmão sólido antes de cirurgia para remoção do mesmo; (h) um sujeito com um tumor de pulmão sólido seguindo a remoção cirúrgica do referido tumor; (i) um sujeito com um tumor de pulmão sólido antes da terapia para o mesmo; e (j) um sujeito com um tumor de pulmão sólido durante ou em seguida à terapia para o mesmo.

28. Método, de acordo com a reivindicação 27, caracterizado pelo fato de que o referido sujeito de referência ou controle (a)-(j) é o mesmo sujeito de teste em um ponto no tempo temporalmente anterior.

29. Método, de acordo com qualquer uma das reivindicações 21 a 28, caracterizado pelo fato de que a amostra é sangue periférico.

30. Método, de acordo com a reivindicação 29, caracterizado pelo fato de que os ácidos nucleicos na amostra foram estabilizados antes da identificação de mudanças nos níveis de expressão de gene.

31. Método para detectar câncer de pulmão em um paciente, caracterizado pelo fato de que compreende:

a. obter uma amostra do paciente; e

b. detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão.

32. Método para diagnosticar câncer de pulmão em um sujeito, caracterizado pelo fato de que compreende:

a. obter uma amostra de sangue de um sujeito;

b. detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão; e

c. diagnosticar o sujeito com câncer quando forem detectadas mudanças na expressão dos genes do sujeito para aquelas da referência.

33. Método para diagnosticar e vencer câncer de pulmão em um sujeito tendo um crescimento neoplásico, caracterizado pelo fato de que compreende:

a. obter uma amostra de sangue de um sujeito;

b. detectar uma mudança na expressão em pelo menos 7 genes selecionados da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX na amostra do paciente em comparação com um controle contatando a amostra com uma composição compreendendo oligonucleotídeos, polinucleotídeos ou ligantes específicos para cada transcrito de gene ou produto de expressão diferente dos pelo menos 7 genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX e detecção de ligação entre o oligonucleotídeo, polinucleotídeo ou ligante e o produto de gene ou produto de expressão;

c. diagnosticar o sujeito com câncer quando forem detectadas mudanças na expressão dos genes do sujeito daquelas da

referência; e

d. remover o crescimento neoplásico.

34. Composição, de acordo com a reivindicação 1, caracterizada pelo fato de que compreende polinucleotídeos ou oligonucleotídeos capazes de hibridizar com cada um dos primeiros 41 genes, fragmentos de genes, transcritos de genes ou produtos de expressão listados na Tabela IX.

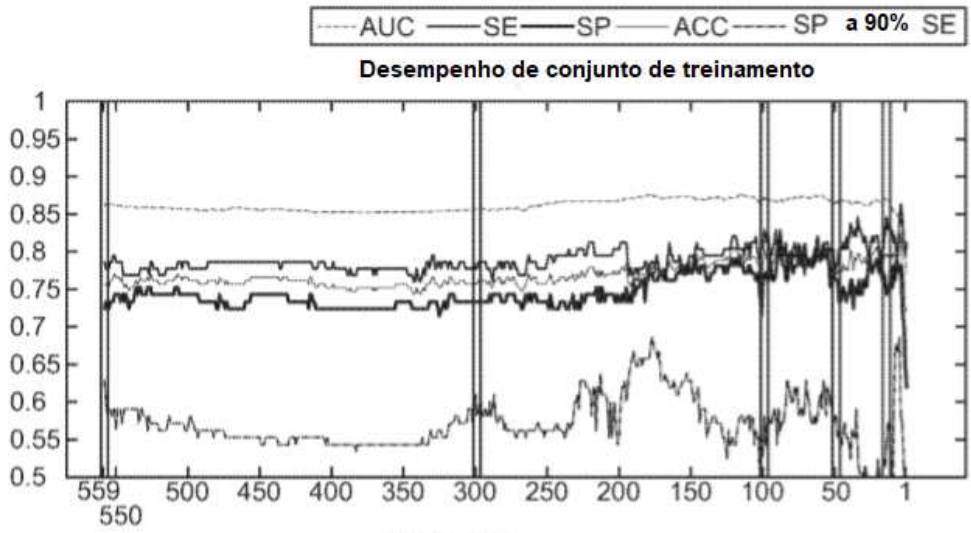

AUC SE SP ACC SP a 90% SE
Desempenho de conjunto de treinamento
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
559
550
500
450
400
350
300
250
200
150
100
50
1

FIG. 1A

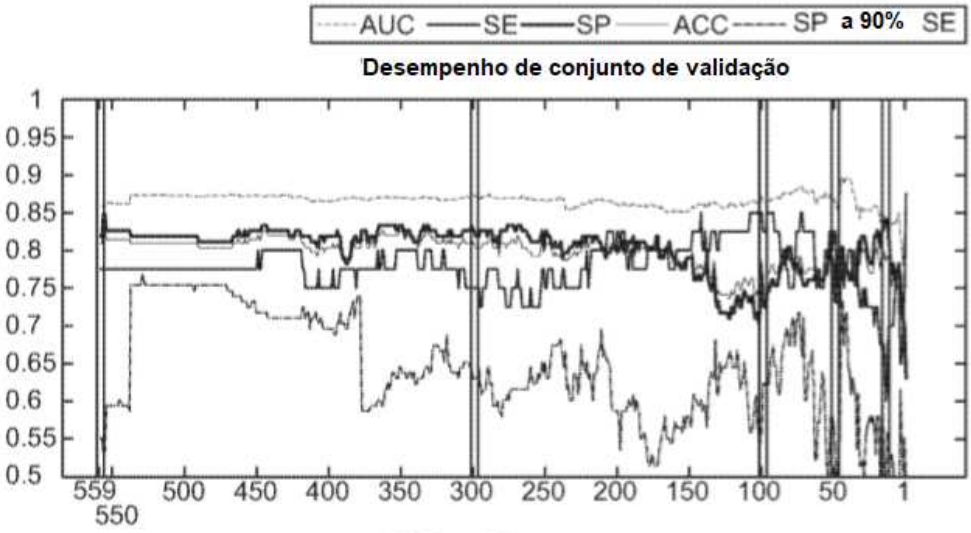

AUC SE SP ACC SP a 90% SE
Desempenho de conjunto de validação
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
559
550
500
450
400
350
300
250
200
150
100
50
1

