(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 397 672**

(51) Int. CI.:

G01N 33/48 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **05.12.2008 E 08859692 (9)**

(97) Fecha y número de publicación de la concesión europea: **31.10.2012 EP 2227691**

(54) Título: **Método de diagnóstico de cánceres de pulmón utilizando perfiles de expresión genética en células mononucleares de sangre periférica**

(30) Prioridad:

05.12.2007 US 5569 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
08.03.2013

(73) Titular/es:

THE WISTAR INSTITUTE OF ANATOMY AND BIOLOGY (50.0%)
3601 SPRUCE STREET
PHILADELPHIA, PA 19104, US y
THE TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA (50.0%)

(72) Inventor/es:

SHOWE, MICHAEL;
SHOWE, LOUISE;
YOUSEF, MALIK;
ALBELDA, STEVEN, M.;
VACHANI, ANIL y
KOSSENKOV, ANDREI V.

(74) Agente/Representante:

DE PABLOS RIBA, Julio

DESCRIPCIÓN

Método de diagnóstico de cánceres de pulmón utilizando perfiles de expresión genética en células mononucleares de sangre periférica

Antecedentes de la invención

El cáncer de pulmón es la causa de mortalidad por cáncer más común en todo el mundo. En los Estados Unidos, el cáncer de pulmón es el segundo cáncer más común tanto en hombres como en mujeres, y asciende a más de 174.000 nuevos casos cada año y más de 162.000 muertes por cáncer. De hecho, el cáncer de pulmón ocasiona más muertes al año que los de pecho, próstata y colo-rectales juntos[2].

La alta mortalidad (80-85% en cinco años), que no ha mostrado ninguna, o muy poca, mejora en los últimos 30 años, pone de relieve el hecho de que se necesitan herramientas nuevas y eficaces que faciliten un pronto diagnóstico con anterioridad a la metástasis en nodos locales o por más allá del pulmón.

Las poblaciones de alto riesgo incluyen los fumadores, los antiguos fumadores, y los individuos con marcadores asociados a predisposiciones genéticas[91-93]. Puesto que la extracción quirúrgica de tumores en fase temprana sigue siendo el tratamiento más eficaz para el cáncer de pulmón, se ha desarrollado un gran interés en proteger los pacientes de alto riesgo con CT espiral de baja dosis (LDCT)[12,14,11,94]. Esta estrategia identifica nódulos pulmonares no calcificados en aproximadamente un 30-70% de individuos de alto riesgo, pero solamente una pequeña proporción de nódulos detectados son finalmente diagnosticados como cánceres pulmonares (0,4 a 2,7%)[16,95,96]. Normalmente, la única manera de distinguir sujetos con nódulos de pulmón de etiología benigna de los sujetos con nódulos malignos es una biopsia invasiva, cirugía u observación prolongada con exploración repetida. Incluso usando los mejores algoritmos clínicos, el 20-55% de los pacientes seleccionados para someterse a biopsia quirúrgica de pulmón para nódulos de pulmón indeterminados, se encuentra que tienen una enfermedad benigna[15] y aquellos que no se someten a biopsia o resección inmediata requieren estudios secuenciales de obtención de imágenes. El uso de CT serie de este grupo de pacientes corre el riesgo de retrasar la potencial terapia curable, junto con los costes de repetición de exploraciones, las dosis de radiación nada despreciables, y la ansiedad del paciente.

Idealmente, una prueba diagnóstica podría ser fácilmente accesible, barata, demostrar una alta sensibilidad y especificidad, y dar como resultado consecuencias mejoradas del paciente (médica y financieramente). Se están realizando esfuerzos por desarrollar diagnósticos no invasivos usando saliva, sangre o suero y analizando sustancias de las células tumorales, ADN[7,8] de tumor metilado, ARN[10] mensajero expresado de polimorfismo de nucleótido simple (SNPs)[9] o proteínas [11]. Esta amplia batería de pruebas moleculares con potencial utilidad para un pronto diagnóstico del cáncer de pulmón, ha sido ya discutida en la literatura. Aunque cada una de estas alternativas tiene sus propios méritos, ninguna de ellas ha superado aún la fase exploratoria en cuanto al esfuerzo por detectar pacientes con cáncer de pulmón en fase temprana, incluso en grupos de riesgo, o pacientes que tienen un diagnóstico preliminar basado en factores radiológicos u otros factores[12] clínicos. Una simple prueba sanguínea, un evento rutinario asociado a visitas regulares a ambulatorios, sería una prueba diagnóstica ideal.

Un método establecido para conseguir el objetivo de diagnóstico genético ha consistido en el uso de signaturas de micro-matriz procedentes de tejido tumoral[20]. Esta alternativa ha sido probada y validada por numerosos investigadores[89]. Un número creciente de estudios ha mostrado que se pueden usar perfiles de células mononucleares de sangre periférica (PBMC) para diagnosticar y clasificar enfermedades sistémicas, incluyendo el cáncer, y para monitorizar la respuesta terapéutica[21]. La validez de usar perfiles de PBMC en pacientes con cáncer ha sido informada previamente en el uso de micro-matrices para comparar PBMC de pacientes con carcinoma de célula renal en fase tardía en comparación con controles normales[20,42]. Una publicación [43] más reciente describe el desarrollo de un clasificador de 37 genes para la detección temprana de un cáncer de pecho a partir de muestras de sangre periférica con un 82% de precisión. Otro estudio identificó perfiles de expresión genética en la PBMC de pacientes de cáncer colo-rectal que podían estar co-relacionados con respuesta a terapia[44]. Algunos de los presentes inventores sugirieron[22] con anterioridad que las quimiocinas y las citocinas liberadas por células malignas podrían imponer una signatura específica de tumor sobre células inmunes de pacientes con cánceres no hematopoyéticos. Se han generado ahora perfiles de expresión genética a partir de PBMC que identifican signaturas asociadas a una diversidad de cánceres, incluyendo el melanoma metastático[23], de pecho[24], renal[25,26] y el cáncer de vejiga[27]. La mayor parte de estos estudios fueron enfocados sobre cánceres en fase posterior o sobre respuesta a terapia y se utilizaron grupos de control sanos más jóvenes a efectos de comparación.

Mientras que el efecto de la enfermedad pulmonar de obstrucción crónica (COPD) sobre la expresión genética de PBMC permanece relativamente sin estudiar hasta la fecha, existen algunos informes limitados acerca del efecto del humo del cigarrillo[33]. La exposición de linfocitos de sangre periférica (PBL) *ex vivo* al humo del cigarrillo indujo muchos cambios en la expresión genética[34]. Los cambios pudieron ser detectados en el transcriptosoma de neutrófilos de la sangre en pacientes de COPD frente a los normales[35]. Un estudio distinguió “entre 85 individuos expuestos y no expuestos al humo del tabaco en base a expresión de mARN en leucocitos periféricos”[36]. No se

encuentra aparentemente ningún dato disponible en cuanto a cambios similares en la sangre que pueda estar presente en antiguos fumadores. Se ha comparado[37] la expresión genética en los epitelios de las vías respiratorias de fumadores, ex-fumadores y no fumadores. Aunque muchas manifestaciones clínicas de fumar se volvieron rápidamente normales después de dejar de fumar, existió un subconjunto de genes cuya expresión permaneció alterada. La hipermetilación[38] genética diferencial y la producción [33] de citocina macrófaga desregulada, han sido vinculadas al humo del cigarrillo. Existen informes de signatura o perfil de expresión genética útil en el diagnóstico del cáncer de pulmón[111].

A pesar de los recientes avances, el reto del tratamiento del cáncer sigue siendo unos regímenes de tratamiento específico objetivo de tipos de tumor patológicamente distintos, y finalmente, personalizar el tratamiento del tumor con el fin de maximizar el resultado. Por ello, existe una necesidad de pruebas que proporcionen simultáneamente información predictiva acerca de respuestas de los pacientes a la diversidad de opciones de tratamiento. En particular, una vez que se diagnostica un cáncer, existe una gran necesidad de métodos que permitan al médico predecir el curso esperado de la enfermedad, incluyendo la probabilidad de recurrencia del cáncer, la supervivencia a largo plazo del paciente, y similares, y seleccionar la opción de tratamiento adecuadamente. También sigue existiendo una necesidad en el estado de la técnica de una prueba menos invasiva que pudiera determinar de manera más precisa el riesgo de enfermedad maligna en pacientes con nódulos pulmonares y que pudiera reducir la cirugía, las biopsias, las exploraciones PET, y/o las exploraciones CT repetidas, que sean innecesarias.

Sumario de la invención

Según un aspecto, se proporciona una composición para evaluar la existencia de un cáncer de pulmón en un mamífero. Esta composición consiste en tres o más polinucleótidos u oligonucleótidos, donde cada polinucleótido u oligonucleótido se hibridiza en un gen, fragmento de gen, transcripción o producto de expresión de gen diferente a partir de células mononucleares de sangre periférica de un mamífero (PMBC). El gen, fragmento de gen, transcripción o producto de expresión de gen se selecciona entre los primeros 29 genes de la Tabla V (mencionados en lo que sigue como “clasificador de 29 genes”, o en un subconjunto de los mismos. Esta realización es particularmente útil para el diagnóstico de un cáncer de pulmón, tal como un NSCLC, y para distinguir entre sujetos con cáncer y sujetos con una enfermedad de pulmón que no sea cáncer.

Según otro aspecto más, la invención se refiere al uso de dicha composición para el diagnóstico *in vitro* de la existencia de un cáncer de pulmón en un sujeto mamífero, que comprende identificar cambios en la expresión de tres o más genes procedentes de células mononucleares de sangre periférica (PBMC) o de la sangre total de dicho sujeto. Los niveles de expresión genética del sujeto de los genes o de la signatura genética seleccionados, se comparan con los niveles de los mismos genes o con el perfil según una referencia o control. Los cambios de expresión de estos genes entre el sujeto y el control están correlacionados con un diagnóstico de cáncer de pulmón.

Otros aspectos y ventajas de estas composiciones y métodos, van a ser descritos con mayor detalle en la descripción que sigue de las realizaciones preferidas de los mismos.

Breve descripción de los dibujos

La Figura 1 es un gráfico de barras que muestra las puntuaciones para 44 muestras (barras oscuras) de pacientes de adenocarcinoma de fase temprana (AC T1T2) y 52 controles no sanos (NHC, indicados mediante barras más claras) que utilizan 15 genes seleccionados mediante SVM-RFE. Véanse los 15 genes de la Tabla IV, columna etiquetada como “AC/NHC”. Las puntuaciones de SVM han sido calculadas como valor promedio a través de todas las puntuaciones de SVM asignadas a una muestra cuando ésta se encuentra en un conjunto de prueba durante validación cruzada. Cada columna representa una muestra. Las barras de error representan la desviación estándar de las clasificaciones sobre los 100 re-muestreos. La curva de ROC para el rendimiento del clasificador de 15 genes produjo una AUC = área bajo curva de 0,92 (curva no representada).

La Figura 2 es un gráfico de barras que muestra la Clasificación SVM de AC + LSCC combinados (NSLC; barras oscuras) y NHC (barras más claras) utilizando los 15 genes seleccionados por medio de SVM-RFE (Tabla IV, columna etiquetada como ALL/NHC). Se han mostrado las puntuaciones discriminantes para las 77 muestras de NSCL y las 52 muestras de NHC. Las barras más claras con puntuaciones positivas son NHC no clasificados, y las barras más oscuras con puntuaciones negativas son muestras de estados no clasificados. La curva ROC para el clasificador de 15 genes produjo una AUC de 0,897 (curva no representada).

La Figura 3 es un gráfico de barras que muestra una comparación por parejas de puntuaciones discriminantes para muestras de pre-cirugía (barras oscuras) y muestras de post-cirugía (muestras claras). Los 15 genes seleccionados por SMV-RFE (véase la Tabla IV, columna etiquetada como “PRE/POST”) fueron usados para asignar puntuaciones discriminantes a las muestras de post-cirugía. Estas puntuaciones han sido mostradas con la puntuación para el mismo paciente dispuesta por parejas de pre-post. Un signo negativo indica que esta muestra es más similar a las muestras de NHC usadas para seleccionar el clasificador de 15 genes.

La Figura 4 es un gráfico de barras que muestra el análisis de SMV-RFE de muestras de pre-cirugía y de post-cirugía. Las 16 muestras de pre-cirugía (barras oscuras) fueron indicadas como de clase positiva y las 16 muestras

de post-cirugía (barras claras) como de clase negativa. El SVM-RFE fue llevado a cabo empezando con los 1.000 genes superiores identificados mediante prueba-t y reducidos después a 1. La conformación clasificadora sobre seis genes (los 6 genes superiores de la Tabla IV, columna etiquetada como PRE/POST, en particular los TSC22D3, CECR4, DNCL1, RPS3, DDIT4, GZMB) dio una precisión global de un 93% y éstos fueron utilizados para generar puntuaciones SVM. La curva ROC para el clasificador de 6 genes produjo una AUC de 0,96 (curva no representada). Se proporcionó una puntuación discriminante a cada muestra (el positivo es indicativo de cáncer de pulmón; el negativo es indicativo de que no existe ningún cáncer). En todas las muestras menos en dos, la puntuación post es más baja que la muestra de pre-cirugía. Estos datos soportan la detección de una signatura de expresión genética relacionada con un tumor que disminuye después de la cirugía. La extensión de esos cambios refleja la posibilidad de recurrencia.

La Figuera 5 es un gráfico que muestra la aplicación del clasificador NSCLC de 29 genes a muestras de PBMC tomadas por resección pre- y post-quirúrgica en 18 pacientes de la Universidad de Pennsylvania.

La Figura 6 es un gráfico que muestra la clasificación de muestras de pre- y post-cirugía con el clasificador de 4 genes (CYUP2R1, MYO5B, DGUOK y DNCL1) ejercitado mediante SVM-RFE con validación cruzada por 10 veces.

Descripción detallada de la invención

Los métodos y composiciones que se describen en la presente memoria aplican tecnología de expresión genética a la selección de sangre para la detección, el diagnóstico y la monitorización de una respuesta al tratamiento del cáncer de pulmón. Las composiciones y los métodos que se describen en la presente memoria permiten el diagnóstico de una enfermedad o de su fase de desarrollo en general, y de cánceres de pulmón en particular, determinando un perfil de ARN característico de los genes de las células mononucleares de sangre periférica (PBMC) o de los linfocitos de sangre periférica (PBL) de un mamífero, con preferencia un sujeto humano. El perfil se establece por comparación de los perfiles de numerosos sujetos de la misma clase (por ejemplo, pacientes con un cierto tipo y fase de desarrollo de cáncer de pulmón, o una mezcla de tipos y fases de desarrollo) con numerosos sujetos de una clase a partir de la cual estos individuos deben ser distinguidos por orden para proporcionar un diagnóstico útil.

Estos métodos de selección de cáncer de pulmón emplean composiciones adecuadas para llevar a cabo una prueba de sangre simple y de bajo coste y no invasiva utilizando un perfilado de expresión genética que podría alertar al paciente y al médico para realizar otros estudios adicionales, tal como una radiografía de pecho o una exploración de CT, de la misma manera que se utiliza el antígeno de próstata específico para ayudar a diagnosticar y seguir los progresos del cáncer de próstata. Los perfiles de expresión genética descritos en la presente memoria proporcionan las bases de una diversidad de clasificaciones relacionadas con este problema de diagnóstico. La aplicación de estos perfiles proporciona diagnósticos solapados y confirmatorios del tipo de la enfermedad de pulmón, empezando con la prueba inicial para enfermedad maligna frente a no maligna.

I. Definiciones

"Paciente" o "sujeto", según se utilizan en la presente memoria, significan un animal mamífero, incluyendo a los humanos, un animal veterinario o de granja, un animal doméstico o una mascota, y los animales usados habitualmente para investigación clínica. En una realización, el sujeto de estos métodos y composiciones es un humano.

"Control" o "sujeto de control", según se utilizan en la presente memoria, se refieren a la fuente de los perfiles de expresión genética de referencia, así como también al panel particular de sujetos de control identificados en los ejemplos que siguen. Por ejemplo, el sujeto de control de una realización puede consistir en controles con cáncer de pulmón, tal como un sujeto que sea un fumador actual o antiguo con enfermedad maligna, un sujeto con un tumor de pulmón sólido con anterioridad a la cirugía para la extracción del mismo; un sujeto con un tumor de pulmón sólido después de la extracción quirúrgica de dicho tumor; un sujeto con un tumor de pulmón sólido con anterioridad a la terapia respecto al mismo; y un sujeto con un tumor de pulmón sólido durante, o después de, la terapia para el mismo. En otras realizaciones, los controles a los efectos de las composiciones y métodos que se describen en la presente memoria, incluyen cualquiera de las clases siguientes de sujeto humano de referencia sin ningún cáncer de pulmón. Tales controles no sanos (NHC) incluyen las clases de fumador con enfermedad no maligna, antiguo fumador con enfermedad no maligna (incluyendo los pacientes con nódulos de pulmón), un no fumador que tiene una enfermedad pulmonar obstructiva crónica (COPD), y un antiguo fumador con COPD. En otras realizaciones más, el sujeto de control es un no fumador sano sin ninguna enfermedad o un fumador sano sin ninguna enfermedad. Todavía en otras realizaciones más, el control o referencia es el mismo sujeto en el que se averiguaron los genes o el perfil genético con anterioridad a la cirugía, o en otro momento anterior para determinar la evaluación de la eficacia quirúrgica o del tratamiento, o la prognosis o el progreso de la enfermedad. La selección de la clase particular de controles depende del uso que el cirujano disponga para los métodos de diagnóstico/monitorización y las composiciones.

En los ejemplos que siguen, el grupo de control seleccionado, con controles no sanos, se elige específicamente de modo que se empareje tan cercanamente como sea posible con los pacientes con enfermedad maligna. El

emparejamiento incluye tanto enfermedades relacionadas con fumar como las relacionadas con el estado del fumador, tal como COPD. Todos los sujetos de ambas clases fueron fumadores, ya sea actuales o ya sea pasados cuando presentaron síntomas de enfermedad. La mayor parte de los genes informativos identificados en lo que sigue pueden distinguir fumadores con enfermedad maligna de fumadores con enfermedad no maligna. Estos genes informativos no incluyen los encontrados previamente para distinguir los fumadores de los no fumadores, por ejemplo CYP1B1, HML2, CCR2, NRG1[36].

"Muestra", según se utiliza en la presente memoria, significa cualquier fluido o tejido biológico que contenga células inmunes y/o células de cáncer. La muestra más adecuada para su uso en la presente invención incluye las células de sangre periférica, y más específicamente las células mononucleares de sangre periférica. Otras muestras biológicas útiles incluyen, sin limitación, la sangre en su conjunto, saliva, orina, fluido sinovial, médula ósea, fluido cerebro-espinal, mucosidad vaginal, mucosidad cervical, secreciones nasales, esputos, semen, fluido amniótico, fluido de lavado bronco-alveolar, y otros exudados celulares procedentes de un paciente que tenga cáncer. Tales ejemplos pueden ser además diluidos con solución salina, solución tampón o con un diluyente fisiológicamente aceptable. Alternativamente, tales muestras son concentradas con medios convencionales.

"Células inmunes", según se utiliza en la presente memoria, significa linfocitos-B, linfocitos-T, células NK, macrófagos, mastocitos, monocitos y células dendríticas.

Según se utiliza en la presente memoria, el término "cáncer" se refiere a, o describe, la condición fisiológica en mamíferos que está típicamente caracterizada por un crecimiento celular no regulado. Más específicamente, el término "cáncer" significa cualquier cáncer de pulmón. En una realización, el cáncer de pulmón es un cáncer de pulmón de células no pequeñas (NSCLC). En una realización específica, el cáncer de pulmón es un adenocarcinoma de pulmón (AC o LAC). En otra realización más específica, el cáncer de pulmón es carcinoma de células escamosas de pulmón (SCC o LSCC). En otra realización, el cáncer de pulmón es un NSCLC de fase I o fase II. En otra realización más, el cáncer de pulmón es una mezcla de fases temprana y tardía y de tipos de NSCLC.

El término "tumor", según se utiliza en la presente memoria, se refiere al crecimiento y la proliferación de células neoplásticas, tanto malignas como benignas, y a todas las células y tejidos pre-cancerosos y cancerosos.

Mediante "diagnosis" o "evaluación" se hace referencia a una diagnosis de cáncer de pulmón, una diagnosis de una fase de desarrollo de cáncer de pulmón, una diagnosis de un tipo o clasificación de un cáncer de pulmón, una diagnosis o detección de una recurrencia de un cáncer de pulmón, una diagnosis o detección de una regresión de un cáncer de pulmón, una prognosis de un cáncer de pulmón, o una evaluación de la respuesta de un cáncer de pulmón a una terapia quirúrgica o no quirúrgica.

Mediante "cambio de expresión" se indica una regulación positiva de uno o más genes seleccionados en comparación con la referencia o control; una regulación negativa de uno o más genes seleccionados en comparación con la referencia o control; o una combinación de ciertos genes de regulación positiva y de regulación negativa.

Mediante "reactivo terapéutico" se indica cualquier tipo de tratamiento empleado en el tratamiento de cánceres con o sin dichos tumores, incluyendo, aunque sin limitación, los productos farmacéuticos quimioterapéuticos, los modificadores de respuesta biológica, la radiación, dieta, terapia vitamínica, terapias hormonales, terapia genética, resección quirúrgica, etc.

Mediante "genes no tumorales", según se utiliza en la presente memoria, se indican genes que están expresados normalmente en otras células, con preferencia células inmunes, de un mamífero sano, y que no son específicamente productos de células tumorales.

Mediante "genes informativos", según se utiliza en la presente memoria, se indican aquellos genes cuya expresión cambia (ya sea de una manera con regulación positiva o ya sea de una manera con regulación negativa), característicamente en presencia de un cáncer de pulmón. Un número estadísticamente significativo de tales genes informativos forma de ese modo perfiles de expresión genética adecuados para su uso en los métodos y composiciones.

El término "número de genes estadísticamente significativo", en el contexto de la presente invención, difiere dependiendo del grado de cambio observado en la expresión genética. El grado de cambio en la expresión genética varía con el tipo de cáncer y con el tamaño o la amplitud del cáncer o del tumor sólido. El grado de cambio varía también con la respuesta inmune del individuo y está sujeto a variación con cada individuo. Por ejemplo, en una realización de la presente invención, un cambio grande, por ejemplo, un incremento o una reducción de 2-3 veces un pequeño número de genes, por ejemplo de 3 a 8 genes característicos, es estadísticamente significativo. Esto es particularmente cierto para cánceres sin tumores sólidos. En otra realización, un cambio relativo más pequeño de alrededor de 10, 20, 24, 29 ó 30 o más genes, es estadísticamente significativo. Esto es particularmente cierto para cánceres con tumores sólidos. También alternativamente, si un gen simple se perfila como expresado o regulado positivamente de manera significativa en células que normalmente no expresan el gen, tal regulación positiva de un

gen simple puede ser por sí solo estadísticamente significativo. A la inversa, si un gen simple se perfila como regulado negativamente o no expresado de manera significativa en células que normalmente expresan el gen, tal regulación negativa de un gen simple puede ser por sí solo estadísticamente significativa. Como ejemplo, un gen simple, que se expresa en torno al mismo en todos los miembros de una población de pacientes, es regulado de forma negativa 4 veces solamente en un 1% de los individuos sin cáncer. Para tales genes regulados independientemente en un individuo, la totalidad de las 4 veces regulado de forma descendente, podría ocurrir con una probabilidad de solamente una vez en 100 millones. Por lo tanto, estos 4 genes son un número estadísticamente significativo de genes para ese cáncer. Alternativamente, si la varianza normal es más alta, por ejemplo una persona sana entre 10 tiene el gen regulado negativamente 4 veces, entonces se requiere un panel más grande para detectar la varianza para un cáncer particular.

De ese modo, los métodos y las composiciones que se describen en la presente memoria contemplan el examen del perfil de expresión de un "número de genes estadísticamente significativo" comprendido en la gama de 1 a alrededor de 100 genes en un perfil simple. En una realización, el perfil de gen se forma mediante un número estadísticamente significativo de 1 o más genes. En otra realización, el perfil de gen se forma mediante un número estadísticamente significativo de 3 o más genes. En otra realización más, el perfil de gen se forma mediante 4 o más genes. En otra realización más, el perfil de gen se forma mediante al menos 5 a 15 o más genes. En otra realización más, el perfil de gen se forma mediante 24 ó 29 o más genes. En otras realizaciones adicionales, los perfiles de gen examinados como parte de estos métodos, particularmente en casos en los que los cánceres están caracterizados por tumores sólidos, contienen, como números de genes estadísticamente significativos, desde 5, 10, 15, 20, 30, 40, 50, 60, 70, 80 ó 90 o más genes en un panel, y números cualesquiera entre ellos.

Las Tablas I a VII que siguen se refieren a colecciones de genes conocidos. Las Tablas I, II y III incluyen los 100 genes superiores de cada clasificación, identificada por los inventores como capaz de formar un perfil de expresión genética para tres clasificaciones de enfermedad distintas. La Tabla I identifica los 100 genes superiores que pueden ser usados en un perfil de expresión genética para identificar la presencia de un cáncer de pulmón, por ejemplo, cualquier NSCLC. La Tabla II identifica los 100 genes superiores que pueden ser usados en un perfil de expresión genética para distinguir la ocurrencia de un cáncer de pulmón, y en una realización son útiles para distinguir AC de cualquier otro NSCLC. La Tabla III identifica los 100 genes superiores que pueden ser usados en un perfil de expresión genética para identificar los cambios consecuentes con una mejora post-quirúrgica de, y/o el mantenimiento de una mejora post-quirúrgica de, un cáncer de pulmón, tal como un NSCLC. Se entiende que esta última colección de genes es útil para una mejora del rastreo durante o después de un tratamiento terapéutico de un cáncer de pulmón, tal como NSCLC (es decir, tomado a partir de la Tabla I), para identificar un AC (es decir, tomado a partir de la Tabla II), y para identificar el estado post-quirúrgico de un sujeto (es decir, tomado a partir de la Tabla III).

La Tabla V identifica 136 genes adicionales útiles para la formación de perfiles genéticos para su uso en el diagnóstico de pacientes con un cáncer de pulmón, tal como un NSCLC, a partir de un control, en particular controles no sanos. Los 29 genes de clasificación más alta en esta tabla, se ha referenciado como "el clasificador de 29 genes" en los Ejemplos 14-18 que siguen. La Tabla IV identifica otro conjunto de 50 genes útiles en un perfil de expresión de gen para identificar los cambios consecuentes con la mejora post-quirúrgica de, y/o con el mantenimiento de la mejora post-quirúrgica de, un cáncer de pulmón. De manera similar, estos genes son útiles como signatura genética para monitorizar la progresión o la regresión del cáncer en un paciente tratado de forma no quirúrgica respecto a un cáncer de pulmón. La Tabla VII identifica un conjunto de 24 genes útiles para la discriminación entre un sujeto que tiene un cáncer de pulmón, por ejemplo un NSCLC, y sujetos que tienen nódulos de pulmón benignos (no malignos).

Los genes identificados en las Tablas I a VII están disponibles públicamente. Un experto en la materia puede reproducir fácilmente los métodos y las composiciones que se describen en la presente memoria con el uso de las secuencias de genes, todas ellas disponibles públicamente a partir de fuentes convencionales, tal como GenBank.

El término "micro-matriz" se refiere a una disposición ordenada de elementos de matriz hibridizables, con preferencia sondas de polinucleótido o de oligonucleótido, sobre un substrato.

El término "polinucleótido" cuando se utiliza de una forma en singular o en plural, se refiere en general a cualquier polirribonucleótido o polidesoxirribonucleótido, que puede ser un ARN o un ADN no modificado o un ARN o un ADN modificado. Así, por ejemplo, los polinucleótidos según se definen en la presente memoria incluyen, aunque sin limitación, ADN de simple y doble cadena, ADN que incluye regiones de simple y doble cadena, ARN de simple y doble cadena, y ARN que incluye regiones de simpe y doble cadena, moléculas híbridas que comprenden ADN y ARN que pueden ser de cadena simple o, más típicamente, de doble cadena o que incluyen regiones de simple y de doble cadena. Adicionalmente, el término "polinucleótido" según se utiliza en la presente memoria, se refiere a regiones de triple cadena que comprenden ARN o ADN o ambos ARN y ADN. Las cadenas de tales regiones pueden proceder de la misma molécula o de moléculas diferentes. Las regiones pueden incluir todas de entre una o más moléculas, pero típicamente incluyen solamente una región de algunas de las moléculas. Una de las moléculas de una región de triple hélice es con frecuencia un oligonucleótido. El término "polinucleótido" incluye específicamente cADNs. El término incluye ADNs (incluyendo los cADNs) y ARNs que contienen una o más bases

modificadas. Así, los ADNs o ARNs con esqueletos modificados por motivos de estabilidad o por otras razones, son "polinucleótidos" tal y como se entiende dicho término en la presente memoria. Además, los ADNs o los ARNs que comprenden bases inusuales, tal como la inosina, o bases modificadas, tal como las bases tritiadas, están incluidas dentro del término "polinucleótidos" según se define en la presente memoria. En general, el término "polinucleótido" abarca todas las formas modificadas química, enzimática y/o metabólicamente de polinucleótidos sin modificar, así como también las formas químicas de ADN y ARN características de virus y células, incluyendo las células simples y las complejas.

El término oligonucleótido se refiere a un polinucleótido relativamente corto, incluyendo, aunque sin limitación, los desoxirribonucleótidos de cadena simple, los ribonucleótidos de cadena simple o doble, los híbridos de ARN:ADN y los ADNs de doble cadena. Los oligonucleótidos, tal como los oligonucleótidos de sonda de ADN de cadena simple, son sintetizados con frecuencia por medio de métodos químicos, por ejemplo utilizando sintetizadores de oligonucleótido automatizados que se encuentran comercialmente disponibles. Sin embargo, los oligonucleótidos pueden ser realizados mediante una diversidad de otros métodos, incluyendo técnicas mediadas por ADN recombinantes in vitro y mediante expresión de ADNs en células y organismos.

Los términos "gen expresado diferencialmente", "expresión de gen diferencial" y sus sinónimos, que son utilizados intercambiablemente, se refieren a un gen cuya expresión está activada a un nivel más alto o más bajo en un sujeto que adolece de una enfermedad, específicamente cáncer, tal como cáncer de pulmón, en relación con su expresión en un sujeto de control. Los términos incluyen también genes cuya expresión es activada a un nivel más alto o más bajo en diferentes fases de desarrollo de la misma enfermedad. También se entiende que un gen expresado diferencialmente puede ser o bien activado o bien inhibido a nivel del ácido nucleico o a nivel de una proteína, o puede estar sujeto a partición alternativa para dar como resultado un producto polipéptido diferente. Tales diferencias pueden ser puestas en evidencia mediante un cambio en los niveles de mARN, en la expresión superficial, en la secreción o en otra partición de un polipéptido, por ejemplo. La expresión genética diferencial puede incluir una comparación de expresión entre dos o más genes o entre sus productos genéticos, o una comparación de las relaciones de la expresión entre dos o más genes o entre sus productos genéticos, o incluso una comparación de dos productos del mismo gen procesados de manera diferente, que difieran entre sujetos normales, controles no sanos y sujetos que sufran una enfermedad, específicamente cáncer, o entre varias fases de desarrollo de la misma enfermedad. La expresión diferencial incluye diferencias tanto cuantitativas como cualitativas en cuanto al patrón de expresión temporal o celular en un gen o en sus productos de expresión entre, por ejemplo, células normales y enfermas, o entre células que han sido sometidas a diferentes eventos de enfermedad o de etapas de desarrollo de enfermedad. A los efectos de esta invención, se considera que la "expresión genética diferencial" está presente cuando existe una diferencia estadísticamente significativa ($p < 0,05$) en expresión de gen entre el sujeto y las muestras de control.

El término "sobre-expresión" con relación a una transcripción de ARN, se utiliza para referirse al nivel de transcripción determinado por normalización respecto a los niveles de mARNs de referencia, que podría estar en todas las transcripciones medidas en la muestra o en un conjunto de referencia particular de mARNs.

La frase "amplificación genética" se refiere a un proceso mediante el que se forman múltiples copias de un gen o de un fragmento de gen en una célula o línea celular particular. La región duplicada (un tramo de ADN amplificado) se denomina con frecuencia como "amplicon". Habitualmente, la cantidad de ARN mensajero (mARN) producida, es decir, el nivel de expresión genética, se incrementa también en la proporción del número de copias realizadas de un gen particular expresado.

El término "prognosis" se utiliza en la presente memoria para referirse a la predicción de la probabilidad de progresión o muerte atribuida a cáncer, incluyendo recurrencia, expansión metastática, y resistencia a los medicamentos, de una enfermedad neoplástica, tal como un cáncer de pulmón. El término "predicción" se utiliza en la presente memoria para referirse a la probabilidad de que un paciente responda ya sea favorablemente o ya sea desfavorablemente a un medicamento o conjunto de medicamentos, y también a la extensión de estas respuestas, o de que un paciente sobreviva, a continuación de la extracción quirúrgica del tumor principal y/o de la quimioterapia durante un cierto período de tiempo sin recurrencia del cáncer. Los métodos predictivos de la presente invención pueden ser usados clínicamente para tomar decisiones de tratamiento eligiendo las modalidades de tratamiento más apropiadas para cualquier paciente particular. Los métodos predictivos descritos en la presente memoria son herramientas valiosas para la predicción de si es probable que un paciente responda favorablemente a un régimen de tratamiento, tal como una intervención quirúrgica, quimioterapia con un medicamente dado o una combinación de medicamentos, y/o terapia de radiación, o de si es probable la supervivencia a largo plazo del paciente, a continuación de de la cirugía y/o de la terminación de la quimioterapia o de otras modalidades de tratamiento.

El término supervivencia "a largo plazo" se utiliza en la presente memoria para referirse a una supervivencia de al menos 1 año, más preferentemente al menos 3 años, más preferentemente al menos 7 años, después de la cirugía o de otro tratamiento.

El "rigor" de las reacciones de hibridación es fácilmente determinable por un experto en la materia, y en general es un cálculo empírico dependiente de la longitud de la sonda, la temperatura de lavado, y la concentración de sal. En general, cuanto más largas sean las sondas más altas son las temperaturas requeridas para una fijación por calor

apropiada, mientras que las sondas más cortas necesitan temperaturas más bajas. La hibridación depende en general de la capacidad del ADN desnaturalizado para re-fijarse por calor cuando se encuentran presentes cadenas complementarias en un entorno por debajo de su temperatura de fusión. Cuanto más alto sea el grado de homología deseada entre la sonda y la secuencia hibridizable, más alta es la temperatura relativa que puede ser usada. Como resultado, se deduce que las temperaturas relativas más altas podrían tender a hacer que las condiciones de reacción sean más rigurosas, mientras que las temperaturas más bajas lo son menos. Diversos textos publicados[69,77] proporcionan detalles adicionales y una explicación de la rigurosidad de las reacciones de hibridación.

"Condiciones rigurosas" o "condiciones de alta rigurosidad", según se definen en la presente memoria, típicamente: (1) emplean una baja resistencia iónica y una alta temperatura para el lavado, por ejemplo cloruro de sodio 0,015 M/citrato de sodio 0,0015 M/0,1% de dodecil sulfato de sodio a 50 ºC; (2) emplean durante la hibridación un agente desnaturalizador, tal como formamida, por ejemplo, un 50% (v/v) de formamida con un 0,1% de albúmina de suero bovino/0,1% de Ficoll/0,1% de polivinilpirrolidona/50 mM de solución tampón de fosfato de sodio a un pH de 6,5 con 750 mM de cloruro de sodio, 75 mM de citrato de sodio a 42 ºC; o (3) emplean un 50% de formamida, 5XSCC (0,75 M NaCl, 0,075 M de citrato de sodio), 50 mM de fosfato de sodio (pH 6,8), 0,1% de pirofosfato de sodio, 5X solución de Denhardt, ADN de esperma de salmón sonicado (50.mu.g/ml), 0,1% de SDS y 10% de sulfato dextrano a 42 ºC, con lavados a 42 ºC en 0,2XSSC (cloruro de sodio/citrato de sodio) y 50% de formamida a 55 ºC, seguido de un lavado de alta rigurosidad consistente en 0,1XSSC que contenía EDTA a 55 ºC.

Las "condiciones moderadamente rigurosas" pueden ser identificadas convencionalmente[70] e incluyen el uso de solución de lavado y condiciones de hibridación (por ejemplo, temperatura, resistencia iónica y % de SDS) menos rigurosas que las descritas en lo que antecede. Un ejemplo de condiciones moderadamente rigurosas consiste en incubación durante la noche a 137 ºC en una solución que comprende: 20% de formamida, 5XSSC (150 mM de NaCl, 15 mM de citrato de sodio), 50 mM de fosfato de sodio (pH 7,6), 5X solución de Denhardt, 10 % de sulfato dextrano, y 20 mg/ml de ADN de esperma de salmón cortador desnaturalizado, seguido de lavado de los filtros en 1XSSC a alrededor de 37-50 ºC. El experto reconocerá cómo ajustar la temperatura, la resistencia iónica, etc., según sea necesario para acomodar factores tales como longitud de la sonda y similares, mediante el uso de las instrucciones del fabricante (véase, por ejemplo, instrucciones del sistema Illumina).

En el contexto de las composiciones y métodos descritos en la presente memoria, la referencia a "tres o más", "al menos cinco", etc., de los genes listados en cualquier conjunto particular de genes (por ejemplo, Tablas I a VII) significa una cualquiera o cualquiera y todas las combinaciones de los genes listados. Por ejemplo, los perfiles de expresión genética adecuados incluyen perfiles que contienen cualquier número entre al menos 3 a 100 genes de esas Tablas. En una realización, los perfiles genéticos formados por genes seleccionados a partir de una tabla son utilizados preferentemente por orden jerárquico, por ejemplo los de la parte superior de la lista demostraron resultados discriminatorios más significativos en las pruebas, y de ese modo pueden ser más significativos en un perfil que los genes de orden más bajo. Sin embargo, en otras realizaciones, los genes que forman un perfil genético útil no tienen que estar en orden jerárquico y puede ser cualquier gen de la tabla respectiva.

Los términos "partición" y "partición de ARN" se utilizan de manera intercambiable y se refieren a procesamiento de ARN que extrae intrones y une exones para producir mARN maduro con secuencia de codificación continua que se mueve en el citoplasma de una célula eucariótica.

En teoría, el término "exón" se refiere a cualquier segmento de un gen interrumpido que esté representado en el producto de ARN maduro[71]. En teoría, el término "intrón" se refiere a cualquier segmento de ADN que esté transcrito pero extraído del interior de la transcripción mediante partición junto con los exones de cualquier lado del mismo. Operativamente, ocurren secuencias de exones en la secuencia de ARN de un gen. Operativamente, las secuencias de intrones son las secuencias intervinientes dentro del ADN genómico de un gen, encorchetado mediante secuencias de exones y que tienen secuencias de consenso de partición GT y AG en sus límites 5' y 3'.

Según se utiliza en la presente memoria, "etiquetas" o "moléculas informadoras" son porciones químicas o bioquímicas útiles para etiquetar un ácido nucleico (incluyendo un nucleótido simple), un polinucleótido, oligonucleótido, o ligando proteínico, por ejemplo un aminoácido o anticuerpo. Las "etiquetas" y los "moléculas informadoras" incluyen agentes fluorescentes, agentes quimioluminiscentes, agentes cromogénicos, agentes de extinción, radionucleótidos, enzimas, substratos, co-factores, inhibidores, partículas magnéticas y otras porciones conocidas en el estado de la técnica. Las "etiquetas" o "moléculas informadoras" son capaces de generar una señal medible y pueden estar unidas covalentemente o monovalentemente a un oligonucleótido o un nucleótido (por ejemplo, un nucleótido no natural) o un ligando.

A menos que se defina de otro modo en la presente descripción, los términos técnicos y científicos usados en la misma tienen el mismo significado que los entendidos habitualmente por un experto en la materia a la que pertenece la presente invención y con referencia a los textos publicados[72,73], los cuales proporcionan a un experto en la materia una guía general respecto a muchos de los términos usados en la presente solicitud.

II. Los perfiles de expresión genética

Los inventores identificaron perfiles de expresión genética de diagnóstico en los linfocitos de sangre periférica de pacientes de cáncer de pulmón. Los inventores han descubierto que los perfiles de expresión genética de las PBMCs de pacientes de cáncer de pulmón difieren significativamente de los observados en controles apropiadamente emparejados (por ejemplo, por edad, sexo, historia de fumador). Por ejemplo, se pueden observar y detectar cambios en los productos de expresión genética de los genes de estos perfiles mediante los métodos de la presente invención en la PBMC circulante normal de pacientes con tumores de pulmón sólidos en fase temprana.

Los perfiles de expresión genética descritos en la presente solicitud proporcionan nuevos marcadores de diagnóstico para la pronta detección de cáncer de pulmón y podrían evitar que los pacientes sean sometidos innecesariamente a procedimientos (es decir, si se descubre un pequeño nódulo de pulmón) o ser usados potencialmente para proteger pacientes de alto riesgo. Puesto que los riesgos son muy bajos, la relación entre el beneficio y el riesgo es muy alta. Los métodos y las composiciones que se describen en la presente memoria pueden ser útiles también en otras poblaciones, es decir, para proteger a ciertas poblaciones de cáncer de pulmón de alto riesgo, tal como los fumadores expuestos a amianto. En otra realización más, los métodos y las composiciones que se describen en la presente memoria pueden ser usados junto con factores de riesgo clínico para ayudar a los médicos a tomar decisiones más precisas respecto a cómo manejar pacientes con nódulos de pulmón. Otra ventaja de la presente invención consiste en que la diagnosis puede realizarse tempranamente puesto que la diagnosis no es dependiente de la detección de células tumorales circulantes que estén presentes en sólo un pequeño número de canceres de pulmón en fase temprana de desarrollo.

Puesto que los efectos de las enfermedades pulmonares obstructivas del tabaco y/o crónicas sobre el perfil de PBMC tienen el potencial de obscurecer los resultados de los métodos de diagnóstico basados en perfiles genéticos, según se detalla en lo que sigue, los efectos de un fumador actual, un antiguo fumador, y una COPD, son direccionados específicamente en las composiciones y métodos de la presente invención mediante el uso de poblaciones apropiadas de controles de emparejamiento. En una realización, la clase de control apropiado para los estudios comparativos consiste en los fumadores de riesgo y en ex-fumadores con enfermedad pulmonar no maligna, de modo que las historias relacionadas con el tabaco de ambos sujeto paciente y sujetos de control, son muy similares. Los datos que se presentan en los ejemplos que siguen indican claramente que los inventores detectan una signatura de cáncer en presencia de una actividad de fumador y/o de COPD.

En una realización, un novedoso perfil o signatura de expresión genética puede identificar y distinguir pacientes en fase de desarrollo temprana (T1/T2 -principalmente Fase de desarrollo I/II) de canceres de células no pequeñas de pulmón (NSCLC) a partir del grupo de control adecuado de fumadores y ex-fumadores en alto riesgo respecto al desarrollo de cáncer de pulmón por la edad, el género y la raza. Véase, por ejemplo, los genes identificados en la Tabla I que pueden formar un perfil adecuado de expresión genética y los de la Tabla IV, columna "TODO/NHC". En otra realización, un novedoso perfil o signatura de expresión genética puede identificar pacientes con tumores de AC (principalmente, Fases de desarrollo I y II) en fase de desarrollo temprana (T1/T2), en comparación con el control de NHC relacionado de forma cercana. Véase la tabla II y la Tabla IV, columna "AC/NHC". La validez de estos métodos y de los perfiles de expresión genética, está soportada en datos experimentales que miden la "puntuación" del cáncer de pulmón en pacientes antes y después de la cirugía. En otra realización, las colecciones de genes de la Tabla III y la Tabla IV, columna PRE/POST, proporcionan un número discreto de genes que forman un perfil adecuado. Estas poblaciones de paciente/control se distinguieron generando una puntuación discriminante basada en diferencias en perfiles de expresión genética según se ha ejemplificado en lo que sigue. En una realización, un clasificador de 15 genes, es decir, un conjunto de genes que forman un perfil de expresión genética, puede distinguir entre tumores de AC en fase de desarrollo temprana y perfiles de control no sanos con una precisión de un 85%. Ese perfil de expresión genética ha sido identificado en la Tabla IV, columna "TODO/NHC", que sigue. Adicionalmente, los inventores han identificado un clasificador de perfil de expresión genética que distingue ambos pacientes de AC y de LSCC de los NHC con una precisión de un 83% necesitando también 15 genes para el perfil. Ese perfil de expresión genética se ha identificado en la Tabla IV, columna "AC/NHC" que sigue. Un perfil de expresión genética similar para distinguir pacientes de pre-cirugía de los de post-cirugía, se encuentra también en la Tabla IV, columna "PRE/POST" que sigue. Los datos mostrados en los ejemplos indican claramente que existe una signatura específica de cáncer en fase de desarrollo temprana compartida que se aparta del patrón que discrimina los tipos de cáncer (AC frente a LSCDC) y que discrimina las fases de desarrollo del cáncer (temprana frente a tardía).

Datos más recientes descritos en los Ejemplos 14-18 que siguen proporcionan una nueva signatura de expresión de 29 genes para diagnosticar sujetos con cáncer de pulmón a partir de controles sanos o no sanos (Tabla V, genes clasificados como 1-29), así como genes adicionales a partir de esa tabla que pueden formar otras signaturas. El panel relativamente pequeño de 29 genes puede distinguir NSCLC en fase de desarrollo temprana (Fase de desarrollo 1A - 1B) de un grupo de control altamente similar con una buena precisión. Adicionalmente, un conjunto de 4 genes procedentes de la selección de 50 genes de la Tabla VI, resulta útil para distinguir y rastrear una mejora post-quirúrgica. Además, un nuevo perfil de expresión de 24 genes para discriminar entre sujetos con cáncer de pulmón y sujetos con nódulos de pulmón benignos, ha sido proporcionado en la Tabla VII. Los datos mostrados en estos ejemplos muestran signaturas genéticas de cáncer de pulmón útiles tanto en diagnosis como en evaluación

del progreso del tratamiento.

Según se describe en detalle en los ejemplos que siguen, comparando expresión genética en PBMC de un grupo amplio de pacientes de NSCLC con un grupo comparable de pacientes con enfermedades de pulmón no malignas, se detectó una signatura tumoral inducida, en fumadores y no fumadores, que puede ser distinguida de los efectos de la enfermedad pulmonar no maligna inducida por fumar. Según se muestra en los ejemplos que siguen, se han identificado signaturas de diagnóstico en PBMC que distinguen pacientes con NSCLC en fase de desarrollo temprana frente a controles en riego con enfermedad de pulmón no maligna uniformizados en cuanto a fumar, la edad, el género, así como la incidencia de COPD. Existían también 14 pacientes de NSCLC en estos ejemplos que no tenían ninguna historia anterior como fumadores. El cáncer de pulmón en individuos que nunca habían fumado ha mostrado tener varias diferencias importantes con los tumores de pulmón asociados al tabaco, y algunos cambios moleculares que ocurrieron han sugerido ser únicos en no fumadores[28,29]. 11 de los 14 que nunca fumaron fueron clasificados correctamente como cáncer por medio del clasificador de 29 genes, sugiriendo que el efecto sobre la expresión genética de PBMC de cánceres de pulmón en fumadores y no fumadores es similar, al menos con respecto a las signaturas de gen de PBMC.

Catorce genes asociados al metabolismo de la nicotina y de la nicotinamida fueron estadísticamente pacientes de NSCLC significativamente más baja cuando se compararon con todos los controles o se compararon solamente con controles con nódulos de pulmón benignos sugiriendo que estas vías de acceso pueden ser suprimidas en pacientes de NSCLC. Las diferencias detectadas en la PBMC entre pacientes antes y después de la resección quirúrgica, fueron numerosas. Sin embargo, 2 de los 4 genes más informativos que distinguen las muestras de pre- frente a las de post- cirugía, tienen funciones mitocóndricas. Los genes mitocóndricos son en general de pre-cirugía más alta sugiriendo que los requisitos de energía incrementada descritos para los tumores están también reflejados en la PBMC cuando el tumor está presente. Las vías de acceso altamente significativas que eran más altas en muestras de pre-cirugía fueron asociadas a la función celular NK, y a la señalización de ceramida. [NK: 29 genes ($p<2{,}08 \times 10^{-8}$), caramida: 17 genes ($p<8{,}83 \times 10.^{-5}$)]. Las vías de acceso con regulación más significativamente descendente incluían genes receptores de apoptosis y de muerte (Apoptosis: 15 genes ($p<1{,}74 \times 10^{-2}$), receptor de muerte: patrones de 13 genes ($p<1{,}37 \times 10^{-3}$) también característicos de tumores[31,32]. La reducción observada de la signatura de cáncer de NSCLC y las diferencias comunes altamente significativas mostradas por post-cirugía de pacientes, soportan la conclusión de que las signaturas descritas en la presente memoria son inducidas por tumores.

Las interacciones específicas entre el tumor, los linfocitos y factores derivados de tumor, contribuyen a los cambios vistos en la expresión genética de PBMC y estos efectos se ven incrementados en la progresión tumoral, según se evidencia mediante la precisión incrementada de nuestro panel genético en la clasificación de NSCLC de fase de desarrollo tardía.

La validez de estas signaturas fue establecida sobre muestras recogidas en diferentes localizaciones por diferentes grupos y en un conjunto de pacientes con nódulos de pulmón sin diagnosticar. Los perfiles de expresión genética identificados en lo que sigue mediante el uso de matrices ILLUMINA proporcionan signaturas de pronóstico global para identificar pacientes con cáncer de pulmón de varios tipos de células, y proporcionan signaturas de diagnóstico específicas del tipo de célula. Además, los perfiles que tienen en cuenta la raza, el género y la historia del fumador. Los inventores han probado también muestras procedentes de un grupo de pacientes antes y después de la cirugía del cáncer, eliminando de ese modo la variabilidad de persona a persona en cuanto a la evaluación del efecto del tumor. La consistencia de la signatura del cáncer de pulmón disminuye o desaparece tras la extracción del tumor. Este resultado, según se discute en los ejemplos que siguen, soporta considerablemente la identificación de una signatura de PBMC para un cáncer de pulmón en fase de desarrollo temprana. Este dato (véase el Ejemplo 12) muestra una reducción consecuente en cada puntuación de cáncer de pulmón del paciente tras la eliminación quirúrgica del cáncer en comparación con esa puntuación con anterioridad a la cirugía.

Las signaturas de cáncer de pulmón o perfiles de expresión genética identificados en la presente memoria y mediante el uso de las colecciones de genes de las Tablas I-VII, pueden ser optimizados para reducir los números de productos de expresión genética necesarios e incrementar la precisión del diagnóstico.

Aunque no se pretende quedar limitados por la teoría, el uso por los inventores de estudios de expresión genética de PBMC en la enfermedad se basa en la proposición de que la PBMC circulante (células mononucleares de sangre periférica - principalmente monocitos y linfocitos) está afectada por procesos localizados que incluyen inflamación y/o tumores. Esto puede ocurrir mediante al menos dos mecanismos. En primer lugar, las células pueden interactuar directamente en los tejidos de la inflamación o tumor. De manera clara, una función clave de los linfocitos consiste en "patrullar" los tejidos del cuerpo, detenerse temporalmente en zonas anormales, salir de los tejidos, interactuar con los tejidos nodales linfoideos, volverse activos, y a continuación re-entrar en la circulación (re-entrando algo en los tejidos). Esta interacción cercana altera claramente su fenotipo. Un segundo proceso, y probablemente igual de importante, consiste en la respuesta de la PBMC a factores circulantes liberados por las células en la respuesta inflamatoria o los tumores. Muchos de esos factores han sido ya descritos, incluyendo factores de estimulación de colonias (tal como G-CSF, GM-CSF), citocinas (es decir, TNF, IL-2, IL-3, IL-4 e IL-, IL-7, IL-15, etc.), quimiocinas (MCP-1, SDF-1), factores de crecimiento (tales como el ligando Fit-3, VEGF), factores de inmunosupresión (tal como IL-10, COX-1, TGF-β), etc. Estos factores afectan a células inmaduras de la médula ósea que son liberadas después

a la circulación, así como también a células que están ya en el compartimento circulante. Este último mecanismo afecta asimismo tanto al fenotipo de células liberadas como al tipo de células liberadas (es decir, muy pronto después de la infección existe un flujo entrante de neutrófilos inmaduros en la circulación).

Aunque las lesiones inflamatorias y los tumores tienen algunas similitudes, existen muchas diferencias, de las que una muy importante es la bien conocida capacidad de los tumores de suprimir respuestas inmunes. Las signaturas del cáncer establecidas por los perfiles de expresión genética descritos en la presente memoria pueden ser diferenciadas de una signatura inflamatoria.

III. Métodos de obtención de perfil de expresión genética

Los métodos de obtención de perfil de expresión genética que se usaron para generar los perfiles útiles en las composiciones y métodos descritos en la presente memoria o en la realización de etapas de diagnóstico utilizando las composiciones descritas en la presente memoria, son conocidos y se encuentran resumidos en la Patente US núm. 7.081.340. Tales métodos de obtención de perfil de expresión genética incluyen métodos basados en análisis de hibridación de polinucleótidos, métodos basados en secuenciación de polinucleótidos, y métodos de base proteínica. Los métodos utilizados más comúnmente, conocidos en la técnica de la cuantificación de expresión de mARN en una muestra, incluyen el uso de la técnica *northern blot* y la hibridación *in situ*[74]; ensayos de protección de ARNs[75]; y métodos basados en PCR, tal como PCR[76]. Alternativamente, se pueden emplear anticuerpos que reconozcan dúplex específicos, incluyendo los dúplex de ADN, los dúplex de ARN, y los dúplex híbridos de ADN-ARN o los dúplex de ADN-proteína. Métodos representativos para análisis de expresión genética basada en secuenciación incluyen Análisis Serie de Expresión Genética (SAGE) y análisis de expresión genética por secuenciación de signatura masivamente paralela (MPSS).

A. Técnicas de reacción de cadena de polimerasa (PCR)

El método cuantitativo más sensible y más flexible es el RT-PCR, el cual puede ser usado para comparar niveles de mARN en diferentes poblaciones de muestra, en tejidos normales y tumorales, con o sin tratamiento con medicamentos, para caracterizar patrones de expresión de gen, para discriminar entre mARNs relacionados de manera cercana, y para analizar estructuras de ARN. La primera etapa consiste en el aislamiento de mARN a partir de una muestra objetivo (por ejemplo, típicamente el ARN total aislado a partir de PBMC humana en este caso). El mARN puede ser extraído, por ejemplo, a partir de muestras de tejido congeladas o incrustadas en parafina archivadas y fijadas (por ejemplo, fijadas en formalina).

Los métodos generales para la extracción de mARN son bien conocidos en el estado de la técnica, tal como en libros de texto de biología molecular[77]. Los métodos para extracción de ARN a partir de tejidos incrustados en parafina son también conocidos[78,79]. En particular, se puede realizar el aislamiento de ARN utilizando el kit de purificación, el conjunto tampón y proteasa procedente de fabricantes comerciales, de acuerdo con las instrucciones del fabricante. Ejemplos de productos comerciales incluyen TRI-REAGENT, Mini-columnas Qiagen RNeasy, Kit de Purificación de ADN y ARN Completa MASTERPURE (EPICENTRE®, Madison, Wis.), Kit de Aislamiento de Bloque de Parafina (Ambion, Inc.), y ARN Stat-60 (Tel-Ensayo). También se pueden emplear técnicas convencionales tales como centrifugación por gradiente de densidad de cloruro de cesio.

La primera etapa en la obtención de perfil de expresión genética mediante RT-PCR consiste en la transcripción reversa de la plantilla de ARN en cADN, seguido de su amplificación exponencial en una reacción de PCR. Las dos transcriptasas reversas utilizadas más habitualmente son la transcriptasa reversa de virus avilo mieloblastosis (AMN-RT) y la transcriptasa reversa de virus de leucemia murina de Moloney (MMLV-RT). La etapa de transcripción reversa es imprimada utilizando imprimadores específicos, hexámeros aleatorios, o imprimadores oligo-dT, dependiendo de las circunstancias y del objetivo de la obtención del perfil de expresión. Véase, por ejemplo, las instrucciones del fabricante que acompañan al producto kit GENEAMP RNA PCR (Perking Elmer, Calif., USA). El cADN derivado puede ser usado entonces como plantilla en la posterior reacción de RT-PCR.

La etapa de PCR utiliza por lo general una polimerasa de ADN dependiente de ADN termoestable, tal como la polimerasa de ADN Taq, la cual tiene actividad nucleasa 4'-3-, pero carece de actividad de endonucleasa de cotejo 3'-5'. De ese modo, la PCR TAQMAN® utiliza típicamente la actividad nucleasa 5' de polimerasa Taq o Th para hidrolizar una sonda de hibridización vinculada a su amplicon objetivo, pero se podría usar cualquier enzima con nucleasa 5' equivalente. Se utilizan dos imprimadores de oligonucleótido para generar un amplicon típico de una reacción de PCR. Un tercer oligonucleótido, o sonda, está diseñado para detectar una secuencia de nucleótido localizada entre los dos imprimadores de PCR. La sonda no es extensible por parte de la enzima de polimerasa de ADN Taq, y está etiquetada con un tinte fluorescente informador y un tinte fluorescente de extinción. Cualquier emisión inducida por láser procedente del tinte informador es extinguida por el tinte extinguidor cuando los dos tintes están situados próximos entre sí según están sobre la sonda. Durante la reacción de amplificación, la enzima de polimerasa de ADN Taq parte la sonda de una manera que depende de la plantilla. Los fragmentos de sonda resultantes se disocian en solución, y la señal procedente del tinte informador liberado está libre del efecto de extinción de segundo fluoróforo. Una molécula de tinte informador es liberada por cada nueva molécula sintetizada, y la detección del tinte informador no extinguido proporciona las bases para la interpretación cuantitativa de los datos.

La RT-PCR TaqMan® puede ser llevada a cabo utilizando un equipo comercialmente disponible. En una realización preferida, el procedimiento de nucleasa 5' se ejecuta en un dispositivo de PCR cuantitativo en tiempo real tal como el ABI PRISM 7900® Sequence Detection System®. El sistema amplifica muestras en un formato de 96 pocillos en un termociclador. Durante la amplificación, la señal fluorescente inducida por láser es recogida en tiempo real por medio de cables de fibra óptica para la totalidad de los 96 pocillos, y detectada en el CCD. El sistema incluye software para hacer funcionar el instrumento y para analizar los dato. Los datos del ensayo de nucleasa 5' son expresados inicialmente como Ct, o ciclo de umbral. Según se ha expuesto en lo que antecede, los valores de fluorescencia son registrados durante cada ciclo y representan la cantidad de producto amplificado en ese punto de la reacción de amplificación. El punto en que la señal fluorescente es registrada en primer lugar como estadísticamente significativa es el ciclo de umbral (C_1).

Para minimizar errores y el efecto de la variación muestra a muestra, la RT-PCR se realiza normalmente utilizando un estándar interno. El estándar interno ideal se expresa a un nivel constante entre diferentes tejidos, y no se ve afectado por el tratamiento experimental. Los ARNs más frecuentemente utilizados para normalizar patrones de expresión genética son mARNs para los genes de limpieza gliceraldehído-3-fosfato-deshidrogenasa (GAPDH) y β-actina.

La PCR en tiempo real es comparable tanto con PCR competitiva cuantitativa, en la que se utiliza el competidor interno para cada secuencia objetivo a efectos de normalización, como con PCR comparativa cuantitativa que utiliza un gen de normalización contenido dentro de la muestra, o un gen de limpieza para RT-PCR[110].

En otro método de PCR, es decir, el método de obtención de perfil de expresión genética basado en MassARRAY (Sequenom, Inc., San Diego, CA), a continuación del aislamiento de ARN y de transcripción reversa, el cADN obtenido es enriquecido con una molécula de ADN sintético (competidor), la cual empareja la región de cADN en todas las posiciones, salvo una base simple, y sirve como estándar interno. La mezcla de cADN/competidor es amplificada en PCR y se somete a un tratamiento de enzima de fosfatasa alcalina de camarón (SAP) post-PCR, lo que da como resultado la desfosforilación de los nucleótidos restantes. Tras la inactivación de la fosfatasa alcalina, los productos de PCR del competidor y el cADN son sometidos a una extensión de imprimador, lo que genera señales en serie distintas para los productos de PCR derivados del competidor y del cADN. Tras la purificación, estos productos son dispensados sobre una matriz de chip, la cual se ha pre-cargado con componentes necesarios para el análisis con análisis de espectrografía de masas de tiempo de vuelo de ionización de desorción de láser asistida por matriz (MALDI-TOF MS). El cADN presente en la reacción es cuantificado a continuación analizando las relaciones de las áreas de pico en el espectro de masas generado[82].

Otras realizaciones adicionales de técnicas basadas en PCR, que son conocidas en el estado de la técnica y que pueden ser usadas para obtención de perfil de expresión de gen incluyen, por ejemplo, visualización diferencial, polimorfismo de longitud de fragmento amplificado (iAFLP), y tecnología de BeadArray ™(Illumina, San Diego, CA) utilizando el sistema Luminex100 LabMAP comercialmente disponible y múltiples microesferas codificadas en color (Luminex Dorp., Austin, Tex.) en un rápido ensayo para expresión genética; y análisis de obtención de perfil de expresión de alta cobertura (HiCEP).

Según se describe con mayor detalle en los ejemplos, en lo que sigue, los perfiles de expresión genética para clasificaciones de cáncer de pulmón fueron recogidas como sigue. Los perfiles de expresión de ARN se obtienen mediante purificación de PBMC a partir de la sangre de sujetos mediante centrifugación utilizando un tubo PCT, un gradiente de Ficoll o separación de densidad equivalente para extraer glóbulos rojos y granulocitos, y la posterior extracción del ARN utilizando el tri-reactivo TRIZOL, el reactivo RNALATER o un reactivo similar para obtener ARN de alta integridad. La cantidad de especies de ARN mensajero individual fue determinada utilizando micro-matrices y/o reacción de cadena de polimerasa cuantitativa.

Tras el análisis de la concentración de ARN, fueron transcritas de forma reversa las etapas de reparación y/o amplificación de ARN utilizando promotores específicos genéticos seguido de RT-PCR. Finalmente, los datos se analizaron para identificar el patrón de expresión genética característico identificado en la muestra de PBMC examinada. Las características de los perfiles de expresión de la enfermedad que va a ser diagnosticada fueron comparados y analizados por parejas con una algoritmo SVM (SV;-RCE)[1] (descrito en los Ejemplos 4 y 5), y con una metodología alternativa descrita en el Ejemplo 14 que sigue. Estos métodos pueden ser demostrados también utilizando un algoritmo similar de aprendizaje de máquina, tal como SVM con Eliminación de Característica Recursiva (SVM-RFE) u otro algoritmo de clasificación tal como Análisis Discriminante Penalizado (PDA) (véase la publicación de la solicitud de Patente Internacional núm. WO 2004/105573, publicada el 9 de Diciembre de 2004) para obtener una función matemática cuyos coeficientes actúan sobre los valores de entrada de expresión genética de ARN y presentan a la salida una "PUNTUACIÓN" cuyo valor determina la clase de individuo y la confianza de la predicción. Al haber determinado esta función mediante análisis de numerosos sujetos que se sabe que son de las clases cuyos miembros han de ser distinguidos posteriormente, se utiliza para clasificar sujetos respecto a sus estados de enfermedad.

En la realización de los ensayos y métodos de la presente invención, se utilizan estas mismas técnicas, el perfil del paciente se compara con el perfil de referencia apropiado, y se selecciona el diagnóstico o la recomendación de

tratamiento en base a esta información.

B. Micro-matrices

La expresión genética diferencial puede ser también identificada, o confirmada, utilizando la técnica de micro-matriz. Así, el perfil de expresión de genes asociados a cáncer de pulmón puede ser medido ya sea en tejido fresco o ya sea incrustado en parafina, utilizando tecnología de micro-matriz. En este método, las secuencias de polinucleótido de interés (incluyendo los cADNs y los oligonucleótidos) son colocadas en placas, o desplegadas, sobre un substrato de microchip. Las secuencias desplegadas son hibridizadas a continuación con sondas de ADN específicas procedentes de células o tejidos de interés. Al igual que en la RT-PCR a los efectos de los métodos y composiciones de la presente invención, la fuente de mARN es el ARN total aislado a partir de la RMBC de los controles y de los pacientes.

En una realización de la técnica de micro-matrices, las inserciones amplificadas por PCR de clones de cADN son aplicadas a un substrato según una matriz densa. Con preferencia, se aplican al menos 10.000 secuencias de nucleótido al substrato. Los genes micro-desplegados, inmovilizados sobre el microchip a razón de 10.000 elementos en cada uno, son adecuados para la hibridación bajo condiciones rigurosas. Las sondas de cADN recién etiquetadas pueden ser generadas mediante incorporación de nucleótidos fluorescentes por transcripción reversa de ARN extraído de tejidos de interés. Las sondas de cADN etiquetadas aplicadas al chip se hibridizan con especificidad en cada mancha de ADN presente sobre la matriz. Tras un lavado riguroso para eliminar las sondas vinculadas de manera no específica, el chip es explorado mediante microscopía láser confocal o mediante otro método de detección, tal como con una cámara CCD. La cuantificación de la hibridación de cada elemento desplegado permite la evaluación de la correspondiente abundancia de mARN. Con fluorescencia de doble color, las sondas de cADN etiquetadas por separado generadas a partir de dos fuentes de ARN, son hibridizadas por parejas en la matriz. La abundancia relativa de las transcripciones a partir de las dos fuentes que corresponden a cada gen especificado, se determina así simultáneamente. La escala miniaturizada de la hibridización proporciona una evaluación conveniente y rápida del patrón de expresión para grandes cantidades de genes. Tales métodos han demostrado tener la sensibilidad necesaria para detectar transcripciones raras, que se expresan mediante unas pocas copias por célula, y para detectar reproduciblemente al menos de forma aproximada el doble de diferencias en los niveles de expresión. El análisis de micro-matriz puede ser llevado a cabo mediante un equipamiento comercialmente disponible, siguiendo los protocolos del fabricante.

Otros métodos útiles resumidos por la Patente US núm. 7.081.340 incluyen el Análisis Serie de Expresión de Gen (SAGE) y la Secuenciación de Signatura Masivamente Paralela (MPSS).

C. Inmunohistoquímica

Los métodos de inmunohistoquímica son también adecuados para detectar los niveles de expresión de los productos de expresión genética de los genes informativos descritos para su uso en los métodos y composiciones de la presente invención. Anticuerpos o antisueros, con preferencia antisueros policlonales, y con mayor preferencia los anticuerpos monoclonales, u otros ligandos de vinculación de proteína específicos para cada marcador, son utilizados para detectar expresión. Los anticuerpos pueden ser detectados por etiquetamiento directo de los propios anticuerpos, por ejemplo con etiquetas radiactivas, etiquetas fluorescentes, etiquetas de hapteno tales como biotina, o una enzima tal como peroxidasa de rábano picante o fosfatasa alcalina. Alternativamente, se utiliza anticuerpo primario sin etiquetas junto con anticuerpo secundario etiquetado, comprendiendo antisueros, antisueros policlonales o un anticuerpo monoclonal específico para el anticuerpo primario. Los protocolos y kits para análisis inmunohistoquimicos son bien conocidos en el estado e la técnica y se encuentran disponibles comercialmente.

D. Proteómica

El término “proteoma” se define como la totalidad de las proteínas presentes en una muestra (por ejemplo, tejido, organismo o cultivo celular) en un cierto instante de tiempo. La proteómica incluye, entre otras cosas, el estudio de los cambios globales de expresión de proteína en una muestra (también mencionada como “proteómica de expresión”). La proteómica incluye típicamente las siguientes etapas: (1) separación de las proteínas individuales de una muestra mediante electroforesis de gel de 2-D (2-D PAGE); (2) identificación de las proteínas individuales recuperadas desde el gel, por ejemplo, mediante espectrometría de masas o secuenciación de terminal N, y (3) análisis de los datos usando bioinformática. Los métodos proteómicos son suplementos válidos para otros métodos de obtención de perfil de expresión genética, y pueden ser usados solos o en combinación con otros métodos, para detectar productos de expresión genética de los perfiles genéticos descritos en la presente memoria.

IY. Composiciones de la invención

Los métodos para el diagnóstico de cáncer de pulmón que utilizan perfiles definidos de expresión genética, permiten el desarrollo de herramientas de diagnóstico simplificado para el diagnóstico de cáncer de pulmón, por ejemplo, BSCLC o el diagnóstico de una fase de desarrollo específica (tempranamente, fase de desarrollo I, fase de desarrollo II o tardíamente) del cáncer de pulmón, diagnosticar un tipo específico de cáncer de pulmón (por ejemplo, AC frente a LSCC) o monitorizar el efecto de la intervención terapéutica o quirúrgica para la determinación del

tratamiento adicional o la evaluación de la probabilidad de recurrencia del cáncer.

De ese modo, una composición para el diagnóstico de cáncer de pulmón de célula no pequeña en un mamífero según se describe en la presente memoria puede ser un kit o un reactivo. Por ejemplo, una realización de una composición incluye un substrato sobre el que son inmovilizados dichos polinucleótidos u oligonucleótidos o ligandos. En otra realización, la composición es un kit que contiene los tres o más polinucleótidos u oligonucleótidos o ligandos relevantes, etiquetas detectables opcionales, substratos de inmovilización, substratos opcionales para etiquetas enzimáticas, así como también otros artículos de laboratorio. En otra realización más, al menos un polinucleótido o un oligonucleótido o un ligando está asociado a una etiqueta detectable.

Una composición de ese tipo contiene, en una realización, tres o más polinucleótidos u oligonucleótidos, de los que cada polinucleótido u oligonucleótido hibridiza en un gen, fragmento de gen, transcripción de gen o producto de expresión diferente a partir de células mononucleares de sangre periférica de mamífero (PBMC), en el que dicho gen, fragmento de gen, transcripción de gen o producto de expresión se elige a partir de (i) los genes de la Tabla I; (ii) los genes de la tabla II; (iii) los genes de la Tabla III, y (iv) los genes de la Tabla IV. En otra realización, una composición de ese tipo contiene tres o más polinucleótidos u oligonucleótidos, en el que cada polinucleótido u oligonucleótido hibridiza en un gen, fragmento de gen, transcripción de gen o producto de expresión diferente procedente de células mononucleares de sangre periférica de mamífero (PBMC), en el que dicho gen, fragmento de gen, transcripción de gen o producto de expresión se elige a partir de (i) los genes de la Tabla V; (ii) los genes de la Tabla VI, o (iii) los genes de la Tabla VII.

En otra realización, una composición de ese tipo contiene tres o más ligandos, en la que cada ligando enlaza con un producto de expresión diferente de células mononucleares de sangre periférica de mamífero (PBMC), en el que el producto de expresión de gen es el producto de un gen seleccionado a partir de (i) los genes de la Tabla I; (ii) los genes de la Tabla II; (iii) los genes de la Tabla III; y (iv) los genes de la Tabla IV. En otra realización más, una composición de ese tipo contiene tres o más ligandos, de los que cada ligando enlaza con un producto de expresión de gen diferente de células mononucleares de sangre periférica de mamífero (PBMC), en el que el producto de expresión genética es el producto de un gen seleccionado a partir de (i) los genes de la Tabla V; (ii) los genes de la Tabla VI; o (iii) los genes de la Tabla VII.

En una realización, una composición para el diagnóstico de cáncer de pulmón en un mamífero incluye tres o más conjuntos de sonda de imprimador de PCR. Cada uno de los conjuntos de sonda de imprimador amplifica una secuencia de polinucleótido diferente procedente de un producto de expresión genética de tres o más genes encontrados en las células mononucleares de sangre periférica (PBMC) del sujeto. Estos genes informativos se eligen de modo que formen un perfil o signatura de expresión genética que sea distinguible entre un sujeto que tiene cáncer de pulmón y un control de referencia seleccionado. Los cambios de expresión de los genes en el perfil de expresión genética con respecto al de un perfil de expresión genética de referencia, están correlacionados con un cáncer de pulmón, tal como un cáncer de pulmón de célula no pequeña (NSCLC).

En una realización de esta composición, los genes informativos se eligen entre los genes identificados en la Tabla I que sigue. La Tabla I contiene aproximadamente los 100 genes superiores identificados por los inventores como representativos de una signatura genómica indicativa de la presencia de algún cáncer de pulmón de NSCLC. Esta colección de genes comprende aquellos respecto a los que la expresión de producto de gen se ve alterada (es decir, incrementada o reducida) frente a la misma expresión de producto de gen en la PBMC de un control de referencia. En una realización, el polinucleótido o los oligonucleótidos, tal como imprimadores de PCR y sondas, son generados respecto a tres o más genes informativos de la Tabla I para su uso en la composición. Un ejemplo de tal composición contiene imprimadores y sondas en una porción objetivada de los tres primeros genes de esa Tabla. En otra realización, Los imprimadores y sondas de PCR son generados respecto a al menos seis genes informativos de la Tabla I para su uso en la composición. Un ejemplo de una composición de ese tipo contiene imprimadores y sondas respecto a una porción objetivada de los seis primeros genes de dicha Tabla. Todavía en otra realización, los imprimadores y sondas de PCR son generados respecto a al menos quince genes informativos de la Tabla I para su uso en la composición. Un ejemplo de composición de ese tipo contiene imprimadores y sondas respecto a una porción objetivadas de los quince primeros genes de dicha Tabla. Según otras realizaciones, se emplean imprimadores y sondas respecto a una porción objetivada de otras combinaciones de los genes de las Tablas. Los genes seleccionados a partir de la Tabla no necesitan estar por orden categórico; por el contrario, cualquier combinación que muestre claramente una diferencia de expresión entre el control de referencia y el paciente enfermo, resulta útil en tal composición.

En una realización específica, los genes informativos de la Tabla I comprenden tres o más genes seleccionados en el grupo consistente en IGSF6, HSPA8(A), LYN, DNCL1, HSPA1A, DPYSL2, HSPA8(I), NFKBIA, FGL2, CALM2, CCL5, RPS2, DDIT4 y Clorf63.

En otra realización de la presente composición, los genes informativos se seleccionan de entre los genes identificados en la Tabla II que sigue. La Tabla II contiene aproximadamente los 100 genes superiores identificados por los inventores como representativos de una signatura genómica indicativa de la presencia de un NSCLC específico, es decir, adenocarcinoma de pulmón. Esta colección de genes comprende aquellos para los que la expresión de producto de gen se ve alterada (es decir, aumentada o disminuida) frente a la misma expresión de

producto de gen de la PBMC de un control de referencia. En una realización, se generan imprimadores y sondas de PCR respecto a tres o más genes informativos de la Tabla II, para su uso en la composición. Un ejemplo de composición de ese tipo contiene imprimadores y sondas para una porción objetivada de los tres primeros genes de la Tabla II. En otra realización, los imprimadores y sondas de PCR son generados en al menos seis genes informativos de la Tabla II para su uso en la composición. Un ejemplo de composición de ese tipo contiene imprimadores y sondas para una porción objetivada de los seis primeros genes de la Tabla II. En otra realización más, los imprimadores y sondas de PCR son generados para al menos quince genes informativos de la Tabla II para su uso en la composición. Un ejemplo de composición de ese tipo contiene imprimadores y sondas respecto a una porción objetivada de los quince primeros genes de la Tabla II. Otras realizaciones emplean imprimadores y sondas respecto a una porción objetivada de otras combinaciones de genes de la Tabla II. Los genes seleccionados a partir de la Tabla II no necesitan estar por orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia de expresión entre el control de referencia y el paciente enfermo, es útil en una composición de ese tipo.

En una realización específica, los genes informativos de la Tabla II comprenden tres o más genes seleccionados a partir del grupo consistente en ETS1, CCL5, DDIT4, CXCR4, DNCL1, MS4ABA, ATP5B, HSPA8(A), ADM PTPN6, ARHGAP9, S100AB, DPYSL2, HSPA1A, y NFKBIA.

En otra realización de la presente composición, los genes informativos se eligen entre los genes identificados en la Tabla III. La Tabla III contiene los 100 genes superiores identificados por los inventores como representativos de una signatura genómica indicativa del efecto de resección quirúrgica del tumor de un paciente con NSCLC. Esta colección de genes contiene aquellos para los que la expresión de producto de gen se ve alterada (es decir, aumentada o disminuida) frente a la misma expresión de producto de gen de la PBMC de un paciente antes o después de la cirugía. En una realización, se generan imprimadores y sondas de PCR respecto a tres o más genes informativos de la Tabla III para su uso en la composición. Un ejemplo de composición de ese tipo contiene imprimadores y sondas respecto a una porción objetivada de los tres primeros genes de la Tabla III. En otra realización, los imprimadores y sondas de PCR son generados respecto a al menos seis genes informativos de la Tabla III para su uso en la composición. Un ejemplo de tal composición contiene imprimadores y sondas para una porción objetivada de los seis primeros genes de la Tabla III. En otra realización más, los imprimadores y sondas de PCR son generados para al menos quince genes informativos de la Tabla III, para su uso en la composición. Un ejemplo de una composición de ese tipo contiene imprimadores y sondas para una porción objetivada de los quince primeros genes de esa Tabla III. Los genes seleccionados a partir de la Tabla III no necesitan estar en orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia de expresión entre el paciente de NSCLC en pre-cirugía en comparación con el paciente de NSCLC en post-cirugía, es útil en la citada composición.

En otra realización de esta composición, los genes informativos se eligen de entre los genes identificados en la Tabla IV. La Tabla IV contiene realizaciones de 15 genes útiles como perfiles o signaturas genómicas representativas para tres usos diagnósticos, es decir, para distinguir entre NSCLC y todos los controles, para distinguir entre NSCLC en general y adenocarcinoma, y para distinguir entre, y con ello la progresión del rastreo de la enfermedad en, sujetos pre- y post-quirúrgicos. En una realización, se generan imprimadores y sondas de PCR para la totalidad de 15 genes informativos a partir de la Tabla IV, columna 1, para su uso en una composición de diagnóstico. En otra realización, se generan imprimadores y sondas de PCR para 15 genes informativos a partir de la Tabla IV, columna 2, para su uso en una composición de diagnóstico. En otra realización más, se generan imprimadores y sondas de PCR para quince genes informativos de la Tabla IV, columna 3, para su uso en una composición de diagnóstico. Otras realizaciones adicionales emplean imprimadores y sondas para una porción objetivada de otras combinaciones de genes de la Tabla IV. Los genes seleccionados a partir de la Tablas IV no necesitan estar en orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia entre el sujeto de prueba y los grupos comparados, resulta útil en tal composición.

En otra realización de la presente composición, los genes informativos se eligen a partir de los genes identificados en la Tabla V. La Tabla V contiene realizaciones de 136 genes útiles como signaturas o perfiles genómicos representativos, para distinguir entre NSCLC y todos los controles, principalmente controles no sanos. En una realización, se generan imprimadores y sondas de PCR respecto a los 29 genes informativos de orden superior de la Tabla V, formando con ello el clasificador de 29 genes de los ejemplos que siguen para su uso en una composición diagnóstica. En otra realización más, se generan imprimadores y sondas de PCR respecto a cualquier número deseado de genes informativos a partir de la Tabla V para su uso en una composición diagnóstica. Los genes seleccionados a partir de la Tabla V no necesitan estar por orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia entre el sujeto de prueba y los grupos comparados, es útil en dicha composición.

En otra realización de la presente composición, los genes informativos se seleccionan entre los genes identificados en la Tabla VI. La Tabla VI contiene realizaciones de 50 genes útiles como signaturas o perfiles genómicos representativos, para distinguir entre sujetos pre-quirúrgicos y post-quirúrgicos. En una realización, se generan imprimadores y sondas de PCR respecto a los 2 genes informativos de clasificación superior, es decir, los CYP2R1 y MYO5B, en la Tabla VI, para su uso en una composición diagnóstica. En otra realización más, se generan

imprimadores y sondas de OCR para los cuatro genes superiores, por ejemplo los CYP2R1, MYO5B, DGUOK y DYNLL1, a partir de la Tabla VI, para su uso en una composición diagnóstica. En una composición adicional, los oligonucleótidos o polinucleótidos, tales como imprimadores y sondas de PCR, que hibridizan o amplifican cualquier número deseado de genes informativos a partir de la Tabla VI, son útiles en una composición diagnóstica. Los genes seleccionados a partir de la Tabla VI no necesitan estar por orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia entre el sujeto de prueba y los grupos comparados, es útil en tal composición.

En otra realización de la presente composición, los genes informativos se seleccionan a partir de los genes identificados en la Tabla VII. La Tabla VII contiene realizaciones de 24 genes útiles como signaturas o perfiles genómicos representativos para distinguir ente sujetos con NSCLC y sujetos con nódulos de pulmón benignos. En una realización, se generan oligonucleótidos o polinucleótidos, tales como imprimadores o sondas de PCR, para la totalidad de los 24 genes informativos de la Tabla VII para su uso en una composición diagnóstica. En otra realización más, se generan imprimadores y sondas de PCR para un pequeño número de genes a partir de la Tabla VII para su uso en una composición diagnóstica. Los genes seleccionados a partir de la Tabla VII no necesitan estar por orden jerárquico; por el contrario, cualquier combinación que muestre claramente una diferencia entre el sujeto de la prueba y los grupos comparados, es útil en tal composición.

En una realización de las composiciones descritas en lo que antecede, el control de referencia es un control no sano (NHC) según se ha descrito con anterioridad. En otras realizaciones, el control de referencia puede ser cualquier clase de control según se ha descrito en lo que antecede en “Definiciones”. Una composición que contenga polinucleótidos u oligonucleótidos que hibridicen miembros del perfil de expresión genética seleccionado, preparados a partir de una selección de genes listados en estas tablas, resulta deseable no sólo a efectos de diagnóstico, sino también para monitorizar los efectos del tratamiento terapéutico quirúrgico o no quirúrgico para determinar si se mantienen los efectos positivos de la resección/quimioterapia durante un período largo después del tratamiento inicial. Estos perfiles permiten también una determinación de recurrencia o la probabilidad de recurrencia de un cáncer de pulmón, por ejemplo NSCLC, si los resultados demuestran un retorno a los perfiles de la pre-cirugía/pre-quimioterapia. De igual modo es probable que estas composiciones puedan ser también empleadas para su uso en la monitorización de la eficacia de las terapias no quirúrgicas para un cáncer de pulmón.