FIG. 1B

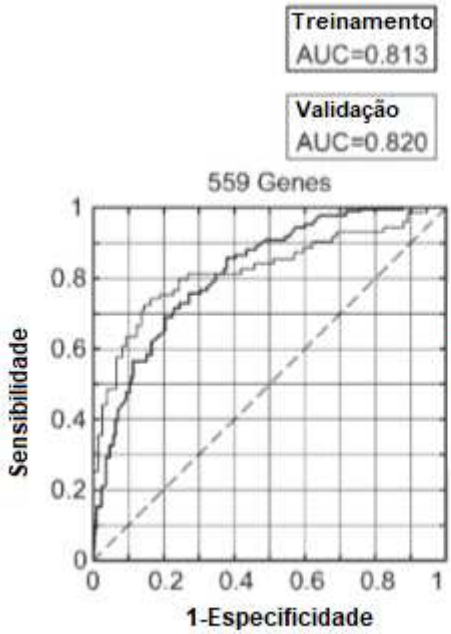


FIG. 2A

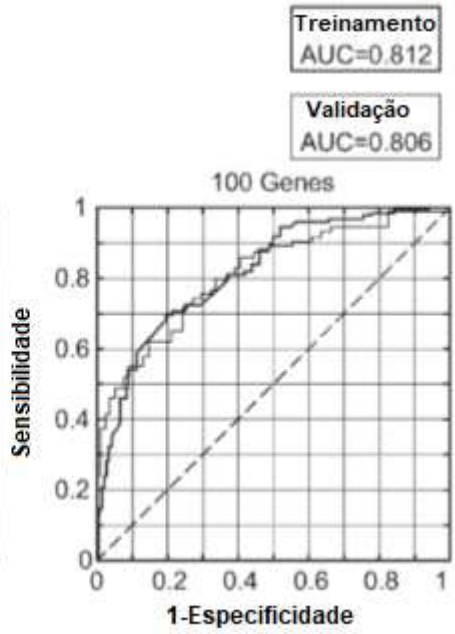


FIG. 2B

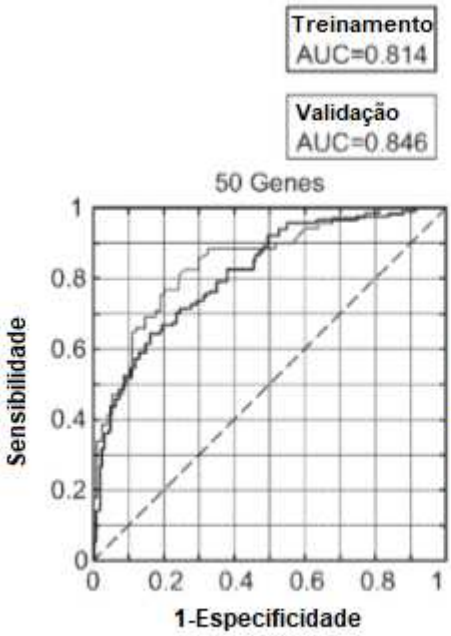


FIG. 2C

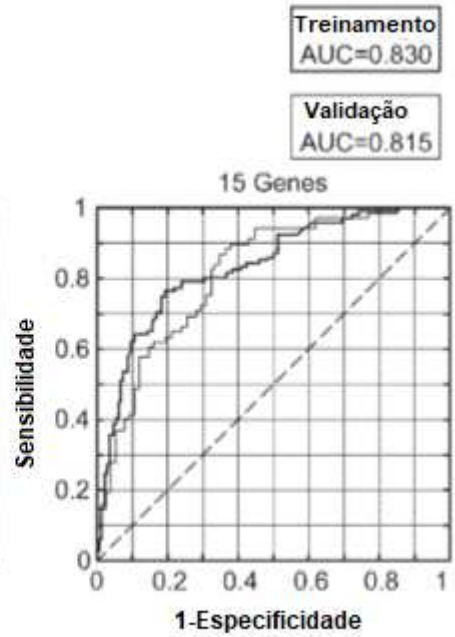


FIG. 2D

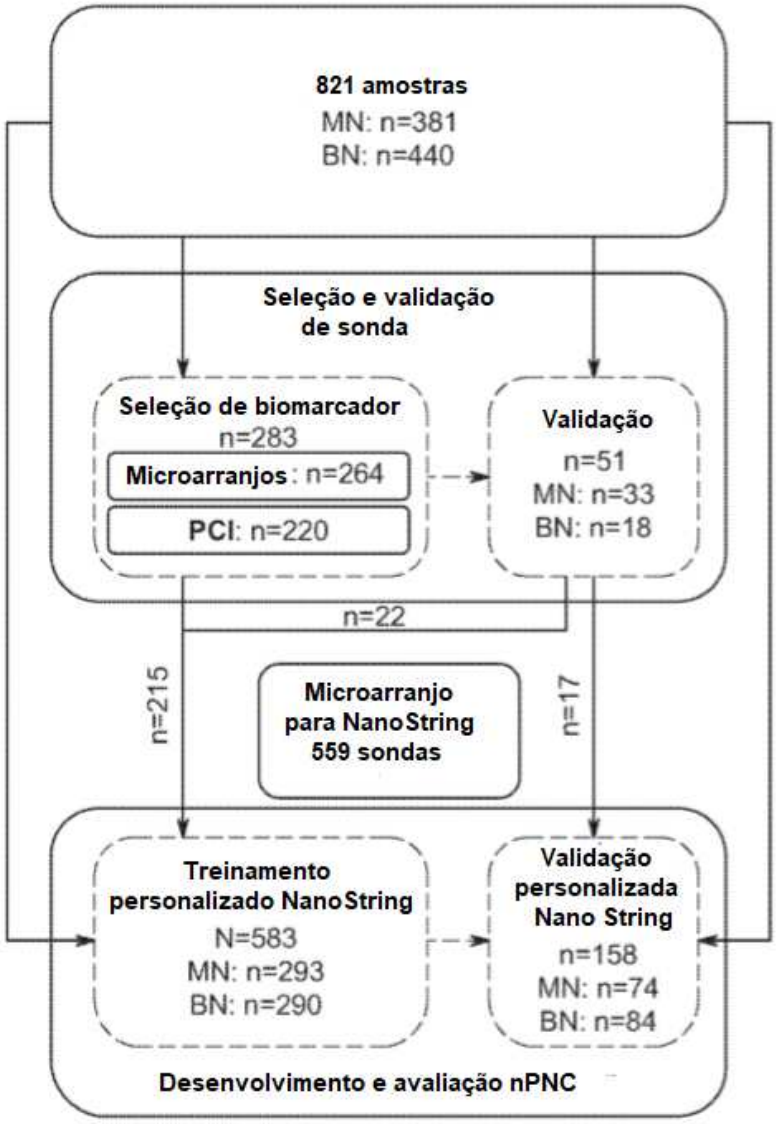

821 amostras
MN: n=381
BN: n=440
Seleção e validação de sonda
Seleção de biomarcador
n=283
Microarranjos: n=264
PCI: n=220
Validação
n=51
MN: n=33
BN: n=18
n=22
n=215
Microarranjo para NanoString 559 sondas
n=17
Treinamento personalizado NanoString
N=583
MN: n=293
BN: n=290
Validação personalizada Nano String
n=158
MN: n=74
BN: n=84
Desenvolvimento e avaliação nPNC