Las composiciones basadas en los genes seleccionados a partir de las Tablas I a VII descritas en la presente memoria, opcionalmente asociadas a etiquetas detectables, pueden ser representadas en formato de una tarjeta, un chip o una cámara microfluídica, o como un kit adaptado para su uso con técnicas de PCR, RT_PCR o QPCR descritas en lo que antecede. En un aspecto, un formato de ese tipo es un ensayo diagnóstico que utiliza matrices TAQMAN® Quantitative de baja densidad de PCR. Los resultados preliminares sugieren que el número de genes requeridos es compatible con estas plataformas. Cuando una muestra de PBMC procedente de un paciente seleccionado se pone en contacto con los imprimadores y sondas de la composición, la amplificación de PCR de los genes informativos objetivados del perfil de expresión genética del paciente permite la detección de cambios de expresión de los genes informativos en la PBMC del paciente a partir del perfil de expresión genética de referencia correlacionado con un diagnóstico de cáncer de pulmón cuando se utilizan composiciones dirigidas a los genes de la Tabla I o V, o de adenocarcinoma de pulmón cuando se utilizan composiciones dirigidas a los genes de la Tabla II. De forma similar, cuando una muestra de PBMC procedente de un paciente post-quirúrgico se pone en contacto con los imprimadores y las sondas de la composición, la amplificación de PCR de genes informativos objetivados elegidos entre los de las Tablas III o IV en el perfil de expresión genética del paciente, permite la detección de cambios de expresión en los genes, en el perfil de expresión genética frente a los del perfil de expresión genética de referencia. En estas circunstancias, el perfil de referencia preferido es el que se obtiene a partir del mismo paciente (o de un paciente similar) con anterioridad a la cirugía. Los cambios significativos en la expresión genética de los genes informativos de la PBMC del paciente frente al perfil de expresión genética de referencia, están correlacionados con el efecto de la cirugía, y/o con el mantenimiento del efecto positivo.

Las Tablas I a VII y la información identificadora de los genes listados en la presente memoria, se describen a continuación.

TABLA I

NOMBRE DEL GEN	Símbolo	Puntuación	Orden
Familia del dominio TSC22, miembro 3 (TSC22D3), variante de transcripción 2, mARN. (A)	TSC22D3	0,9522	1
quimiocina (motivo C-X-C) receptor 4 (CXCR4), variante de transcripción 1, mARN. (A)	CxCR4	0,0444	2
polipéptido claro 1, citoplásmico, de dineína, (DNCL1), mARN. (S)	DNCL1	0,8668	3
proteína ribosómica S3 (ROS3), mARN. (S)	RPS3	0,8556	4
transcripción inducible por daños de ADN 4 (DDIT4), mARN. (S)	DDIT4	08502	5
granzima B (granzima 2, estearasa de serina asociada a linfocito T	GZMB	0,8148	6

citotóxica 1) (GZMB), mARN. (S)			
gen de translocación de célula B 1, anti-proliferativo (BTG1), mARN. (S)	BTG1	0,8	7
Proteína 8 de 70 kDa de choque térmico (HSPA8), variante de transcripción 1, mARN. (I)	HSPA8	0,793	8
proteína ribosómica L12 (RPL 12), mARN. (S)	RPL12	0,7564	9
adaptador a modo de Sr (SLA), mARN. (S)	SLA	0,7322	10
factor de transcripción relacionado con el tallo (RUNX3), variante de transcripción 2, mARN. (I)	RUNX3	0,7306	11
gen HGFL (MGC 17330), mARN. (S)	MGC17330	0,6982	12
Proteína 1A de 70 kDa de choque térmico (HSPA1A), mARN. (S)	HSPA1A	0,684	13
proteína accesoria receptora de interleuquina 18 (IL18RAP), mARN. (S)	IL18RAP	0,6728	14
proteína de enlace de ARN inducible por frío (CIRBP), mARN. (S)	CIRBP	0,67	15
adrenomedulina (ADM), mARN. (S)	ADM	0,662	16
proteína de enlace de CCAAT/potenciadora (E/EBP), beta (CEBPB), mARN. (S)	CEBPB	0,654	17
PREDICADO: similar a ribonucleoproteína A1 heterogénea nuclear (LOC 645385), mARN. (S)	LOC645385	0,654	18
proteína de enlace de CCAAT/potenciadora (C/EBP), delta (CEBPD), mARN. (S)	CEBDP	0,6416	19
Factor 9 de modo Kruppel (KLF9), mARN. (S)	KLF9	0,6392	20
PRONOSTICADO: proteína hipotética LOC440345, variante de transcripción 6 (LOC440345), mARN. (I)	LOC440345	0,6358	21
inhibidor de ligante de ADN 2, proteína hélice-bucle-hélice dominante (ID2), mARN. (S)	ID2	0,617	22
receptor similar a Ig de célula asesina, dos dominios, cola citoplásmica larga, 3 (KIR2DL3), variante de transcripción 2, mARN. (A)	KIR2DL3	0,6126	23
proteína de activación de araquidona 5-lipoxigenasa (ALOX5AP), mARN. (S)	ALOX5AP	0,6106	24
superfamilia de inmunoglobulina, miembro 6 (IGSF6), mARN. (S)	IGSF6	0,6068	25
Proteína 8 de 70 kDa de choque térmico (HSPA8), variante de transcripción 2, mARN. (A)	HSPA8	0,6032	27
Tubulina, alfa, ubicua (K-ALPHA-1), mARN. (S)	K-ALPHA-1	0,6002	28
proteína kinasa C, delta (PRKCD), variante de transcripción 2, mARN. (A)	PRKCD	0,5992	29
dominio de PR que contiene 1, con dominio de ZNF (PRDM1), variante de transcripción 1, mARN. (A)	PRDM1	0,594	30
antígeno CD55, factor de aceleración de decadencia para complemento (grupo sanguíneo Cromer) (CD55), mARN. (S)	CD55	0,5722	31
cistatina F (leucocistatina) (CST7), mARN. (S)	CST7	0,5698	32
marcador de diferenciación asociado a mieloide (MYADM), variante de transcripción 4, mARN. (A)	MYADM	0,568	33
complejo de histocompatibilidad mayor, clase I, F (HLA-F), mARN. (S)	HLA-F	0,568	34
proteína 2A de dominio SH2 (SH2D2A), mARN. (S)	SH2D2A	0,5656	35
dominio de tetramerización de canal de potasio que contiene 12 (KCDT12), mARN. (S)	KCDT12	0,5638	36
proteína de enlace de dominio SH3, proteína de activación de Ras-GRTPasa (G3BP), variante de transcripción 1, mARN. (A)	G3BP	0,5636	37
tipo fibrinógeno 2 (FGL2), mARN. (S)	FGL2	0,552	38
CCAAT/proteína ligante potenciadora (C/EBP), alfa (CDBPA), mARN. (S)	CEBPA	0,5368	39
homólogo DnaJ (Hsp40), subfamilia A, miembro 1 (ADN1A1), mARN. (S)	DNAJA1	0,5306	40
proteína de nivelación (filamento de actina) línea Z de músculo, alfa 2 (CAPZA2), mARN. (S)	CAPZA2	0,5244	41
Factor de transcripción general IIIA (GTF3A), mARN. (S)	GTF3A	0,523	42
dominio de IBR que contiene 2 (IBRDC2), mARN. (S)	IBRDC2	0,5228	43

gen de exonucleasa estimulado por interferón 20 kDa (ISG20), mARN. (S)	ISG20	0,5208	44
PRONOSTICADO: similar a proteína ribosómica L13a, variante de transcripción 4 (LOC649564), mARN. (A)	LOC649564	0,5134	45
receptor acoplado a proteína G 171 (GPR171), mARN. (S)	GPR171	0,5124	46
receptor similar a inmunoglobulina de célula asesina, dos dominios, cola citoplásmica larga, 4 (KIR2DL4), mARN. (S)	KIR2DL4	0,5044	47
polipéptido asociado a sin3, 30 kDA (SAP30), mARN. (S)	SAP30	0,4972	48
PRONOSTICADO: meteorina, similar a regulador de diferenciación de célula glial (METRNL), mARN. (I)	METRNL	0,4936	49
canal intracelular cloruro 3 (CLIC3), mARN. (S)	CLIC3	0,4926	50
factor 3 de iniciación de traslación eucariótica, subunidad 12 (EIF3S 12), mARN. (S)	EIF3S12	0,4912	51
Substrato 2 de receptor de insulina (IRS2), mARN. (S)	IRS2	0,4824	52
receptor celular 2 de virus de hepatitis A (HAVCR2), mARN. (S)	HAVCR2	0,4758	53
dominio de HD que contiene 2 (HDDC2), mARN. (S)	HDDC2	0,4754	54
factor 1 de exportación de ARN nuclear (NXF1), mARN. (S)	NXF1	0,468	55
perforina 1 (proteína de formación de poro) (PRF1), mARN. (S)	PRF1	0,4642	56
dominio SAM, dominio SH3 y señales de localización nuclear, 1 (SAMSN1), mARN. (S)	TSAMSN1	0,4614	57
TERF1 (TRF1)-factor de nuclear de interacción 2 (TINF2), mARN. (S)	TINF2	0,4604	58
compartimento intermedio de golgi-retículo endoplásmico (ERGIC) 1 (ERGIC 1), variante de transcriptasa 1, mARN. (I)	ERGIC1	0,4554	59
factor de necrosis tumoral, proteína inducida alfa 2 (TNFAIP2), mARN. (S)	TNFAIP2	0,455	60
factor de transcripción de gancho-AT (AKNA), mARN. (S)	AKNA	0,4548	61
proteína relacionada con diferenciación adiposa (ADFP), mARN. (S)	ADFP	0,4546	62
piruvato deshidrogenasa kinasa, isozima 4 (PDK4), mARN. (S)	PDK4	0,4538	63
factor de activación de peptidasa apoptótica (APAF1), variante de transcripción 5, mARN. (A)	APAF1	0,4486	64
transductor de señal y activador de transcripción 4. (STAT4), mARN. (S)	STAT4	0,4478	65
familia 1 de aldo-keto reductasa, miembro C3 (3-alfa hidroesteroide deshidrogenasa, tipo II), mARN. (S)	AKR1C3	0,4454	66
dominio SH2 que contiene 3C (SH2D3C), variante de transcripción 2, mARN. (I)	SH2D3C	0,4444	67
proteína 1 de 105 kDa/110 kDa de choque térmico (HSPH1), mARN. (S)	HSPH1	0,4396	68
fosfoinositida-3-kinasa, subunidad reguladora 1 (p85 alfa) (PIK3R1), variante de transcripción 2, mARN. (A)	PIK3R1	0,4312	69
presenilina asociada, similar a romboide (PSARL), mARN. (S)	PSARL	0,4284	70
desoxiguanosina kinasa, gen nuclear que codifica proteína mitocóndrica, variante de transcripción 1, mARN. (A)	DGUOK	0,4272	71
homología de pleckstrina, SEC7 y dominios arrollados en espiral, proteína ligante (PSCDBP), mARN. (S)	PSCDBP	0,4206	72
fosforilasa de uridina 1 (UPP1), variante de transcripción 2, mARN. (A)	UPP1	0,4188	73
familia portadora de soluto 35 (transportador de ácido siálico – CMP), miembro A1 (SLC35a!), mARN. (S)	SLC35A1	0,4176	74
proteína kinasa kinasa kinasa 8 activada por mitógeno (MAP3K8), mARN. (S)	MAP3K8	0,4162	75
cuadro 39 de lectura de cromosoma 15 abierto (C15orf39), mARN. (S)	C15orf39	0,411	76
proteína ribosómica L35 (RPL35), mARN. (S)	RPL35	0,4106	77
factor de intercambio de nucleótido de guanina rho/rac (GEF) 2 (ARHGEF2), mARN. (S)	ARHGEF2	0,4074	78
cuadro 37 de lectura de cromosoma 19 abierto (C19orf37), mARN. (S)	C19orf37	0,4072	79
proteína de motivo ligante de ARN 14 (RBM14), mARN. (S)	RBM14	0,4068	80

proteína hipotética MGC7036 (MGC7036), mARN. (S)	MGC7036	0,4056	81
poli(A) polimerasa alfa (PAPOLA), mARN. (S)	PAPOLA	0,4044	82
RAB 10, familia oncógena de miembro RAS (RAB10), mARN. (S)	RAB10	0,403	83
cuadro 28 de lectura de cromosoma 2 abierto (C2orf28), variante de transcripción 2, mARN. (A)	C2orf28	0,403	84
dominio LIM solamente 2 (similar a rombotina 1) (LMO2), mARN. (S)	LMO2	0,3972	85
polimerasa(ARN) III (ADN dirigido) polipéptido G (32 kDa) similar (POLR3GL), mARN. (S)	POLR3GL	0,3968	86
dedo de zinc y dominio BTB que contiene 16 (ZBTB16), variante de transcripción 1, ARN. (A)	ZBTB16	0,3948	87
factor de iniciación de traducción eucariótica 3, subunidad 5 épsilon, 47 kDa (EIF3S5), mARN. (S)	EIF3S5	0,3924	88
proteína HSCARG (HSCARG), mARN. (S)	HSCARG	0,3916	89
similar a sinaptotagmina 3 (SYTL3), mARN. (S)	SYTL3	0,3896	90
proteína hipotética FLJ32028 (FLJ32028), mARN. (S)	FLJ32028	0,3886	91
repetición rica en leucina que contiene 33 (LRRC33), mARN. (S)	LRRC33	0,3862	92
cuadro de lectura de cromosoma 1 abierto 162 (C1orf162), mARN. (S)	C1orf162	0,3846	93
citocromo P450, familia 2, subfamilia R, polipéptido 1 (CYP2R1), mARN. (S)	CYP2R1	0,3846	94
proto–oncógeno jun D (JUND), mARN. (S)	JUND	0,381	95
familia D de antígeno de melanoma, 1 (MAGED1), variante de transcripción 1, mARN. (A)	MAGED1	0,3806	96
susceptibilidad de autismo candidato 2 (AUTS2), mARN. (S)	AUTS2	0,3806	97
factor de transcripción de oligodendrocito 1 (OLIG1), mARN. (S)	OLIG1	0,379	98
factor de elongación de traducción eucariótica 1 delta (proteína de intercambio de nucleótido de guanina), variante de transcripción 1, mARN. (A)	EEF1D	0,3776	99
subfamilia K de receptor similar a lectina de célula asesina, miembro 1 (KLRK1), mARN. (S)	KLRK1	0,3736	100

TABLA II

NOMBRE DE GEN	**Símbolo**	**Puntuación**	**Orden**
homólogo 1 de oncogén E26 de virus de eritroblastosis v-ets (aviar) (ETS1), mARN. (S)	ETS1	0,9612	1
ligando 5 de quimiocina (motivo C-C) (CCL5), mARN. (S)	CCL5	0,9438	2
transcripción inducible por daños de ADN 4 (DDIT4), mARN. (S)	DDIT4	0,9024	3
receptor de quimiocina (motivo C-X-C) 4 (CXCR4), variante de transcripción 1, mARN. (A)	CXCR4	0,8098	4
polipéptido ligero, citoplásmico, de dineína 1 (DNCL1), mARN. (S)	DNCL1	0,8058	5
subfamilia A, de 4 dominios que abarca membrana, miembro 6A (MS4A6A), variante de transcripción 2, mARN. (I)	MS4A6A	0,796	6
polipéptido beta, complejo F1 mitocóndrico, transportador de H^+, de ATP sintasa (ATP5B), proteína mitocóndrica de codificación de gen nuclear, mARN. (S)	ATP5B	0,7754	7
proteína de 70 kDa de choque térmico 8 (HSPA8), variante de transcripción 1, mARN. (I)	HSPA8	0,7718	8
adrenomedulina (ADM), mARN. (S)	ADM	0,7708	9
proteína tirosina fosfatasa, tipo no receptor 6 (PTPN6), variante de transcripción 3, mARN. (A).	PTPN6	0,7576	10
proteína de activación de GTPasa Rho 9 (ARHGAP9), mARN. (S)	ARHGAPg	0,7548	11

S100 proteína ligante de calcio A8 (calgranulina A) (S 100ª8), mARN. (S)	S100AB	0,7336	12
similar a dihidropirimidinasa 2 (DPYSL2), mARN. (S)	DPYSL2	0,724	13
proteína de 70 kDa de choque térmico 1A (HSPA1A). (S)	HSPA1A	0,7156	14
factor nuclear de potenciador genético de polipéptido ligero kappa en inhibidor de células B, alfa (NFKBIA), mARN. (S)	NFKBIA	0,7132	15
N-acetilglucosamina kinasa (NAGK(, mARN. (S)	NAGK	0,7098	16
superfamilia de inmunoglobulina, miembro 6 (IGSF6), mARN. (S)	IGSF6	0,7088	17
complejo de histocompatibilidad principal, clase II, DM beta (HLA-DMB), mARN. (S)	HLA-DMB	0,704	18
familia con similitud de secuencia 100, miembro B (FAM100B), mARN. (S)	FAM100B	0,7016	19
miosina, polipéptido ligero 6, álcali, músculo liso y no músculo, variante de transcripción 1, mARN. (A)	MYL6	0,6962	20
familia portadora de soluto 2 (transportador de glucosa facilitado), miembro 3 (SLC2A3), mARN. (S)	SLC2A3	0,6738	21
Proteína de 70 kDa de choque térmico 8 (HSPA8), variante de transcripción 2, mARN. (A)	HSPA8	0,653	22
H2A familia de histona, miembro Z (HsAFZ), mARN. (S)	H2AFZ	0,6422	23
factor similar de Kruppel 9 (KLF9), mARN. (S)	KLF9	0,6354	24
factor de necrosis tumoral, proteína 3 alfa inducida (TNFAIP23), mARN. (S)	THFAIP3	0,6312	25
selenoproteína W, 1 (SEPW1), mARN. (S)	SEPW1	0,6164	26
nexina de clasificación 2 (SNX2), mARN. (S)	SNX2	0,609	27
fosfatasa de doble especificidad 1 (DUSP1), mARN. (S)	DUSP1	0,6076	28
cistatina F (leucocistatina) (CST7), mARN. (S)	CST7	0,5858	29
PRONOSTICADO: similar a proteína ribosómica acídica 60S P1, variante de transcripción 4 (LOC440927), mARN. (A)	LOC440927	0,5844	30
dominio PR que contiene 1, con dominio ZNF (PRDM1), variante de transcripción 1, mARN. (A)	PRDM1	0,581	31
proteína de enlace de ARN inducible por frío (CIRBP), mARN. (S)	CIRBP	0,5786	32
región de cromosoma de síndrome de ojo de gato, candidato 1 (CERC1), variante de transcripción 1, mARN. (A)	CECR1	0,575	33
complejo F1 mitocóndrico, transportador de H^+, de ATP sintasa, subunidad alfa 1, músculo cardíaco (ATP5A1), proteína mitocóndrica de codificación genética nuclear, variante de transcripción 1, mARN. (A)	ATP5A1	0,5664	34
dominio LIM solamente 2 (similar a rombotina 1) (LMO2), mARN. (S)	LMO2	0,5608	35
simulador de disociación de nucleótido de guanina ral (RALGDS), mARN. (S)	RALGDS	0,5572	36
receptor acoplado a proteína G 171 (GOR171), mARN. (S)	GPR171	0,5536	37
proteína de motivo ligante de ARN 5 (RBM5), mARN. (S)	RBM5	0,5532	38
	ITK	0,545	39
kinasa de célula T inducible de IL2 (ITK), mARN. (S)	CTDSP2	0,542	40
fosfatasa de CDT (dominio de terminal carboxi, polimerasa II de ARN, polipéptido A) pequeña 2, mARN. (S)	GTF3A	0,5394	41
Factor IIIA de transcripción general (GTF3A), mARN. (S)	MYADM	0,5394	42
marcador de diferenciación asociado a mieloide (MYADM), variante de transcripción 4, mARN. (A)	NALP1	0,5384	43
NACT, repetición rica en leucina y PYD (dominio de pirina) que contiene 1, variante de transcripción 4, mARN. (I)	DDX17	0,5304	44
MUERTE polipéptido de cuadro 17 (Asp-Glu-Ala-Asp) (DDX17), variante de transcripción 2, mARN. (A)	THBS1	0,5278	45
trombospondina 1 (THBS1), mARN. (S)	ALOX5	0,523	46
araquidonato 5-lipoxigenasa (ALOX5), mARN. (A)	SPOCK2	0,5186	47
proteiglicano de dominios similar a cwcv y kazal, sparc/osteonectina	MGC7036	0,5182	48

(testican) 2 (SPOCK2), mARN. (S)			
Proteína hipotética MGC7036 (MGC7036), mARN. (S)	PIK3R1	0,5176	49
fosfoinositida-3-kinasa, subunidad reguladora 1 (p85 alfa) (PIK3R1) variante de transcripción 2, mARN. (A)	MNDA	0,5158	50
antígeno de diferenciación nuclear de célula de mieloide (MNDA), mARN. (S)	SLC35A1	0,5142	51
familia portadora de soluto 35 (CMP-transportador de ácido siálico), miembro A1 (SLC35A1), mARN. (S)	C19orf37	0,514	52
cuadro de lectura de cromosoma 19 abierto 37 (C19orf37), mARN. (S)	GZMM	0,5066	53
granzima M (linfocito met-ase 1), mARN. (S)	TRFC	0,5024	54
receptor de transferrina (p90, CD71) (TFRC), mARN. (S)	MLKL	0,501	55
similar a dominio de kinasa de linaje mixto (MLKL), mARN. (I)	COMMD3	0,4976	56
dominio COMM que contiene 3 (COMMD3), mARN. (S)	RAB24	0,497	57
RAB24, miembro RAS familia de oncogén (RAB24), variante de transcripción 2, mARN. (A)	LOC645385	0,4966	58
PRONOSTICADO: similar a ribonucleoproteína nuclear heterogénea A1 (LOC645385), mARN. (S)	RBM14	0,4948	59
proteína de motivo de enlace de ARN 14 (RBM14), mARN. (S)	PSCD4	0,4928	60
homología de pleckstrina, Sec7 y dominios arrollados en espiral 4 (PSCD4), mARN. (S)	ZDHHC7	0,489	61
dedo de zinc, tipo DHHC que contiene 7 (ZDHHC7), mARN. (S)	PRKCH	0,4886	62
proteína hipotética MGC11257 (MGC11257), mARN. (S)	MGC11257	0,4854	63
proteína de 105 kDa/110 kDa de choque térmico 1 (HSPH1), mARN. (S)	HSPH1	0,4812	64
receptor de retinoide X, alfa (RXRA), mARN. (S)	RXRA	0,481	65
homólogo de bicaudal D 2 (Drosofila) (BICD2), variante de transcripción 1, mARN. (A)	BICD2	0,4756	66
familia portadora de soluto 27 (transportador de ácido graso), miembro 3 (SLC27A3), mARN. (S)	SLC27A3	0,47	67
antígeno CD96 (CD96), variante de transcripción 1, mARN. (A)	CD96	0,4688	68
proteína ribosómica S2 (RPS2), mARN. (S)	RPS2	0,4662	69
substrato receptor de insulina 2 (IRS2), mARN. (S)	IRS2	0,4654	70
fosfatasa de tirosina de proteína, substrato de tipo no receptor I (PTPNS 1), mARN. (S)	PTPNS1	0,4612	71
similar a estimulador de disociación de nucleótido de guanina ral 2 (RGL2), mARN. (S)	RGL2	0,457	72
PRONOSTICADO: similar a proteína de tumor controlado traslacional-mente (TCTP) (p23) (factor de liberación de histamina) (HRF) (Fortilina) (LOC643870), mARN. (S)	LO643870	0,4566	73
MID1 proteína de interacción 1 (similar a G12 específica de gastrulación (pez cebra)) (MID1IP1), mARN. (S)	MID1IP1	0,454	74
familia portadora de soluto 7 (transportador de aminoácido catiónico, sistema y^+), miembro 7 (SLC7A7), mARN. (S)	SLC7A7	0,4502	75
proteína de enlace FK506 11, 19 kDa (FKBP11), mARN. (S)	FKBP11	0,4492	76
dominio SH2 que contiene 3C (SHD3C), variante de transcripción 2, mARN. (S)	SH2D3C	0,4454	77
factor de intercambio de nucleótido de guanina rho/rac (GEF) 2 (ARHGEF2), mARN. (S)	ARHGEF2	0,4444	78
nucleoporina 62 kDa (NUP62), variante de transcripción 1, mARN. (A)	NUP62	0,4424	79
proteína hipotética FLJ20186 (FL120186), variante de transcripción 1, mARN. (I)	FLJ20186	0,438	80
ATPasa, transportador de H^+, 56/58 kDa lisosómico, V1 subunidad B, isoforma 2 (ATP6V1B2), mARN. (S)	ATP6V1B2	0,436	81
homólogo de oncogén relacionado con sarcoma viral de Yamaguchi v-yes-	LYN	0,4358	82

1 (LYN), mARN. (S)			
factor de necrosis tumoral, proteína inducida alfa 2 (TNFAIP2), mARN. (S)	TNFAIP2	0,433	83
ST3 beta-galactosida alfa-2,3-sialiltransferasa 1 (ST3GAL1), variante de transcripción 2, mARN. (A)	ST3GAL1	0,4318	84
similar a proteína asociada a receptor GABA(A) 1 (GABARAPL1), mARN. (S)	GABARAPL1	0,4276	85
homólogo de enzima descubierta DCP2 (S. cerevisiae) (DCP2), mARN. (S)	DCP2	0,4272	86
familia con similitud de secuencia 46, miembro A (FAM46A), mARN. (S)	FAM46A	0,4266	87
proteína ribosómica mitocóndrica L51 (MRPL51), proteína mitocóndrica que codifica gen nuclear, mARN. (S)	MRPL51	0,4256	89
similar a ligando 4 de quimiocina (motivo C-C) 1 (CCL4L1), mARN. (S)	CCL4L1	0,4208	90
proteína mitocóndrica que codifica gen nuclear, desoxiguanosina kinasa, variante e transcripción 1, mARN. (A)	DGUOK	0,4204	91
linfomas de célula T frecuentemente repetidos por adelantado 2 (FRAT2), mARN. (S)	FRAT2	0,4202	92
proteína ligante de kinasa de dominio SH3 1 (SH3KBP1), variante de transcripción 1, mARN. (I)	SH3KBP1	0,4172	93
fosfatasa de especificidad dual 2 (DUSP2), mARN. (S)	DUSP2	0,4172	94
factor de iniciación de traducción eucariótica 2B, subunidad 4 delta, 67 kDa, variante de transcripción 1, mARN. (A)	EIF2B4	0,4136	95
similar a fibrinógeno 2 (FGL2), mARN. (S)	FGL2	0,4126	96
glucosidasa, alfa; AB neutra (GANAB), variante de transcripción 2, mARN. (A)	GANAB	0,4112	97
proteína ligante de CCAAT/potenciador (C/EBP), alfa (CEBPA), mARN. (S)	CEBPA	0,41	98
policarboxipeptidasa (angiotensinasa C) (PRCP), variante de transcripción 2, mARN. (A)	PRCP	0,4046	99
succinato–CoA ligasa, formador de GDP, subunidad beta (SUCLG2), mARN. (S)	SUCLG2	0,4012	100

TABLA III

NOMBRE DE GEN	**Símbolo**	**Puntuación**	**Orden**
familia de dominio TSC22, miembro 3 (TSC22D3), variante de transcripción 2, mARN. (A)	TSC22D3	0,9522	1
receptor de quimiocina (motivo C-X-C) 4, variante de transcripción 1, mARN. (A)	CXCR4	0,9444	2
polipéptido ligero, citoplásmico, de dineína 1 (DNCL1), mARN. (S)	DNCL1	0,8668	3
proteína ribosómica S3 (RPS3), mARN. (S)	RPS3	0,8556	4
transcripción inducible por daños de ADN 4 (DDIT4), mARN. (S)	DDIT4	0,8502	5
granzima B (granzima 2, estearasa de serina asociada a linfocito T citotóxica 1) (GZMB), mARN. (S)	GZMB	0,8148	6
gen de traslocación de célula B, anti-proliferativa (BTG1), mARN. (S)	BTG1	0,8	7
proteína de 70 kDa de choque térmico 8 (HSPA8), variante de transcripción 1, mARN. (I)	HSPA8	0,793	8
proteína ribosómica L12 (RPL12), mARN. (S)	RPL12	0,7564	9
adaptador similar a Src (SLA), mARN. (S)	SLA	0,7322	10
factor de transcripción relacionado con runt 3 (RUNX3), variante de transcripción 2, mARN. (I)	RUNX3	0,7306	11
gen HGFL (MGC17330), mARN. (S)	MGC17330	0,6982	12
proteína de 70 kDa de choque térmico 1ª (HSPA1A), mARN. (S)	HSPA1A	0,684	13

proteína accesoria de receptor de interleukina 18 (IL18RAP), mARN. (S)	IL18RAP	0,6728	14
proteína de enlace de ARN inducible por frío (CIRBP), mARN. (S)	CIRBP	0,67	15
adrenomedulina (ADM), mARN. (S)	ADM	0,662	16
proteína de enlace de CCAAT/potenciador (E/EBP), beta (CEBPB), mARN. (S)	CEBPB	0,654	17
PRONOSTICADO: similar a ribonucleoproteína nuclear heterogénea A1 (LOC645385), mARN. (S)	LOC645385	0,654	18
Proteína de enlace de CCAAT/potenciador (C/EBP), delta (CEPBD), mARN. (S)	CEBPD	0,6416	19
factor similar de Kruppel 9 (KLF9), mARN. (S)	KLF9	0,6392	20
PRONOSTICADO: proteína hipotética (LOC440345), variante de transcripción 6 (LOC440345), mARN. (I)	LOC440345	0,6358	21
inhibidor de ligante de ADN 2, proteína de hélice-bucle-hélice negativa dominante (ID2), mARN. (S)	ID2	0,617	22
receptor similar a Ig de célula asesina, dos dominios, cola citoplásmica larga 3, variante de transcripción 2, mARN. (A)	KIR2DL3	0,6126	23
proteína de activación de araquidonato 5-lipoxigenasa (ALOX5AP), mARN. (S)	ALOX5AP	0,6106	24
superfamilia de inmunoglobulina, miembro 6 (IGSF6), mARN. (S)	IGSF6	0,6068	25
proteína de 70 kDa de choque térmico 8 (HSPA8), variante de transcripción 2, mARN (A)	HSPA8	0,6032	27
tubulina, alfa, ubicua (K-ALPHA-1), mARN. (S)	K-ALPHA-1	0,6002	28
proteína kinasa C, delta (PRKCD), variante de transcripción 2, mARN. (A)	PRKCD	0,5992	29
dominio de PR que contiene 1, con dominio de ZNF (PRDM1), variante de transcripción 1, mARN. (A)	PRDM1	0,594	30
antígeno CD55, factor de aceleración de decadencia para complemento (grupo sanguíneo de Cromer) (CD55), mARN. (S)	CD55	0,5722	31
cistatina F (leucocistatina) (CST7), mARN. (S)	CST7	0,5698	32
marcador de diferenciación asociado a mieloide (MYADM), variante de transcripción 4, mARN. (A)	MYADM	0,568	33
complejo de histocompatibilidad principal, clase I, F (HLA-F), mARN. (S)	HLA-F	0,568	34
proteína de dominio SH2 2ª (SH2D2A), mARN. (S)	SH2D2A	0,5656	35
dominio de tetramerización de canal de potasio que contiene 12 (KCTD12), mARN. (S)	KCTD12	0,5638	36
proteína de activación de Ras-GTPasa proteína ligante de dominio SH3 (G3BP), variante de transcripción 1, mARN. (A)	G3BP	0,5636	37
similar a fibrinógeno 2 (FGL2), mARN. (S)	FGL2	0,5552	38
proteína de enlace de CCAAT/potenciador (C/EBP), alfa (CEBPA), mARN. (S)	CEBPA	0,5368	39
homólogo de DnaJ (Hsp40), subfamilia A, miembro 1 (DNAJAI), mARN. (S)	DNAJA1	0,5306	40
proteína de nivelación (filamento de actina) línea músculo Z, alfa 2 (CAPZA2), mARN. (S)	CAPZA2	0,5244	41
factor de transcripción general IIIA (GTF3A), mARN. (S)	GTF3A	0,523	42
dominio de IBR que contiene 2 (IBRDC2), mARN. (S)	IBRDC2	0,5228	43
gen de exonucleasa estimada por interferon 20 kDa (ISG20), mARN. (S)	ISG20	0,5208	44
PRONOSTICADO: similar a proteína ribosómica L13a, variante de transcripción 4 (LOC649564), mARN. (A)	LOC649564	0,5134	45
receptor acoplado a proteína G 171 (GPR171), mARN. (S)	GPR171	0,5124	46
receptor similar a inmunoglobulina de célula asesina, dos dominios, cola citoplásmica larga, 4 (KIR2DL4), mARN. (S)	KIR2DL4	0,5044	47
Polipéptido asociado a sin3, 30 kDa (SAP30), mARN. (S)	SAP30	0,4972	48
PRONOSTICADO: similar a regulador de diferenciación de célula glial,	METRNL	0,4936	49

meteorina (METRNL), mARN. (I)			
canal intracelular cloruro 3 (CLIC3), mARN. (S)	CLIC3	0,4926	50
factor de iniciación de traducción eucariótica 3, subunidad 12 (EIF3S12), mARN. (S)	EIF3S12	0,4912	51
substrato receptor de insulina (IRS2), mARN. (S)	IRS2	0,4824	52
receptor celular de virus de hepatitis A 2 (HAVCR2), mARN. (S)	HAVCR2	0,4758	53
dominio de HD que contiene 2 (HDDC2), mARN. (S)	HDDC2	0,4754	54
factor de exportación de ARN nuclear 1 (NXF1), mARN. (S)	NXF1	0,468	55
perforina 1 (proteína de formación de poro) (PRF1), mARN. (S)	PRF1	0,4642	56
señales de localización nuclear y de dominio SH3, de dominio SAM (SAMSN1), mARN. (S)	SAMSN1	0,4614	57
TERF1 (TRF1)-factor nuclear de interacción 2 (TINF2), mARN. (S)	TINF2	0,4604	58
retículo endoplásmico-compartimento intermedio de golgi (ERGIC) 1, variante de transcripción 1, mARN. (I)	ERGIC1	0,4554	59
factor de necrosis tumoral, proteína inducida alfa 2 (TNFAIP2), mARN. (S)	TNFAIP2	0,455	60
factor de transcripción de gancho AT (AKNA), mARN. (S)	AKNA	0,4548	61
proteína relacionada con diferenciación adiposa (ADFP), mARN. (S)	ADFP	0,4546	62
piruvato deshidrogenasa kinasa, isozoma 4 (PDK4), mARN. (S)	PDK4	0,4538	63
factor de activación de peptidasa apoptótica (APAF1), variante de transcripción 5, mARN. (A)	APAF1	0,4486	64
transductor de señal y activador de transcripción 4 (STAT4), mARN. (S)	STAT4	0,4478	65
familia de aldo-keto reductasa 1, miembro C3 (3-alfa hidroxiesteroide deshidrogenasa, tipo II), mARN. (S)	AKR1C3	0,4454	66
dominio SH2 que contiene 3C (SH2D3C), variante de transcripción 2, mARN. (I)	SH2D3C	0,4444	67
proteína de 105 kDa/110 kDa de choque térmico 1 (HSPH1), mARN. (S)	HSPH1	0,4396	68
fosfoinositida-3-kinasa, subunidad reguladora 1 (p85 alfa) (PIK3R1), variante de transcripción 2, mARN. (A)	PIK3R1	0,4312	69
presenilina asociada, similar a romboide (PSARL), mARN. (S)	PSARL	0,4284	70
desoxiguanosina kinasa. Proteína mitocóndrica que codifica un gen nuclear, variante de transcripción 1, mARN. (A)	DGUOK	0,4272	71
homología de pleckstrina, Sec7 y dominios arrollados en espiral, proteína de enlace (PSCDBP), mARN. (S)	PSCDBP	0,4206	72
uridina fosforilasa 1 (UPP1), variante de transcripción 2, mARN. (A)	UPP1	0,4188	73
familia portadora de soluto 35 (CMP-transportador de ácido siálico), miembro A1 (SLC35A1), mARN. (S)	SLC35A1	0,4176	74
proteína kinasa kinasa kinasa activada por mitógeno 8 (MAP3K8), mARN. (S)	MAP3K8	0,4162	75
cuando de lectura de cromosoma 15 abierto 39 (C15orf39), mARN. (S)	C15orf39	0,411	76
proteína ribosómica L35 (RPL35), mARN. (S)	RPL35	0,4106	77
factor de intercambio de nucleótido de guanina rho/rac (GEF) 2 (ARHGEF2), mARN. (S)	ARHGEF2	0,4074	78
cuadro de lectura de cromosoma 19 abierto 37 (C19orf37), mARN. (S)	C19orf37	0,4072	79
proteína de motivo de enlace de ARN 14 (RBM14), mARN. (S)	RBM14	0,4068	80
proteína hipotética MGC7036 (MGC7036), mARN. (S)	MGC7036	0,4056	81
poli(A) polimerasa alfa (PAPOLA), mARN. (S)	PAPOLA	0,4044	82
RAB 10, familia de oncogén de miembro RAS (RAB10), mARN. (S)	RAB10	0,403	83
cuadro de lectura de cromosoma 2 abierto 28 (C2orf28), variante de transcripción 2, mARN. (A)	C2orf28	0,403	84
dominio LIM solamente 2 (similar a rombotina 1) (LMO2), mARN. (S)	LMO2	0,3972	85
polimerasa (ARN) III (ADN dirigido) polipéptido G (32KD) similar (POLR3GL), mARN. (S)	POLR3GL	0,3968	86

dedo de zinc y dominio de BTB que contiene 16 (ZBTB16), variante de transcripción 1, mARN. (A)	ZBTB16	0,3948	87
factor 3 de traducción eucariótica, HSCARG subunidad 5 épsilon, 47 kDa (EIF3S5), mARN. (S)	EIF3S5	0,3924	88
proteína HSCARG (HSCARG), mARN. (S)	HSCARG	0,3916	89
similar a sinaptotagmina 3 (SYTL3), mARN. (S)	SYTL3	0,3896	90
proteína hipotética FLJ32028 (FLJ32028), mARN. (S)	FLJ32028	0,3886	91
repetición rica en leucina que contiene 33 (LRRC33), mARN. (S)	LRRC33	0,3862	92
cuadro de lectura de cromosoma 1 abierto 162 (Clorf162), mARN. (S)	Clorf162	0,3846	93
citocromo P450, familia 2, subfamilia R, polipéptido 1 (CYP2R1), mARN. (S)	CYP2R1	0,3846	94
proto–oncogén jun D (JUND), mARN. (S)	JUND	0,381	95
familia D de antígeno de melanoma, 1 (MAGED1), variante de transcripción 1, mARN. (A)	MAGED1	0,3806	96
candidato de susceptibilidad de autismo 2 (ATS2), mARN. (S)	AUTS2	0,3806	97
factor de transcripción de oligodendrocito 1 (OLIG1), mARN. (S)	OLIG1	0,379	98
factor de elongación de traducción eucariótica 1 delta (proteína de intercambio de nucleótido de guanina), variante de transcripción 1, mARN. (A)	EEF1D	0,3776	99
subfamilia K de receptor similar a lectina de célula asesina, miembro 1 (KLRK1)m mARN. (S)	KLRK1	0,3736	100

TABLA IV

Clasificadores superiores de 15 genes			
Orden	**TODOS/NHC**	**AC/NHC**	**PRE/POST**
1	IGSF6	ETS1	TSC22D3
2	**HSPA8(A)**	**CCL5**	**CXCR4**
3	LYN	**DDIT4**	**DNCL1**
4	**DNCL1**	**CSCR4**	RPS3
5	**HSPA1A**	**DNCL1**	**DDIT4**
6	**DPYSL2**	MS4A6A	GZMB
7	NAGK	ATP5B	BTG1
8	**HSPA8(I)**	**HSPA8(A)**	**HSPA8(I)**
9	**NFKBIA**	**ADM**	**RPL12**
10	**FGL2**	**PTPN6**	**SLA**
11	**CALM2**	**ARHGAP9**	**RUNX3**
12	**CCL5**	**S100AB**	**MGC17330**
13	**RPS2**	**DPYSL2**	**HSPA1A**
14	**DDIT4**	**HSPA1A**	**IL18RAP**
15	**C12orf63**	**NFKBIA**	**CIRBP**

TABLA V

ID mancha	Núm. Adhes.	NOMBRE DE GEN	Símbolo	Orden	Fold Chg
5490167	Nm_016578	proteína asociada a virus x de hepatitis B (HBXAP), mARN. (S); o alternativamente, denominado factor de remodelación y partición 1	HBXAP o RSF1	1	1,27

3890735	Num_003583	Kinasa 2 regulada por fosforilación de tirosina-(Y) de doble especificidad (DYRK2), variante de transcripción 1, mARN. (A)	DYRK2	2	-1,34
3840377	NM_033430	factor de transcripción de YY1 (YY1), mARN. (S)	YY1	3	-1,08
1470605	NM_001031726	lectura de cromosoma 19 abierto 12, variante de transcripción 1, mARN. (I)	C19orf12	4	1,36
4230709	NM_018473	miembro 2 de superfamilia de tioesterasa (THEM2), mARN. (S)	THEM2	5	-1,13
1430678	NM_007118	dominio funcional triple (interacción de PTRF) (TRIO), mARN. (S)	TRIO	6	-1,16
1340086	NM_001020820	marcador de diferenciación asociado a mieloide, variante de transcripción 4, mARN. (A)	MYADM	7	-1,34
2940370	NM_017450	proteína asociada a BAI1 2 (BAIAP2), variante de trascripción 1, mARN. (I)	BAIAP2	8	-1,34
6400075	NM_024589	mARN de proteína de dominio de cremallera de leucina. (S); o alternativamente homólogo de Rogdi (Drosofila)	FLJ22386 o ROGDI	9	-1,18
20196	NM_024920	Homólogo DnaJ (Hsp40), subfamilia B, miembro 14 (DNAJB14), variante de transcripción 2, mARN. (I)	DNAJB14	10	-1,14
7330360	NM_199191	cerebro y órgano reproductivo expresado (modulador de TNFRSF1A) (BRE), variante de transcripción 3, mARN. (A)	BRE	11	1,04
240280	NM_080652	proteína de transmembrana 41 A (TNMM41A), mARN. (S)	TNEM41A	12	1,15
3940687	NM_032307	cuadro de lectura de cromosoma 9 abierto 64 (C9orf64), mARN. (S)	Corf64	13	-1,14
4150253	NM_031424	cuadro de lectura de cromosoma 20 abierto 55, variante de transcripción 1, mARN. (A); o alternativamente, familia con similitud de secuencia 110, miembro A	C20orf55 o FAM110A	14	-1,14
1660445	NM_014801	similar a pecanex 2 (Drosofila), variante de transcripción 1, mARN. (I)	PCNXL2	15	1,21
4120187	NM_005612	factor de transcripción de silenciamiento de RE1 (REST), mARN. (S)	REST	16	1,29
7610494	NM_014173	proteína HSPC142 (HSPC142), variante de transcripción 2, mARN. (A); o alternativamente, cuadro de lectura de cromosoma 19 abierto 62	HSPC142 o C19orf62	17	1,10
4250121	NM_138779	proteína hipotética BC015148 (LOC93081), mARN. (S); o alternativamente, cuadro de lectura de cromosoma 13 abierto 27	LOC93081 o C13orf27	18	-1,18
4810674	NM_022091	subunidad 3 compleja de co-integrador de señal de activación 1(ASCC3), variante de transcripción 2, mARN. (A)	ASCC3	19	1,83
3460224	NM_005628	familia portadora de soluto 1 (transportador de aminoácido neutro), miembro 5 (SLC1A5), mARN. (S)	SLC1A5	20	-1,16
1110110	NM_016395	dominio de fosfatasa-A de tirosina de proteína que contiene 1, mARN. (A)	PTPLAD1	21	-1,22

2630397	NM_005590	homólogo A de recombinación meiótica 11 de MRE11 (S. cerevisiae) (MRE11A), variante de transcripción 2, mARN. (A)	MRE11A	22	-1,18
1400541	NM_033107	proteína hipotética (DKFZP688A10121), mARN. (S); o alternativamente, proteína de enlace de GTP 10 (putativa), variante de transcripción 2	DKFZP686A 10121 o GTPBP10	23	-1,27
4390100	BX118737	cADN de hígado bazo fetal 1NFLS de Soares BX118737, clon IMACrp998K18127, secuencia de mARN (S)	NaN	24	-1,40
1500246	NM_006217	inhibidor de peptidasa de serpina, subtipo I (pancpina), miembro 2 (SERPINI2), variante de transcripción 2, mARN. (S)	SERPINI2	25	-1,41
6590377	AK 126342	FLJ44370 fis de cADN, clon TRACH3008902; o alternativamente, proteína de enlace de elemento de respuesta de CAMP 1	NaN o CREB1	26	-1,45
3710754	NM_016053	dominio arrollado en espiral que contiene 53 (CCDC53), mARN. (S)	CCDC53	27	-1,07
990112	NM_032236	peptidasa específica de ubiquitina 48 (USP48), variante de transcripción 1, mARN. (I)	USP48	28	-1,17
2640255	NM_001007072	dedo de zinc y dominio SCAN que contiene 2, variante de transcripción 3, mARN (I)	ZSCAN2	29	1,18
2370482	NM_024754	dominio de repetición de pentatricopéptido 2 (PTCD2), mARN. (S)	PTCD2	30	
6380040	NM_025201	dominio de homología de pleckstrina que contiene familia Q miembro 1 mARN. (S)	PLEKHQ1	31	
6370338	AW191734	secuencia (S) de mARN, cADN de visualización diferencial de cADN de isleta humana de HIMC10.07.00	NaN	32	
5340544	NM_002616	homólogo de período 1 (Drosofila) (PER1), mARN. (S)	PER1	33	
5910367	NM_012154	factor 2C de iniciación de traducción eurocariótica, 2 (EIF2C2), mARN. (S)	EIF2C2	34	
2570440	NM_022128	ribokinasa (RBKS), mARN. (S)	RBKS	35	
6100707	NM_002419	kinasa kinasa kinasa de proteína activada por mitogén, maRN. (S)	MAP3K11	36	
2490615	NM_207443	proteína FLJ45244 (FLJ5244), mARN. (S)	FLJ5244	37	
6580368	NM_006611	subfamilia A de receptor similar a lectina de célula asesina, mARN. (S)	KLRA1	38	
4570553	NM_016282	adenilato kinasa 3 (AK3), mARN. (S)	AK3	39	
5130500	BG741535	IMAGEN de clon de cADN 602635144F1 NCI_CGAP_Skn3: 4780090 5, secuencia de mARN. (S)	NaN	40	
1240026	NM_001003941	oxoglutarato (alfa-ketoglutarato) deshidrogenasa (lipoamida), proteína mitocóndrica de codificación de gen nuclear, variante de transcripción 2, mARN. (I)	OGDH	41	

2680593	NM_006582	proteína de enlace de elemento modulador de glucocorticoide 1 (GMEB1), variante de transcripción 1, mARN. (A)	GMEB1	42	
130403	NM_006567	fenilalanina–tARN sintetasa 2 (mitocóndrica) (FARS2), proteína mitocóndrica de codificación de gen nuclear, mARN. (S)	FARS2	43	
1710338	NM_170768	homólogo de proteína 91 de dedo de zinc (ratón), variante de transcripción 2, mARN. (A)	ZFP91	44	
150021	NM_013285	similar a proteína de enlace de nucleótido de guanina (nucleolar) (GNL2), mARN. (S)	GNL2	45	
4250703	XM_498909	PTRONOSTICADO: hipotética LOC440900 (LOC440900), mARN. (S)	LOC440900	46	
7000731	NM_020453	ATPasa, Clase V, tipo 10D (ATP10D), mARN, (S)	ATP10D	47	
4590563	XM_942240	PRONOSTICADO: similar a antígeno de histocompatibilidad HLA clase II, DQ (W1.1) precursor de cadena beta (DQB1 *0501), variante de transcripción 1 (LOC650557), mARN. (A)	LOC650557	48	
3310446	NM_018169	Cuadro de lectura de cromosoma 12 abierto 35 (C12orf35), mARN. (S)	C12orf35	49	
3460066	XM_932088	PRONOSTICADO: proteína hipotética LOC642788, variante de transcripción 2 (LOC642788), mARN. (A)	LOC642788	50	
160152	NM_003789	TNFRSF1A asociada a través de dominio de muerte, variante de transcripción 1, mARN. (A)	TRADD	51	
840379	NM_031212	familia portadora de soluto, miembro 28 (SLC25A28), mARN. (S)	SLC25A28	52	
405402	B459101	BX459101 cADN de placenta clon CSODE012YP17 5-PRIME, secuencia (S) de mARN	NaN	53	
3440441	AK124002	cADN FLJ42008 fis, clon SPLEN2031724 (S)	NaN	54	
5390504	NM_001165	IAP baculovírico repetición que contiene 3, variante de transcripción 1, maRN. (I)	BIRC3	55	
5490564	XM_940798	PRONOSTICADO: similar a factor 1 de transición asociado a Bcl-2 (Btf), variante de transcripción 1 (LOC650759), mARN. (I)	LOC650759	56	
1940220	XM_940538	PRONOSTICADO: similar a fosfatasa de tirosina de proteína, dominio A que contiene 1 (PTPLAD1), mARN. (A)	PTPLAD1	57	
770221	NM_005950	metalotioneína 1G (MTG1G), mARN. (S)	MT1G	58	
1500647	NM_005665	sitio 5 de integración vírica ecotrópica (EVI5), mARN. (S)	EVI5	59	
5900730	NM_005813	proteína kinasa D3 (PRKD3), mARN. (S)	PRKD3	60	
1980689	NM_024029	familia de dominio Yip1, miembro 2 (YIPF2), mARN. (S)	YIPF2	61	
770253	NM_024076	canal de potasio que contiene dominio	KCTD15	62	

		de tetramerización 15, mARN. (S)			
2260484	NM_022070	amplificado en cáncer de pecho 1 (ABC1), mARN. (S)	ABC1	63	
380561	NM_020773	familia de dominio TBC1, miembro 14 (TBC1D14), mARN (S)	TBC1D14	64	
780576	NM_014238	supresor de kinasa de ras I, mARN. (S)	KSR1	65	
240292	BG564169	clon de cADN 602590145F1 NIH_MGC_76, IMAGEN: 4724074 5, mARN secuencia (S)	NaN	66	
6590021	NM_024804	proteína de dedo de zinc 669 (ZNF669), mARN. (S)	ZNF669	67	
6330471	NM_005337	cuadro de lectura de cromosoma 8 abierto 1 (C8orf1), mARN. (S)	C8orf1	68	
3170398	NM_000747	receptor colinérgico, nicotínico, beta 1 (músculo) (CHRNB1), mARN. (S)	CHRNB1	69	
3170477	NM_001004489	receptor olfativo, familia 2, subfamilia AG, miembro 1, mARN. (S)	OR2AG1	70	
2510563	NM_024874	similar a KIAA0319 (KIAA0319L), variante de transcripción 1, mARN. (I)	KIAA0319L	71	
2510280	NM_015106	similar a RAD54 2 (S. cerevisiae) (RAD54L2), mARN. (S)	RAD54L2	72	
670685	NM_003557	fosfatidilinositol-4- fosfato 5-kinasa, tipo I, alfa, mARN. (S)	PIP5K1A	73	
4230736	NM_001329	proteína de enlace de terminal-C 2 (CTBP2), variante de transcripción 1, mARN. (I)	CTBP2	74	
7510164	XM_938545	PRONOSTICADO: similar a proteína 3 de enlace de formina 3 (proteína de enlace de formina 11) (FBP 11), variante de transcripción 1, (LOC648039), mARN. (I)	LOC648039	75	
4210576	NM_022490	factor asociado a polimerasa I (RNA) 1 (PRAF1), mARN. (S)	PRAF1	76	
5910376	NM_003246	trombospondina 1 (THBS 1), mARN. (S)	THBS1	77	
2480202	NM_006933	familia portadora de soluto 5 (transportadores de inositol), miembro 3, mARN. (S)	SLC5A3	78	
5960035	NM_170699	receptor de ácido biliar acoplado a proteína G 1 (GPBAR1), mARN. (S)	GPBAR1	79	
5290192	CR616845	clon de cADN de longitud completa CS0DF020Y de cerebro fetal de (humano) (S)	NaN	80	
1170301	NM_014572	LATS, supresor de tumor grande, homólogo 2 (Drosofila), mARN. (S)	LATS2	81	
2340224	NM_181724	proteína de transmembrana 119 (TMEM119), maRN. (S)	TMEM119	82	
4210008	NM_022168	interferón inducido con dominio de helicasa C 1 (IFIH1), mARN. (S)	IFIH1	83	
3060563	CD639673	clon de cADN AGENCOURT_14534956 NIH_MGC_191, IMAGEN: 30418908 5, mARN secuencia (S)	NaN	84	
7320600	AK123531	cADN FLJ41537 fis, clon BRTHA2017985 (S)	NaN	85	
520097	NM_003541	histona 1, H4k (HIST1H4K), mARN. (S)	HIST1H4K	86	

5270315	NM_001240	ciclina T1 (CCNT1), mARN. (S)	CCNT1	87	
2690008	BC025734	Homo sapiens, clon IMAGEN: 5204729, mARN. (S)	NaN	88	
110044	NM_001001795	similar a clon C030006K11 de RIKEN cADN (MGC70857), mARN. (S)	MGC70857	89	
2030487	BX118124	clon de cADN de BX118124 Soares_parathyroid_tumor_Nb HPA IMAGp998P234189, mARN secuencia (S)	NaN	90	
1170139	NM_033141	proteína activada por mitógeno kinasa kinasa kinasa 9 (MAP3K9, mARN. (S)	MAP3K9	91	
1190300	NM_15353	Canal potasio que contiene dominio de tetramerización 2, mARN. (I)	KCTD2	92	
4760543	NM_153719	nucleoporina 62 kDa (NUP62), variante de transcripción 1, mARN. (A)	NUP62	93	
7150564	NM_003171	Similar a supresor de var1, 3 (S. cerevisiae) 1 (SUPV3L1), mARN. (S)	SUPV3L1	94	
5820475	NM_002690	polimerasa (dirigida a ADN) beta (POLB), mARN. (S)	POLB	95	
870563	N_014710	proteína de clasificación de receptor acoplado a proteína G 1, mARN. (S)	GPRASP1	96	
4640202	AW962976	re-secuencias EST375049 MAGE, MAGH cADN, mARN secuencia (S)	NaN	97	
4250332	XM_932676	PRONOSTICADO: similar a Gamma-glutamiltranspeptidasa 1 precursor (Gamma-glutamiltransferasa 1) (CD224 antígeno), variante de transcripción 3 (LOC645367), mARN. (I)	LOC645367	98	
2570017	NM_023034	similar a candidato I de síndrome de Wolf-Hirschhom 1 (WHSC1L1), variante de transcripción larga, mARN. (I)	WHSCIL1	99	
3390458	NM_002243	rectificación interior de canal de potasio, subfamilia J, miembro 15 (KCNJ15), variante de transcripción 1, mARN. (A)	KCNJ15	100	
5360053	XM_926644	PRONOSTICADO: similar a componente de 240 kDa de complejo de proteína asociado a receptor de hormona de tiroides (Trap240) (proteína asociada a receptor de hormona de tiroides 1) (componente DRIP250 complejo de proteína de interacción con receptor de vitamina D3) (DRIP 250) (co-factor reclutado por activador ... (LOC643296)), mARN. (S)	LOC643298	101	
6760653	XM_935750	PRONOSTICADO: similar a proteína Elk1 de dominio de ETS (LOC641976), mARN. (S)	LOC641976	102	
3800615	NM_080549	proteína tirosina fosfatasa, no-receptor tipo 6 (PTPN6), variante de transcripción 3, mARN. (I)	PTPN6	103	
5310452	NM_153645	Nucleoporina 50 kDa (NUP50), variante de transcripción 3, mARN. (A)	NUP50	104	
3850288	XM_934211	PRONOSTICADO: similar a homólogo proteína de biogénesis de ribosoma BMS 1, variante de transcripción 2 (LOC653471), mARN. (I)	LOC653471	105	

7560538	NM_153209	miembro 19 de familia de kinesina (KIF19), mARN. (S)	KIF19	106	
6250338	NM_152371	cuadro de lectura de cromosoma 1 abierto 93 (C1orf93), maRN. (S)	C1orf93	107	
3360382	NM_001625	adenilato kinasa 2 (AK2), variante de transcripción AK2A, maRN. (A)	AK2	108	
6960564	NM_030934	cuadro de lectura de cromosoma 1 abierto 25 (C1orf25), mARN. (S)	C1orf25	109	
1820131	XM_945571	PRONOSTICADO: familia de dominio 13 de repetición de anquirina, miembro D, variante de transcripción 7 (ANKRD13D), mARN. (I)	ANKRD13D	110	
3850255	NM_001238	ciclina E1 (CCNE1), variante de transcripción 1, mARN. (A)	CCNE1	111	
990523	NM_006799	proteasa, serina, 21 (testisina), variante de transcripción 1, mARN. (A)	PRSS21	112	
4280577	NM_006749	Familia portadora de soluto 20 (transportador de fosfato), miembro 2, mARN. (S)	SCL20A2	113	
7160368	BC039681	Homo sapiens, clon IMAGEN: 5218705, mARN. (S)	NaN	114	
6020500	NM_024923	nucleoporina 210 kDa (NUP210), mARN. (S)	NUP210	115	
2360253	NM_007041	arginiltransferasa 1 (ATE1), variante de transcripción 2, mARN. (I)	ATE1	116	
160372	NM_006761	proteína de activación de tirosina 3-monooxigenasa/triptofan 5-monooxigenasa, polipéptido épsilon (YWHAE), mARN. (I)	YWHAE	117	
3370170	BX093763	cADN BX093763 Soares_fetal_heart_NbHH 19 W, clon IMAGp998N10870, mARN secuencia (S)	NaN	118	
60546	AK057981	cADN FLJ25252 fis, clon STM03814 (S)	NaN	119	
1710411	XM_374029	PRONOSTICADO: LOC389089 hipotética (LOC389089), mARN (S)	NaN	120	
6900315	NM_017958	familia B (evectinas), que contiene dominio de homología de pleckstrina, miembro 2 (PLEKHB2), variante de transcripción 2, mARN. (I)	PLEKHB2	121	
1240603	NM_000887	integrina, alfa X (subunidad de componente 3 de complemento, receptor 4), mARN. (I)	ITGAX	122	
60707	NM_001119	aducina 1 (alfa) (ADD1), variante de transcripción 1, maRN. (A)	ADD1	123	
7160707	NM_198285	proteína hipotética LOC349136 (LOC349136), mARN. (S)	LOC349136	124	
2970332	NM_006328	proteína de motivo de enlace de ARN 14 (RBM14), mARN. (S)	RBM14	125	
2760433	NM_173564	proteína hipotética FLJ37538 (FLJ37538), mARN. (S)	FLJ37538	126	
580041	NM_001252	superfamilia de factor de necrosis de tumor (ligando), miembro 7, mARN. (S)	TNFS7	127	
4120133	NM_022827	espermatogénesis asociada 20 (SPATA20), mARN. (S)	SPATA20	128	
6560647	NM_018696	homólogo elaC, 1 (E. coli) (ELAC1),	ELAC1	129	

		mARN. (S)			
4180195	NM_001001520	proteína relacionada con factor de crecimiento derivado de hepatoma 2 (HDGF2), variante de transcripción 1, ARN. (A)	HDGF2	130	
6650020	NM_001124	adrenomedulina (ADM), mARN. (S)	ADM	131	
2570364	NM_020847	repetición de nucleótido que contiene 6A, variante de transcripción 2, mARN. (I)	TNRC6A	132	
1850682	NM_015530	proteína de apilamiento de montaje de golgi 2, 55 kDa (GORASP2), mARN. (S)	GORASP2	133	
50414	NM_006973	proteína de dedo de zinc 32 (KOX 30) /ZNF32), variante de transcripción 1, mARN. (A)	ZNF32	134	
7200373	NM_194310	proteína hipotética LOC284837 (LOC284837), mARN. (S)	LOC284837	135	
3940215	NM_015453	dominio que contiene THUMP 3 (THUMPD3), mARN. (S)	THUMPD3	136	

TABLA VI

Orden	ID-Mancha Illumina	Núm. Acc.	Nombre	Símbolo	valor-p	POS/PRE fold chg
1	3370291	NM_024514	Citocromo P450, familia 2, subfamilia R, polipéptido 1	CYP2R1	0,00000	-1,39
2	6660437	NM_006111	Acetil-Coenzima A acetiltransferasa 2	MYO5B	0,00001	-1,34
3	6380402	NM_080915	desoxiguanosina kinasa (DGUOK), proteína mitocóndrica de codificación de gene nuclear, variante de transcripción 5, mARN. (I)	DGUOK	0,00000	-1,82
4	1990500	NM_003746	Dineína, cadena ligera, tipo LC8 1	DYNLL1	0,00002	1,38
5	150048	NM_052873	Cuadro de lectura de cromosoma 14 abierto 179	C14orf179	0,00001	-1,30
6	2230731	NM 017745	co-represor BCL6	BCOR	0,00002	1,35
7	270070	BF448693	clon de cADN 7N93B04.X1 NCI.. CGAP_Ov18 IMAGEN:3571927 3, mARN secuencia (S)	NaN	0,00001	-1,57
8	6560482	NM_001280	Proteína de enlace de ARN inducible en frío	CIRBP	0,00000	-1,35
9	2970332	NM_006328	Proteína de motivo de enlace de ARN 14	RBM14	0,00006	1,25
10	3890682	NM_003975	Proteína de dominio de SH2 2A	SH2D2A	0,00000	-1,66
11	6560349	NM_018425	Fosfatidilinositol 4-kinasa tipo 2 alfa	P14K2A	0,00005	1,37
12	1710411	XM_374029	PRONOSTICADO: LOC389089 hipotética, mARN. (S)	NaN	0,00007	-1.4.4
13	1660019	NM_001876	Palmitoiltransferasa de carnitina 1A (hígado)	CPT1A	0,00003	-1,33
14	2680161	NM_006584	TCP1 que contiene chaperonina, subunidad 6B	CCT6B	0,00002	-1,58

			(zeta 2)			
15	4060270	BC009563	Homo sapiens, clon IMAGEN: 3901628, mARN. (S)	NaN	0,00006	-1,38
16	2650152	NM_020698	Transmembrana y dominio arrollado en espiral, familia 3	TMCC3	0,00014	-1,86
17	20451	NM_148976	Subunidad de proteasoma (prosoma, macropain), tipo alfa, 1	PSMA1	0,00040	-1,49
18	6220672	NM_001031711	Retículo endoplásmico-compartimento intermedio de golgi 1	ERGIC1	0,00055	-1,35
19	6840017	XM_941287	PRONOSTICADO: familia portadora de soluto 25 (carnitina/acilcarnitina translocasa), miembro 20 (SLC25A20), mARN. (A)	SLC25A20	0,00159	-1,24
20	870709	NM_006133	Diacilglicerol lipasa, alfa	DAGLA	0,00086	1,40
21	5860148	NM 007320	Proteína de enlace de ARN 3	RANBP3	0,00179	-1,38
22	20707	NM_207584	Receptor de interferón (alfa, beta y omega) 2	IFNAR2	0,00025	-1,25
23	5900156	NM 006082	Tubulina, alfa 1b	TUBA1B	0,00268	1,13
24	6480170	NM_001005333	Familia D antígena de melanoma, 1	MAGED1	0,00001	-1,27
25	4010605	NM_001008739	Similar a RIKEN cADN 2310039H09	LOC441150	0,00007	-1,24
26	7210192	NM_003123	Sialoforina (leukosialina, CD43)	SPN	0,00014	-1,89
27	4260148	X_371534	PRONOSTICADO: similar a CG10806-PB, isoforma B, mARN. (A)	LOC389000	0,00075	-1,33
28	6560020	NM_017651	Sitio de integración asistente de Abelson 1	AHI1	0,00379	-1,33
29	6480661	NM_002255	Receptor similar a Ig de célula asesina, dos dominios, cola citoplásmica larga, 4	KIR2DL4	0,00117	-2,02
30	650753	NM_006712	Serina/treonina kinasa activada por Fas	FASTK	0,00003	-1,40
31	1230528	NM_006644	Proteína de 105 kDa/110 kDa de choque térmico 1	HSPH1	0,00006	1,47
32	6420086	NM_001539	Homólogo de DnaJ(Hsp40), subfamilia A, miembro 1	DNAJA1	0,00009	1,26
33	4120092	NM_018244	Chaperona compleja de c reductasa de ubiquinol-citocromo, homólogo de CBP3 (levadura)	UQCC	0,00286	-1,40
34	4250438	NM_M5267	Cuadro de lectura de cromosoma 6 abierto 57	6orf57	0,00188	-1,15
35	5860477	NM_005226	Receptor de espingosina-1-fosfato 3	S1PR3	0,00017	1,69
36	5910037	NM_182757	Dedo anular 144B	RNF144B	0,00000	-1,97
37	6020707	NM_003416	Proteína de dedo de zinc 7	ZNF7	0,00023	-1,14
38	4260497	NM_018179	Activación de proteína de interacción de factor 7 de transcripción	ATF7IP	0,00092	1,40
39	2760068	NM_005489	Dominio SH2 que contiene 3C	SH2D3C	0,00007	1,34
40	6250056	NM_152832	Familia con similitud de secuencia 89, miembro B	FAM89B	0,00043	1,21

41	6040273	BX115698	Clon de cADN BX115698 Soares_testis_NHT IMAGo998M211829, mARN secuencia (S)	NaN	0,00031	-1,37
42	1990100	XM_930024	PRONOSTICADO: proteína hipotética LOC132241, variante de transcripción 2 (LOC32241), mARN. (A)	LOC 132241	0,00005	-1,21
43	264066	NM_001008910	Serina/treonina kinasa 16	STK16	0,00000	-1,90
44	770605	NM_145271	Proteína de dedo de zinc 688	ZNF688	0,00000	-1,56
45	7200356	NM 001008541	Interactor MAX 1	MXI1	0,00192	1,55
46	1690709	NM_024815	Motivo de tipo Nudix (difosfato nucleósido enlazado a porción X) 18	NUDE18	0,00167	-1,20
47	100743	NM_004089	Familia de dominio TSC22, miembro 3	TSC22D3	0,00003	-1,40
48	2100201	NM_015558	Gen de translocación de sarcoma sinovial sobre cromosoma 18 similar 1 (SS18L1), variante de transcripción 2, maRN. (A)	SS18L1	0,00008	1,19
49	1820209	NM_001659	Factor de ribosilación de ADP 3	ARF3	0,00090	1,19
50	1780762	NM_032847	Cuadro de lectura de cromosoma 8 abierto 76	C8orf76	0,00037	-1,15

TABLA VII

#	Orden	ID	Núm. Acc.	Descripción Nombre de Gen	Símbolo	Valor-p	Fold Chg
1		4880431	NM_181738	Peroxirredoxina 2	PRDX2	0,00000	1,42
2	16	4120187	NM_005612	Factor de transcripción de silenciamiento de RE-1	REST	0,00034	1,40
3		4590563	XM_942240	PRONOSTICADO: similar a antígeno de histocompatibilidad clase II, DQ(W1.1) precursor de cadena beta (DQB1*0501), variante de transcripción 1	LOC650557	0,00042	-2,41
4		7210129	NM_178025	Similar a gamma-glutamiltransferasa 3 (GGTL3), variante de transcripción 2	GGTL3	0,00018	1,35
5	19	4810674	NM_022091	Activar subunidad 3 compleja de co-integrador de señal 1	ASCC3	0,00274	1,73
6		4280722	NM_005481	Subunidad compleja de mediador 16	MED16	0,00027	1,22
7	23	1400541	NM_033107	Proteína de enlace de GTP 10 (putativa)	GTPBP10	0,00559	-1,26
8		1190022	NM_176895	Fosfatasa de ácido fosfatídico tipo 2A	PPAP2A	0,00355	1,20
9		3060692	NM_001010935	RAP1A, miembro de	RAP1A	0,00018	-1,35

				la familia de oncogén RAS			
10		2570440	NM_022128	Cerebro y órgano reproductivo expresado (modulador de TNFRSF1A)	BRE	0,00282	1,10
11		4060138	XM_941904	PRONOSTICADO: similar a regulador transcripcional ATRX (helicasa II con enlace X) (proteína nuclear con enlace X) (XNP)	LOC652455	0,00029	1,47
12		6180296	NM_001017969	KIAA2026	KIAA2026	0,00028	-1,15
13		1430292	NM_000578	Familia portadora de soluto 11 (transportadores de ion de metal divalente acoplado a protón), miembro 1	SLC11A1	0,00006	-1,41
14		110112	NM_005701	Snurportin 1	SNUPN	0,00033	-1,17
15		6330471	NM_004337	Familia de inhibidor de crecimiento inducido de fatiga oxidativa, miembro 2	QSGIN2	0,00204	-1,09
16		5050019	XM_945607	PRONOSTICADO: paraplejia espástica 21 (recesiva autosómica, síndrome de Mast), variante de transcripción 3 (SPG21), mARN	SPG21	0,2419	1,13
17	4	1470605	NM_001031726	Cuadro de lectura de cromosoma 19 abierto 12	C19orf12	0,00382	1,43
18		6620224	NM_001024662	Proteína ribosómica L6	RPL6	0,00350	-1,03
19		4250133	NM_005188	Secuencia de transformación retroviral ecotrópica de Cas-Br-M (murina)	CBL	0,00001	-1,18
20	9	6400075	NM_024589	Homólogo de Rogdi (Drosofila)	ROGDI	0,00023	1,32
21		6580419	NM_001015880	3'-fosfoadenosina 5'-fosfosulfato sintasa 2	PAPSS2	0,00341	-1,34
22	8	2940370	NM_017450	Proteína asociada a BAI1, 2	BAIAP2	0,00046	-1,38
23		3360026	NM_017911	Familia con similitud de secuencia 118, miembro A	FAM118A	0,01598	1,94
24	6	1430678	NM_007118	Dominio funcional triple (interacción de PTPRF)	TRIO	0,00001	-1,31

Para su uso en las composiciones indicadas en lo que antecede, los imprimadores y sondas de PCR están preferentemente diseñados en base a secuencias de intrón presentes en el (los) gen(es) que va(n) a ser

amplificado(s), seleccionado(s) a partir del perfil de expresión genética. El diseño de las secuencias de imprimador y de sonda está dentro del alcance del experto en la materia una vez que ha sido seleccionado el objetivo genético particular. Los métodos particulares seleccionados para el diseño de imprimador y de sonda y las secuencias particulares de imprimador y de sonda no son características limitadoras de estas composiciones. Una explicación rápida de técnicas de diseño de imprimador y de sonda disponibles para los expertos en la materia está resumida en la Patente US núm. 7.081.340, con referencia a herramientas públicamente disponibles tales como el software DNA BLAST, el programa Repeat Masker (Colegio Baylor de Medicina), Primer Express (Biosistemas Aplicados); MGB assay-by-design (Biosistemas Aplicados); Primer3 (Steve Rozen y Helen J. Skaletsky (2000) Primer3 en la WWW para usuarios en general y para programadores biólogos[85] y otras publicaciones[86, 87, 88].

En general, los imprimadores y sondas de PCR óptimos usados en las composiciones descritas en la presente memoria son en general de 17-30 bases de longitud, y contienen aproximadamente un 20-80%, tal como por ejemplo aproximadamente un 50-60%, de bases G+C. Las temperaturas de fusión de entre 50 y 80 ºC, por ejemplo de alrededor de 50 a 70 ºC, son típicamente las preferidas.

En un aspecto, una composición para el diagnóstico de cáncer de pulmón en un mamífero contiene una pluralidad de polinucleótidos inmovilizados sobre un substrato, en el que la pluralidad de sondas genómicas hibridizan a tres o más productos de expresión genética de tres o más genes informativos seleccionados a partir de un perfil de expresión genética de las células mononucleares de sangre periférica (PBMC) del sujeto, comprendiendo el perfil de expresión genética genes seleccionados a partir de la Tabla I hasta la Tabla VII. Este tipo de composición se basa en el reconocimiento de los mismos perfiles genéticos que se han descrito en lo que antecede para las composiciones de PCR, pero emplea las técnicas de una matriz de cADN. La hibridación de los polinucleótidos inmovilizados en la composición respecto a los productos de expresión de gen presentes en la PBMC del paciente, se emplea para cuantificar la expresión de los genes informativos seleccionados a partir de los genes identificados en las Tablas I a VII para generar un perfil de expresión genética para el paciente, el cual se compara a continuación con el de una muestra de referencia. Según se ha descrito en lo que antecede, dependiendo de la identificación del perfil (es decir, el de los genes de la Tabla I, II, III, IV, V, VI o VII o subconjuntos de los mismos), esta composición permite la diagnosis y la prognosis de cánceres de pulmón de NSCLC. De nuevo, la selección de las secuencias de polinucleótido, sus longitudes y etiquetas utilizadas en la composición, son determinaciones rutinarias realizadas por un experto en la materia a la vista de las enseñanzas de qué genes pueden formar los perfiles de expresión genética adecuados para la diagnosis y la prognosis de cánceres de pulmón.

La composición, que puede ser presentada en el formato de una tarjeta micro-fluídica, una micro-matriz, un chipo o una cámara, emplea las técnicas de hibridización de polinucleótido descritas en la presente memoria. Cuando una muestra de una PBMC procedente de un sujeto seleccionado se pone en contacto con las sondas de hibridización de las composiciones, la amplificación de PCR de los genes informativos objetivados en el perfil de expresión genética del paciente permite la detección y la cuantificación de cambios de expresión en los genes del perfil de expresión genética respecto a la de un perfil de expresión genética de referencia. Los cambios significativos en la expresión genética de los genes informativos en la PBMC del paciente respecto a los de un perfil de expresión genética de referencia, están correlacionados con una diagnosis de cáncer de pulmón de célula no pequeña (NSCLC).

Según otro aspecto más, una composición o un kit útiles en los métodos descritos en la presente memoria, contienen una pluralidad de ligandos que enlazan tres o más productos de expresión genética de tres o más genes informativos seleccionados a partir de un perfil de expresión genética en las células mononucleares de sangre periférica (PBMC) del sujeto. El perfil de expresión genética contiene los genes de cualquiera de las Tablas I a VII, según se ha descrito en lo que antecede para las otras composiciones. Esta composición permite la detección de las proteínas expresadas por los genes en las Tablas indicadas. Aunque preferentemente los ligandos son anticuerpos para las proteínas codificadas por los genes en el perfil, puede resultar evidente para un experto en la materia que se pueden usar varias formas de anticuerpo, por ejemplo, policlonal, monoclonal, recombinante, quimérico, así como fragmentos y componentes (por ejemplo, CDRs, regiones variables de cadena simple, etc.), en lugar de los anticuerpos. Tales ligandos pueden estar inmovilizados sobre substratos adecuados para su contacto con la PBMC del paciente, y analizados de una forma convencional. En ciertas realizaciones, los ligandos están asociados a etiquetas detectables. Estas composiciones permiten también la detección de cambios en proteínas codificadas por los genes en el perfil de expresión genética respecto a los de un perfil de expresión genética de referencia. Tales cambios se correlacionan con cáncer de pulmón, por ejemplo NSCLC, o con la diagnosis de la fase de desarrollo o del tipo de cáncer, o con el estado y la prognosis pre/post quirúrgica, de una manera similar a la de la PCR y las composiciones que contienen nucleótido descritas en lo que antecede.

Según un aspecto adicional más, una composición útil puede contener una pluralidad de productos de expresión genética de tres o más genes informativos seleccionados a partir del perfil de expresión genética en las células mononucleares de sangre periférica (PBMC) del sujeto inmovilizadas sobre un substrato para la detección o la cuantificación de anticuerpos en las proteínas codificadas por los genes de los perfiles en la PBMC de un sujeto. Los perfiles de expresión genética incluyen genes seleccionados a partir de cualquiera de las Tablas I a VII, o subconjuntos de los mismos, tal como el clasificador de 29 genes de la Tabla V (genes clasificados de 1-29). Este tipo de composición, dirigida a la detección de anticuerpos con respecto a los productos de los genes, es también útil

para identificar y detectar cuantitativamente cambios en la expresión de los genes en el perfil de expresión genética respecto a los de un perfil de expresión genética de referencia por las mismas razones identificadas en lo que antecede para las composiciones PCR/contenedoras de nucleótido. Al igual que con las otras composiciones, este tipo de composición correlaciona los niveles de expresión de las proteínas codificadas por los genes informativos de las PBMCs del paciente con los de un control de referencia. Los cambios significativos que son indicativos de una diagnosis de cáncer de pulmón, son útiles para la monitorización de una intervención quirúrgica/terapéutica en la enfermedad, y/o para proporcionar una diagnosis de la misma.

Para todas las formas anteriores de composiciones de diagnóstico/pronóstico, el perfil de expresión de gen puede, en una realización, incluir al menos los primeros 5 de los genes informativos de las Tablas I a VII o subconjuntos de los mismos. En otra realización para todas las formas anteriores de composiciones de diagnóstico/pronóstico, el perfil de expresión de gen puede incluir 10 o más de los genes informativos de cualquiera de las Tablas I a VII o subconjuntos de los mismos. En otra realización para todas las formas anteriores de composiciones de diagnóstico/pronóstico, el perfil de expresión genética puede incluir 15 o más de los genes informativos de cualquiera de las Tablas I a VII o subconjuntos de los mismos. En otra realización de todas las formas anteriores de composiciones de diagnóstico/pronóstico, el perfil de expresión de gen puede incluir 24 o más de los genes informativos de cualquiera de las Tablas I a III, y V-VII, o subconjuntos de los mismos. En otra realización para todas las formas de composiciones de diagnóstico/pronóstico, el perfil de expresión genética puede incluir de 30 a 50 o más de los genes informativos de cualquiera de las Tablas I-III, V y VII, o subconjuntos de los mismos.

Estas composiciones pueden ser utilizadas para diagnosticar cánceres de pulmón, tal como NSCLC en fase de desarrollo I o en fase de desarrollo II. Además, estas composiciones son útiles para proporcionar una diagnosis suplementaria u original en un sujeto que tenga nódulos de pulmón de etiología desconocida. Los perfiles de expresión genética formados por genes seleccionados a partir de cualquiera de las Tablas I-VII o por subconjuntos de los mismos, son distinguibles de un perfil de expresión genética inflamatoria. Además, varias realizaciones de estas composiciones pueden utilizar perfiles de expresión genética de referencia que incluyan tres o más genes informativos de cualquiera de las Tablas I-VII o subconjuntos de los mismos a partir de la PBMC de una, o de una combinación de clases de sujetos humanos de referencia. Las clases de los sujetos de referencia pueden incluir un fumador con enfermedad maligna, un fumador con una enfermedad no maligna, un antiguo fumador con una enfermedad no maligna, un no fumador sano sin ninguna enfermedad, un no fumador que tenga enfermedad pulmonar obstructiva crónica (COPD), un antiguo fumador con COPD, un sujeto con un tumor de pulmón sólido con anterioridad a la cirugía para la extracción del mismo; un sujeto con un tumor de pulmón sólido después de la extracción quirúrgica del tumor; un sujeto con un tumor de pulmón sólido con anterioridad a la terapia para el mismo; y un sujeto con un tumor de pulmón sólido durante o después de la terapia para el mismo. La selección de la clase apropiada depende del uso de la composición, es decir, para diagnosis original, para diagnosis posterior a la terapia o cirugía, o para diagnosis específica del tipo de enfermedad, por ejemplo, AC frente a LSCC.

IV. Métodos de diagnóstico de la invención

Todas las composiciones descritas en lo que antecede proporcionan una diversidad de herramientas de diagnóstico que permiten una evaluación no invasiva, basada en sangre, de un estado de enfermedad en un sujeto. El uso de estas composiciones en pruebas diagnósticas, que pueden ser complementadas con otras pruebas de protección, tal como una exploración de CT o rayos X del pecho, incrementa la precisión del diagnóstico y/o direcciona pruebas adicionales. En otros aspectos, las composiciones y herramientas de diagnóstico descritas en la presente memoria permiten la prognosis de la enfermedad, monitorizando la respuesta a terapias específicas, y la evaluación regular del riesgo de recurrencia. Los métodos y el uso de las composiciones descritas en la presente memoria permiten la evaluación de cambios en signaturas de diagnóstico presentes en muestras pre-cirugía y post-cirugía, e identifica un perfil de expresión genética que refleja la presencia de tumor y que puede ser usado para evaluar la probabilidad de recurrencia. Los resultados sobre cáncer de pulmón post-cirugía identificados en los ejemplos que siguen, soportan un efecto detectable similar del tumor sobre expresión genética en PBMCs de pacientes.

De ese modo, según un aspecto, se proporciona un método para diagnosticar cáncer de pulmón en un mamífero. Este método incluye identificar un perfil de expresión genética en las células mononucleares de sangre periférica (PBMC) de un mamífero, con preferencia un sujeto humano. El perfil de expresión genética incluye tres o más productos de expresión de tres o más genes informativos que tienen expresión incrementada o reducida en cáncer de pulmón. Los perfiles de expresión genética se forman mediante la selección de tres o más genes informativos a partir de los genes de cualquiera de las Tablas I-VII o de subconjuntos de los mismos. La comparación del perfil de expresión genética del sujeto con un perfil de expresión genética de referencia, permite la identificación de cambios en la expresión de los genes informativos que se correlacionan con un cáncer de pulmón (por ejemplo, NSCLC). Este método puede ser llevado a cabo utilizando cualquiera de las composiciones descritas en lo que antecede.

En una realización, el método permite la diagnosis de adenocarcinoma específicamente. A este efecto, el perfil de expresión genética se selecciona deseablemente a partir de genes de la Tabla II. En otra realización, el método permite la diagnosis de NSCLC en fase de desarrollo I o II. A este efecto, el perfil de expresión genética está formado deseablemente por tres o más genes de la Tabla I o de la Tabla V, incluyendo el clasificador de 29 genes.

Según se ha descrito en lo que antecede para las composiciones, los perfiles genéticos incluyen opcionalmente 5, 6,

10, 15, 25, y una cantidad mayor de 30 genes informativos procedentes de las tablas respectivas, y puede utilizar uno cualquiera de los formatos de método a los que se ha hecho referencia en la presente memoria.