FIG. 3

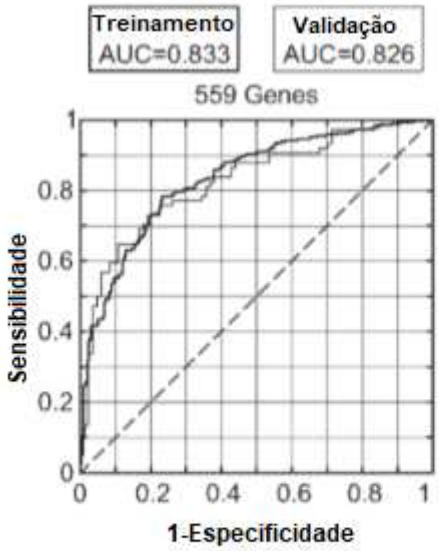


FIG. 4A

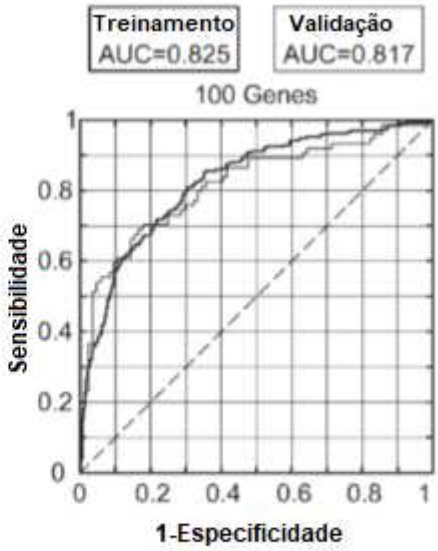


FIG. 4B

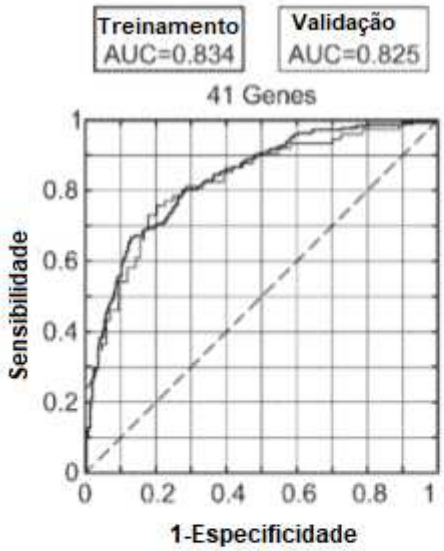


FIG. 4C

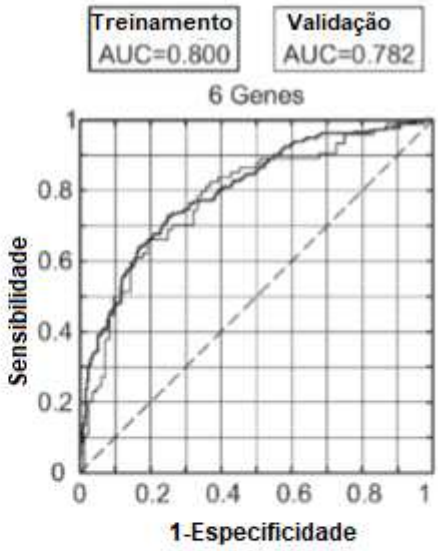


FIG. 4D

Escore de classificação	-14	-10	-6.6	-4.1	-2	-0.1	1.7	3.8	6.3	10	14
Probabilidade de câncer	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%

FIG. 4E

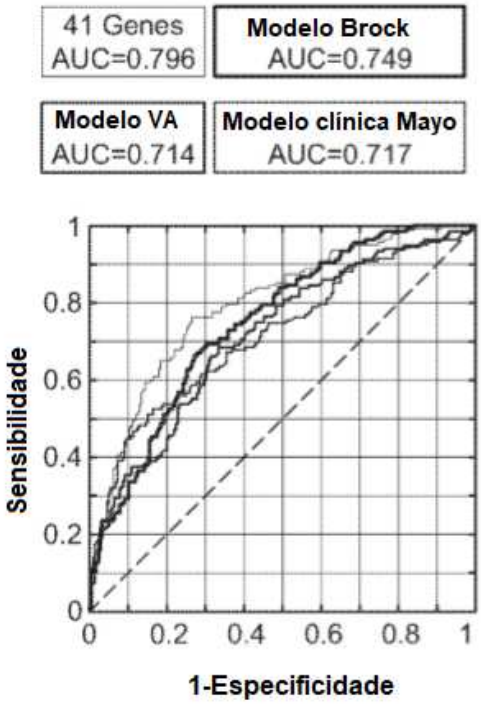

41 Genes
AUC=0.796
Modelo Brock
AUC=0.749
Modelo VA
AUC=0.714
Modelo clínica Mayo
AUC=0.717
Sensibilidade
1-Especificidade
0
0.2
0.4
0.6
0.8
1