Según otro aspecto más, se proporciona un método para pronosticar la probabilidad de recurrencia de cáncer de pulmón en un mamífero. Este método incluye identificar un perfil de expresión genética en las células mononucleares de sangre periférica (PBMC) del sujeto después de la resección de un tumor sólido o de la quimioterapia. A este efecto, el perfil de expresión genética incluye tres o más productos de expresión de tres o más genes informativos de la Tabla II o la Tabla VI. En otra realización, los productos de expresión genética incluyen los 2 ó 4 genes de categoría superior de la Tabla VI. En una realización, los productos de expresión genética son los seis genes superiores de la Tabla III o VI. En otra realización, los productos de expresión genética incluyen al menos 10 ó 15 genes de categoría superior de la Tabla III o VI. Otras combinaciones de genes de la Tabla III o VI son útiles para la formación de un perfil de expresión genética para este propósito. El perfil de expresión genética post-quirúrgica o post-terapéutica del sujeto se compara con el perfil de expresión genética pre-quirúrgica o pre-terapéutica del sujeto. Los cambios significativos en la expresión de dichos genes informativos están correlacionados con una probabilidad reducida de recurrencia. El mantenimiento de la expresión de perfil genético cambiado con el tiempo, es indicativo de una baja recurrencia post-cirugía o post-terapia. Según se indica en los ejemplos que siguen, este cambio es identificable en la PBMC de un sujeto que tenga un pasado de fumador y/o que tenga enfermedad pulmonar obstructiva crónica (COPD). Según se ha expuesto anteriormente, este método puede ser llevado a cabo utilizando composiciones de diagnóstico y metodologías generales descritas a lo largo de la presente descripción.

Las composiciones y los métodos de diagnóstico descritos en la presente memoria proporcionan una diversidad de ventajas sobre los métodos de diagnóstico actuales. Entre tales ventajas se encuentran las que siguen. Según se ha ejemplificado en la presente memoria, los sujetos con adenocarcinoma o con carcinoma de célula escamosa del pulmón, los dos tipos más comunes de cáncer de pulmón, se distinguen de los sujetos con enfermedades de pulmón no malignas incluyendo la enfermedad de pulmón obstructiva crónica (COPD) o el granuloma u otros tumores benignos. Estos métodos y composiciones proporcionan una solución al problema práctico de diagnóstico sobre si un paciente que presenta una clínica pulmonar con un nódulo pequeño, puede tener una enfermedad maligna. Los pacientes con un nódulo de riesgo intermedio se podrán beneficiar claramente de una prueba no invasiva que puede llevar al paciente ya sea a una categoría de probabilidad muy baja o ya sea de probabilidad muy alta de riesgo de enfermedad. Una estimación precisa de la malignidad basada en un perfil genómico (es decir, estimación de que un paciente dado tenga una probabilidad de un 90% de tener cáncer frente a la estimación de que el paciente tenga solamente un 5% de probabilidad de tener cáncer) podría dar como resultado menor número de cirugías para enfermedad benigna, más tumores en fase temprana extraídos en una fase de desarrollo curable, un menor número de exploraciones de CT de seguimiento, y una reducción significativa de los costes sicológicos de la inquietud acerca de un nódulo. El impacto económico podría ser también igualmente significativo, tal como reduciendo el coste normal estimado de los cuidados adicionales de salud asociados con la protección de CT para el cáncer de pulmón, es decir, 116.000 $ por año de vida ganado con calidad añadida. Una prueba genómica de PBMC no invasiva que tenga una sensibilidad y una especificada suficientes, podría alterar significativamente la probabilidad de malignidad post-prueba, y de ese modo, los cuidados clínicos posteriores.

Una ventaja deseable de estos métodos sobre los métodos existentes consiste en que los mismos están capacitados para caracterizar el estado de enfermedad a partir de un procedimiento mínimamente invasivo, es decir, con la toma de una muestra de sangre. Por el contrario, la práctica actual para la clasificación de tumores cancerígenos a partir de perfiles de expresión genética depende de una muestra de tejido, normalmente una muestra procedente de un tumor. En el caso de tumores muy pequeños, una biopsia es problemática y, claramente, sino es un tumor conocido o visible, una muestra del mismo resulta imposible. No se requiere ninguna purificación del tumor, como en el caso en que se analizan las muestras del tumor. Un método recientemente publicado depende de recortar células epiteliales del pulmón durante una broncoscopia, un método que es también considerablemente más invasivo que el hecho de tomar una muestra de sangre, y solamente aplicable a cánceres de pulmón, mientras que los métodos descritos en la presente memoria son generalizables a cualquier cáncer. Las muestras de sangre tienen una ventaja adicional, que consiste en que el material es preparado y estabilizado fácilmente para análisis posteriores, lo que es importante cuando se va a analizar ARN mensajero.

En una realización de los métodos descritos en la presente memoria, es ventajoso el uso de nuevos algoritmos para analizar los perfiles de expresión genética, que son superiores en cuanto a clasificación respecto a los algoritmos existentes especialmente en el análisis de datos ruidosos o de baja relación señal/ruido. Cuando se compara una enfermedad generalizada con una no enfermedad generalizada, es probable que los datos sean ruidosos debido a que se están combinando en la comparación muchas subclases diferentes. Este método podría ser usado como un añadido a la diagnosis de enfermedad de pulmón existente, en cualquier clínica pulmonar.

V. Ejemplos

La invención va a ser descrita ahora con referencia a los ejemplos que siguen. Estos ejemplos se proporcionan con fines de ilustración solamente, y la invención no debe considerarse en modo alguno como limitada a estos ejemplos sino que por el contrario está construida de modo que abarca cualquiera y todas las variaciones que resulten

evidentes como resultados de las enseñanzas proporcionadas por la misma.

Ejemplo 1: Sujeto de paciente y sujetos de control para muestras de PBMC

Se recogieron muestras de PBMC e información clínica procedente de 300 pacientes de cáncer de pulmón y de 150 controles, incluyendo muestras procedentes de 16 pacientes recogidas en pre- y post- cirugía. Los sujetos de paciente y los sujetos de control tienen ambos el factor de riesgo clave para cáncer de pulmón, es decir, fumar, y muchos de los sujetos de paciente y de los controles no sanos (NHCs) tienen enfermedades relacionadas con fumar tal como la COPD. La diferencia principal entre las 2 clases es la presencia de un nódulo maligno en la clase de pacientes.

A. Sujetos de pacientes

Las poblaciones de pacientes útiles en la provisión de datos para el desarrollo de los perfiles de expresión genética descritos en la presente memoria incluyen pacientes machos y hembras recién diagnosticados con cáncer de pulmón en fase de desarrollo temprana. Los criterios de inclusión para la selección de estos pacientes fueron pacientes en un número significativo de pacientes Afro-Americanos (en torno al 15%), Hispánicos (el 5%) y ningún isleño del Pacífico. La gama de edad fue de 50-80 años. Éstos tenían una salud (ambulatoria) moderadamente buena. Aunque con enfermedad médica. De entre éstos fueron excluidos si habían tenido cánceres, quimioterapia, radiación o cirugía de cáncer con anterioridad. Éstos debían tener una diagnosis de cáncer de pulmón dentro de los 6 meses anteriores, confirmación histológica, y ninguna terapia sistémica, tal como quimioterapia, terapia de radiación o cirugía de cáncer puesto que los niveles de los biomarcadores pueden cambiar con la terapia. Así, la mayor parte de los pacientes de cáncer estaban en fase de desarrollo temprana (es decir, Fase de desarrollo I y Fase de desarrollo II).Otro grupo de pacientes era el de pacientes de cáncer en el que se obtuvo sangre con anterioridad a la cirugía y después de nuevo en un intervalo post-cirugía razonable (~ 2-6 meses) para asegurar que ninguno de los cambios quirúrgicos/inflamatorios se hubiera resuelto. Esto permite que cada paciente sirva como su "propio control". Los criterios de inclusión fueron pacientes con una diagnosis de cáncer de pulmón en Fase de desarrollo I o II, que es reseccionable quirúrgicamente. Fueron excluidos si habían tenido cánceres, quimioterapia, radiación o cirugía de cáncer con anterioridad. Se recogieron datos sobre 16 pares de muestras de pre- frente a post- cirugía que fueron analizadas sobre la plataforma Illumina. Estos estudios mostraron una pérdida de signatura de tumor post cirugía en 13 de los 16 pares comprobados que soportan la detección de una signatura inducida por tumor en las muestras de sangre periférica monitorizadas.

B. Sujetos de control

En vez de usar controles sanos emparejados (no fumadores o fumadores "sanos"), el grupo de control fue extraído principalmente de pacientes con riesgo pulmonar emparejados (fumadores y ex-fumadores) con enfermedad pulmonar no maligna y pacientes con nódulos pulmonares benignos (por ejemplo, granulomas o hamartomas). El grupo de control se menciona en la presente memoria como "controles no sanos" (NHC). Estos pacientes fueron evaluados en clínicas pulmonares, o se sometieron a cirugía torácica por un nódulo de pulmón. Todas las muestras fueron recogidas con anterioridad a la cirugía. Los criterios de inclusión para los pacientes fueron pacientes entre 50-80 años de edad, con un uso del tabaco > de 10 años, y con una exploración por CT o rayos X del pecho dentro de los últimos seis meses que no mostrara ninguna evidencia de cáncer de pulmón ni de ningún otro cáncer en los 5 años anteriores. Los sujetos de control fueron emparejados con los sujetos de paciente en base a la edad, la raza, el género, y el estado de fumador. De ese modo, la mayoría de los controles eran fumadores o ex-fumadores mayores de 50 años de edad. Otro grupo de control incluía pacientes que se sometieron a cirugía para nódulos de pulmón en los que el nódulo se comprobó que era benigno. Los NCHs son una población que pudo beneficiarse significativamente de la monitorización regular debido a su riesgo incrementado en cuanto al desarrollo de cáncer de pulmón.

Ejemplo 2: Protocolos de recogida de muestra y procesamiento

Las muestras de sangre fueron recogidas en la clínica por parte de un técnico en la adquisición de tejido. La sangre fue recogida en dos tubos CPT® (Becton-Dickenson). Los tubos CPT eran tubos vacíos de recogida de sangre que contenían reactivo FICOLL por debajo de un inserto de gel y un anticoagulante por encima del gel. Ésta es una forma muy fácil y eficaz de aislar directamente la PBMC. La sangre se recogió a partir de los mismos pacientes durante su visita de 2-6 meses a la clínica tras la cirugía. Las muestras de sangre fueron recogidas en tubos PAXgene a partir de un subconjunto de pacientes y de sujetos de control. Todas las muestras codificadas, incluyendo bloques de tejido y componentes sanguíneos (PBMC, suero y plasma), fueron almacenadas en base a la identificación del sujeto en cajas de almacenamiento marcadas de un congelador, a -80 ºC.

Las muestras recogidas fueron procesadas mediante una diversidad de etapas rutinarias que han sido altamente estandarizadas. Las muestras fueron procesadas como lotes (normalmente 20-50 muestras) de ambos casos y controles en vez de cómo las muestras individuales que fueron recogidas. En cada etapa, fueron aleatorizadas de modo que ninguna clase particular de pacientes o controles fuera procesada como grupo separado. La purificación de ARN se llevó a cabo utilizando TRI-REAGENT (Molecular Research) según lo recomendado. El ADN y el ARN fueron extraídos a partir de cada muestra, y el ADN fue archivado para estudios futuros. Las muestras de ARN

fueron controladas en cuanto a calidad utilizando el Bioanalizador, y solamente se utilizaron muestras con relaciones 28S/16S de >0,75 para estudios adicionales. Las muestras con relaciones más baja fueron archivadas puesto que las mismas eran aún adecuadas para estudios de validación de PCR. La misma cantidad (250 ng del ARN total) fue amplificada (aARN) utilizando el kit de amplificación de ARN (Ambion). Esto proporcionó material amplificado suficiente (5-10 µg) para múltiples repeticiones de las matrices y para estudios de validación de PCR. Todas las muestras fueron amplificadas una sola vez.

Un esquema alternativo de recogida de muestras emplea el Sistema PAXgene de ARN sanguíneo (Preanalytix – una compañía Qiagen/BD) para estabilizar el ARN en muestras de sangre total. Puesto que el PAXgene no requiere ningún procesamiento especial de las muestras de sangre, permite el desarrollo más fácil de estándares para la recogida de muestras. Para optimizar la recogida uniforme de muestras recogidas en múltiples sitios de un ensayo clínico, el Sistema PAXgene de ARN Sanguíneo (Preanalytix- una compañía de Qiagen/BD) integra las etapas clave de recogida global de sangre, estabilización de ácido nucleico y purificación de ARN. Éste utiliza tecnología BD Vacunatainer™ estándar que contiene un reactivo del propietario que estabiliza inmediatamente el ARN intracelular durante días a temperatura ambiente, semanas a 4 ºC, y que puede ser almacenado al menos un año a menos 80 ºC con anterioridad a la purificación del ARN. Los tubos PAXgene pueden ser expedidos durante la noche y almacenados a -80 ºC hasta su utilización. Todos los tubos se mantienen a temperatura ambiente durante 2-4 horas antes de la congelación puesto que ello aumenta la producción de ARN. La capacidad de minimizar la urgencia del procesamiento aumenta considerablemente la eficacia de laboratorio. Para más detalles, véase http://www.preanalytix.com/RAN.asp.

De muchos modos, éste es el mejor método para conservar inmediatamente las poblaciones de mensaje de ARN presentes en el instante de la recogida. Sin embargo, la gran cantidad de mensaje de globina presente en las tres muestras interfirió con la determinación del mensaje sobre micro-matrices, a pesar de los esfuerzos por superar este problema. Si se emplea un ensayo de PCR para los perfiles de expresión genética descritos en la presente memoria, se prefiere el uso de PaxGene, puesto que el mensaje de globina no interfiere con los ensayos de PCR.

Ejemplo 3: Métodos de procesamiento de datos para obtención de perfil de expresión genética

El BeadChip ILLUMINA es un método relativamente nuevo de llevar a cabo múltiples análisis genéticos. El elemento esencial de la tecnología de BeadChip es la fijación de oligonucleótidos a perlas de sílice. Las perlas fueron depositadas a continuación aleatoriamente en pocillos sobre un substrato (por ejemplo, una placa de vidrio). La matriz resultante fue descodificada para determinar qué combinación de oligonucleótido-perla se encontraba en cada pocillo. Las matrices descodificadas pueden ser usadas para un número de aplicaciones, incluyendo análisis de expresión genética. Estas matrices tienen la misma cobertura de gen que las matrices Affymetrix (47.000 sondas para 27.000 genes incluyendo las variantes de partición) pero utilizan oligonucleótidos de 50-mer en vez de 25-mer y de ese modo proporcionan una mayor especificidad.

Los procedimientos de línea de análisis de datos que utilizan funciones de Matlab, el PDA y SVM codificados con RFE y SVM-RCE, fueron utilizados rutinariamente y exitosamente como había sido puesto de manifiesto por las publicaciones y según se ha descrito en la presente memoria.

A. Pre-procesamiento de datos y control de calidad de matriz

Los datos fueron procesados según ha sido descrito de manera general[1] y los niveles de expresión para la señal y las sondas de control fueron exportados. Se utilizó un conjunto de sondas de control negativas para calcular el nivel medio de fondo y para determinar el umbral de detección de señal. Los datos de expresión de sonda fueron normalizados utilizando normalización cuantil. Los datos fueron comprobados en cuanto a resultados atípicos calculando una puntuación atípica para cada una de las muestras. En primer lugar, se calcularon los coeficientes de correlación de Spearman para cada par de muestras. Se calculó la correlación media para cada muestra (Ms), la correlación media para todos los pares de muestras (Mp) y la desviación media absoluta a partir de Mp (MADp). La puntuación atípica (de manera similar a la puntuación-Z) para la muestra *i* fue calculada a continuación como (Msi – MP)/MADp. Las puntuaciones atípicas fueron estudiadas para elegir un umbral que marque los potenciales resultados atípicos. Normalmente, las muestras con puntuaciones atípicas de más de 5, fueron consideradas como resultados técnicos atípicos. La identificación adicional de resultados atípicos se hace mediante estadísticas multi-variantes tal como diagramas de componentes principales (PCA), escalado multidimensional, y PCA robusto.

Con el fin de reducir el ruido experimental, los datos se filtran eliminando sondas no informativas, es decir, sondas que no fueron detectadas en la mayoría de las muestras (más del 95%) o sondas que no cambian en al menos 1,2 veces entre al menos dos muestras. Si una muestra tenía réplicas, se tomó la última réplica para el análisis.

B. Revisión no supervisada

Cuando fue apropiado, se aplicó agrupamiento jerárquico utilizando correlación o distancia Euclídea, y se utilizó escalado multidimensional para inspeccionar conjuntos de datos en cuanto a evidencia de resultados atípicos o subclases. Se utilizó VISDA (53) para este propósito con buen éxito.

C. Clasificación supervisada

La Máquina de Vector de Soporte (SVM) puede ser aplicada a conjuntos de datos de expresión de gen para descubrir y clasificar la función genética. Se ha encontrado que la SVM es la más eficiente para distinguir los casos y controles relacionados de forma más próxima que residen en los márgenes. Principalmente, la SVM-REF (48, 54) fue utilizada para desarrollar clasificadores de expresión genética que distinguen clases definidas clínicamente de pacientes a partir de clases definidas clínicamente de controles (fumadores, no fumadores, COPD, granuloma, etc.). La SVM-RFE es un modelo basado en SVM utilizado en el estado de la técnica que retira genes, basada recursivamente en su contribución a la discriminación, entre las dos clases que van a ser analizadas. Los genes de puntuación más baja por pesos de coeficientes, fueron retirados y los genes restantes fueron puntuados de nuevo y el procedimiento fue repetido hasta que solamente permanecieron unos pocos genes. Ese método ha sido usado en varios estudios para realizar tareas de clasificación y de selección de gen. Sin embargo, eligiendo valores apropiadas de los parámetros del algoritmo (parámetro de penalización, función de núcleo) se puede con frecuencia influir en el rendimiento.

La SVM-RCE es un modelo relacionado basado en SVM, ya que el mismo, al igual que la SVM-RFE, evalúa las contribuciones relativas de los genes al clasificador. La SVM-RCE evalúa las contribuciones de grupos de genes interrelacionados en vez de genes individuales. Adicionalmente, aunque ambos métodos retiran los genes menos importantes en cada etapa, la SVM-RCE puntúa y retira grupos de genes, mientras que la SVM-RFE puntúa y retira un único, o unas pequeñas cantidades de genes en cada recorrido del algoritmo.

El método de SVM-RCE se va a describir aquí brevemente. Los genes de expresión baja (expresión media menor de 2 x fondo) fueron retirados, se realizó normalización cuantil, y a continuación se retiraron las matrices “de resultados atípicos” cuyos valores de expresión medios difieren en más de 3 sigma del valor medio del conjunto de datos. Las muestras restantes fueron sometidas a SVM-RCE utilizando diez repeticiones de validación cruzada de 10 pliegues del algoritmo. Los genes fueron reducidos mediante prueba-t (aplicada al conjunto ejercitante) respecto a un valor óptimo determinado experimentalmente que produce una precisión más alta en el resultado final. Estos genes de partida fueron agrupados mediante K-medios en grupo de genes correlacionados cuyo tamaño medio es de 3-5 genes. Se llevó a cabo la puntuación de la clasificación de SVM en cada grupo utilizando un re-muestreo de 3 pliegues repetido 5 veces, y los grupos de peor puntuación fueron eliminados. Se determinó la precisión sobre la agrupación de genes supervivientes utilizando el 10% de muestras de la izquierda (conjunto de prueba) y se registraron los 100 genes de puntuación superior. El procedimiento fue repetido a partir de la etapa de agrupamiento hasta un punto final de 2 grupos. El panel genético óptimo fue tomado como el número mínimo de genes que proporciona la máxima precisión a partir del gen más frecuentemente seleccionado. La identidad de los genes individuales de este panel no es fija, puesto que el orden refleja el número de veces que un gen dado fue seleccionado entre los 100 genes informativos superiores, y este orden está sujeto a alguna variación.

Utilizando la SVM-RCE, la evaluación inicial del rendimiento de cada grupo de gen individual, como característica separada, permitió la identificación de aquellos grupos que contribuyeron como mínimo a la clasificación. Éstos fueron retirados del análisis aunque aquellos grupos que presentaron un rendimiento de clasificación relativamente mejor fueron extraídos. Se permitió que el re-agrupamiento de genes después de cada etapa de eliminación permitiera la formación de nuevos grupos, potencialmente más informativos. Los grupos de genes más informativos fueron retenidos durante rondas adicionales de averiguación hasta que se identificaron los grupos de genes con la mejor precisión de clasificación.

La utilización del método que usa grupos de genes, en vez de genes individuales, aumentó la precisión de clasificación supervisada de los mismos datos en comparación con la precisión cuando se usó SVM o bien Análisis Discriminante Penalizado (PDA) con eliminación de característica recursiva (SVM-RFE y PDA-RFE) para extraer genes en base a sus pesos discriminantes individuales. El método permitió también la determinación arbitraria del número de grupos y del tamaño del grupo al principio del análisis por el investigador y, según avanzaba el algoritmo, los grupos menos significativos fueron progresivamente retirados. El método proporcionó además los n grupos superiores requeridos para diferenciar de la manera más precisa las dos clases predefinidas. Estos dos métodos están mejor definidos en los ejemplos que siguen.

D. Selección de biomarcador

Los genes que puntúan más alto (mediante SVM) en la discriminación de pacientes a partir de controles, fueron examinados en cuanto a su utilidad para pruebas clínicas. Los factores considerados incluyen diferencias más altas en niveles de expresión entre clases, y variabilidad baja dentro de las clases. Cuando se seleccionan biomarcadores para su validación, se ha realizado un esfuerzo para seleccionar genes con perfiles de expresión distintos para evitar la selección de genes correlacionados (55) y para identificar genes con niveles de expresión diferencial que fueran robustos mediante técnicas alternativas que incluyen PCR y/o inmuno-histoquímica.

E. Validación

Se consideraron tres métodos de validación.

Validación cruzada: Para minimizar el sobre-ajuste dentro de un conjunto de datos, se utilizó validación cruzada K veces (K es normalmente igual a 10), cuando el conjunto de datos se partió en K partes aleatoriamente y se usaron K-1 partes para la ejercitación y 1 para la prueba. Así, para K = 10, el algoritmo fue ejecutado sobre una selección aleatoria del 90% de pacientes y el 90% de los controles, y a continuación se probó el 10% restante. Esto se repitió hasta que todas las muestras habían sido empleadas como sujetos de muestra y el clasificador acumulado hizo uso de todas las muestras, pero no se probó ninguna muestra usando un conjunto de ejercitación del que la misma formara parte. Para reducir el impacto de la aleatorización, la separación de K pliegues se repitió M veces produciendo combinaciones diferentes de pacientes y controles en cada uno de las K pliegues, cada vez. Por lo tanto, para el conjunto de base de datos individuales, se utilizaron M*K rondas de selección permutada de ejecución y de conjuntos de prueba por cada conjunto de genes.

Validación independiente: Para estimar la reproducibilidad de los datos y la generalidad del clasificador, se necesita examinar el clasificador que se construyó utilizando un conjunto de datos y se probó utilizando otro conjunto de datos para estimar el rendimiento del clasificador. Para estimar el rendimiento, se llevó a cabo una validación en el segundo conjunto utilizando el clasificador desarrollado con el conjunto de datos original.

Re-muestreo (permutación): Para demostrar la dependencia del clasificador del estado de enfermedad, los pacientes y los controles del conjunto de datos fueron elegidos de manera aleatoria (permutada) y se repitió la clasificación. La exactitud de la clasificación utilizando muestras aleatorizadas fue comparada con la exactitud del clasificador desarrollado para determinar el valor p del clasificador, es decir, la posibilidad de que el clasificador pudiera haber sido elegido por azar. Con el fin de probar la generalidad de un clasificador desarrollado de esa manera, se utilizó el hecho de clasificar conjuntos independientes de muestras que no fueron usadas en el desarrollo del clasificador. Las precisiones de validación cruzada del clasificador permutado y original, fueron comparadas sobre conjuntos de prueba independientes para confirmar su validez en cuanto a la clasificación de nuevas muestras.

F. Rendimiento de clasificador

El rendimiento de cada clasificador fue estimado mediante diferentes métodos y se usaron varias mediciones de rendimiento para comparar clasificadores entre sí. Estas mediciones incluyen la precisión, el área bajo curva de ROC, la sensibilidad, la especificidad, la tasa positiva verdadera y la tasa negativa verdadera. En base a las propiedades requeridas de la clasificación de interés, se pueden usar diferentes mediciones de rendimiento para elegir el clasificador óptimo, por ejemplo el clasificador a usar en la protección de la población total podría requerir una mejor especificidad para compensar una pequeña prevalencia (~1%) de la enfermedad y evitar por tanto un gran número de falsos aciertos positivos, mientras que un clasificador de diagnóstico de pacientes en hospital debería ser más sensible.

G. Aplicación del clasificador

Una construcción de clasificador lineal por parte de la SVM para un conjunto de genes en base a un conjunto de ejercitamiento, puede ser usado para asignar una puntuación de SVM a cualquier muestra. Matemáticamente, el clasificador es un conjunto de g+1 coeficientes, en el que g es un número de genes presentes en el conjunto. Si E_1, ..., E_0 son valores de expresión de estos genes para una muestra, y C_1, ..., C_{g+1} son los coeficientes correspondientes, entonces la puntuación de SVM para la muestra se calcula fácilmente como $C_1E_1 + ... + C_gE_g + C_{g+1}$.

H. Análisis de ROC

El análisis de ROC fue realizado con vistas a estimar la eficacia de cada clasificador que toma en consideración tanto la sensibilidad como la especificidad. La curva de ROC para un clasificador se construye variando el punto de corte de la puntuación de SVM y calculando la sensibilidad y la especificidad correspondientes. Se calculó el área bajo la curva de ROC (AUC) para usarla como medición de rendimiento de clasificador. Puesto que el clasificador aleatorio de muestras podría tener una AUC de 0,5 y el clasificador perfecto podría tener una AUC de 1,0, el valor de AUC calculado puede ser usado e informado como expresión de porcentaje de la eficacia del clasificador.

I. Valores predictivos positivos y negativos

El cálculo de valores predictivos positivos (PPV) y de valores predictivos negativos (NPV) tiene en cuenta no solo la especificidad y la sensibilidad, sino también una prevalencia *p* de la enfermedad:

$$PPV = \frac{sens \cdot p}{sens \cdot p + (1 - spec) \cdot (1 - p)}$$

$$NPV = \frac{spec(1 - p)}{spec(1 - p) + (1 - sens) \cdot p}$$

Así, PPV es similar a la tasa positiva cierta y muestra una fracción de sujetos que realmente tienen enfermedad entre muestras clasificadas positivamente, mientras que NPV es similar a la tasa negativa cierta, y muestra una fracción de sujetos que realmente no tienen las muestras de enfermedad clasificadas negativamente.

Los valores PPV y NPV fueron calculados a partir de cada punto de corte de puntuación de SVM posible para diversos valores de prevalencia (1%, 5% y 50%). Adicionalmente a direccionar el uso de los valores PPV y NPV, esto permite identificar un punto de corte de puntuación de SMV para su uso en la clasificación con el fin de conseguir un valor predictivo de clasificador especificado.

Ejemplo 4: Clasificaciones supervisadas de SVM

(i) El proceso de SVM-RFE fue aplicado a un conjunto de ejercitación de muestras, como sigue. Se realizó una prueba-T sobre genes procedentes del conjunto de ejercitamiento para determinar los 1000 mejores genes que separan dos clases de muestras, Para cada etapa de reducción de gen se ejecutó SVM utilizando el número de genes restantes. Los coeficientes para estos genes procedentes del clasificador ejercitado, fueron comparados a continuación para eliminar genes con el impacto menor en la puntuación discriminante. El diez por ciento de los genes menos significativos fueron retirados y el proceso fue repetido hasta que se dejó solamente 1 gen. El rendimiento de los clasificadores para cada número de genes fue calculado utilizando el clasificador correspondiente en el conjunto de prueba. Cada gen recibió una puntuación que se corresponde con la etapa de iteración en la que el gen fue eliminado. Para eliminar el sobre-ajuste dentro de un conjunto de datos, se utilizó validación cruzada K pliegues (normalmente, K es igual a 10). Los datos fueron divididos en K partes (pliegues) y el algoritmo fue ejercitado sobre k-1 pliegues de los grupos de caso y de control, y a continuación probado sobre el resto 1 vez. Esto garantizó que cada muestra fuera empleada como sujeto de prueba. La partición aleatoria sobre K-pliegues fue repetida 10 veces, dando como resultado 100 pares diferentes de subconjuntos de ejercitamiento-prueba. Cada partición de datos de ejercitamiento-prueba fue analizada mediante SVM-RFE por separado. A continuación se calculó una puntuación final de gen para todos los genes que estuvieron involucrados en el ejercitamiento de al menos un clasificador. La puntuación fue igual a la puntuación genética media a través de todas rondas de re-muestreo, dividida por el número de iteraciones de eliminación, De ese modo, el gen hipotético que alcance la etapa de iteración de eliminación máxima en la totalidad de las 100 ejecuciones de SVM-RFE, recibirá una puntuación de 1, mientras que el gen que fue siempre eliminado en la primera etapa recibirá una puntuación de 0. Se utilizaron diferentes cantidades de genes superiores con puntuaciones más altas para calcular el rendimiento de la formación clasificadora de estos genes. El clasificador con el mejor rendimiento indica el número óptimo de genes a usar para la clasificación.

(ii) El algoritmo central del método de SVM-RCE ha sido descrito como diagrama de flujo (en la Figura 3 de referencia 1), el cual consiste en tres etapas principales aplicadas sobre la parte ejercitante de los datos:

Etapa de agrupamiento para el agrupamiento de los genes; etapa de puntuación de SMV para calcular la *Puntuación*($X(s_i)$, *f*, *r*) de cada agrupamiento de genes, y etapa de RCE para retirar grupos con baja puntuación. El método de SVM-RCE fue realizado de acuerdo con lo siguiente:

Se ha supuesto que el conjunto de datos *D* tiene *S* genes (todos los genes son genes *n_g* superiores mediante prueba-t) y que los datos fueron divididos en dos partes: una para el ejercitamiento (90% de las muestras) y la otra (10% de las muestras) para la prueba. *X* indica un conjunto de datos de ejercitamiento de dos clases que consiste en muestras y en *S* genes. Se definió la medición de puntuación para cualquier lista *S* de genes como la capacidad para diferenciar las dos clases de muestras mediante aplicación de SVM lineal. La puntuación fue calculada realizando una partición aleatoria sobre el conjunto *X* de muestras de ejercitamiento en *f* subconjuntos no solapantes de iguales tamaños (*f*-pliegues). La SVM lineal fue ejercitada sobre *f*-1 subconjuntos y el subconjunto restante fue utilizado para calcular el rendimiento. El procedimiento fue repetido *r* veces para tomar en consideración un particionado diferente posible.

La *Puntuación (X(S), f, r)* fue definida como la precisión media de la SVM lineal sobre los datos *X* representados por los *S* genes calculados como validación cruzada de *f*-pliegues repetida *r* veces. Los valores por defecto son $r = 3$ y $r = 5$. Si los *S* genes están agrupados en sub-grupos de genes $S_1, S_2, ..., s_n$, se definió la *Puntuación*($X(s_i)$, *f*, *r*) para cada sub-grupo mientras que los datos *X* estaban representados por los genes de S_i. *n* = número inicial de agrupamientos, *m* = número final de agrupamientos, *d* = parámetro de reducción. Mientras que ($n \leq m$) hacer que: 1. Agrupar los *S* genes dados en *n* agrupamientos $S_1, S_2, ..., S_n$ utilizando K-medios (etapa de agrupamiento); 2. Para cada grupo $i = 1, ..., n$, calcular su *Puntuación*($X(s_i)$, *f*, *r*) (etapa de cálculo de puntuación de SVM); 3. Retirar el d% de grupos con puntuación más baja (etapa de RCE); 4. Reunir genes supervivientes de nuevo en un grupo *S*; 5. Reducir *n* en un d%.

La alternativa básica de la SVM-RCE fue agrupar en primer lugar los perfiles de expresión genética en *n* agrupamientos, utilizando K-medios. Se asignó una puntuación (*Puntuación* ($X(s_i)$, *f*, *r*)) a

cada uno de los grupos mediante SVM lineal, indicando su éxito en la separación de muestras en la tarea de clasificación. El *d*% de grupos (o *d* grupos) con las puntuaciones más bajas, fueron apartados a continuación del análisis. Las etapas 1 a 5 fueron repetidas hasta que el número *n* de grupos se redujo a *m*. Se indica con *Z* el conjunto de datos de prueba. En la etapa 4, se formó un clasificador de SVM a partir del conjunto de datos de ejercitamiento utilizando los genes S supervivientes. Este clasificador fue probado sobre *Z* para estimar el rendimiento. Véase la Figura 3 referenciada con anterioridad de (1), el panel de "Prueba" en el lado derecho.

Para la versión actual, las opciones de *n* y *m* fueron determinadas por el investigador. En esta implementación, el valor por defecto de *m* fue 2, lo que indica que se necesitó el método para capturar los 2 agrupamientos (grupos) significativos superiores de genes. Sin embargo, la precisión fue determinada después de cada ronda de eliminación de agrupamiento y un grupo más alto de agrupamientos pudo ser más preciso que los dos finales. El paquete de gist-svm fue utilizado para la implementación de SVM-RFE, con función de núcleo lineal (producto escalar), con parámetros por defecto. En gist-svm, la SVM emplea un margen suave de dos normas con C = 1 como parámetro de penalización. La SVM-RCE fue codificada en MATLAB mientras que la emisión de Bioinformatics Toolbox 2.1 fue utilizada para la implementación de SVM lineal con margen suave de dos normas con C = 1 como parámetro de penalización. La puntuación de PDA-RFE fue implementada en lenguaje de programación C utilizando una interfaz de usuario JAVA.

Con el fin de asegurar una comparación razonable y para reducir el tiempo de cálculo, los 300 genes superiores (*n_g* = 300) fueron seleccionados mediante prueba-t en el conjunto de ejercitamiento para todos los métodos. Sin embargo, el uso de estadísticas-t para reducir el número de de genes iniciales sometidos a SVM-RFE no sólo fue eficiente, sino que también aumentó el rendimiento del clasificador. Para todos los resultados presentados, se usó un 10% (*d* = 0,1) para la reducción de agrupamiento de gen para SVM-RCE y el 10% de los genes con SVM-RFE y PDA-RFE. Para la SVM-RCE, el experimento se inició utilizando 100 agrupamientos (*n* = 100) y cesó cuando quedaban 2 agrupamientos (*m* = 2). Se utilizaron 3 pliegues (*f* = 3) repetidos 5 veces (*r* = 5) en el método de SVM-RCE para evaluar la puntuación de cada agrupamiento (etapa de puntuación de SVM en la Figura 3 de referencia 1). Se pueden usar parámetros de evaluación más rigurosos incrementando el número de validaciones cruzadas repetidas, mientras se incrementa simultáneamente el tiempo de cálculo.

(iii) Para evaluar el rendimiento global de SVM-RCE y SVM-RFE (y de PDA-RFE), se empleó validación cruzada de 10 pliegues (9 pliegues para ejercitamiento y 1 pliegue para prueba), repetida 10 veces. Después de cada ronda de reducción de característica o de agrupamiento, se calculó la precisión sobre el conjunto de prueba extendido. Para cada muestra del conjunto de prueba, una puntuación asignada mediante SVM indicó su distancia desde el híper-plano de discriminación generado a partir de las muestras de ejercitamiento, en el que un valor positivo indicaba un miembro de la clase positiva y un valor negativo indicaba un miembro de la clase negativa. La etiqueta de clase para cada muestra de prueba fue determinada mediante el promediado de un total de 10 de sus puntuaciones de SVM, y la muestra fue clasificada en base a este valor. Este método para calcular la precisión proporcionó una medición más precisa del rendimiento, puesto que no sólo capturó el hecho de que una muestra específica esté clasificada positivamente (+1) o negativamente (-1), sino también cómo ha sido clasificada en cada categoría, según se determina mediante una puntuación asignada a cada muestra individual. La puntuación sirvió como medición de la confianza de la clasificación. La gama de puntuaciones proporcionó un intervalo de confianza.

Los métodos de agrupamiento son técnicas no supervisadas en las que no están asignadas las etiquetas de las muestras. K-medios[67] es un algoritmo de agrupamiento ampliamente utilizado. Éste constituye un método iterativo que agrupa genes con perfiles de expresión correlacionados en *k* agrupamientos mutuamente exclusivos, siendo *k* un parámetro que necesita ser determinado al comienzo. El punto de partida del algoritmo K-medios consiste en iniciar *k* agrupamientos activos generados aleatoriamente. Cada perfil de gen está asociado al agrupamiento con distancia mínima (se podrían usar diferentes métricas para definir la distancia) a su "centroide". El centroide de cada agrupamiento es re-calculado a continuación como la media de todos los perfiles de los miembros de gen del agrupamiento. El procedimiento se repite hasta que no se detecten cambios en los centroides, para varios agrupamientos. Finalmente, este algoritmo tiene por objeto minimizar una *función objetiva* con *k* agrupamientos;

$$\boldsymbol{F(datos, k)} = \sum_{j=1}^{k}\sum_{i=1}^{t}||\boldsymbol{g_i^j} - \boldsymbol{c_j}||^2$$, donde *t* es el número de genes, donde || ||^2 es la distancia medida entre el perfil g_i de gen y el centroide c_j del agrupamiento. La medición de distancia de "correlación" fue utilizada como métrica para la alternativa de SVM-RCE. La distancia de correlación entre los genes g_r y g_s se define como:

$$d_{rs} = 1 - \frac{(g_r - \bar{g}_r)(g_s - \bar{g}_s)'}{\sqrt{(g_r - \bar{g}_r)(g_r - \bar{g}_r)'}\sqrt{(g_s - \bar{g}_s)(g_s - \bar{g}_s)'}}$$

donde

$$\bar{g}_r = \frac{1}{t}\sum_j g_{rj} \text{ and } \bar{g}_s = \frac{1}{t}\sum_j g_{sj}$$

K-medios es sensible a la elección de los agrupamientos activos (centroides iniciales) y se pueden considerar diferentes métodos para la elección de los agrupamientos activos. En la etapa de K-medios, es decir, la etapa de agrupamiento de la Figura (3) de (1), de SVM-RCE, *k* genes son seleccionados *aleatoriamente* para formar los agrupamientos activos, y este proceso se repite varias veces (*u* veces) con el fin de alcanzar el valor óptimo, con el valor más bajo de la función *F*(*datos; k*) objetiva.

El método de SVM-RCE difiere de los métodos de clasificación relacionados en el estado de la técnica en que el método de SVM-RCE agrupa en primer lugar genes en agrupamientos de gen correlacionados mediante K-medios, y a continuación evalúa las contribuciones de cada uno de esos agrupamientos a la tarea de clasificación por SVM.

Ejemplo 5: Uso de algoritmos de Máquina de Vector de Soporte (SVM) y de Eliminación Recursiva de Agrupamiento (RCE) para seleccionar genes significativos para expresión genética comparativa en cáncer de pulmón

En este ejemplo, el algoritmo de SVM-RCE para la selección y clasificación de gen fue mostrado utilizando dos (2) conjuntos de datos. Según se ha indicado anteriormente, este novedoso algoritmo combina el algoritmo K-medios para agrupamiento genético y el algoritmo de aprendizaje de máquina (SVM) para identificar y clasificar (ordenar) esos agrupamientos de genes con el propósito de clasificación y categorización de agrupamiento de gen. La eliminación recursiva de grupo (RCE) fue aplicada a continuación para extraer iterativamente esos agrupamientos de genes que contribuyen menos al rendimiento de la clasificación.

Este algoritmo fue llevado a cabo utilizando la versión Matlab™ del algoritmo de SVM-RCE que puede ser descargado en http://showelab.wistar.upenn.edu bajo la lengüeta “Tools->SVM-RCE”. En resumen, el algoritmo SVM-RCE fue evaluado en este ejemplo usando conjuntos de datos de tumores de cabeza y cuello (I) y (II) según se expone en lo que sigue.

Para el Conjunto de datos (I), la obtención del perfil de expresión genética se llevó a cabo sobre un panel de 18 muestras de tumor de cabeza y cuello (HN) y 10 cánceres de pulmón (LC) utilizando matrices Affymetrix® U133A, según se describe en Vachani et al., Accepted Cli. Cancer Res., 2001.

Para el Conjunto de datos (II), la obtención del perfil de expresión genética fue llevada a cabo sobre un panel de 52 pacientes ya sea con carcinomas de pulmón primarios (21 ,muestras) o de cabeza y cuello primarios (31 muestras), utilizando la micro-matriz[68] de nucleótido Affymetrix® HG_U95Av2 de alta densidad.

Se utilizaron tres algoritmos, a saber SVM-RCE, PDA-RFE y SVM-RFE para reducir iterativamente el número de genes desde el valor de partida en estos conjuntos de datos (I) y (II) utilizando precisión de clasificación intermedia como métrica. En resumen, se determinó la precisión del algoritmo de SVM-RCE en los 2 grupos finales de genes, y dos niveles intermedios, normalmente los agrupamientos 8 y 32, que corresponden a 8 genes, 32 genes y 102 genes, respectivamente. Para los algoritmos SVM-RFE y PDARFE, se determinó también la precisión para números de genes comparables. Véase la Tabla VIII.