FIG. 5A

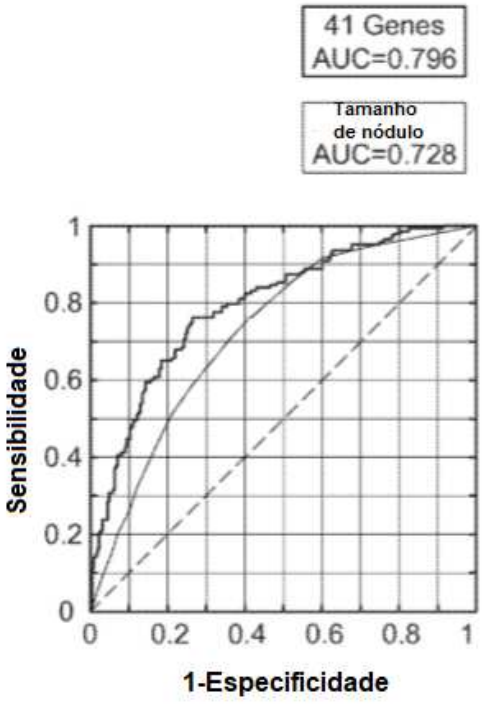

41 Genes
AUC=0.796
Tamanho de nódulo
AUC=0.728
Sensibilidade
1-Especificidade
0
0.2
0.4
0.6
0.8
1

FIG. 5B

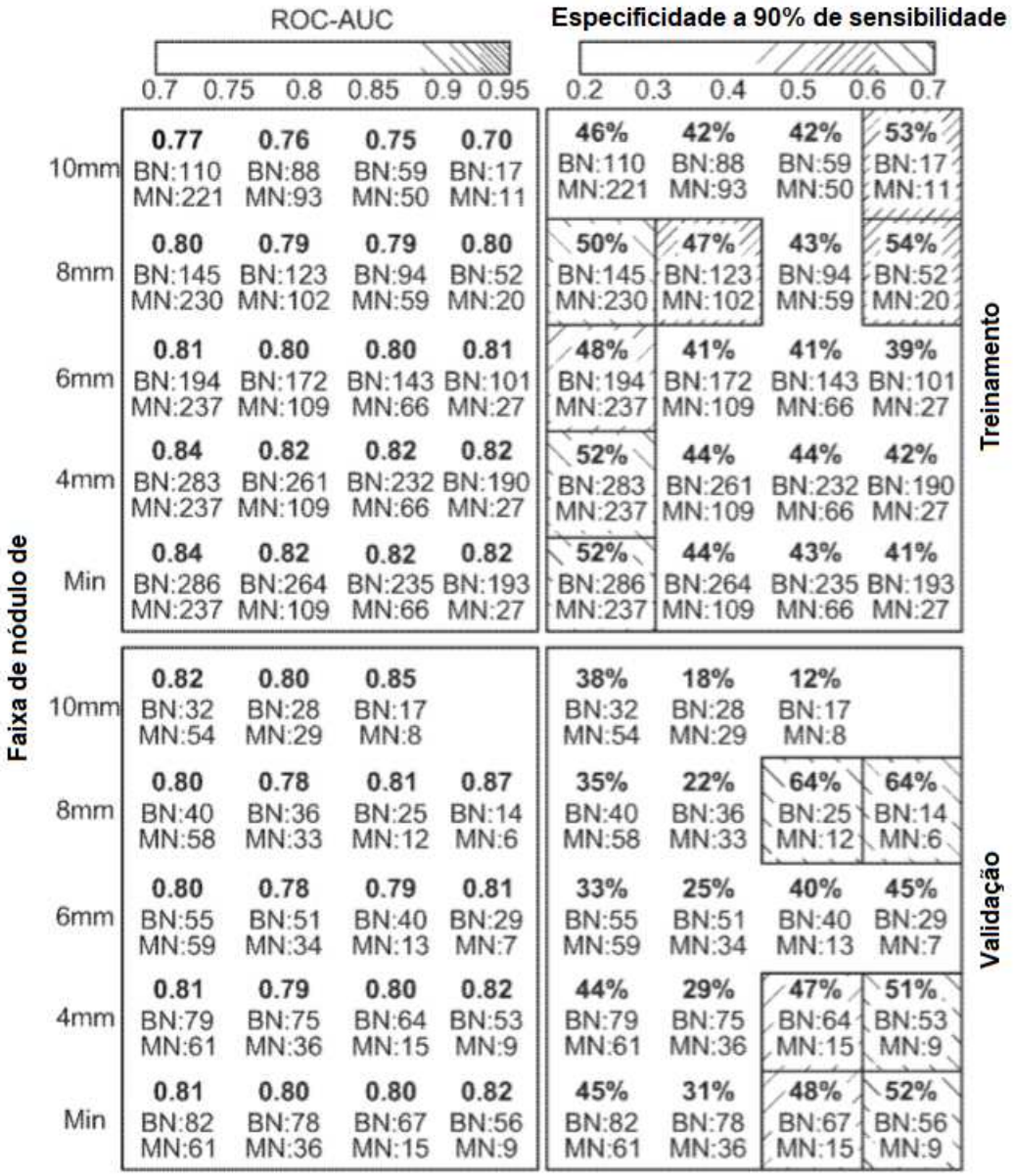

ROC-AUC
0.7 0.75 0.8 0.85 0.9 0.95
Especificidade a 90% de sensibilidade
0.2 0.3 0.4 0.5 0.6 0.7
Faixa de nódulo de
Treinamento
Validação
10mm
8mm
6mm
4mm
Min
0.77 0.76 0.75 0.70
BN:110 BN:88 BN:59 BN:17
MN:221 MN:93 MN:50 MN:11
46% 42% 42% 53%
BN:110 BN:88 BN:59 BN:17
MN:221 MN:93 MN:50 MN:11
0.80 0.79 0.79 0.80
BN:145 BN:123 BN:94 BN:52
MN:230 MN:102 MN:59 MN:20
50% 47% 43% 54%
BN:145 BN:123 BN:94 BN:52
MN:230 MN:102 MN:59 MN:20
0.81 0.80 0.80 0.81
BN:194 BN:172 BN:143 BN:101
MN:237 MN:109 MN:66 MN:27
48% 41% 41% 39%
BN:194 BN:172 BN:143 BN:101
MN:237 MN:109 MN:66 MN:27
0.84 0.82 0.82 0.82
BN:283 BN:261 BN:232 BN:190
MN:237 MN:109 MN:66 MN:27
52% 44% 44% 42%
BN:283 BN:261 BN:232 BN:190
MN:237 MN:109 MN:66 MN:27
0.84 0.82 0.82 0.82
BN:286 BN:264 BN:235 BN:193
MN:237 MN:109 MN:66 MN:27
52% 44% 43% 41%
BN:286 BN:264 BN:235 BN:193
MN:237 MN:109 MN:66 MN:27
0.82 0.80 0.85
BN:32 BN:28 BN:17
MN:54 MN:29 MN:8
38% 18% 12%
BN:32 BN:28 BN:17
MN:54 MN:29 MN:8
0.80 0.78 0.81 0.87
BN:40 BN:36 BN:25 BN:14
MN:58 MN:33 MN:12 MN:6
35% 22% 64% 64%
BN:40 BN:36 BN:25 BN:14
MN:58 MN:33 MN:12 MN:6
0.80 0.78 0.79 0.81
BN:55 BN:51 BN:40 BN:29
MN:59 MN:34 MN:13 MN:7
33% 25% 40% 45%
BN:55 BN:51 BN:40 BN:29
MN:59 MN:34 MN:13 MN:7
0.81 0.79 0.80 0.82
BN:79 BN:75 BN:64 BN:53
MN:61 MN:36 MN:15 MN:9
44% 29% 47% 51%
BN:79 BN:75 BN:64 BN:53
MN:61 MN:36 MN:15 MN:9
0.81 0.80 0.80 0.82
BN:82 BN:78 BN:67 BN:56
MN:61 MN:36 MN:15 MN:9
45% 31% 48% 52%
BN:82 BN:78 BN:67 BN:56
MN:61 MN:36 MN:15 MN:9

FIG. 6-1

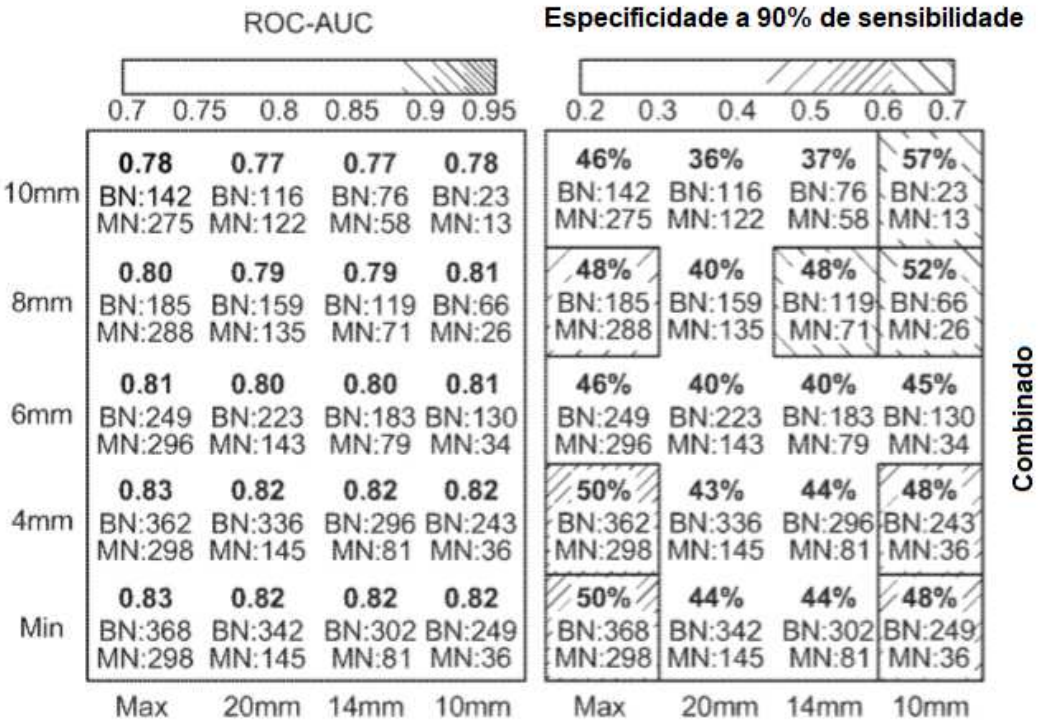

ROC-AUC
0.7 0.75 0.8 0.85 0.9 0.95
Especificidade a 90% de sensibilidade
0.2 0.3 0.4 0.5 0.6 0.7
10mm
8mm
6mm
4mm
Min
0.78 BN:142 MN:275
0.77 BN:116 MN:122
0.77 BN:76 MN:58
0.78 BN:23 MN:13
0.80 BN:185 MN:288
0.79 BN:159 MN:135
0.79 BN:119 MN:71
0.81 BN:66 MN:26
0.81 BN:249 MN:296
0.80 BN:223 MN:143
0.80 BN:183 MN:79
0.81 BN:130 MN:34
0.83 BN:362 MN:298
0.82 BN:336 MN:145
0.82 BN:296 MN:81
0.82 BN:243 MN:36
0.83 BN:368 MN:298
0.82 BN:342 MN:145
0.82 BN:302 MN:81
0.82 BN:249 MN:36
46% BN:142 MN:275
36% BN:116 MN:122
37% BN:76 MN:58
57% BN:23 MN:13
48% BN:185 MN:288
40% BN:159 MN:135
48% BN:119 MN:71
52% BN:66 MN:26
46% BN:249 MN:296
40% BN:223 MN:143
40% BN:183 MN:79
45% BN:130 MN:34
50% BN:362 MN:298
43% BN:336 MN:145
44% BN:296 MN:81
48% BN:243 MN:36
50% BN:368 MN:298
44% BN:342 MN:145
44% BN:302 MN:81
48% BN:249 MN:36
Max 20mm 14mm 10mm
Max 20mm 14mm 10mm
Combinado

FIG. 6-2

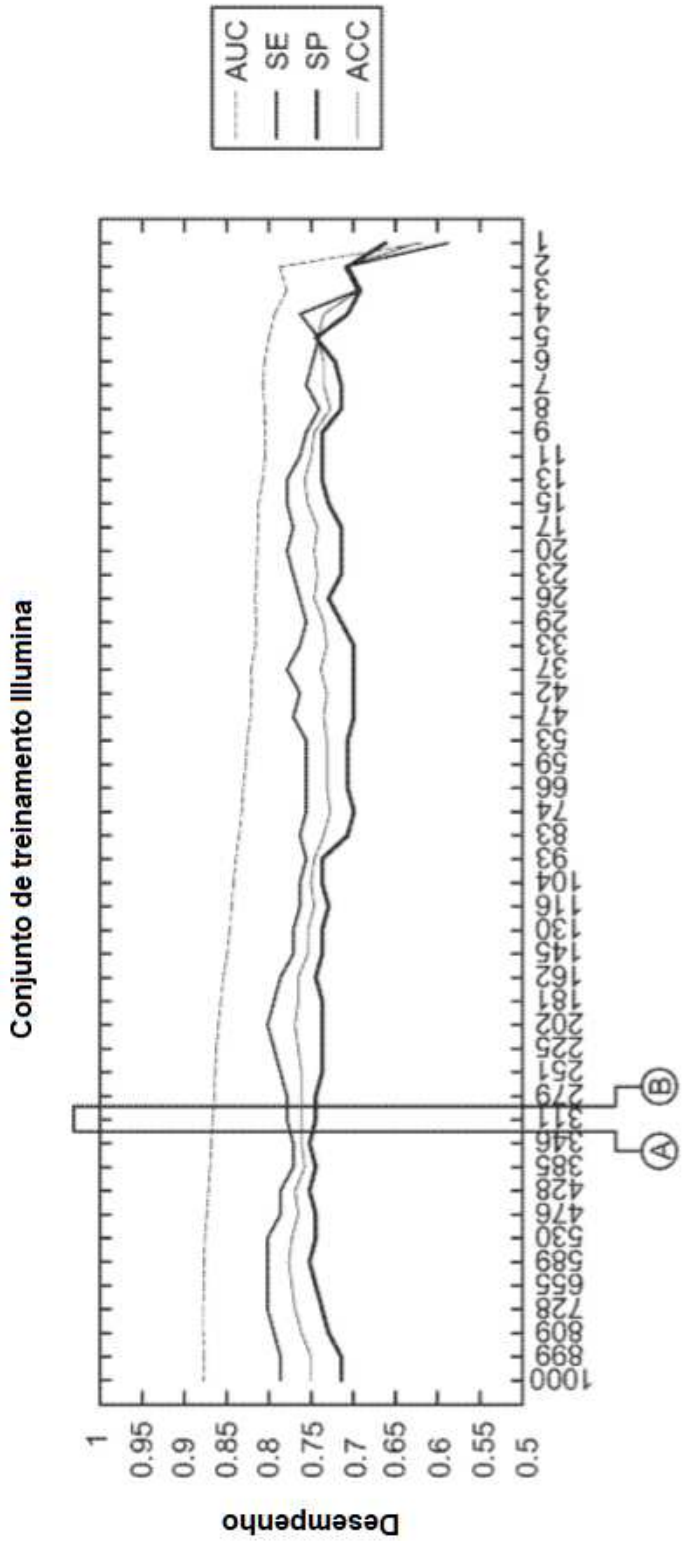

AUC
SE
SP
ACC
Conjunto de treinamento Illumina
Desempenho
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
A
B