Tabla VIII

	Tumores de Cabeza & Cuello frente a Tumores de Pulmón (I)			**Tumores de Cabeza & Cuello frente a Tumores de Pulmón (II)**		
Algoritmo	# grupos (#c)	# genes (#g)	precisión (ACC) (%)	# grupos (#c)	# genes (#g)	precisión (ACC) (%)
SVM-RCE	2	8	100	2	9	100
	8	32	100	6	32	100
	28	103	100	25	103	100
SVM-RFE		8	92		8	98
		32	90		32	98

		102	90		102	98
PDA-RFE		8	89		8	70
		31	96		32	98
		109	96		102	98

Los resultados que comparan el uso independiente de los algoritmos de SVM-RCE y de SVM-RFE en el conjunto de datos (I), ilustraron que el algoritmo de SVM-RCE tuvo un incremento de precisión sobre el algoritmo de SVM-RFE. Específicamente, se obtuvo un incremento de precisión de un 8%, un 10% y un 10% con alrededor de 8, alrededor de 32 y alrededor de 100 genes, respectivamente. De una manera similar, los resultados con la utilización de estos algoritmos sobre el conjunto de datos (II), mostraron un incremento de alrededor de un 2% con el algoritmo SVM-RCE, utilizando alrededor de 8, alrededor de 32 y alrededor de 102 genes (100% ACC). El algoritmo SVM-RFE, sin embargo, mostró un ACC de alrededor de un 98%. Estos resultados muestran claramente la superioridad del algoritmo SVM-RCE sobre el algoritmo SVM-RFE.

También se apreció que el tiempo de ejecución para el algoritmo SVM-RCE utilizando el código MATLAB fue mayor que el tiempo de ejecución para el algoritmo SVM-RFE, que utiliza el lenguaje de programación C. Por ejemplo, cuando se aplicó el SVM-RCE en un ordenador personal con un procesador P4-Duo-core de 3.0 GHz y 2 GB de RAM sobre el conjunto de datos (I), se obtuvieron los resultados en aproximadamente 9 horas para 100 iteraciones (10 pliegues repetidos 10 veces). Los mismos resultados fueron obtenidos utilizando el algoritmo SVM-RFE (con el paquete svm-gist) en 4 minutos. Para determinar la fiabilidad de estos resultados, se ejecutó de nuevo el algoritmo SVM-RCE sobre el conjunto de datos (I), mientras se rastreaba simultáneamente el resultado de cada iteración y sobre cada nivel de grupos de genes. Los resultados obtenidos utilizando el algoritmo SVM-RCE, con independencia de las iteraciones, tuvieron una desviación estándar de 0,04 a 0,07. Los resultados obtenidos utilizando el algoritmo SVM-RFE tuvieron una desviación estándar de 0,2 a 0,23. Estos resultados muestran que el algoritmo SVM-RCE fue más robusto y más estable que el algoritmo SVM-RFE.

La misma superioridad del algoritmo SVM-RCE fue observada cuando se comparó el algoritmo SVM-RCE con el algoritmo PDA-RFE. Véase la Tabla 1^{1} publicada y la Figura 1^{1} utilizando agrupamiento jerárquico y escalado multidimensional (MDS) para ayudar a ilustrar la precisión mejorada de la clasificación del algoritmo SVM-RCE para el conjunto de datos (I). Los genes seleccionados mediante el algoritmo SVM-RCE separaron claramente las dos clases mientras que los genes seleccionados mediante el algoritmo SVM-RFE colocaron una o dos muestras en el lado erróneo del margen de separación.

También se apreció que el tiempo de ejecución para el algoritmo SVM-RCE que utiliza el código MATLAB fue mayor que para el algoritmo PDA-RFE, que utiliza el lenguaje de programación C.

Se demostró también la convergencia del algoritmo en la solución óptima, y que proporciona una ilustración más visual del algoritmo SVM-RCE, En resumen, se calculó el rendimiento medio sobre todos los grupos de cada nivel de reducción para el conjunto de datos (I).Véase la Figura 1^{1} publicada, en la que ACC es la precisión, TP es la sensibilidad, y TN es la especificidad de los genes restantes, determinados sobre el conjunto de prueba. Avd es la precisión media de los genes albergados por los agrupamientos.

En resumen, se seleccionaron 1000 genes mediante prueba-t a partir del conjunto de ejercitamiento, distribuidos en 300 agrupamientos (número inicial de agrupamientos (n) = 300, número final de agrupamientos (m) = 2, parámetro de reducción (d) = 0,3, n-g = 1000) y a continuación se redujo recursivamente a 2 agrupamientos. El rendimiento medio de clasificación sobre el conjunto de prueba por agrupamiento en cada nivel de reducción (Figura 1^{1} publicada, AVG lineal) mejoró drásticamente desde alrededor de un 55% a alrededor de un 95% según se redujo el número de agrupamientos. La precisión media también se incrementó según se eliminaron los agrupamientos bajos en información. Estos resultados soportan la sugerencia de que los agrupamientos menos significativos fueron retirados mientras que los agrupamientos informativos fueron conservados según se empleó el algoritmo RCE.

El algoritmo SVM-RCE fue también útil en cuanto a estimación de estabilidad, como fue evidenciado por los resultados sobre el conjunto de datos (I). La estabilidad fue estimada mediante obtención de valores de u (u = número de veces que se repite el proceso) de 1, 10 y 100 repeticiones, y comparando estos valores con los 20 genes más informativos devueltos de cada experimento. Alrededor de un 80% de los genes fueron comunes a las tres ejecuciones, lo que sugirió que los resultados del algoritmo SVM-RCE fueron robustos y estables.

En resumen, estos datos ilustran que el algoritmo SVM-RCE proporciona información importante que no puede ser obtenida utilizando algoritmos del estado de la técnica que evalúan la contribución de cada gen individualmente. Aunque las observaciones iniciales estaban basadas en los 2 agrupamientos superiores necesarios para la separación de conjuntos de datos con 2 clases conocidas de muestras, es decir, los conjuntos de datos (I) y (II), el análisis puede ser extendido a la captura de, por ejemplo, los 4 agrupamientos superiores de genes.

Los resultados sugieren que la selección de genes significativos para su clasificación, utilizando el algoritmo SVM-

RCE, fue más fiable que los algoritmos SVM-RFE o PDA-RFE. El algoritmo SVM-RFE utiliza el coeficiente de peso, el cual aparece en la fórmula de SVM, para indicar la contribución de cada gen al clasificador. El éxito del algoritmo SVM-RCE sugirió que las estimaciones basadas en la contribución de genes, que compartían un perfil similar (genes correlacionados), eran importantes y proporcionaban a cada grupo de genes el potencial para ser categorizados como un grupo. Además, los genes seleccionados por el algoritmo SVM-RCE tenían la garantía de ser útiles para la clasificación global puesto que la medición de los genes (grupo de genes) de mantenimiento o de retirada estaba basada en su contribución al rendimiento del clasificador. La formación de agrupamientos sin supervisar por el algoritmo SVM-RCE es también útil para identificar sub-grupos de muestras biológica o clínicamente importantes.

Ejemplo 6: Formatos de ensayo

Para proporcionar una signatura biomarcadora que pueda ser usada en la práctica clínica para diagnosticar cáncer de pulmón, se proporcionó un perfil de expresión genética con número *más pequeño* de genes que conservan una precisión satisfactoria, mediante el uso de tres o más genes identificados en la Tabla I, II, III o IV. Estos perfiles o signaturas genéticas permiten pruebas más simples y más prácticas que son fáciles de usar en un laboratorio clínico estándar. Puesto que el número de genes discriminantes es bastante pequeño, las plataformas PCR cuantitativas en tiempo real se desarrollan utilizando estos perfiles de expresión genética.

A. PCR cuantitativa en tiempo real (RT-PCR)

Un ensayo diagnóstico según se describe en la presente memoria, puede emplear Matrices de Baja Densidad TAQMAN® (TLDA). Los perfiles de expresión genética descritos en la presente memoria sugieren que el número requerido de genes es compatible con estas plataformas. La RT-PCR ha sido considerada como el "estándar de oro" para validar resultados de matriz. Sin embargo, al constituir un diagnóstico basado en PCR, se incrementan los problemas de reproducibilidad con el número de genes requeridos para incrementar la diagnosis y, de forma más crítica, si las diferencias en los niveles de expresión son pequeñas.

Inicialmente, se usó una tarjeta micro-fluídica de Matriz de Baja Densidad TAQMAN® diseñara para ensayar 24 genes por duplicado utilizando ensayos TAQMAN® Multiplexados. Esta configuración particular ensaya 8 muestras diferentes que han sido cargadas en los puertos numerados de la parte superior de la tarjeta. Un perfil de 24 genes fue probado por duplicado con 8 muestras por tarjeta. Cada muestra fue ensayada por duplicado en pocillos precargados con los ensayos de gen específicos, reduciendo la variabilidad asociada a los ensayos de pocillo simple. Las reacciones de transcripción reversa para cada una de las 8 muestras fueron cargadas en los pocillos con la parte superior etiquetada como 1-8. Esta plataforma es útil tanto para validación de resultados de la matriz como para el desarrollo de una plataforma de diagnóstico que va a ser probada sobre nuevas muestras. La utilización de las tarjetas TLDA simplifica significativamente la validación de expresión de matriz, y también proporciona una alternativa razonable a las plataformas de matriz StaRT PCR y Enfocada, a efectos de validación de clasificador.

B. StaRT PCR

StaRT PCR (Expresión de Gen) es esencialmente una PCR competitiva con estándares internos tanto para el gen de interés como para el (los) gen(es) de limpieza. Disponer de controles internos para los genes experimentales y de limpieza, tiene la ventaja de proporcionar una referencia conocida en cada muestra, y una cuantificación directa de números de copia de mensaje en vez de un número de copia relativa, según se ha referenciado en una curva estándar con un ARN de referencia. Esta técnica es en la actualidad la única tecnología que cumple las guías de la FDA respecto a Métodos de Ensayo Multi-Gen para Farmacogenómica. La alta precisión absoluta de este método se sustituye en los métodos descritos en la presente memoria por el uso de múltiples genes y controles internos. Sin embargo, la matriz de diagnóstico puede ser probada frente a StaRT PCR para comparar la precisión y el coste.

C. Matriz de gen de diagnóstico enfocada

Según se desarrollaron los perfiles de diagnóstico, los resultados procedentes de matrices Illumina fueron comparados con datos RT-PCR procedentes de las TLDAs. Cualquiera de entre una matriz ILLUMINA habitual o una TDLA habitual, puede estar diseñada para uso clínico.

Ejemplo 7: Estudios que utilizan un diagnóstico de matriz y una herramienta PCR para diagnosticar cáncer de pulmón en muestras procedentes de pacientes con nódulos de pulmón pequeños, no diagnosticados

Se validó la utilidad diagnóstica de los ensayos clínicos descritos en lo que antecede. La población del estudio consistió en sujetos en los que había sido identificado un nódulo de pulmón mediante exploración ya sea de rayos X o ya sea de CT. Este grupo de pacientes representaba una población ideal para el uso de un biomarcador por dos razones principales. En primer lugar, el riesgo global de cáncer de pulmón era relativamente alto (18-50%) en este grupo, dependiendo del tamaño del nódulo (>0,8 cm). En segundo lugar, existían riesgos y costes importantes asociados a la evaluación diagnóstica de estos pacientes, lo que en general incluye exploraciones CT serie, exploraciones PET, procedimientos de biopsia invasiva, y, en algunos casos, cirugía.

Los sujetos del estudio eran pacientes con un nódulo pulmonar solitario, no calcificado (>0,8 cm y <3 cm de

diámetro) detectado mediante exploración de rayos X o de CT. Solamente los sujetos sin síntomas específicos que fueran sugerentes de malignidad (por ejemplo, hemoptisis, pérdida significativa de peso) fueron incluidos (es decir, pacientes asintomáticos). Síntomas no específicos (por ejemplo, disnea o tos) son frecuentemente comunes en fumadores actuales o antiguos, y por lo tanto los sujetos con estos síntomas fueron incluidos. Los pacientes en los que se descubrió que tenían un nódulo de pulmón no calcificado, fueron evaluados normalmente en base a la probabilidad clínica de la malignidad. De ese modo, todos los sujetos del grupo fueron finalmente identificados como un caso de cáncer de pulmón o bien como un sujeto de control en base a los criterios patológicos y clínicos específicos que se han expuesto en lo que antecede. Los sujetos del caso utilizados para este objetivo, eran similares a los sujetos del caso que se ha descrito en los ejemplos que anteceden.

La población de control (sujetos con nódulos benignos) era diferente de la población de control descrita en los ejemplos anteriores dado que solamente se incluyeron pacientes con nódulos de alto riesgo. Los controles fueron confirmados mediante análisis radiográfico o estabilidad radiográfica durante más de dos años.

Los datos procedentes de ensayos cuantitativos de RT-PCR o de matrices de gen enfocadas, fueron evaluados como pruebas diagnósticas. El análisis principal estimó la sensibilidad y especificidad de los perfiles de expresión de gen descritos en lo que antecede. Puesto que la sensibilidad y la especificidad dependen del valor del punto de corte del valor cuantitativo de RT-PCR (para un biomarcador simple) o de la puntuación discriminante lineal (para una matriz de biomarcadores), se ejecutó un análisis de características operativas del receptor (ROC) que esquematizó la sensibilidad y la especificidad como una función del valor del punto de corte. El área bajo la curva de ROC fue estimada mediante métodos convencionales[59].

Se estimó el valor predictivo positivo (PPV) y el valor predictivo negativo (NPV), es decir, la posibilidad de que un sujeto con una prueba positiva tenga realmente cáncer (el PPV) o la probabilidad de que un sujeto con una prueba negativa no tenga cáncer (el NPV). Puesto que estas cantidades dependen de la prevalencia de cáncer en el grupo que está siendo probado, así como de la sensibilidad y especificidad de la prueba, estas cantidades fueron calculadas respecto a que se mantuviera una gama de probabilidades de prevalencias posibles en diferentes poblaciones clínicas. Se llevó a cabo un análisis de subgrupo para determinar el efecto de la raza, del género, de la edad y del estado de fumar sobre la precisión de la puntuación discriminante.

Se efectuó un análisis de regresión logística (virtualmente equivalente al análisis discriminante lineal (LDA)) de los marcadores objetivo, y se evaluaron determinadas variables clínicas utilizando la alternativa de rutina de carga[60] para corregir el sobre-ajuste en la estimación de tales índices de predicción como el área bajo la curva de ROC y la estadística alfa de Cronbach. Las variables clínicas importantes (tamaño de nódulo, años de paquete, años desde que dejó de fumar, edad y género), fueron usadas para crear un modelo predictivo básico. El valor de los biomarcadores de expresión genética para predecir cáncer de pulmón fue calculado creando modelos adicionales que incorporen la puntuación discriminante lineal. Este análisis estableció el valor incremental de los biomarcadores de expresión genética como parte de la evaluación clínica de pacientes con nódulos de pulmón asintomáticos.

Para determinar si el biomarcador es útil como activador para el cambio en la intervención en un ensayo, las estimaciones de tamaño de muestra estuvieron basadas en valores objetivo en cuanto a especificidad de 0,9 y sensibilidad de 0,9. Para asegurar estos intervalos de confianza para que la sensibilidad y la especificidad no se extiendan más que un 5% de los valores estimados, se usaron al menos 138 casos y 138 controles.

Ejemplo 8: Determinación de valores predictivos positivos (PPV) y negativos (NPV) para el NSCLC frente al perfil de NHC

Los valores para el PPV y el NPV calculados respecto a la sensibilidad y la especificidad alcanzadas probando los cánceres de NSCLC combinados frente a las muestras de NHC, han sido mostrados en la Tabla IX que sigue. Los valores de prevalencia sugeridos por el Grupo Biomarcador de Cáncer de Pulmón (LCBG) de EDRN disponibles en (http//edm.nci.nih.gov/resources/simple-reference-sets) fueron adoptados con fines de protección. El valor de prevalencia es 0,01 para una población en riesgo de edad >50 y un estado de fumador >30 años, y 0,05 para un individuo que presenta una exploración de CT anormal, con un nódulo no calcificado de entre 0,5 y 3 cm. Estos valores de PPV y NPV fueron comparados con los valores considerados como útiles para su estudio adicional por el LCBG, y con los valores determinados a partir de un reciente estudio que utiliza un perfil de 80 genes obtenido a partir de raspados bronquiales, suponiendo la misma prevalencia. El clasificador de 15 genes (véase la Tabla IV, col. NSCLC/NHC) supera ya el rendimiento sugerido mediante LCBG para un buen candidato biomarcador, y también excedía el de la especificidad de biomarcador de cáncer de pulmón publicado más recientemente.

Tabla IX

Valores Predictivos Positivos y Negativos para NSCLC de 15 genes frente a perfil de NCH					
Sujeto	**Sensibilidad**	**Especificidad**	**Prevalencia**	**PPV**	**NPV**
Clasificador de 80 genes (Spira et al., 51)	0,83	0,76	1%	0,034	0,998
			5%	0,154	0,988

Biomarcador propuesto de LCBG	0,80	0,70	1%	0,026	0,997
			5%	0,123	0,985
NSCLC frente a clasificador de 15 genes NCH	0,86	0,79	1%	0,040	0,998
			5%	0,177	0,991

Ejemplo 9: Cálculos de potencia

Con el fin de estimar el número de muestras necesarias para conseguir una precisión especificada a partir de las clasificaciones, se utilizó el método conocido como Mujherjee[52]. La estimación se hizo construyendo una curva de aprendizaje empírica que expresaba una tasa de error de clasificación *e* como una función *n* del tamaño del conjunto de ejercitamiento, de acuerdo con: $e(n) = an^{-\alpha} + b$, donde *a, α, b* han de ser encontrados mediante ajuste de la curva a las tasas de error observadas cuando se utiliza una gama de tamaños de conjunto de ejercitamiento elaborados a partir de un conjunto de datos preliminares. El conjunto de datos preliminares consistió, en este caso, en 78 NSCLC de tipos de células mezcladas y 52 muestras de NHC, dando como resultado 130 muestras disponibles para cálculos de potencia. Éste fue el conjunto de clasificación más dificultoso. Las tasas de error fueron registradas tomando subconjuntos de ejercitamiento de tamaños de 25, 32, 38, 45, 51, 58, 64, 70 y 71 muestras (corresponden aproximadamente a un 20% a 60% de las muestras), conservando la proporción original de casos de NSCLC y de NHC.

La SVM fue ejecutada 50 veces por cada tamaño utilizando muestras aleatorias cada vez, muestras de clasificación que utilizan los 500 mejores genes seleccionados mediante prueba-t entre casos y controles. Las tasas de error medio, junto con cantidades porcentuales de un 25% y 75% por cada tamaño de conjunto de ejercitamiento, fueron utilizadas para ajustar la curva de aprendizaje. La tasa de error para esta formación clasificadora que utiliza 117 muestras (un 90%) como conjunto de ejercitamiento, puede ser observada sobre la curva de ROC con una AUC de 0,867 (no representada). La precisión de un 83% (error de un 17%) está basada en la curva calculada (no en la representada). La tasa de error real fue de 0,17 observada para el tamaño de ejercitamiento máximo disponible a partir de datos preliminares. Se detectaron aproximaciones de de tasa de error de un 25% y un 75% en un conjunto (datos no representados).

Ejemplo 10: Clasificación de adenocarcinoma (AC) de pulmón en fase de desarrollo temprana y de carcinoma de célula escamosa de pulmón (LSCC) a partir de PBMC usando matrices de cADN

Para determinar si era posible detectar una signatura de expresión de gen en la sangre periférica que pudiera estar correlacionada con NSCLC en fase temprana, se usaron muestras procedentes de pacientes de AC y de LSCC puesto que éstos representan alrededor de un 85% de todos los NSCLC. Las formas menos comunes de NSCLC (por ejemplo, un carcinoma de célula grande) pueden ser detectadas también mediante la formación de un clasificador sobre los tipos más comunes de NSCLC.

El procesamiento de todas las muestras para purificación de ARN se llevó a cabo bajo condiciones estandarizadas.

Los inventores generaron un clasificador mediante obtención de ARN de PBMC a partir de conjuntos de pacientes de control “no sanos” (NHC) y pacientes con varios tipos y fases de desarrollo de NSCLC, y realizando análisis de micro-matriz utilizando una plataforma de cADN, es decir, matrices de cADN de nailon fabricadas por la Wistar Genomics Core.

El análisis fue llevado a cabo utilizando Máquinas de Vector de Soporte con Eliminación de Característica Recursiva (SVM-RFE), según ha sido descrito en los ejemplos 4 y 5, y en otras publicaciones[48]. En algunos casos se utilizó el algoritmo de las Máquinas de Vector de Soporte con Eliminación de Grupo Recursivo (SVM-RCE) (Publicación de solicitud de Patente Internacional núm. WO 2004/105573). Los intentos iniciales por clasificar patrones a partir de controles procedentes de PBMC usando SVM-RFE, dieron como resultado tasas de error para algunas de las comparaciones, en particular todas las de cáncer frente a NHC, demasiado altas para ser útiles (precisión media de alrededor de un 70%). Para direccionar la baja relación señal/ruido, se desarrolló un nuevo algoritmo SVM-RCE que agrupa los genes (mediante agrupamiento de K-medios) en grupos cuya expresión diferencial está correlacionada, y que elimina recursivamente los agrupamientos menos informativos en vez de genes individuales. Esto da como resultado una selección final de grupos de genes cuya expresión diferencial cambia en conjunto. Sobre 6 conjuntos de datos[1] publicados, este método demostró ser más preciso en cuanto a clasificación que la SVM-RFE o el análisis discriminante penalizado (PDA-RFE), y en algunos casos también dio como resultado un agrupamiento biológicamente significativo de muestras. Éste es el más útil en cuanto a datos con baja relación señal/ruido o alta varianza puesto que la utilización de grupos de genes como variables minimiza los efectos de ambos aspectos mencionados de los datos.

Si se aplicó SVM-RFE o SVM-RCE, con el fin de eliminar en sobre-ajuste dentro de un conjunto de datos, se utilizó validación cruzada de M-pliegues (siendo M igual a 10). El algoritmo fue ejercitado sobre M-1 pliegues del caso y del grupo de control, y a continuación probado sobre el restante 1 pliegue. Esto garantiza que cada muestra se emplee

como sujeto de prueba. Se calculó la puntuación media para cada paciente y también la puntuación media para cada gen. El (los) gen(es) menos informativo(s) fue (fueron) eliminado(s), y se repitió el proceso. Las Tablas XA y XB muestran la precisión de clasificación y la sensibilidad (tasa positiva verdadera) y la especificidad (tasa negativa verdadera) frente al número de genes utilizados para la clasificación. Las alternativas analíticas están descritas con detalle en lo que sigue.

Los datos para los 208 pacientes y controles listados en la Tabla XA, han sido mostrados en la Tabla XB. Estos datos fueron procesados en tres "lotes" diferentes denominados conjuntos 3, 4 y 5. Según se describe en la Tabla XA, las muestras fueron agrupadas como adenocarcinomas en fase de desarrollo temprana (AC T1T2), adenocarcinomas en fase de desarrollo tardía (AC T3T4), cáncer de pulmón de célula escamosa en fase de desarrollo temprana (LSCC T1T2) y controles no sanos (NHCs). Tanto los casos como los controles eran normalmente fumadores antiguos o ex-fumadores.

En segundo lugar, aunque la clasificación de los NSCLCs en fase de desarrollo temprana fue muy difícil, se pudo conseguir una precisión muy buena comparando los ACs con los NHCs o bien los LSCCs con los NHCs solos o respecto a un clasificador combinado AC+LSCC (Tabla XB- 3 líneas superiores). La comparación de AC+LSCC con los NHCs necesitó inicialmente 287 genes para clasificar muestras combinadas en fase de desarrollo temprana con NHCs con una precisión de un 80% (línea 1). Sin embargo, estos resultados sugirieron que podía ser posible desarrollar un clasificador más general que pudiera detectar tanto ACs como LSCCs. Cuando los ACs y los LSCCs fueron segregados y clasificados por separado, se necesitaron inicialmente 160 genes para distinguir ACs en fase de desarrollo temprana de los NHCs con una precisión de un 85%, y solamente 56 genes para identificar el LSCC con la misma precisión. Se encontró entonces que la comparación entre las muestras de ACs y LSCCs en fase de desarrollo temprana requería solamente 21 genes para la discriminación, confirmando las observaciones previas de los inventores diferencias significativas entre estos 2 tipos de células de NSCLC. Finalmente, según se muestra en la Tabla IV, columna "AC/NHC", un perfil genético de 15 genes puede distinguir AC respecto a otras formas de NSCLC. Se prevé un análisis adicional para demostrar que solamente se necesita una cantidad tan baja como 6 genes para este perfil, al igual que con el perfil de pre/post cirugía formado por los 6 genes superiores de la Tabla IV, col. Pre/Post.

TABLA XA

Sumario de Muestras Analizadas sobre Matrices de cADN	
AC T1T2	59
AC T3T4	18
LSCC T1T2	36
LSCC T3T4	12
NHC	95

TABLA XB

Clases de muestras comparadas	**# Genes Requer. Para Clasif./# grupos**	**Precisión de la Clasificación**	**Sensibilidad**	**Especificidad**
[1]AC+LSCC T1T2 frente a NHC	287/22	0,8	0,82	0,78
[1]AC T1T2 frente a NHC	160/11	0,85	0,83	0,85
[2]LSCC T1T2 frente a NHC	105	0,87	0,72	0,93
[2]LSCC T1T2 frente a NHC	56/2	0,85	0,90	0,84
[2]AC T1T2 frente a LSCC T1T2	21	0,88	0,92	0,81
AC T1T2 frente a LSCC T1T2	3	0,85	0,86	0,83
AC T1T2 frente a AC T3T4	10	0,92	0,98	0,72
[1]AC T3T4 frente a NHC	15/2	0,88	0,77	0,94
[1] Se utilizó SVM-RCE para estos análisis [2] Se informaron dos precisiones en las que resulto un pequeño descenso de precisión a partir de un gran descenso en el número de genes				

Puesto que las diferencias de expresión genética detectadas entre los casos y los controles podían estar causadas por un cambio en alguna fracción de la población de PBMC, se llevó a cabo un estudio de una pequeña citometría de flujo comparando fracciones de PBMC procedentes de linfocitos de 14 NHC con linfocitos de 14 pacientes con AC, 15 pacientes con LSCC, y otros 6 con NSCLC. En concordancia con los recientes hallazgos[49] para pacientes con melanoma maligno, no existió diferencia estadísticamente significativa en proporciones de Células-T, Células-B

o monocitos de CD4 o CD8 entre los casos y los controles.

Ejemplo 11: Clasificación de NSCLCs en fase de desarrollo temprana (T1/T2) a partir de NHCs sobre matrices Q-PRC de ILLUMINA

Los resultados de la matriz de cADN necesitaron 287 genes para distinguir las clases combinadas de muestras de NSCLC de los NHCs (véase la Tabla XB, línea 1, que antecede). Los datos Illumina permitieron, sin embargo, el desarrollo de un clasificador más preciso y global para clasificación de AC/NHC con muchos menos genes. Los datos de Illumina disponibles para este análisis incluían 78 muestras de NSCLCs (incluyendo 51 ACs, 15 LSCCs, 12 NSCLCs sin clasificar) y 52 muestras de NHC. El análisis de SVM-RFE indicó que 15 genes podían clasificar este conjunto de datos con una precisión de un 83%. Véase la Tabla IV anterior. Las puntuaciones de SVM para los pacientes individuales y los controles, mostradas en la Figura 3, fueron generadas a partir del rendimiento del clasificador de 15 genes de la Tabla IV, columna NSCLC/NHC. Estos resultados muestran que se puede usar un clasificador más general para clasificar los dos tipos de células principales de NSCLC.

En un experimento, se utilizó la PBMC de 44 pacientes con AC pequeño (tumores de tamaño T1 o T2) frente PBMC procedente de 95 controles emparejados por la edad, el género y la acción de fumar. Se generaron puntuaciones discriminantes utilizando matrices de nailon y SVM-RCE según se ha descrito con anterioridad. Los resultados se proporcionan en la Figura 2. Una puntuación positiva indica cáncer de pulmón, y una puntuación negativa indica que no hay cáncer. Cada columna representa un paciente o muestra de control únicos. La altura de la columna es una medición de lo bien clasificada que está una muestra individual. Las muestras de control están a la derecha y se les ha dado una puntuación negativa. Los pacientes están a la izquierda. Las barras más claras con puntuación positiva son controles mal clasificados y las barras más oscuras con una puntuación negativa son casos mal clasificados. Las muestras del margen con puntuaciones cercanas a cero, podrían estar sin clasificar. Solamente se han mostrado las muestras de AC T1T2. Las muestras del centro donde las columnas cambian de positivas a negativas, o viceversa, están mal clasificadas. Utilizando este clasificador que emplea 15 genes de la Tabla IV, col. AC/NHC, se identificó la presencia de cáncer de pulmón en fase de desarrollo temprana con una precisión de un 85%.

En otro experimento, se compararon cuarenta y cuatro (44) muestras de pacientes con AC T1T2 en fase de desarrollo temprana con 52 NHC. Los genes fueron filtrados mediante prueba-t y a continuación se aplicó SVM-RFE (véase el ejemplo 4 ó 5) y los 15 genes seleccionados por SVM-RFE fueron utilizados (Tabla IV, columna AC/NHC). Las precisiones de clasificación fueron analizadas con eliminación progresiva de gen (desde 2781 genes a 1) mediante SVM-RFE[48] (datos no mostrados), midiendo Positivos Verdaderos, es decir, el número de pacientes a los que el clasificador asignó correctamente una puntuación de SVM positiva, y Negativos Verdaderos, es decir, el número de controles a los que el clasificador asignó correctamente una puntuación de SVM negativa. La puntuación fue representada gráficamente como (TP+TN)/n (n = número total de muestras). La relación s/n favorable y la varianza más baja utilizando las matrices Illumina, hicieron que el uso del algoritmo SVM-RCE fuera innecesario. Se utilizó SVM-RFE para todos los estudios Illumina puesto que la SVM-RCE necesita tiempos de ejecución mucho más largos que la SVM-RFE. El clasificador óptimo s elige en base a la mejor precisión con el número más pequeño de genes. Se encontró que niveles de expresión de sólo 15 genes (por ejemplo, los 15 genes superiores de la Tabla IV, columna etiquetada como TODOS/NHC) discriminan los ACs T1T2 en fase de desarrollo temprana de los NHCs con una precisión global de un 85%. Esta misma precisión fue encontrada con matrices de cADN, pero se necesitaron inicialmente 160 genes para este grado de separación. Estos resultados confirman que el perfil de expresión genética no es una plataforma específica. El descubrimiento del perfil de expresión genética original de los inventores fue confirmado sobre una segunda plataforma totalmente diferente.

Ejemplo 12: Cambios en signaturas asociadas a tumor en PBMC tras la extracción del tumor

Para identificar una signatura que refleje la presencia de tumor y que sea útil para la evaluación de la probabilidad de recurrencia, se compararon perfiles de PBMC procedentes del subconjunto de pacientes con cánceres de pulmón en fase temprana de los que se habían tomado muestras de sangre con anterioridad y recientes (2-6 meses) después de la cirugía “curativa”. Esto minimizó el “ruido” de fondo de modo que una signatura de expresión genética correlacionada con la presencia de un tumor puede ser identificada más fácilmente. La reversión del perfil de PBMC a un perfil de “cáncer de pulmón”, pronostica así la recurrencia.

A. Efecto de presencia de tumor

Con el fin de determinar si la diferencia en los perfiles de expresión genética apreciada entre los casos y los controles dependía de la presencia del tumor, los inventores examinaron cómo unas muestras de PBMC tomadas en el mismo paciente de NSCLC antes de la cirugía y de nuevo ~2-6 tras meses después de la cirugía fueron clasificadas con el clasificador de 15 genes que fue seleccionado, en comparación con 78 pacientes de NSCLC y 52 NHCs (véanse las Figuras 3 y 4). Los genes seleccionados en esta comparación como muestras de pre-post fueron extraídos de pacientes con AC, con LSCC o bien con NSCLC indeterminados. Las muestras de NSCLC pre-cirugía fueron incluidas en el análisis mostrado en las Figuras 3 y 4, pero las muestras post-cirugía no fueron incluidas. Las muestras post-cirugía comprenden un conjunto de prueba independiente. La razón era determinar si las muestras de paciente recogidas en la post-cirugía conservaban la signatura de tumor, lo que en este caso se ha indicado mediante una puntuación predictiva positiva, o si la extracción del tumor podía reducir la signatura de tumor y

aquellas deberían puntuar ahora más como los controles. Las probabilidades de que esto ocurriera por casualidad son <0,01.

13 de 16 pares de pacientes presentaron una reducción en la puntuación predictiva de tumor tras la cirugía. Seis de los casos tienen puntuaciones pre-cirugía positivas y una puntuación post cirugía que es negativa, situándolos claramente en la clase de control, mientras que 4 muestras adicionales tuvieron caídas significativas en las muestras post-cirugía llevándolas a valores cercanos a cero. Dos de los casos no tuvieron ningún cambio en la puntuación de tumor tras la cirugía, y 1 caso tuvo un incremento en la puntuación de tumor. Dos de los casos tienen una puntuación pre-cirugía negativa pero incluso en este caso se hace más negativa en la muestra post-cirugía. Un seguimiento adicional del paciente determina la medida en que las puntuaciones post-cirugía son un pronóstico en cuanto a recurrencia. La observación de que la signatura de tumor disminuyó después de la extracción de la malignidad, soportó que el perfil o la signatura de expresión genética fuera una respuesta a la presencia del tumor. Véanse las Figuras 5 y 6.

B. Comparación de muestras de pre- y post- cirugía

Las muestras de pre-cirugía fueron comparadas con las muestras de post-cirugía para determinar si las 2 clases de muestras podían ser separadas en base a las diferencias intrínsecas que fueron mostradas por los análisis por parejas en la Figura 3. Las 16 muestras pre-cirugía fueron comparadas con las 16 muestras post-cirugía. Se llevó a cabo SVM-RFE a partir de los 1.000 genes identificados mediante prueba-t utilizando validación cruzada de 10 pliegues repetida 10 veces. Solamente se determinaron seis genes para distinguir las muestras pre- de las post- con una precisión de un 93%. Este clasificador de 6 genes (los genes superiores identificados en la Tabla IV (columna Pre/Post) fueron utilizados entonces para generar puntuaciones discriminantes para las muestras de pre- y post-cirugía como se muestra en la Figura 3. Las muestras de pre-cirugía (sombreado oscuro) están todas clasificadas correctamente aunque una muestra tiene una puntuación cercana a cero. Una de las muestras post-cirugía tiene una puntuación negativa cercana a cero y 2 están mal clasificadas. Este resultado sugiere que se podría desarrollar un clasificador que pudiera ser eficaz en cuanto a la protección de pacientes post-cirugía respecto a recurrencia debido a que podría proporcionar la posibilidad de comparar puntuaciones post-cirugía con la puntuación pre-cirugía inicial del mismo paciente con el tiempo. Las muestras que siguen proporcionan un indicador sensible de recurrencia.

En otro estudio, utilizando datos de matriz Illumina, se seleccionaron genes por comparación de muestras de cáncer de pulmón pre-cirugía con controles de fumador con NHC. Se utilizaron cincuenta y cuatro (54) genes para clasificar las muestras post-cirugía. Se dio una puntuación discriminante a cada muestra (el positivo es indicativo de cáncer de pulmón; el negativo es indicativo de que no hay ningún cáncer). En el análisis de fase temprana (no representado) en todas menos en una comparación la puntuación post- es más baja que la puntuación de la muestra pre-cirugía, lo cual es adyacente. En tres casos, la puntuación de la muestra post-cirugía es negativa, clasificando estas muestras con los controles de COPD. Este dato soporta la detección de una signatura de expresión de gen relacionada con tumor que disminuye tras la cirugía. La extensión de estos cambios refleja la posibilidad de recurrencia.

Dados los resultados positivos del estudio piloto sobre 16 muestras emparejadas presentadas en el estudio, la utilidad de esta prueba se basa en su aplicación junto con la presencia de un nódulo de pulmón detectado mediante otros procedimientos tal como mediante exploraciones de CT. Además, se pueden diferenciar NSCLCs de diferentes tipos de células (ACs y LSCCs) mediante una signatura diseñada para hacer esta distinción.

Ejemplo 13: Comparación de los 15 genes superiores según estén clasificados, mediante SVM-RFE para los 3 clasificadores de SVM-RFE

Los 15 genes superiores por orden de SVM-RFE a partir de los 3 estudios Illumina, han sido listados en la Tabla IV que antecede. Los órdenes para cada uno de los genes según se han asignado en los estudios individuales por SVM, se mantienen en la Tabla IV. Para la comparación de AC/NHC y la comparación de todos los tipos de células de NSCLC con NHC (TODOS/NHC), los 15 genes listados son los genes utilizados para asignar las puntuaciones de SVM mostradas en las Figuras 2, 4 y 6. Los 15 genes para la comparación de TODOS/NHC fueron $p<32x10^{-5}$. Los 15 genes de AC/NHC fueron $p<2x10^{-4}$ y los genes de Pre/Post fueron $p<6x10^{-43}$. Los primeros 6 genes de la columna PRE/POST fueron usados para generar las puntuaciones para la Figura 4. Los genes mostrados con tipos en negrita son comunes para 2 ó 3 comparaciones. Los genes que no son comunes a los 3 clasificadores no son necesariamente únicos para esa comparación sino que pueden aparecer simplemente en una posición de orden más bajo en las listas de genes ampliadas. Ocho de los 15 genes de categoría superior para la AC/NHC y la de TODOS/NHC aparecen en ambas listas. De los 6 genes superiores utilizados para la clasificación de PRE/POST, 3 están listados en cualquiera, o en ambas de las otras listas. Dos sondas para HSPA8 están listadas. La (**A**) indica que todos los isotipos de HSPA8 son detectados por esta sonda, (**I**) indica que un isotipo específico (en este caso, la variante 1 de transcripción) es detectado por la segunda sonda de HSPA8.

Los datos sobre la clasificación informada de plataforma de matriz de cADN proporcionan precisiones para las comparaciones de NSCLCs de diferentes tipos de células y de fases de desarrollo T y en NHCs, y de unos con otros. Los datos preliminares de los inventores sobre la plataforma Illumina estaban limitados a aquellos pacientes con AC en fase de desarrollo temprana frente a NHCs o NSCLCs combinados frente a NHCs. Esto fue por decisión propia, puesto que los ACs son el tipo más común de NSCLCs y era importante minimizar la heterogeneidad

histológica en las muestras iniciales que iban a ser analizadas en la nueva plataforma. Un clasificador más general incluye un conjunto de muestra de casos más diversificados que incluyen LSCCs y NSCLCs indeterminados. Las muestras adicionales ensayadas sobre matrices Illumina demuestran si los subtipos particulares de cáncer de pulmón (es decir, AC frente a LSCC) tienen sus propios patrones de expresión distintos según sugieren las matrices de cADN y/o si existe alguna signatura de PBMC que pueda identificar de manera precisa todos los NSCLCs en fase de desarrollo temprana.

En una realización, la columna de TODOS/NHC de la Tabla IV muestra el perfil de 15 genes para identificar un NSCLC a partir de controles. En otra realización, la columna de AC/NHC de la Tabla IV muestra el perfil de 15 genes para identificar un AD. En otra realización más, la columna de PRE/POST muestra el perfil de 15 genes para identificar la eficacia de la resección quirúrgica del tumor y la prognosis respecto al futuro. Según se ha descrito anteriormente, este perfil de gen ha sido reducido con éxito a solamente los 6 genes superiores de esa columna. Se entiende que selecciones más pequeñas de genes podrán ser identificadas para los otros dos perfiles indicados también. En otra realización, las signaturas específicas del tipo de célula que utilizan genes que están presentes en las tres signaturas, se prevé que aumenten la potencia predictiva de estas puntuaciones informadas.

Ejemplo 14: Signatura de expresión de 29 genes

Para identificar una signatura de expresión genética de PBMCs que pudiera distinguir de forma precisa pacientes con cáncer de pulmón respecto a controles sin cáncer con factores de riesgo similares (es decir, emparejados por edad, género, raza, historia como fumador), se compararon perfiles de expresión genética de células mononucleares de sangre periférica (PBMC) de pacientes con NSCLC, con un grupo de control con enfermedad de pulmón no maligna relacionada con el fumar. Se encontró una signatura genética distintiva y se validó sobre 2 conjuntos independientes de muestras no usadas para selección de gen. Los cambios de expresión genética fueron comparados también entre muestras de pre- y post- cirugía de 18 pacientes.

Se encontró una novedosa signatura de diagnóstico de 29 genes (genes categorizados como 1-29 en la Tabla V), que distingue individuos con NSCLC de controles con enfermedad de pulmón no maligna con una sensibilidad de un 91%, una especificidad de un 79% y una AUC de ROC de un 92%. Las precisiones sobre conjuntos independientes de 18 muestras de NSCLC a partir de la misma posición y de 27 muestras desde una posición independiente, fueron de un 74% y un 79%, respectivamente. La signatura de 29 genes fue reducida significativamente tras la extracción del tumor en el 83% de un subconjunto de 18 pacientes en los que se midió la expresión genética antes y después de la resección quirúrgica.

Aunque ambas situaciones de fumar y COPD pueden afectar, cada una de ellas, a la expresión genética de PBMC, se puede identificar la respuesta adicional a la presencia de un tumor, permitiendo la diagnosis de pacientes con cáncer de pulmón a partir de controles con alta precisión. La signatura de PBMC es particularmente útil en el algoritmo de diagnóstico para esos pacientes con un nódulo de pulmón no calcificado. La observación de que la signatura de 29 genes disminuye tras la resección quirúrgica, soporta que está relacionada con el tumor.