FIG. 7A

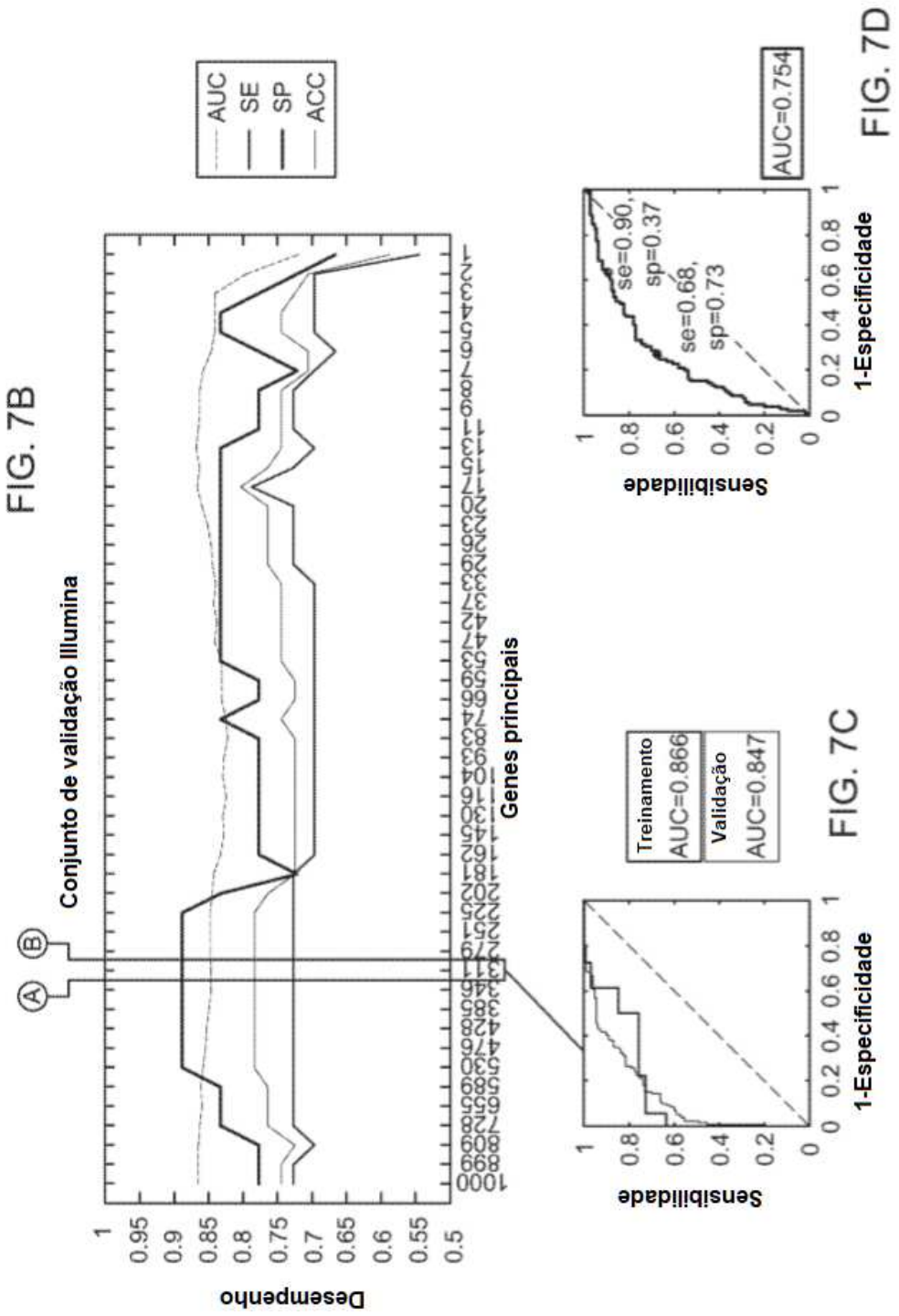


FIG. 7B

FIG. 7C

FIG. 7D

FIG. 8A

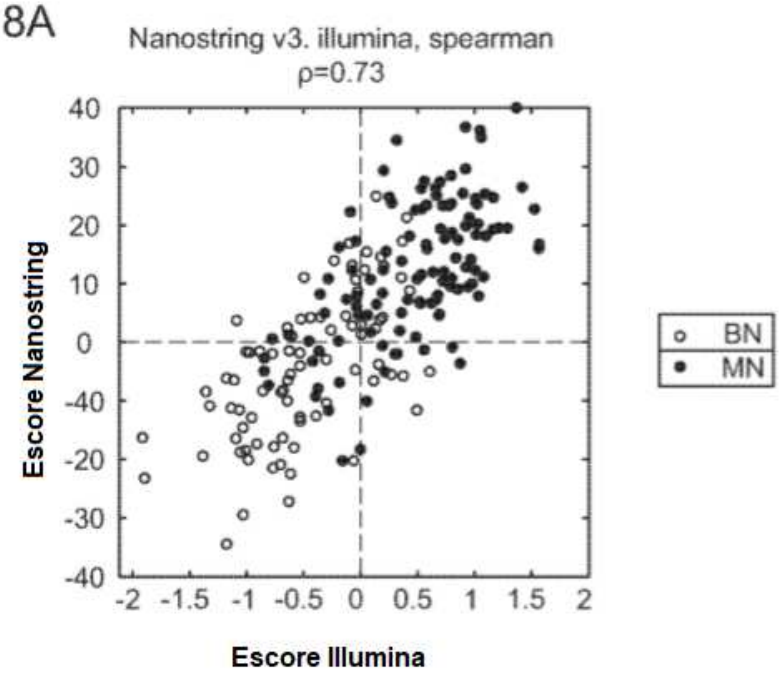

Nanostring v3. illumina, spearman
ρ=0.73
Escore Nanostring
Escore Illumina
BN
MN

FIG. 8B

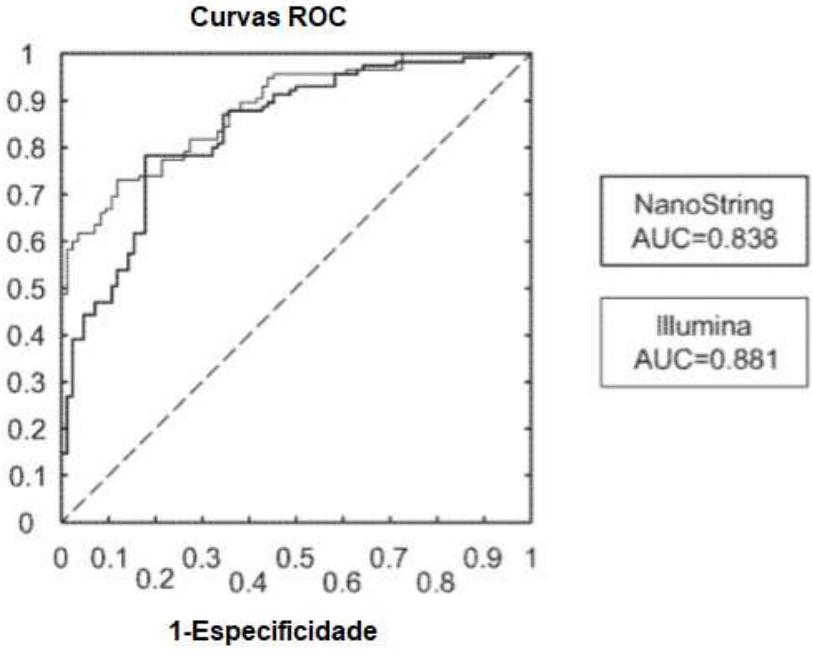

Curvas ROC
1-Especificidade
NanoString
AUC=0.838
Illumina
AUC=0.881

FIG. 9A

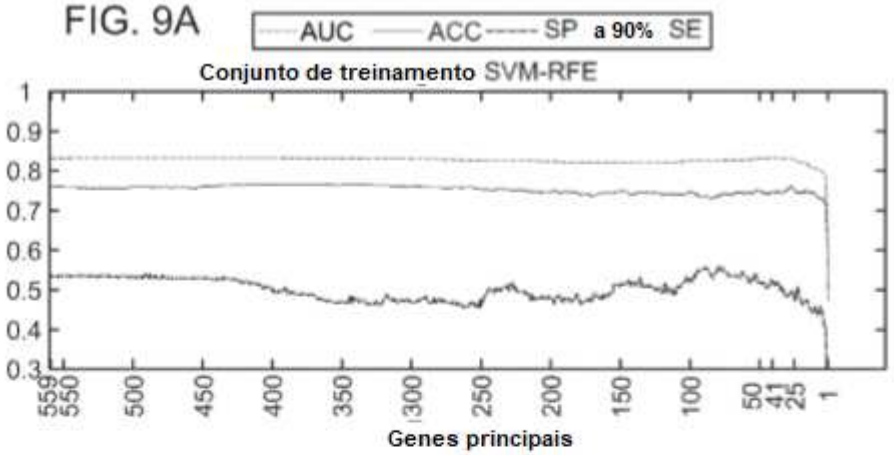


FIG. 9B

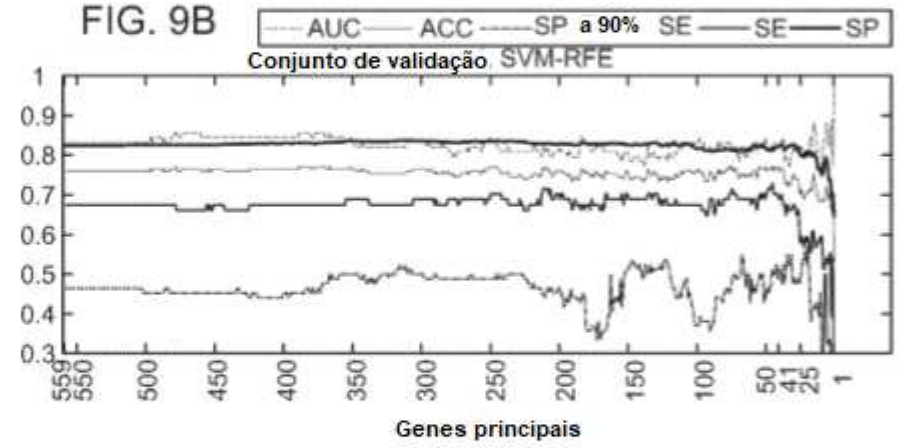
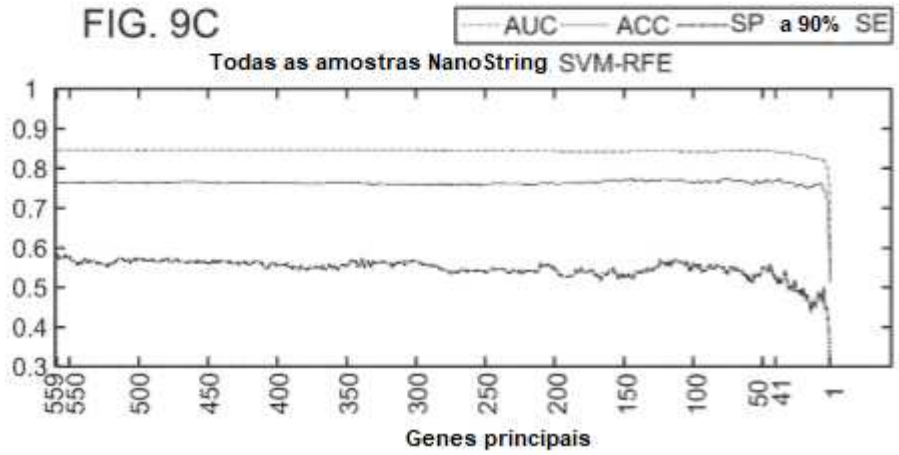