Poblaciones de estudio: Los participantes en el estudio (Tabla XVI) para los conjuntos iniciales de ejercitamiento y validación fueron reclutados en el Centro Médico de la Universidad de Pennsylvania (Penn) durante el período de 2003 a 2007: 91 sujetos con una historia de uso de tabaco sin cáncer de pulmón que incluyen 41 sujetos que tenían un nódulo de pulmón no calcificado diagnosticado como benigno después de una biopsia, y 155 pacientes con cáncer de pulmón de célula no pequeña confirmado histopatológicamente, recién diagnosticado. Los sujetos con cualquier historia anterior de cáncer o en tratamiento de cáncer salvo cáncer de piel no melanoma, fueron excluidos. El estudio fue aprobado por el Penn Institutional Review Boards. 27 pacientes y controles adicionales, fueron recopilados en el Centro Médico de la Universidad de Nueva York (NYU) bajo la aprobación de IRB y también aparecen listados en la Tabla XVI.

Tabla XVI: Resumen de datos demográficos

Categoría	Número de pacientes
Todos NSCLC frente NHC	muestras experimentos
Total	228
Controles	91
Pacientes	137
tenían COPD	128
sin COPD	82
COPD desconocida	18
sin COPD	82
Fumadores	34
Dejaron de fumar	170
Nunca fueron fumadores	24

Pacientes de NSCLC frente a NHC	experimento
Total	137
AC	85
LSCC	42
NSCLC	10
tiene COPD	63
sin COPD	65
COPD desconocida	9
Fase 1A	48
Fase 1A+1B	75
Fase 4	5

Pacientes de NSCLC frente NHC	experimento
Total	137
Fase$^{1/2}$	93
Fase$^{3/4}$	44
Fase 2/3/4	62
AC1A	30
AC1	48
AC 2/3/4	37
LSCC 1 A	16
LSCC 1	24
LSCC 2/3/4	18
Fumadores	26
Dejaron de fumar	102
Nunca fumaron	9

Controles de NSCLC frente NHC	experimento
Total	91
COPD pura (nada más)	38
GI/NM	41
tiene COPD	65
sin COPD	17
COPD desconocida	9
Fumadores	8
Dejaron de fumar	68
Nunca fumaron	15
Parejas Pre-post	
Total	18
AC	10
LSCC	6
NSCLC	2
Muestras NYU	
Total	27
AC	12
NHC	15

Recogida y Procesamiento de PBMC: A los pacientes con cáncer de pulmón y los pacientes con enfermedad de pulmón no maligna, se les extrajo sangre con anterioridad a la cirugía y/o con anterioridad al tratamiento con quimioterapia. A los pacientes de control se les extrajo sangre en una visita clínica. Las muestras de sangre fueron elaboradas en dos tubos “CPT” (BD). La PBMC fue aislada dentro de los 90 minutos siguientes a la extracción de la sangre, lavada en PBS, transferida a solución RNA-Later (Ambion) y almacenada a continuación a 4 ºC durante la

noche con anterioridad a su transferencia a -80 ºC. Un subconjunto de PBMCs de pacientes, fueron analizadas mediante citometría de flujo con anticuerpos anti-CD3, CD4, CD8, CD14, CD16, CD19 o CD-56 o controles de isotipo (BC Biosciences) y analizadas usando software Flo-Jo. Las muestras procedentes de NYU fueron procesadas dentro de las 2 horas siguientes a la recogida, las PBMC fueron transferidas a Trizol (Invitrogen) y almacenadas a -80 ºC. El ARN extraído fue transferido a Wistar para su procesamiento adicional.

Procesamiento de Muestra: La purificación de ARN del primer conjunto de muestras "Penn" fue llevada a cabo utilizando TriReagent (Molecular Research) según se recomienda, y controlado en calidad utilizando el Bioanalizador. Solamente se usaron muestras con relaciones de 28S/16S que eran >0,75 para estudios adicionales. Una cantidad constante (400 ng) de ARN total fue amplificada según la recomendación de Illumina. El segundo conjunto de muestras "NYU" eran ADNs tratados con anterioridad a la hibridación. Las muestras fueron procesadas como lotes mezclados de pacientes y controles, e hibridizadas en matrices de perlas de genoma completo humano WG-6v2 de Illumina (http://www.illumina.com/paginas.ilmn?ID=197).

Control de calidad y pre-procesamiento de matriz: Todas las matrices fueron revisadas en cuanto a valores atípicos mediante cálculo a modo de gen entre correlación media de matriz y su comparación con la correlación para cada matriz. Se retiraron las sondas no informativas si su intensidad era baja en relación con el fondo en la mayoría de las muestras o si la relación máxima entre 2 muestras cualesquiera no era de al menos 1,2. Las matrices fueron a continuación normalizadas en cuanto a valor cuantil y el fondo fue substraído de los valores de expresión.

Análisis: Se realizó clasificación utilizando un Máquina de Vector de Soporte con eliminación de característica recursiva (SVM-RFE)[19] utilizando validación cruzada de 10 pliegues repetida 10 veces. Las puntuaciones de clasificación para cada una de las muestras fueron registradas en cada etapa de reducción, descendente hasta un solo gen. Se calculó la precisión media para cada etapa de reducción y todos los genes en los puntos de máxima precisión formaron el discriminador inicial que se sometió después a reducción adicional para formar el discriminador final según se describe en lo que sigue.

PCR cuantitativa en tiempo real: La validación de RT-PCR de los resultados de la matriz fue llevada a cabo utilizando el Sistema TaqMan ABI según lo recomendado, en un Sistema PCR 7900HT de ABI. Cada muestra fue analizada por duplicado, y se repitieron las muestras con CVs entre réplicas que eran más de 0,5 delta de Ct.

Los resultados se informan en lo que sigue:

Las variables clínicas y demográficas de las muestras del estudio (caso y control) se han resumido en la Tabla XVI anterior para 155 pacientes de caso y 91 controles clínicos incluyendo los diagnosticados clínicamente como nódulos benignos. Los grupos fueron similares en términos de edad, raza, género e historia de fumador. El 84% del grupo de control clínico y el 93% del grupo de NSCLC eran fumadores actuales o antiguos. Estas muestras fueron recopiladas todas en el Centro Médico de la Universidad de Pennsylvania. Se utilizaron 12 pacientes adicionales y 15 controles para validación externa. Se llevó a cabo citometría de flujo sobre 35 casos de cáncer y 14 controles. No existieron diferencias significativas en los porcentajes de células-T, células CD4, células-B, monocitos, o células NK (datos no representados). El grupo de tumor tenía un porcentaje ligeramente más bajo de células CD8 (18,9%) que los controles (24,5%), lo que llegó a ser significativo.

Los perfiles de expresión genética en muestras de PBMC de 137 pacientes con NSCLC fueron comparados con 91 controles con enfermedad de pulmón no maligna (controles no sanos, NHC) para determinar si se podían detectar diferencias coherentes en cuanto a expresión genética a través de un gran conjunto de datos. Se encontró que la expresión de gen en PBMC identificaba individuos con un cáncer de pulmón, por ejemplo, NSCLC. Sobre 4.500 de 48.000 sondas (9%) fueron cambiadas significativamente (prueba-t de dos colas, $p<0,5$, tasa de descubrimiento falsa del 8%) entre casos y control. A efectos de comparación, los datos informados sobre tumores de pulmón identificaron 1.649 de 12.600 transcripciones (13%) que distinguen adenocarcinomas respecto al tejido pulmonar normal, y 1.886 (15%) que distinguen carcinoma de célula escamosa del pulmón de la misma importancia. La fracción de genes cambiados en la PBMC del paciente medio de NSCLC, es similar a la fracción informada de genes cambiados entre el tumor y su equivalente de tejido normal[20].

Una máquina de vector de soporte con eliminación de característica recursiva (SVM-RFE) y validación cruzada de 10 pliegues, fue utilizada a continuación para encontrar el número mínimo de genes que podían distinguir los grupos de cáncer y de control de su expresión genética de PBMC. El proceso de selección de los 29 genes por SVM-RFE se describe con detalle como sigue.

Pre-Procesamiento de datos/Niveles de expresión y normalización: Se procesaron muestras como lotes mezclados (un total de 12 lotes) de pacientes y controles, y se hibridizaron en matrices de perlas de genoma completo humano WG-6v2 de Illumina. Los datos en bruto fueron procesados mediante software Bead Studio v. 3.0. Los niveles de expresión fueron exportados a sondas de control negativo y de señal. El conjunto de sondas de control negativo fue utilizado para calcular el nivel de fondo medio para un filtrado adicional y etapas de substracción de fondo. Los valores medios de los datos de expresión de sonda de señal para las matrices de muestra de los 137 pacientes (NSCLC) y los 91 controles (NHC) (resultados atípicos eliminados, véase lo que sigue), fueron usados como base para normalización y todas las matrices, incluyendo 18 muestras PRE/18 muestras POST y muestras NYU, fueron

normalizadas en valor cuantil respecto a esta base.

Control de calidad de matriz. Tras cada lote de hibridación, se calculó la correlación global de tipo gen como una correlación media de Spearman a través de todos los pares de micro-matrices procedentes de todos los lotes utilizando niveles de expresión de todas las sondas de señal (>48K). También se calculó la desviación absoluta media de la correlación global. A continuación, se calculó para cada micro-matriz una correlación media de Spearman respecto a todas las demás matrices. Las matrices cuya correlación media difiere de la correlación global en más de 8 desviaciones absolutas (el umbral fue determinado empíricamente) fueron marcadas como resultados atípicos y no fueron usadas para análisis adicionales. Se encontraron 22 resultados atípicos en varias fases de desarrollo, pero 11 de éstos proporcionaron datos válidos sobre matrices repetidas y éstos fueron incluidos en el análisis.

Substracción de fondo. Tras la normalización cuantil, el valor medio de fondo (60, según se determinó para estos datos) fue restado de cada uno de los datos de expresión de la sonda, lo que fue a continuación redondeado en descenso hasta una desviación estándar del fondo (15 para nuestros datos), el valor de expresión mínima usado en cualquier cálculo.

Filtraje de sonda. En base a las 137 matrices de muestra de paciente y las 91 de control, se definieron sondas no informativas al objeto de que fueran sondas no expresadas al menos 1,5 veces el fondo (corresponde a un valor de expresión de 30 para los datos de fondo substraídos) en más de un 25% (57) de las muestras o sondas que no cambiaron al menos 1,2 veces entre al menos dos muestras. Los datos de todas las matrices fueron filtrados eliminando estas sondas no informativas, dando como resultado datos de expresión de 15.227 sondas para el análisis. Estos procedimientos dan como resultado un valor cuantil normalizado, resultado atípico eliminado, fondo substraído, datos filtrados de sonda no informativa, que fueron analizados como sigue:

La alternativa principal incluyó un clasificador para un conjunto de datos ejercitado con la utilización del algoritmo de SVM. Se utilizó la estrategia de Eliminación de Característica Recursiva (RFE) para reducir el número de genes requeridos para la clasificación. A continuación, se empleó validación cruzada de 10 pliegues para evitar el sobre-ajuste de datos y proporcionar una estimación imparcial de la precisión del clasificador. El clasificador ejercitado aplicado a una muestra, proporcionó una puntuación discriminante que fue usada para pronosticar una de dos clases (enfermedad malina o no maligna, pre- o post-, etc.) para la muestra.

Validación cruzada. Se utilizó validación cruzada de 10 pliegues con 10 re-muestreos en las clasificaciones de NSCLC frente a NHC (incluyendo validaciones de exclusión y de permutación) y conjuntos de datos de PRE frente a POST. En cada una de las 10 etapas de re-muestro, los datos fueron divididos aleatoriamente en 10 partes (pliegues) mientras se conservaba la relación original de las dos clases. Cada pliegue fue usado como subconjunto de prueba una vez mientras que las otras 9 partes fueron usadas como subconjuntos de ejercitamiento. Esto dio como resultado 10 conjuntos únicos de ejercitamiento-prueba por cada re-muestreo, y se combinaron con 10 etapas de re-muestreo, 100 combinaciones únicas de 90% de muestras usadas para el ejercitamiento y 10% de muestras usadas para prueba. Esto aseguró también que cada muestra estaba involucrada en la prueba exactamente 10 veces. La prueba se hizo utilizando clasificadores que no estaban ejercitados sobre la muestra en modo alguno. Una puntuación discriminante para cada muestra fue calculada como promedio de 10 puntuaciones pronosticadas por clasificadores que no estaban ejercitados sobre un subconjunto que incluye la muestra.

RFE: Cada una de las 100 particiones únicas de ejercitamiento-prueba proporcionadas por validación cruzada, fue utilizada por SVM-RFE de manera independiente. A partir del subconjunto de ejercitamiento, se recuperaron los 1000 genes (características) superiores clasificados por medio de un valor p de la prueba-t entre las dos clases. El clasificador fue ejercitado utilizando un núcleo lineal para distinguir entre las clases utilizando niveles de expresión de esos genes. El clasificador fue aplicado a continuación a cada una de las muestras a partir del subconjunto de prueba y se registraron las puntuaciones discriminantes. A continuación, la SVM-RFE eliminó el 10% de los genes restantes que tenían coeficientes absolutos más pequeños en la función de puntuación del clasificador, es decir, aquellos genes menos importantes que afectaban a la puntuación final en lo más mínimo. El proceso se repitió (50 veces) hasta que quedó un gen para el ejercitamiento.

Rendimiento: 100 etapas de validación cruzada del proceso de SVM-RFE produjeron para cada muestra 10 puntuaciones de predicción en cada iteración de eliminación de característica. Se calculó una puntuación de muestra final como un valor medio de esas puntuaciones de predicción para cada conjunto de genes probados, desde 1000 a 1. La precisión, sensibilidad y especificidad de la clasificación fueron calculadas en base a las puntuaciones finales de las muestras, utilizando 0 como el umbral de clasificación, es decir, se clasificaron las muestras con puntuaciones ≥ 0 como de clase positiva, mientras que las muestras con puntuaciones <0 se clasificaron como negativas. Se seleccionaron los clasificadores ejercitados en tal característica de iteración de eliminación de característica que proporcionaron la mejor precisión, y un clasificador global para todas las muestras que consistía en genes de cada uno de los 100 clasificadores óptimos. Por ejemplo, 100 etapas de validación cruzada, cada una de ellas con una precisión máxima en alrededor de 8 genes, produjeron un clasificador global de 136 genes para un experimento de NSCLC frente a NHC (Tabla V anterior). Se construyó una curva de ROC de variación del umbral de clasificación desde el máximo entre puntuaciones de muestra, hasta el mínimo.

Minimización de clasificador: Para reducir el número de genes utilizados por clasificadores en todas las etapas de validación cruzada, sin re-ejercitamiento y con la condición de no reducción de precisión, los genes únicos que estuvieron involucrados en la clasificación para una iteración de RFE dada a través de todas las etapas de validación cruzada estaban clasificados por sus coeficientes absolutos promediados en la función de puntuación del clasificador. Los genes menos importantes fueron retirados, uno cada vez, de todas las funciones de puntuación. La precisión fue registrada para cada extracción y se utilizó un número mínimo de genes N que proporcionaron la misma precisión M de clasificación final. Se utilizó la notación "clasificador de N-genes que tiene una precisión de un M%" basada en estos resultados.

Aplicación de clasificador: Para nuevas muestras no utilizadas en validación cruzada, se aplicó un clasificador seleccionado en el máximo de precisión y a continuación se minimizó en genes. Este clasificador fue construido a partir de 100 sub-clasificadores recibidos en cada etapa de la validación cruzada para la iteración de RFE seleccionada. La puntuación final de la muestra fue un promedio de 100 puntuaciones proporcionadas por esos clasificadores. Obsérvese que cuando se aplica a una muestra que fue utilizada en la validación cruzada, solamente se usaron 10 de los 100 sub-clasificadores que no estaban ejercitados sobre la muestra.

137 muestras de NSCLC y 91 de NHC fueron repartidas en 5 partes. Se usó 1 parte como conjunto de exclusión y 4 partes fueron usadas como conjunto de datos que fue analizado utilizando SVM-RFE con validación cruzada de 10 re-muestreos, de 10 pliegues. El mejor clasificador final de N genes fue aplicado a continuación a la parte de exclusión. Se compararon las precisiones de la validación cruzada y de exclusión. Se generaron 10 conjuntos de datos de permutación. Las etiquetas de los 137 NSCLC y 91 HNC fueron mezcladas aleatoriamente y los datos fueron analizados utilizando SVM-RFE con validación cruzada de 10 re-muestreos, de 10 pliegues. Se seleccionó el clasificador de N genes con mejor precisión final para cada permutación, y se registró la precisión. Se calculó la precisión de permutación media a través de las 10 rondas.

El rendimiento medio de validación cruzada de SVM-RFE (figura no representada) indicó que, por término medio, se necesitaron 8 genes para obtener la mejor precisión en cada etapa durante las 100 etapas de validación cruzada. Las 100 etapas dieron como resultado 136 genes distintos informados en la Tabla V anterior. Los 136 genes que proporcionaron la mejor precisión fueron reducidos adicionalmente para la extracción por filtrado en tantos genes como fuera posible sin pérdida de precisión. Un polinomio de potencia 5 fue ajustado a la precisión para detectar el número de genes donde la precisión empieza a declinar (por ejemplo, a los 29 genes). Los genes de la Tabla V están clasificados por orden en base a su contribución a la puntuación de la clasificación final (el gen más importante está clasificado como el primero, etc.). Se han referenciado nombres y símbolos alternativos y el símbolo "NaN" indica que no se encuentra aún disponible un símbolo para el gen.

Se asignaron puntuaciones de clasificación mediante el clasificador de 29 genes a los 137 pacientes de NSCLC y los 91 pacientes con enfermedad de pulmón no maligna. Una puntuación positiva indicaba clasificación como cáncer, una puntuación negativa como enfermedad no maligna. La Tabla XI lista el número de ID del paciente, la clase de enfermedad (AC-adenocarcinoma, LSCC-carcinoma de célula escamosa de pulmón, NSCLC- sin caracterización adicional, pacientes de muestras de control No Sanos (NHC) con enfermedad de pulmón no maligna: COPD; enfermedad pulmonar obstructiva crónica solamente, Nódulos Benignos (determinados por biopsia). Otros: diversos tipos de enfermedades de pulmón sin diagnosis de COPD definida), la puntuación de clasificación de cada paciente, el error estándar del medio, el diagnóstico, y la fase de desarrollo del cáncer, si lo hay.

Tabla XI

Puntuaciones de SVM de pacientes individuales mediante clasificador de NSCLC de 29 genes					
ID	**Clase**	**Puntuación**	**Error**	**DX**	**Fase**
NSCLC. 1519	NSCLC	1,77	0,21	AC	3A
NSCLC. 1138	NSCLC	1,65	0,07	LSCC	3B
NSCLC. 1471	NSCLC	1,64	0,32	NSCLC	3A
NSCLC. 1282	NSCLC	1,54	0,26	AC	3B
NSCLC. 1154	NSCLC	1,54	0,23	AC	3A
NSCLC. 1222	NSCLC	1,51	0,24	AC	1B
NSCLC. 1175	NSCLC	1,48	0,21	AC	1A
NSCLC. 1352	NSCLC	1,45	0,31	AC	1B
NSCLC. 1600	NSCLC	1,40	0,29	NSCLC	3B
NSCLC. 1647	NSCLC	1,39	0,23	LSCC	3B
NSCLC. 1280	NSCLC	1,38	0,30	LSCC	38
NSCLC. 1311	NSCLC	1,36	0,15	AC	1A

NSCLC. 1200	NSCLC	1,35	0,26	AC	3A
NSCLC. 1602	NSCLC	1,35	0,22	LSCC	1A
NSCLC. 1192	NSCLC	1,34	0,19	LSCC	1B
NSCLC. 1177	NSCLC	1,32	0,11	AC	1B
NSCLC. 1583	NSCLC	1,32	0,22	LSCC	3A
NSCLC. 1397	NSCLC	1,32	0,34	AC	1A
NSCLC. 1362	NSCLC	1,30	0,11	AC	3B
NSCLC. 1403	NSCLC	1,30	0,18	AC	3B
NSCLC. 1307	NSCLC	1,29	0,30	AC	1A
NSCLC. 1559	NSCLC	1,27	0,14	AC	3A
NSCLC. 1589	NSCLC	1,26	0,19	AC	28
NSCLC.1155	NSCLC	1,25	0,17	AC	3A
NSCLC. 1211	NSCLC	1,23	0,23	AC	1A
NSCLC. 1631	NSCLC	1,23	0,18	AC	2B
NSCLC. 1475	NSCLC	1,21	0,17	LSCC	1A
NSCLC. 1437	NSCLC	1,20	0,28	LSCC	3A
NSCLC. 1484	NSCLC	1,15	0,17	LSCC	3A
NSCLC. 1166	NSCLC	1,15	0,35	AC	1B
NSCLC. 1674	NSCLC	1,14	0,09	AC	3A
NSCLC. 1454	NSCLC	1,13	0,19	LSCC	28
NSCLC. 1361	NSCLC	1,12	0,28	AC	1B
NSCLC. 1569	NSCLC	1,11	0,21	NSCLC	3A
NSCLC. 1339	NSCLC	1,07	0,27	LSCC	28
NSCLC. 1264	NSCLC	1,06	0,29	LSCC	4
NSCLC. 1325	NSCLC	1,05	0,12	NSCLC	3B
NSCLC. 1632	NSCLC	1,05	0,15	AC	2A
NSCLC. 1473	NSCLC	1,03	0,30	AC	1B
NSCLC. 1402	NSCLC	1,02	0,24	AC	4
NSCLC. 1557	NSCLC	1,01	0,23	NSCLC	1B
NSCLC. 1183	NSCLC	0,98	0,25	AC	1A
NSCLC. 1455	NSCLC	0,97	0,16	LSCC	1A
NSCLC. 1194	NSCLC	0,97	0,17	AC	4
NSCLC. 1193	NSCLC	0,96	0,20	AC	1B
NSCLC. 1224	NSCLC	0,96	0,13	AC	2A
NSCLC. 1573	NSCLC	0,94	0,14	AC	3B
NSCLC. 1375	NSCLC	0,94	0,25	NSCLC	1A
NSCLC. 1214	NSCLC	0,93	0,32	LSCC	1B
NSCLC. 1630	NSCLC	0,92	0,22	NSCLC	3A
NSCLC.1343	NSCLC	0,92	0,20	AC	3A
NSCLC. 1561	NSCLC	0,91	0,21	LSCC	2A
NSCLC. 1435	NSCLC	0,89	0,25	AC	1A
NSCLC. 1221	NSCLC	0,88	0,32	AC	3A
NSCLC. 1449	NSCLC	0,87	0,14	LSCC	1A
NSCLC. 1413	NSCLC	0,85	0,21	LSCC	18
NSCLC. 1287	NSCLC	0,84	0,20	AC	1B
NSCLC. 1387	NSCLC	0,84	0,21	AC	3A
NSCLC. 1140	NSCLC	0,83	0,21	AC	38
NSCLC. 1598	NSCLC	0,83	0,31	AC	1A
NSCLC. 1415	NSCLC	0,78	0,20	AC	1A

NSCLC. 1369	NSCLC	0,77	0,21	AC	1B
NSCLC. 1591	NSCLC	0,75	0,10	AC	1A
NSCLC. 1469	NSCLC	0,75	0,25	AC	1A
NSCLC. 1141	NSCLC	0,75	0,23	AC	1B
NSCLC. 1340	NSCLC	0,74	0,37	AC	1A
NSCLC. 1178	NSCLC	0,73	0,13	LSCC	3B
NSCLC. 1604	NSCLC	0,73	0,21	AC	2B
NSCLC. 1429	NSCLC	0,70	0,15	LSCC	1A
NSCLC. 1681	NSCLC	0,67	0,26	NSCLC	3B
NSCLC. 1542	NSCLC	0,67	0,24	AC	1A
NSCLC. 1572	NSCLC	0,66	0,26	AC	1A
NSCLC. 1143	NSCLC	0,66	0,31	AC	1A
NSCLC. 1439	NSCLC	0,66	0,36	AC	38
NSCLC. 1189	NSCLC	0,61	0,27	LSCC	3A
NSCLC. 1189	NSCLC	0,61	0,27	LSCC	3A
NSCLC. 1312	NSCLC	0,61	0,27	AC	2B
NSCLC. 1323	NSCLC	0,61	0,32	AC	4
NSCLC. 1466	NSCLC	0,61	0,30	LSCC	2B
NSCLC. 1643	NSCLC	0,59	0,21	AC	3B
NSCLC. 1550	NSCLC	0,58	0,21	AC	2B
NSCLC. 1423	NSCLC	0,55	0,26	LSCC	1B
NSCLC. 1468	NSCLC	0,54	0,19	LSCC	1A
NSCLC. 1167	NSCLC	0,54	0,31	AC	1A
NSCLC. 1436	NSCLC	0,54	0,31	AC	1A
NSCLC. 1368	NSCLC	0,53	0,16	AC	1A
NSCLC. 1158	NSCLC	0,52	0,41	AC	1A
NSCLC. 1137	NSCLC	0,51	0,26	AC	2B
NSCLC. 1656	NSCLC	0,51	0,12	AC	3A
NSCLC. 1592	NSCLC	0,50	0,20	LSCC	1B
NSCLC. 1489	NSCLC	0,48	0,29	AC	2A
NSCLC. 1566	NSCLC	0,47	0,21	LSCC	3B
NSCLC. 1284	NSCLC	0,45	0,25	LSCC	1A
NSCLC. 1204	NSCLC	0,43	0,31	LSCC	1A
NSCLC. 1400	NSCLC	0,43	0,33	LSCC	1A
NSCLC. 1622	NSCLC	0,42	0,42	NSCLC	1A
NSCLC. 1482	NSCLC	0,42	0,19	LSCC	1A
NSCLC. 1390	NSCLC	0,41	0,11	LSCC	2B
NSCLC. 1597	NSCLC	0,39	0,11	AC	3A
NSCLC. 1388	NSCLC	0,36	0,27	NSCLC	3B
NSCLC. 1444	NSCLC	0,35	0,23	AC	3A
NSCLC. 1463	NSCLC	0,35	0,22	LSCC	1A
NSCLC. 1586	NSCLC	0,34	0,29	LSCC	1A
NSCLC. 1233	NSCLC	0,30	0,28	LSCC	2A
NSCLC. 1713	NSCLC	0,29	0,22	AC	3B
NSCLC. 1344	NSCLC	0,29	0,28	AC	1B
NSCLC. 1171	NSCLC	0,27	0,35	LSCC	1A
NSCLC. 1590	NSCLC	0,25	0,18	AC	3A
NSCLC. 1196	NSCLC	0,25	0,26	LSCC	2B
NSCLC. 1451	NSCLC	0,24	0,22	AC	1B

NSCLC. 1709	NSCLC	0,24	0,23	LSCC	3B
NSCLC. 1560	NSCLC	0,23	0,30	AC	3A
NSCLC. 1584	NSCLC	0,19	0,44	AC	1A
NSCLC. 1269	NSCLC	0,18	0,23	LSCC	1A
NSCLC. 1595	NSCLC	0,17	0,23	LSCC	1B
NSCLC. 1286	NSCLC	0,16	0,25	AC	1A
NSCLC. 1202	NSCLC	0,14	0,31	AC	1B
NSCLC. 1292	NSCLC	0,13	0,22	LSCC	1B
NSCLC. 1491	NSCLC	0,12	0,17	AC	1B
NSCLC. 1373	NSCLC	0,09	0,23	AC	1B
NSCLC. 1303	NSCLC	0,09	0,20	LSCC	1A
NSCLC. 1614	NSCLC	0,08	0,28	LSCC	1B
NSCLC. 1337	NSCLC	0,05	0,31	AC	1A
NSCLC. 1453	NSCLC	0,02	0,15	AC	4
NSCLC. 1227	NSCLC	0,01	0,32	AC	1A
NSCLC. 1216	NSCLC	-0,01	0,38	AC	1A
NSCLC. 1254	NSCLC	-0,09	0,30	LSCC	1A
NSCLC. 1136	NSCLC	-0,13	0,32	AC	1A
NSCLC. 1346	NSCLC	-0,15	0,21	AC	2A
NSCLC. 1445	NSCLC	-0,32	0,35	AC	2A
NSCLC. 1431	NSCLC	-0,34	0,29	AC	1A
NSCLC. 1582	NSCLC	-0,38	0,17	AC	1B
NSCLC. 1427	NSCLC	-0,43	0,24	AC	1A
NSCLC. 1430	NSCLC	-0,45	0,23	AC	1A
NSCLC. 1153	NSCLC	-0,51	0,27	AC	1A
NSCLC. 1262	NSCLC	-0,51	0,29	AC	1A
NSCLC. 1548	NSCLC	-0,61	0,31	AC	1B
NSCLC. 1386	NSCLC	-0,65	0,32	AC	5B
NHC. 1218	NHC	1,13	0,36	GI	0
NHC. 1588	NHC	0,96	0,31	GI	0
NHC. 1146	NHC	0,80	0,23	HAM	0
NHC. 10062	NHC	0,77	0,33	COPD	0
NHC. 1554	NHC	0,72	0,20	NM	0
NHC. 10027	NHC	0,60	0,30	COPD	0
NHC. 1474	NHC	0,59	0,19	NM	0
NHC. 1628	NHC	0,51	0,37	GI	0
NHC. 10010	NHC	0,48	0,29	HTN	0
NHC. 1263	NHC	0,48	0,21	NM	0
NHC. 1619	NHC	0,45	0,10	GI	0
NHC. 1361	NHC	0,42	0,27	NM	0
NHC. 1575	NHC	0,38	0,19	GI	0
NHC. 1522	NHC	0,21	0,12	GI	0
NHC. 1562	NHC	0,11	0,27	NM	0
NHC. 10047	NHC	0,11	0,31	COPD	0
NHC. 1424	NHC	0,04	0,21	GI	0
NHC. 10037	NHC	0,02	0,32	COPD	0
NHC. 10063	NHC	-0,01	0,22	COPD	0
NHC. 1677	NHC	-0,05	0,15	GI	0
NHC. 10044	NHC	-0,16	0,23	SARC	0

NHC. 1260	NHC	-0,16	0,25	NM	0
NHC. 1182	NHC	-0,23	0,38	PN	0
NHC. 10043	NHC	-0,25	0,31	COPD	0
NHC. 10064	NHC	-0,29	0,29	COPD	0
NHC. 1148	NHC	-0,30	0,35	GI	0
NHC. 1184	NHC	-0,30	0,26	NM	0
NHC. 1618	NHC	-0,33	0,20	GI	0
NHC. 10046	NHC	-0,33	0,15	COPD	0
NHC. 1657	NHC	-0,37	0,25	SARC	0
NHC. 10034	NHC	-0,44	0,24	COPD	0
NHC. 10036	NHC	-0,45	0,21	COPD	0
NHC. 10058	NHC	-0,47	0,23	COPD	0
NHC. 10054	NHC	-0,49	0,20	COPD	0
MHC. 10028	NHC	-0,50	0,14	COPD	0
NHC. 10004	NHC	-0,52	0,32	PS	0
NHC. 10040	NHC	-0,53	0,20	COPD	0
NHC. 1442	NHC	-0,56	0,32	NM	0
NHC. 1438	NHC	-0,61	0,25	NM	0
NHC. 10038	NHC	-0,63	0,20	COPD	0
NHC. 1488	NHC	-0,64	0,16	GI	0
NHC. 10042	NHC	-0,65	0,22	COPD	0
NHC. 1594	NHC	-0,66	0,17	GI	0
NHC. 1186	NHC	-0,66	0,36	NM	0
NHC. 1399	NHC	-0,66	0,29	GI	0
NHC. 1191	NHC	-0,68	0,27	NM	0
NHC. 10048	NHC	-0,69	0,30	COPD	0
NHC. 10061	NHC	-0,69	0,35	COPD	0
NHC. 10049	NHC	-0,70	0,28	COPD	0
NHC. 10055	NHC	-0,70	0,25	COPD	0
NHC. 10023	NHC	-0,74	0,17	CR	0
NHC. 1242	NHC	-0,74	0,27	NM	0
NHC. 10003	NHC	-0,77	0,34	HTN	0
NHC. 10039	NHC	-0,80	0,22	COPD	0
NHC. 1697	NHC	-0,84	0,14	GI	0
NHC. 1309	NHC	-0,86	0,25	NM	0
NHC. 1305	NHC	-0,92	0,19	GI	0
NHC. 1185	NHC	-0,93	0,21	NM	0
NHC. 1289	NHC	-0,94	0,28	NM	0
NHC. 1277	NHC	-0,94	0,27	NM	0
NHC. 10029	NHC	-0,95	0,21	COPD	0
NHC. 10053	NHC	-0,97	0,18	COPD	0
NHC. 1616	NHC	-1,00	0,11	NM	0
NHC. 10030	NHC	-1,03	0,25	SARC	0
NHC. 10019	NHC	-1,07	0,10	NHC	0
NHC. 10035	NHC	-1,07	0,14	COPD	0
NHC. 10051	NHC	-1,08	0,19	COPD	0
NHC. 10013	NHC	-1,08	0,28	COPD	0
NHC. 1251	NHC	-1,09	0,19	GI	0
NHC. 10008	NHC	-1,11	0,28	GI	0

NHC. 10018	NHC	-1,13	0,15	COPD	0
NHC. 10012	NHC	-1,21	0,21	COPD	0
NHC. 1342	NHC	-1,22	0,21	GI	0
NHC. 10052	NHC	-1,25	0,25	COPD	0
NHC. 10041	NHC	-1,27	0,18	COPD	0
NHC. 10031	NHC	-1,32	0,27	COPD	0
NHC. 1490	NHC	-1,34	0,15	NM	0
NHC. 1250	NHC	-1,37	0,26	NM	0
NHC. 10005	NHC	-1,40	0,13	CR	0
NHC. 1267	NHC	-1,43	0,12	NM	0
NHC. 10057	NHC	-1,52	0,27	COPD	0
NHC. 1450	NHC	-1,56	0,34	GI	0
NHC. 10001	NHC	-1,56	0,16	HTN	0
NHC. 10022	NHC	-1,57	0,20	COPD	0
NHC. 10059	NHC	-1,65	0,15	COPD	0
NHC. 1328	NHC	-1,65	0,14	NM	0
NHC.1314	NHC	-1,68	0,20	GI	0
NHC.10050	NHC	-1,82	0,19	COPD	0
NHC. 10033	NHC	-1,83	0,20	COPD	0
NHC. 10032	NHC	-1,89	0,15	COPD	0
NHC. 10056	NHC	-2,45	0,10	COPD	0

Ejemplo 15: Estudios de validación independientes sobre muestras de exclusión

Para direccionar cuestiones de sobre-ajuste de datos y para probar la generalidad del modelo de clasificación antes de aplicarlo a nuevas muestras, el análisis volvió a ser llevado a cabo, dejando a un lado el 20% de los pacientes y muestras de control, incluyendo representantes de cada una de las subclases para validación y ejercitamiento sobre el 80% restante. 5 conjuntos de exclusión separados y no solapantes fueron sometidos a esta validación. La precisión media sobre los 5 conjuntos de validación fue del 81% en comparación con una precisión media de un 82% para los 5 conjuntos de ejercitamiento (datos no representados). La precisión similar de los conjuntos de ejercitamiento y de validación demostró la capacidad del algoritmo para clasificar nuevas muestras con precisión pronosticada. La precisión ligeramente más baja con los conjuntos de exclusión en comparación con la validación cruzada utilizando todos los datos (81% frente a 86%) fue un reflejo del número más pequeño de muestras disponibles para ejercitamiento. Por el contrario, la precisión media del análisis con etiquetas de muestra permutadas fue sólo del 58% a través de 10 ejecuciones de permutación. Se concluyó que la signatura de 29 genes de la Tabla V puede distinguir pacientes con cualquiera de los dos tipos principales de NSCLC y cualquiera de las cuatro fases de desarrollo del tumor, a partir de pacientes con enfermedades de pulmón relacionadas con el tabaco pero no malignas.

Ejemplo 16: Precisión de la clasificación para subclases de paciente y de control utilizando 29 genes

La precisión del clasificador de 29 genes fue examinada para los diferentes tipos de pacientes y controles en el conjunto de datos. La Tabla XII que sigue relaciona las precisiones para los 29 genes en la identificación de las diversas clases de pacientes y controles así como las crecientes fases de desarrollo de tumor patológico. Las precisiones de clasificación individual para AC o LSCC solos fueron del 86% y 98% respectivamente en comparación con el 91% para los pacientes combinados. Había como mucho la mitad de LSCC en el conjunto de datos, pero estaban clasificados con una precisión significativamente más alta.

Las líneas 7-12 de la Tabla XII mostraron un aumento incremental en la precisión de la clasificación desde la fase de desarrollo 1A (83%) hasta las fases de desarrollo 3 y 4 (100%), lo que supone que la signatura de cáncer de PBMC se vuelve más pronunciada con la enfermedad progresiva. Si solamente se consideraron los controles con COPD y sin ninguna evidencia de nódulos de pulmón, éstos se clasificaron con una precisión del 89%, mientras que los pacientes con nódulos benignos confirmados (con independencia del estado de COPD) tuvieron una precisión de clasificación del 71%. De ese modo, la precisión de clasificación estuvo influenciada por la fase de desarrollo del cáncer.

Tabla XII

Rendimiento del clasificador de 29 genes sobre subclases de pacientes y controles			
#	**Subclase**	**Precisión por clase**	**Número de muestras**
1	NSCLC	91%	137
2	NHC	80%	91
3	AC	86%	85
4	LSCC	98%	42
5	Nódulos	71%	41
6	COPD	89%	38
7	Fase 1A	83%	48
8	Fase 1B	89%	27
9	Fase 1	85%	75
10	Fase 2	89%	18
11	Fase 3	100%	39
12	Fase 4	100%	5

Aunque 29 genes fueron suficientes para distinguir las clases de paciente y de control, muchos genes estadísticamente más significativos fueron expresados diferencialmente (véase la Tabla V). Las funciones moleculares más altamente representadas incluían regulación de expresión de gen, muerte celular y crecimiento celular y diferenciación. Los genes asociados a la generación de células T de memoria, acumulación de célula T y movilización de células NK estaban en su mayoría en el cáncer, mientras que las vías de señalización de receptor de célula B eran descendentes. Los genes asociados a la activación o la quimiotaxia de células de mieloide y de genes de señalización de receptor gluco-corticoide, eran en su mayoría descendentes en los pacientes de cáncer.

La aplicación clínica de la signatura de expresión genética de PBMC es clara. Suponiendo una prevalencia de cáncer de pulmón del 5% para pacientes con un nódulo de pulmón entre 0,5 y 3,0 cm, el clasificador de 29 genes (con un valor de punto de corte de cero), se entiende que consigue un valor predictivo positivo (PPV) y negativo (NPV) de 0,19 y 0,99, respectivamente, según se muestra en la Tabla XIII que sigue. Estos valores exceden los establecidos por el Grupo Biomarcador de Cáncer de Pulmón EDRN que determina si un biomarcador debe ser considerado útil para un estudio adicional. Éstos son similares a los valores para el panel de expresión de 80 genes procedentes de los raspados bronquiales recientemente descritos[18]. De forma importante, se podría conseguir incluso una utilidad clínica más alta en muchos pacientes sacando ventaja del valor real de la puntuación predictiva en vez de usar un punto de corte de puntuación estricta positiva o negativa. En el gran conjunto de datos mostrado en la Tabla XI anterior, ningún sujeto con una puntuación SMV menor de -0,65 tenía cáncer de pulmón y solamente de 5 a 91 pacientes de control de no cáncer que tuvieron una puntuación SMV >+0,65 fueron clasificados como cáncer de pulmón. De ese modo, el valor real de la puntuación SVM es útil para determinar qué pacientes requieren una intervención invasiva en oposición a una alternativa más conservadora, tal como una serie de imágenes CT.

Tabla XIII

Valor predictivo positivo y valor predictivo negativo para clasificador de NSCLC de 29 genes					
Estudio	**Sensibilidad**	**Especificidad**	**Prevalencia**	**PPV**	**NPV**
NSCLC vs. NHC ---------- clasificador 29 genes	0,91	0,8	1% ---------- 5%	0,044 ---------- 0,193	0,999 ---------- 0,004
Spira et al., 2007 ---------- clasificador 80 genes	0,8	0,84	1% ---------- 5%	0,048 ---------- 0,208	0,998 ---------- 0.998
LCBG ---------- Biomarcador propuesto	0,8	0,7	1% ---------- 5%	0,026 ---------- 0,123	0,997 ---------- 0,985

Ejemplo 17: Clasificación de muestras de paciente y de control desde un sitio independiente

Todas las muestras utilizadas para desarrollar y validar el panel de 29 genes fueron recogidas en el Hospital de la Universidad de Pennsylvania. Para una validación adicional de la utilidad del clasificador, se analizaron 27 muestras recogidas en el Centro de Biomarcadores de Cáncer de Pulmón NYU, una Clínica de Red de Investigación de Detección Temprana (EDRN) y en el Centro de Validación Epidemiológica. Las 27 muestras incluían 12 NSCLC de Fase 1 (5 de los cuales nunca habías sido fumadores), y 15 controles de fumadores y ex-fumadores, incluyendo 6 controles diagnosticados mediante exploraciones de serie CT como que tenían Opacidades de Vidrio Deslustrado (GGO)[21] no malignas. Ninguna de las muestras de GGO fue incluida en nuestro conjunto de ejercitamiento original.