FIG. 9C



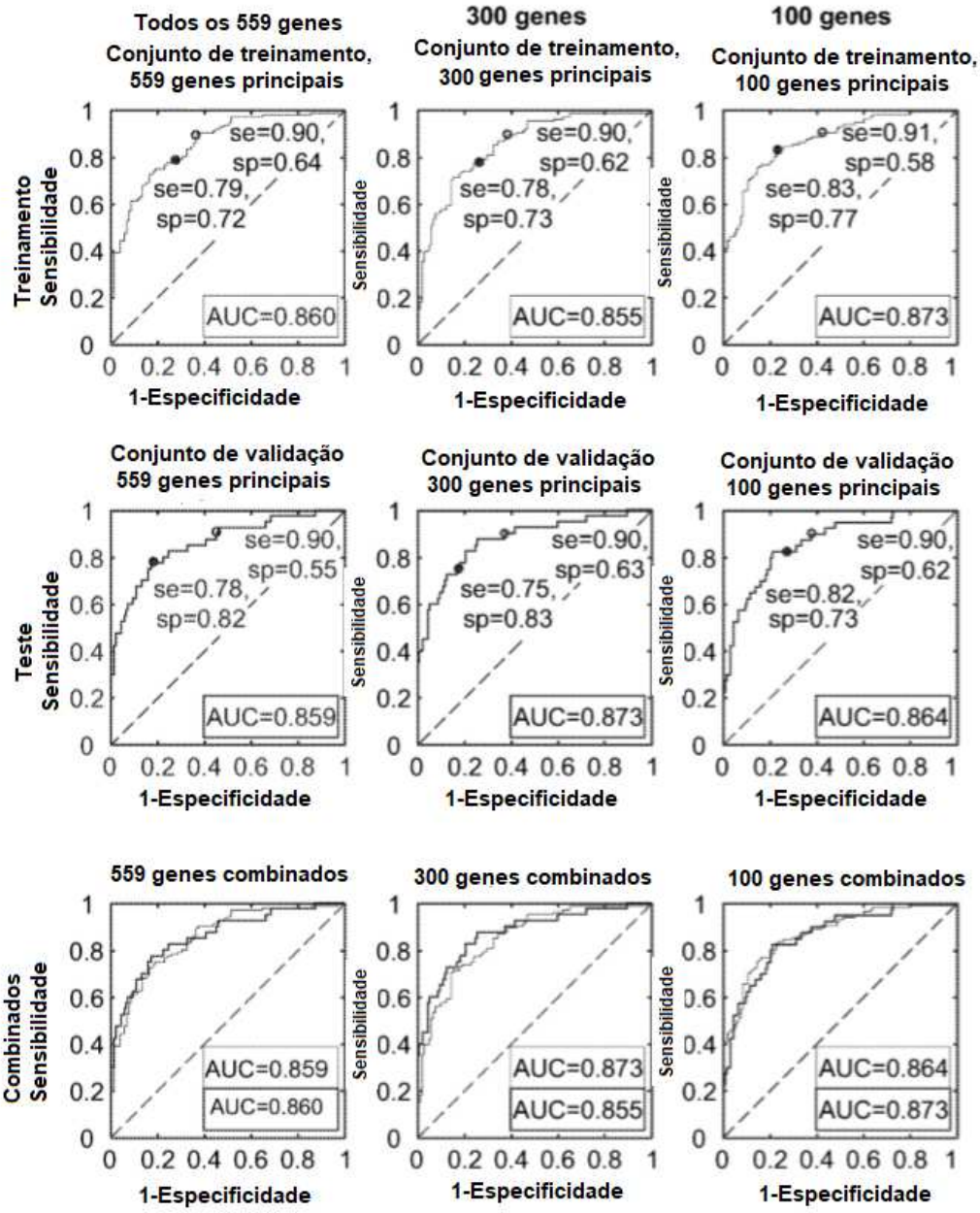

Todos os 559 genes
Conjunto de treinamento, 559 genes principais
se=0.90, sp=0.64
se=0.79, sp=0.72
AUC=0.860
300 genes
Conjunto de treinamento, 300 genes principais
se=0.90, sp=0.62
se=0.78, sp=0.73
AUC=0.855
100 genes
Conjunto de treinamento, 100 genes principais
se=0.91, sp=0.58
se=0.83, sp=0.77
AUC=0.873
Treinamento Sensibilidade
Conjunto de validação 559 genes principais
se=0.90, sp=0.55
se=0.78, sp=0.82
AUC=0.859
Conjunto de validação 300 genes principais
se=0.90, sp=0.63
se=0.75, sp=0.83
AUC=0.873
Conjunto de validação 100 genes principais
se=0.90, sp=0.62
se=0.82, sp=0.73
AUC=0.864
Teste Sensibilidade
559 genes combinados
AUC=0.859
AUC=0.860
300 genes combinados
AUC=0.873
AUC=0.855
100 genes combinados
AUC=0.864
AUC=0.873
Combinados Sensibilidade
Sensibilidade
1-Especificidade
0 0.2 0.4 0.6 0.8 1

FIG. 10A

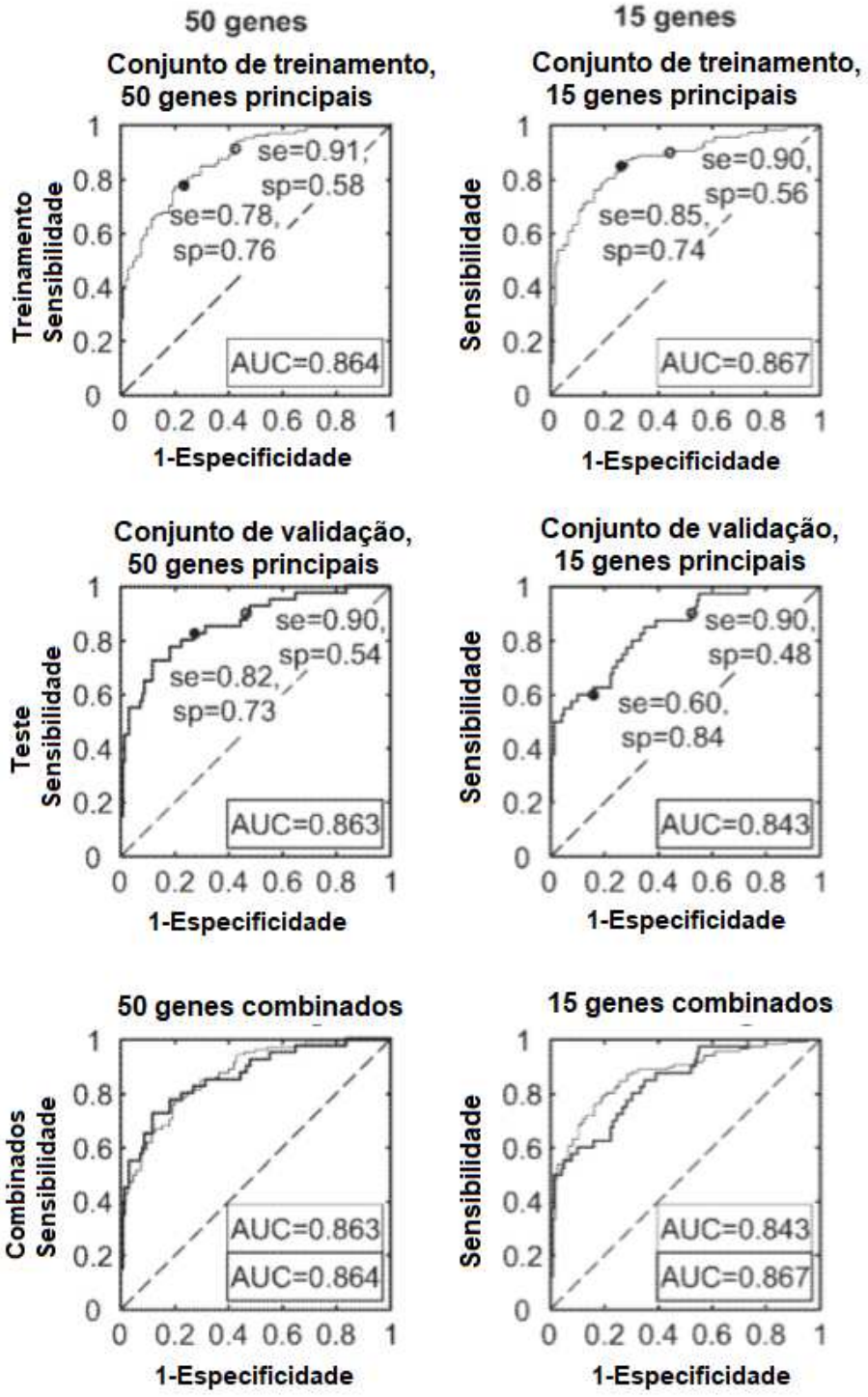

50 genes
Conjunto de treinamento, 50 genes principais
Treinamento Sensibilidade
se=0.91, sp=0.58
se=0.78, sp=0.76
AUC=0.864
1-Especificidade
15 genes
Conjunto de treinamento, 15 genes principais
Sensibilidade
se=0.90, sp=0.56
se=0.85, sp=0.74
AUC=0.867
1-Especificidade
Conjunto de validação, 50 genes principais
Teste Sensibilidade
se=0.90, sp=0.54
se=0.82, sp=0.73
AUC=0.863
1-Especificidade
Conjunto de validação, 15 genes principais
Sensibilidade
se=0.90, sp=0.48
se=0.60, sp=0.84
AUC=0.843
1-Especificidade
50 genes combinados
Combinados Sensibilidade
AUC=0.863
AUC=0.864
1-Especificidade
15 genes combinados
Sensibilidade
AUC=0.843
AUC=0.867
1-Especificidade

FIG. 10B

RESUMO

Patente de Invenção: **"COMPOSIÇÕES E MÉTODOS PARA DIAGNOSTICAR CÂNCERES DE PULMÃO USANDO PERFIS DE EXPRESSÃO DE GENE".**

A presente invenção refere-se a métodos e composições para diagnosticar câncer de pulmão em um sujeito mamífero pelo uso de 7 ou mais genes selecionados, por exemplo, um perfil de expressão de gene do sangue do sujeito que é característico da doença. O perfil de expressão de gene inclui 7 ou mais genes da Tabela I, Tabela II, Tabela III, Tabela IV ou Tabela IX aqui.