A pesar de las diferencias en cuanto a los sitios de recogida, al procesamiento de muestra y a la distinta población de control, las 27 muestras fueron clasificadas con una precisión global de un 74% (20 de 27), una sensibilidad del 67% (8 de 12) y una especificidad del 80% (12 de 15). La clasificación SMV se muestra con detalle en la Tabla XIV que sigue.

Tabla XIV

Puntuaciones de clasificación SMV mediante clasificador de NSCLC para muestras de validación NYU				
ID	**Clase**	**Puntuación**	**Error**	**Dx**
NYU. 1	NSCLC	1,07	0,06	AC
NYU. 2	NSCLC	1,01	0,07	AC
NYU. 3	NSCLC	0,95	0,06	AC
NYU. 4	NSCLC	0,81	0,07	AC
NYU. 5	NSCLC	0,71	0,08	AC
NYU. 6	NSCLC	0,48	0,06	AC
NYU. 7	NSCLC	0,29	0,08	AC
NYU. 8	NSCLC	0,18	0,09	AC
NYU. 9	NSCLC	-0,25	0,09	AC
NYU. 10	NSCLC	-0,29	0,10	AC
NYU. 11	NSCLC	-0,37	0,10	AC
NYU. 12	NSCLC	-0,94	0,08	AC
NYU. 13	NHC	1,16	0,10	GGO
NYU. 14	NHC	0,70	0,11	N
NYU. 15	NHC	0,69	0,10	GGO
NYU. 16	NHC	-0,12	0,08	N
NYU. 17	NHC	-0,13	0,09	GGO
NYU. 18	NHC	-0,26	0,08	N
NYU. 19	NHC	-0,39	0,09	GGO
NYU. 20	NHC	-0,39	0,08	N
NYU. 21	NHC	-0,46	0,10	N
NYU. 22	NHC	-0,52	0,10	N
NYU. 23	NHC	-0,58	0,07	N
NYU. 24	NHC	-0,73	0,09	N
NYU. 25	NHC	-0,75	0,10	N
NYU. 26	NHC	-0,84	0,09	GGO
NYU. 27	NHC	-0,94	0,08	GGO
Abreviaciones Dx: AC = adenocarcinoma, N = normal, GGO = Opacidades vidrio deslustrado				

Dos de los pacientes mal clasificados nunca fueron fumadores y 2 de los controles fueron GGOs. La reducida precisión en el conjunto de validación externa resultó más probable debido a las diferencias en el procesamiento de las muestras (datos no representados).

Ejemplo 18: Clasificación de 29 genes de muestras independientes antes y después de la extracción del tumor

El clasificador de 29 genes fue probado sobre un conjunto independiente de 26 muestras procedentes de 18 pacientes de NSCLC que incluían tanto muestras de pre- como de post- resección. En primer lugar, como validación adicional, cuando se usa este clasificador, catorce de las 18 muestras pre-cirugía fueron clasificadas correctamente como cáncer, para una sensibilidad del 78%. En segundo lugar, las puntuaciones SVM para 13 de las 14 (92%) mostraron reducciones significativas en la puntuación de clasificación tras la resección quirúrgica. Siete de las muestras post-resección tuvieron puntuaciones SMV que eran negativas y fueron clasificadas como no cáncer en este análisis (datos no representados). No existió correlación obvia entre el cambio en las puntuaciones de SVM y el tiempo de recogida de PBMC de post-resección, aunque el conjunto de datos es relativamente pequeño.

Los perfiles de expresión genética cambian en la PBMC después de la extracción del tumor, según se demuestra a continuación. El análisis mostrado en la Figura 5 de las muestras pre/post emparejadas, fue realizado para determinar si el clasificador de 29 genes desarrollado sobre pacientes con enfermedad maligna frente a no maligna podía detectar una diferencia en la expresión de gen tras la extracción del tumor. Dada la observación de que esto fue cierto para la mayoría de las muestras, se examinó la magnitud de las diferencias en las clases de muestra. Los pares de muestra fueron comparados directamente para evaluar mejor los cambios en la expresión de gen que pudieran resultar de la extracción del tumor. Se encontró un efecto significativo sobre la expresión de gen de PBMC; se encontró que 2060 genes estaban expresados de manera diferente a través de los pares (prueba-t de dos colas emparejadas, $p<0,05$ con una tasa de descubrimiento falsa del 28%).

Se generó un clasificador SVM separado para los pacientes de pre- y post- cirugía y los 50 genes que forman ese conjunto clasificador fueron informados en la Tabla VI anterior. Un clasificador seleccionado a partir de los genes de Tabla VI fue capaz de separar perfectamente las dos clases con sólo cuatro genes. Los cuatro genes de jerarquía superior de este clasificador incluyen CYP2R1 (una hidroxilasa de vitamina D microsómica), MYO5B (3-oxoacil-Coenzima A tiolasa mitocondrial), DGUOK (Desoxiguanosina Kinasa Mitocondrial), todas ellas post-cirugía reguladas descendentemente, y DNCL1 (Doineína, citoplásmica, cadena ligera 1) que está regulada ascendentemente tras la cirugía. Dos (CYP2R1 y DGUOK) de los 4 genes fueron también validados mediante PCR Cuantitativa en Tiempo Real sobre 10 pares de muestras. Los resultados se han indicado en la Figura 6 y en la Tabla XV que sigue.

Tabla XV

Relaciones de expresión PRE/POST cirugía de PBMC para 10 pacientes según se determinó mediante matrices de expresión de gen Illumina y análisis QPCR				
Paciente		**CYP2R1**		**DGUOK**
	Matrices Illumina	***PCR***	***Matrices Illumina***	***PCR***
4	1,13	1,33	1.33	1,21
5	1,55	1,28	1,12	1,01
6	1,49	1,73	1,21	1,41
7	1,33	1,06	1,12	1,17
11	1,44	1,58	1,38	1,09
14	1,37	1,29	1,30	1,25
15	1,15	2,65	1,14	2,14
16	1,42	0,96	1,19	0,76
17	1,60	1,57	1,21	1,56
18	1,10	1,09	1,31	1,15
PROMEDIO	1,36	1,45	1,23	1,27

Ejemplo 19: Signatura de expresión de gen para diferenciación de pacientes con nódulos de pulmón benignos

Puesto que los pacientes con diagnosis de un nódulo benigno son la clase de control más importante para la diferenciación, se desarrolló un clasificador separado utilizando solamente los controles con nódulos benignos y se evaluó su precisión. Utilizando los 41 controles con nódulos y un grupo seleccionado aleatoriamente de 54 muestras de NSCLC, se aplicó SVM-RFE con validación cruzada, según se ha descrito con anterioridad. El clasificador resultante (Tabla VII, genes 1-24) tuvo una precisión de un 79%, con una especificidad de un 80% para los nódulos y sólo requirió 24 genes, 7 de los cuales estaban incluidos en el panel de 29 genes. La Tabla VII relaciona el rango del gen "RANGO" en NSCLC frente al clasificador NHC, el ID de Mancha Illumina "ID", el Núm. de Adhesión "Núm. Acc.", la descripción del gen, su símbolo, el NSCLC frente al valor p de GI.NM "valor-p", y el cambio de pliegue de NSCLC/GI.NM "Fold Chg".

VI. REFERENCIAS

1. Yousef, M., et al., 2007 BMC Bioinformatics, 8; p. 144.
2. Jemal, A., et al., 2006 J Clin 56(2): p. 106-30.
3. Marcus, P.M., et al., 2000 J Natl Cancer Inst., 92(16): p. 1308-16.
4. Palmisano, W.A., et al., 2000 Cancer Res, 60(21): p. 5954-8.
5. Patz, E.F., Jr., et al., 2000 N Engl J Med. 343(22): p. 1627-33.
6. Hirsch, F.R., et al., 2001 Clin Cancer Res. 7(1): p. 5-22.
7. Burczynski M.E., et al., 2005 Clin Cancer Res., 11 (1181-9).
8. Burczynski, M.E., et al., 2005 Curr Mol Med, 5(1): p. 83-102.
9. Chang, H.Y., et al., 2002 Proc Natl Acad Sci U S A, 99(20): p. 12877-82.
10. Borczuk, A.C., et al., 2003 Am J Pathol, 163(5): p. 1949-60.
11. Gao, C., et al., 2005 Nitric Oxide, 12(2): p. 121-6.
12. Mulshine, J.L., 2005 Oncology (Williston Park), 19(13): p. 1724-30; disc. 30-1.
13. Haiman, C.A. et al., 2006 N Engl J Med, 354(4): p. 333-42.
14. Diederich, S. y D. Wormanns, 2004 Lung Cancer 45 Suppl 2: p. S13-9.
15. Jett, J.R., 2005 Clin Cancer Res, 11(13 Pt 2): p. 4988s-4992s.
16. Deppermann, K.M., 2004 Lung Cancer, 45 Suppl. 2: p. S39-42.
17. MacMahon, H., et al., 2005 Radiology, 237(2): p. 395-400.
18. Berger, M., et al., 2003 AJR Am J Roentgenol, 2003, 181(2): p. 359-65.
19. Mulshine, J.L., 2005 Clin Cancer Res, 11(13 Pt 2): p. 4993s-4998s.
20. Bhattacharjee, A., et al., 2001 Proc. Natl. Acad. Sci., USA, 98: 13790-13795.
21. Burczynski, M.E. y A.J. Comer, 2006 Pharmacogenomics, 7(2): p. 187-202.
24. Deng MC, et al., 2006 Am J Transplant., 6: p. 150-160.
25. Achiron, A., et al., 2005 Breast Cancer Res Treat, 89(3): p. 265-70.
26. Achiron, A. y M. Gurevich, 2006 Autoimmun Rev, 5(8): p. 517-22.
27. Goronzy, J.J., et al., 2004 Arthritis Rheum, 2004. 50(1): p. 43-54.
28. Bull TM, et al., 2006 Am J Respir Crit Care Med., 4(170): p. 911-919.
29. Achiron, A., et al., 2007 Ann N Y Acad Sci, 1107: p. 155-67.
30. Sharp, F.R., et al., 2006 Arch Neurol, 63(11): p. 1529-1536.
31. Forrest, M.S., et al., 2005 Environ Health Perspect, 113(6): p. 801-7.
32. Theodoro, T.R., et al., 2007 Neoplasia, 9(6): p. 504-10.
33. Karimi, k., et al., 2006 Respir Res, 7: p. 66.
34. van Leuwen, D.M., et al., 2007 Carcinogenesis, 28(3): p. 691-7.
35. Oudijk, E. J:, et al., 2005 Thorax, 60(7): p. 538-44.
36. Lampe, J.W., et al., 2004 Cancer Epidemiol Biomarkers Prev, 13(3): p. 445-53.
37. Spira, A., et al., 2004 Proc Natl Acad Sci U S A, 101(27): p. 10143-8.
38. Russo, A.L., et al., 2005 Clin Cancer Res, 11(7): p. 2466-70.
39. Kari, L., et al., 2003 J Exp Med, 197(11): p. 1477-88.
40. Talmadge, J.E., et al., 1996 Bone Marrow Transplant, 17(1): p. 101-9.
41. Redente, E.F., et al., 2007 Am J Pathol, 170(2): p. 693-708.
42. Twine, N., et al., 2003 Cancer Res., 6: p. 6069-75.
43. Sharma, P., et al., 2005 Breast Cancer Res, 7: p. 634-44.
44. DePrimo, S.E., et al., 2003 BMC Cancer, e: p. http://www.biomedcentral.com/1471-2407/3/3.
45. Eady, J.J., et al., 2005 Physiol Genomics, 22(3): p. 402-11.
46. Whitney, A.R., et al., 2003 Proc Natl Acad Sci U S A, 100(4): p. 1896-901.
47. Loboda, A., et al., 2003 Proc. Eur. Conf. On Computational Biology, GE-10: p. p383-84.
48. Guyon, I., et al., 2002 Machine Learning, 46(1-3): p. 389-422.
49. Critchley-Thome, R.J., et al., 2007 PLoS Med, 4(5): p. 3176.
50. Vachani, A., et al., 2007 Clin. Canc. Res., 13(10): p. 2905-2915.

51. Spira, A., et al., 2007 Nat Med, 13(3): p. 361-6.

52. Mukherjee, S., et al., 2003 J Comp Biol, 10(2): p. 119-42.

53. Wang, J. et al., 2007 Bioinformatics, 23(15): p. 2024-7.

54. Vapnik, V., 1999. The Nature of Statistical Learning Theory, Springer-Verlag, 1999, ISBN 0-387-98780-0.

55. Nebozhyn, M., et al., 2006 Blood, 107(8): p. 3189-96.

56. Marron, J. y M. Todd (2003) Distance Weighted Discrimination School of Operations Research and Industrial Engineering, Cornell University.

57. Virok, D., et al., 2003 J Infect Dis, 188(9): p. 1310-21.

58. Pepe, M.S., et al., 2003 Biometrics, 59(1): p. 133-42.

59. DeLong, E.R., et al., 1988 Biometrics, 44(3): p. 837-45.

60. Harrell, F.E., Jr., et al., WHO/ARI Young Infant Multicentre Study Group. Stat Med, 1998. 17(8): p. 909-44.

61. Benito, M., et al., 2004 Bioinformatics, 20(1): 105-114.

62. Chung, GT., et al., 1995 Oncogene, 11: 2591-2598.

63. Hirano, T., et al., 1994 Am J. Pathol., 144: 296-302.

64. Kishimoto, Y., et al., J Natl Cancer Inst, 1995 87: 1224-1229.

65. Tibshirani, R. et al., Proc Natl Acad Sci USA, 2002 99: 6567-6572.

66. Tonon, G, et al., Proc Natl Acad Sci, 2005 102: 9625-9630.

67. MacQueen, J. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press; 1967. Algunos métodos para la clasificación y análisis de observaciones multivariadas; pp. 281-297.

68. Talbot, SG., et al., Cancer Res. 2005; 65: 3063-3071.

69. Ausubel et al., Current Protocols in Molecular Biology, Wiley Interscience Publishers, (1995).

70. Sambrook, et al., Clonación Molecular: Un Manual de Laboratorio, New York: Cold Spring Harbor Press, 1989.

71. B. Lewin, Genes IV Cell Press, Cambridge Mass. 1990.

72. Singleton et al., Diccionario de Microbiología y Biología Molecular, 2ª edic., J. Wiley & Sons (New York, N.Y. 1994).

73. March, Reacciones de Química Orgánica Avanzada, Mecanismo y Estructura, 4ª edic., John Wiley & Sons (New York, N.Y. 1992).

74. Parker & Barnes, 1999 Métodos en Biología Molecular 106: 247-283.

75. Hod, 1992 Biotécnica 13: 852-854.

76. Weis et al., 1992, Tendencias en Genética 8: 263-264.

77. Ausubel, et al., Protocolos Corrientes de Biología Molecular, John Wiley e Hijos (1997).

78. Rupp y Locker, 1987 Lab Invest. 56: A67.

79. De Andrés et al., 1995 Bio Técnicas 18: 42044.

80. T.E. Godfrey et al., 2000 J Molec. Diagnostics 2: 84-91

81. K. Specht et al., 2001 Am. J. Pathol. 158: 419-29.

82. Ding y Cantor, 2003 Proc. Natl. Acad. Sci. USA 100: 3059-3064.

83. Patente US núm. 7.081.340.

84. Publicación de solicitud de Patente Internacional núm. WO 2004/105573, publicada el 9 de Diciembre de 2004

85. Krawetz S., Misener S (eds) Métodos y Protocolos Bioinformáticos: Métodos en Biología Molecular. Humana press, Totowa, N.J., pp. 365-386.

86. Dieffenbach, C.W. et al., “Conceptos Generales para Diseño de Imprimador de PCR” en: Imprimador de PCR: Un Manual de Laboratorio, Cold Spring Harbor Laboratory Press, New York, 1995, pp. 133-155.

87. Innis y Gelfand, “Optimización de PCRs” en “Protocolos de PCR, Una Guía para Métodos y Aplicaciones, CRC Press, London, 1994, pp. 5-11.

88. Plasterer, T.N. 1997 Methods Mol. Biol. 70: 520-527.

89. Golub TR, et al., 1999 Science. 286: 531-537.

90. ACS. Cancer Facts and Figures 2007. Atlanta: American Cancer Society, 2008.

91. Amos CI, et al., 2008 Nat Genet, 40: 616-22.

92. Kang JU, et al., 2008 Cancer Genet Cytogenet, 184: 31-7.

93. Thorgeirsson TE, et al., 2008 Nature, 452: 638-42.

94. Henschke CI, et al., 2006 N Engl J Med, 355: 1763-71.

95. Bach PB, 2007 JAMA, 297: 953-61.
96. Ikeda K, et al., 2007 Chest, 132: 984-90.
97. Machida EO, et al., 2006 Cancer Res, 66: 6210-8.
98. Patz EF, Jr., et al., 2007 J Clin Oncol, 25: 5578-83.
99. Yanagisawa, K, et al., 2003 Lancet, 362: 433-9.
100. Brichory FM, et al., 2001 Proc Natl Acad Sci U S A, 98: 9824-9.
101. Pontes ER, et al., 2006 Próstata, 66: 1463-73
102. Belinsky SA, et al., 2006 Cancer Res, 66: 3338-44.
103. Ohta, Y, et al., 2006 Ann Thorac Surg, 81: 1194-7.
104. Osman I, et al., 2006 Clin Cancer Res, 12: 3374-80.
105. Subramanian J, Govindan R. 20071 J Clin Oncol, 25: 561-70.
106. Sun S, et al., 2007 Nat Rev Cancer, 7: 778-90.
107. Hung RJ, et al., 2008 Nature, 452: 633-7.
108. Mashima T, Tsuruo T. 2005 Drug Resist Updat, 8: 339-43.
109. Ozoren N, El-Deiry WS. 2003 Semin Cancer Ciol, 13: 135-47.
110. Held et al., Genome Research 6: 986-994 (1996).
111. US 2007 026424 A

REIVINDICACIONES

1.- Una composición para evaluar la existencia de un cáncer de pulmón en un sujeto mamífero, consistiendo dicha composición en: tres o más polinucleótidos u oligonucleóticos, en la que cada polinucleótido u oligonucleótido hibridiza un gen, fragmento de gen o transcripción de gen o producto de expresión diferente a partir de células mononucleares de sangre periférica (PBMC) o de sangre total del mamífero, y

en la que cada uno de dichos gen, fragmento de gen, transcripción de gen o producto de expresión se elige a partir de los genes 1-29 de la Tabla V.

2.- La composición de acuerdo con la reivindicación 1, la cual es un reactivo que comprende un substrato sobre el que se inmovilizan dichos polinucleótidos u oligonucleótidos.

3.- La composición de acuerdo con la reivindicación 1, que comprende una micro-matriz, una tarjeta micro-fluídica, un chip o una cámara.

4.- La composición de acuerdo con la reivindicación 1, la cual es un kit que contiene dichos tres o más polinucleótidos u oligonucleótidos, en la que dichos polinucleótidos u oligonucleótidos son cada parte de un conjunto de imprimador-sonda, y dicho kit comprende tanto imprimadores como sondas, en la que conjunto de imprimador-sonda citado amplifica un gen, fragmento de gen o producto de expresión de gen diferente.

5.- La composición de acuerdo con la reivindicación 1, en la que uno o más polinucleótidos u oligonucleótidos está(n) asociado(s) a una etiqueta detectable.

6.- La composición de acuerdo con la reivindicación 1, en la que dichos genes seleccionados comprenden 4 o más genes.

7.- La composición de acuerdo con la reivindicación 1, en la que dichos genes seleccionados comprenden 15 o más genes.

8.- La composición de acuerdo con la reivindicación 1, en la que dichos genes seleccionados comprenden 20 a 29 genes.

9.- La composición de acuerdo con la reivindicación 1, en la que dichos genes seleccionados consisten en los genes 1 a 29 de la Tabla V.

10.- Uso de una composición de una cualquiera de las reivindicaciones 1 a 9 para el diagnóstico *in vitro* de la existencia de un cáncer de pulmón, que comprende identificar cambios en la expresión de tres o más genes procedentes de células mononucleares de sangre periférica (PBMC) o de la sangre total de un sujeto, en el que los cambios de expresión de los genes del sujeto con respecto a una referencia están correlacionados con un diagnóstico de un cáncer de pulmón.

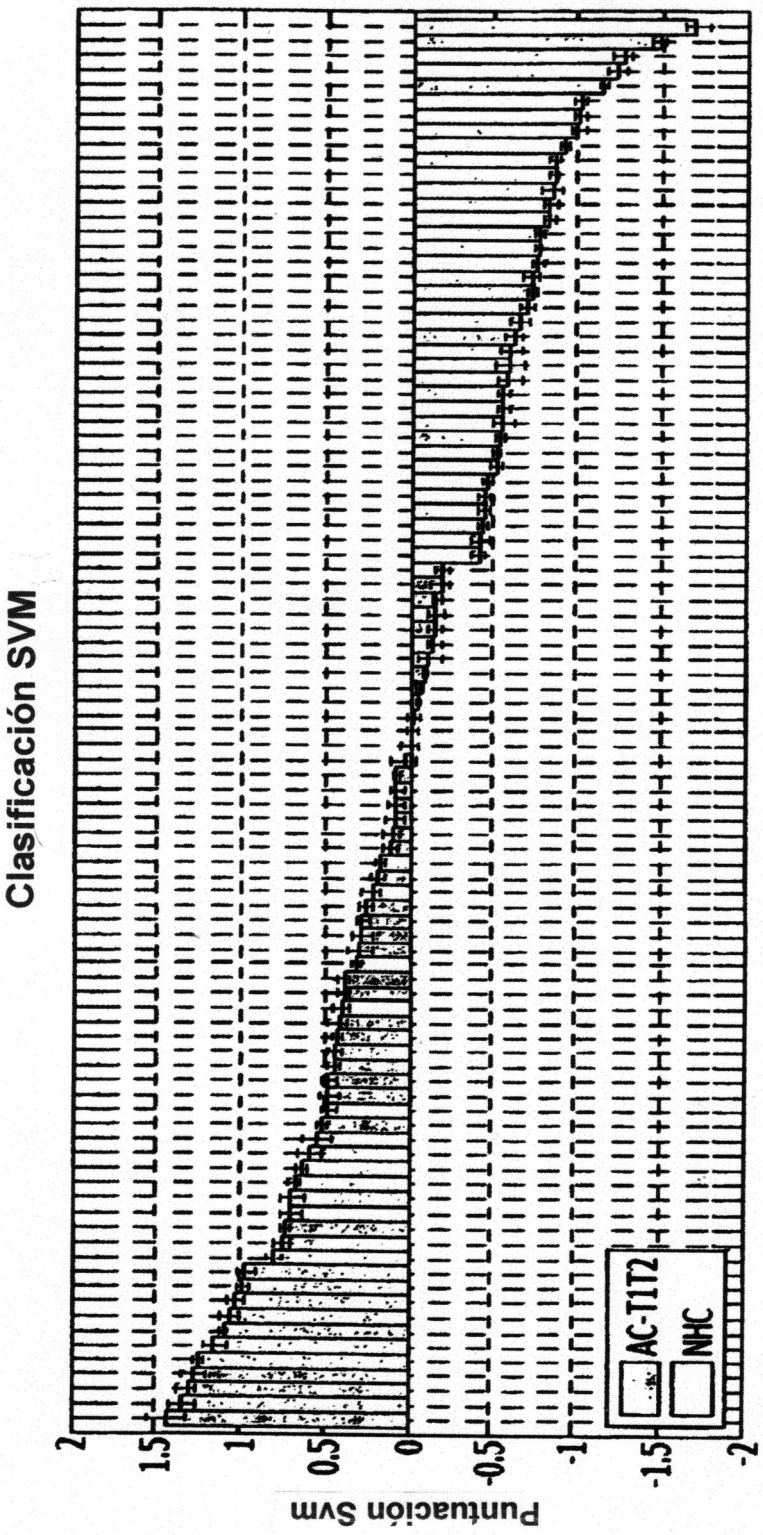
Clasificación SVM
Puntuación Svm
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
AC-T1T2
NHC
Muestras

FIG. 1

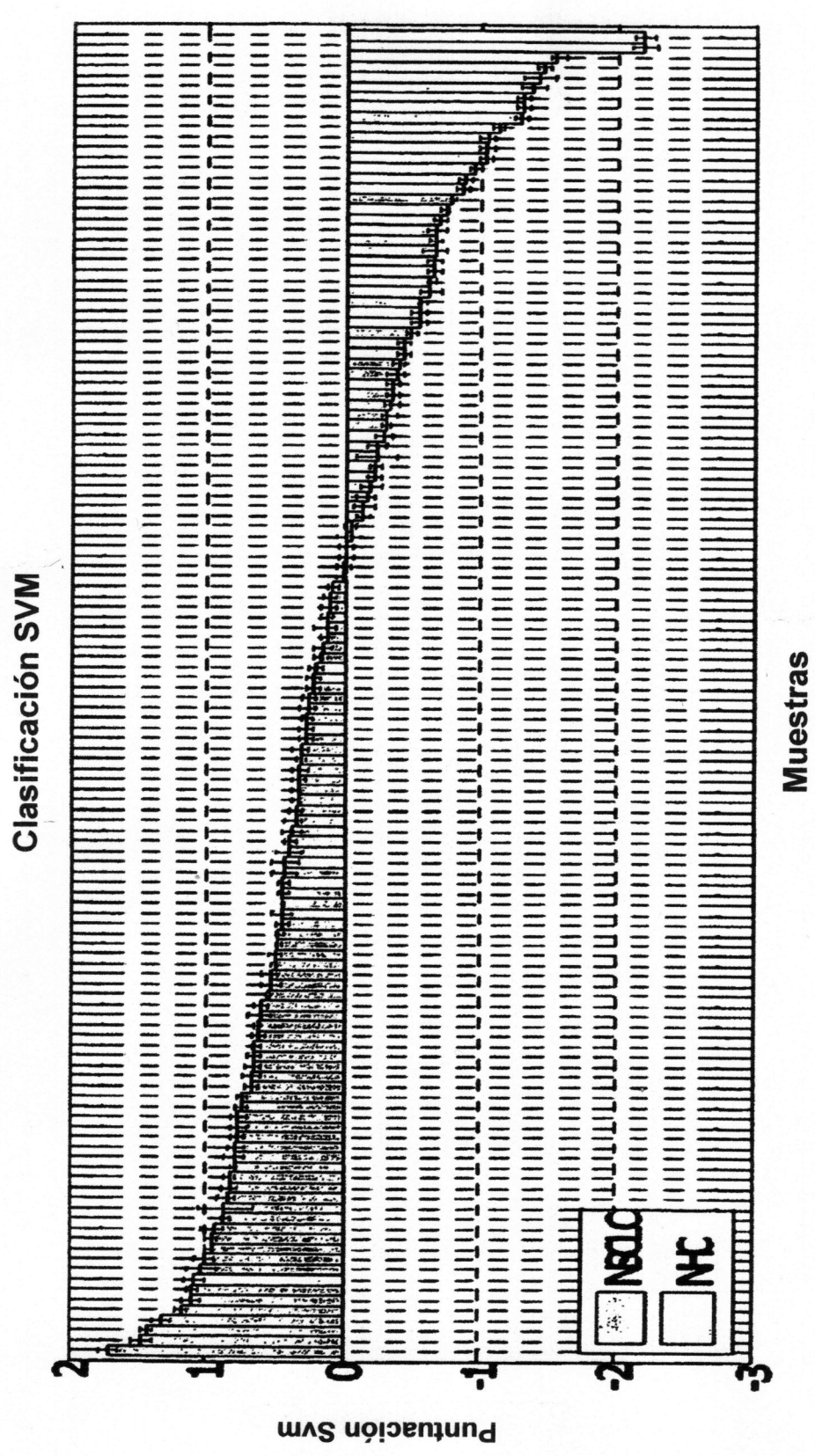
Clasificación SVM
Puntuación Svm
2
1
0
-1
-2
-3
NSCLC
NHC
Muestras

FIG. 2

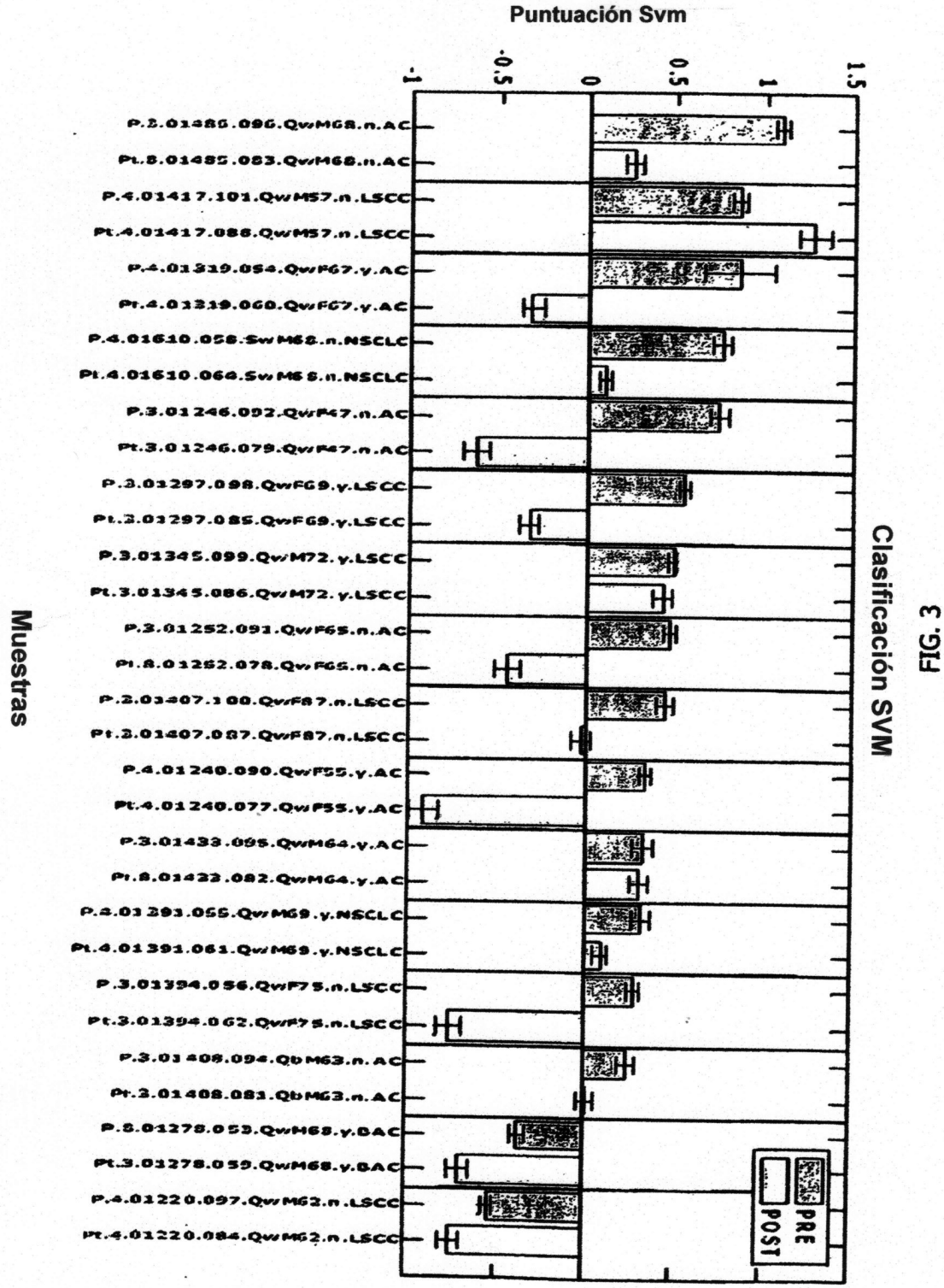
Puntuación Svm
-1
-0.5
0
0.5
1
1.5
P.2.01486.096.QwM68.n.AC
Pt.8.01485.083.QwM68.n.AC
P.4.01417.101.QwM57.n.LSCC
Pt.4.01417.086.QwM57.n.LSCC
P.4.01319.054.QwF67.y.AC
Pt.4.01319.060.QwF67.y.AC
P.4.01610.058.SwM68.n.NSCLC
Pt.4.01610.064.SwM68.n.NSCLC
P.3.01246.092.QwP47.n.AC
Pt.3.01246.079.QwP47.n.AC
P.3.01297.098.QwFG9.y.LSCC
Pt.3.01297.085.QwFG9.y.LSCC
P.3.01345.099.QwM72.y.LSCC
Pt.3.01345.086.QwM72.y.LSCC
P.3.01252.091.QwF65.n.AC
Pt.8.01252.078.QwF65.n.AC
P.2.01407.100.QwF87.n.LSCC
Pt.3.01407.087.QwF87.n.LSCC
P.4.01240.090.QwF55.y.AC
Pt.4.01240.077.QwF55.y.AC
P.3.01433.095.QwM64.y.AC
Pt.8.01433.082.QwM64.y.AC
P.2.01391.055.QwM69.y.NSCLC
Pt.4.01391.061.QwM69.y.NSCLC
P.3.01394.056.QwF75.n.LSCC
Pt.3.01394.062.QwF75.n.LSCC
P.3.01408.094.QbM63.n.AC
Pt.3.01408.081.QbM63.n.AC
P.8.01278.053.QwM68.y.BAC
Pt.3.01278.059.QwM68.y.BAC
P.4.01220.097.QwM62.n.LSCC
Pt.4.01220.084.QwM62.n.LSCC
PRE
POST
Muestras
Clasificación SVM

FIG. 3

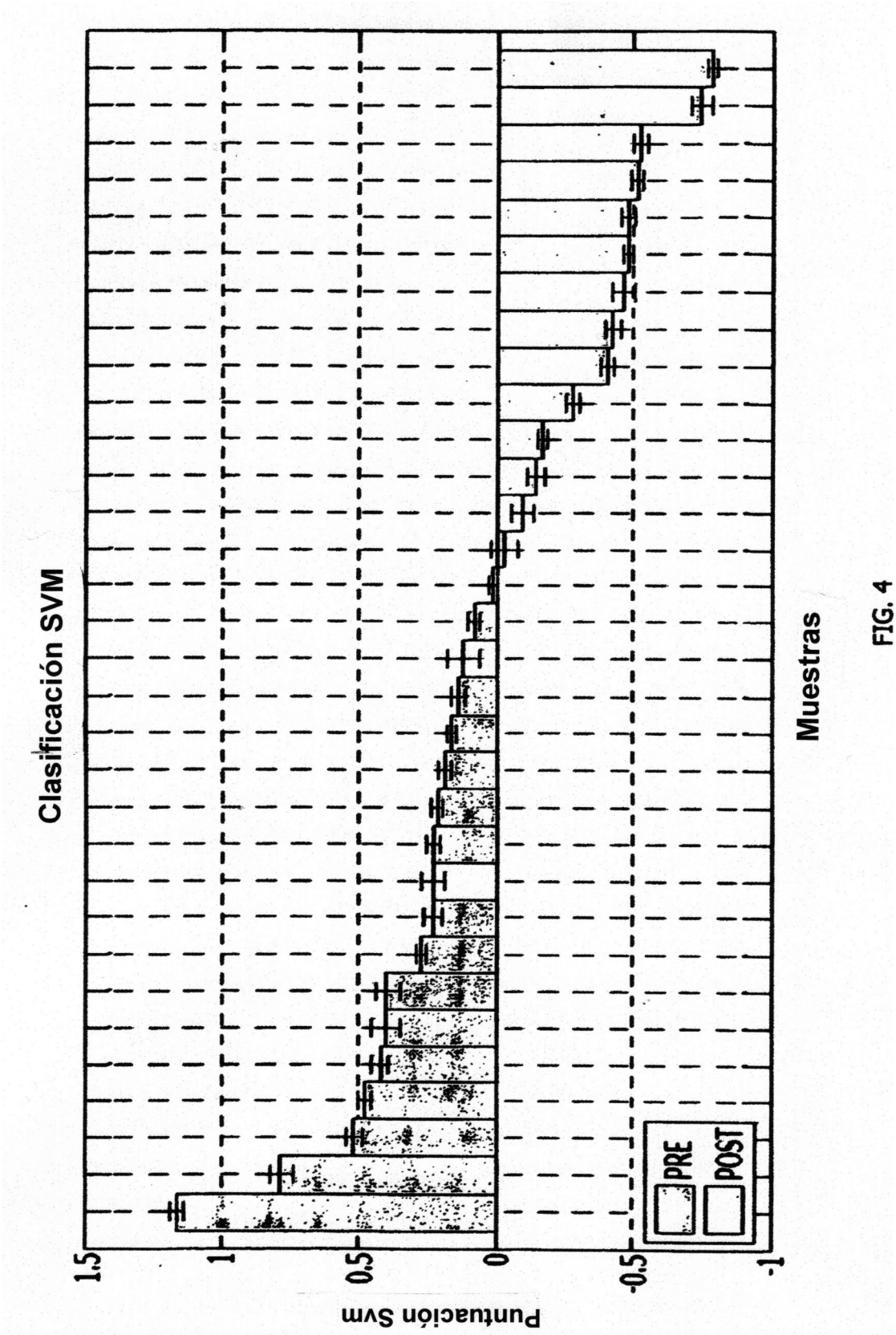
Clasificación SVM
Puntuación Svm
1.5
1
0.5
0
-0.5
-1
PRE
POST
Muestras

FIG. 4

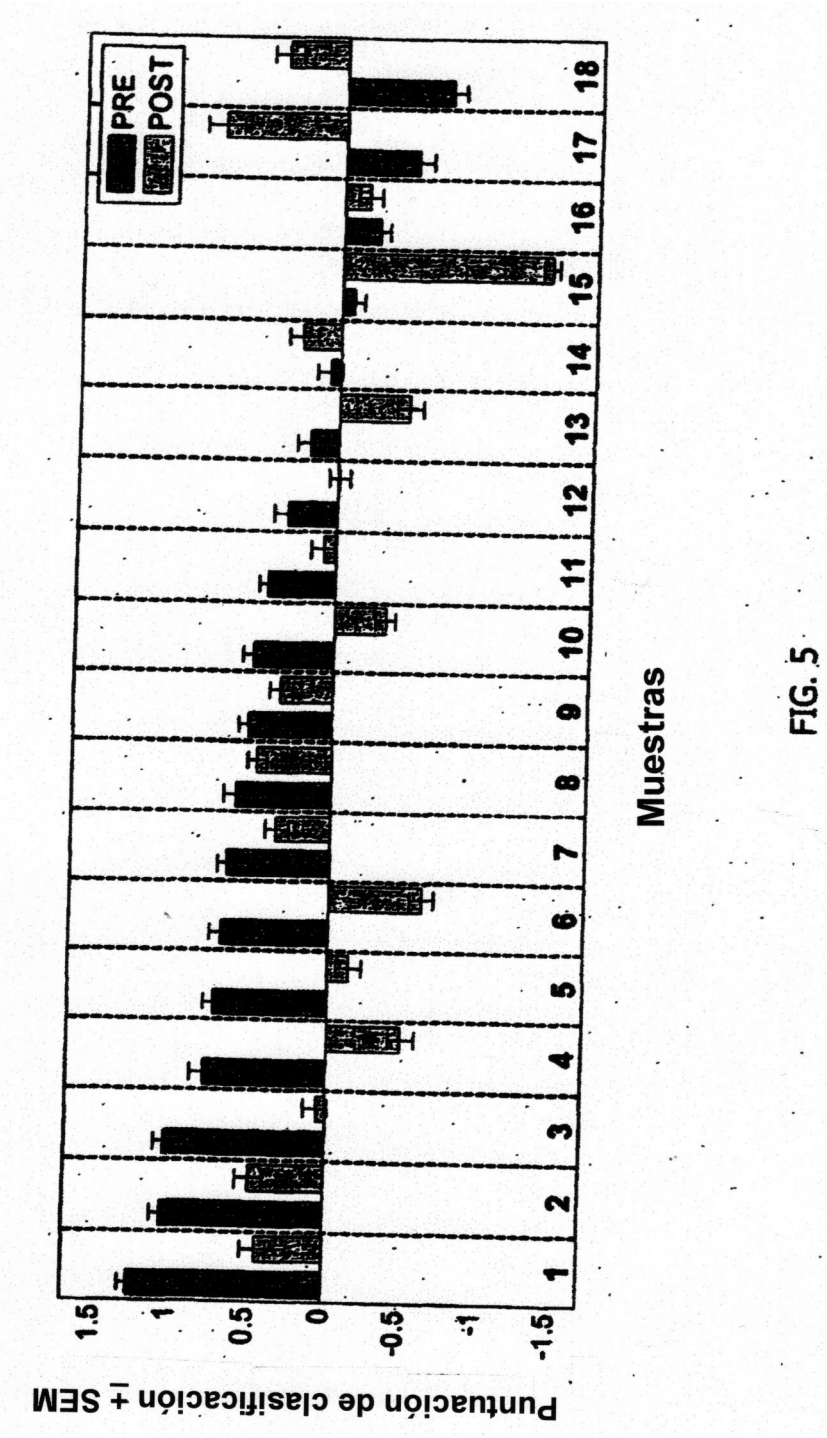
PRE
POST
1.5
1
0.5
0
-0.5
-1
-1.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Puntuación de clasificación ± SEM
Muestras

FIG. 5

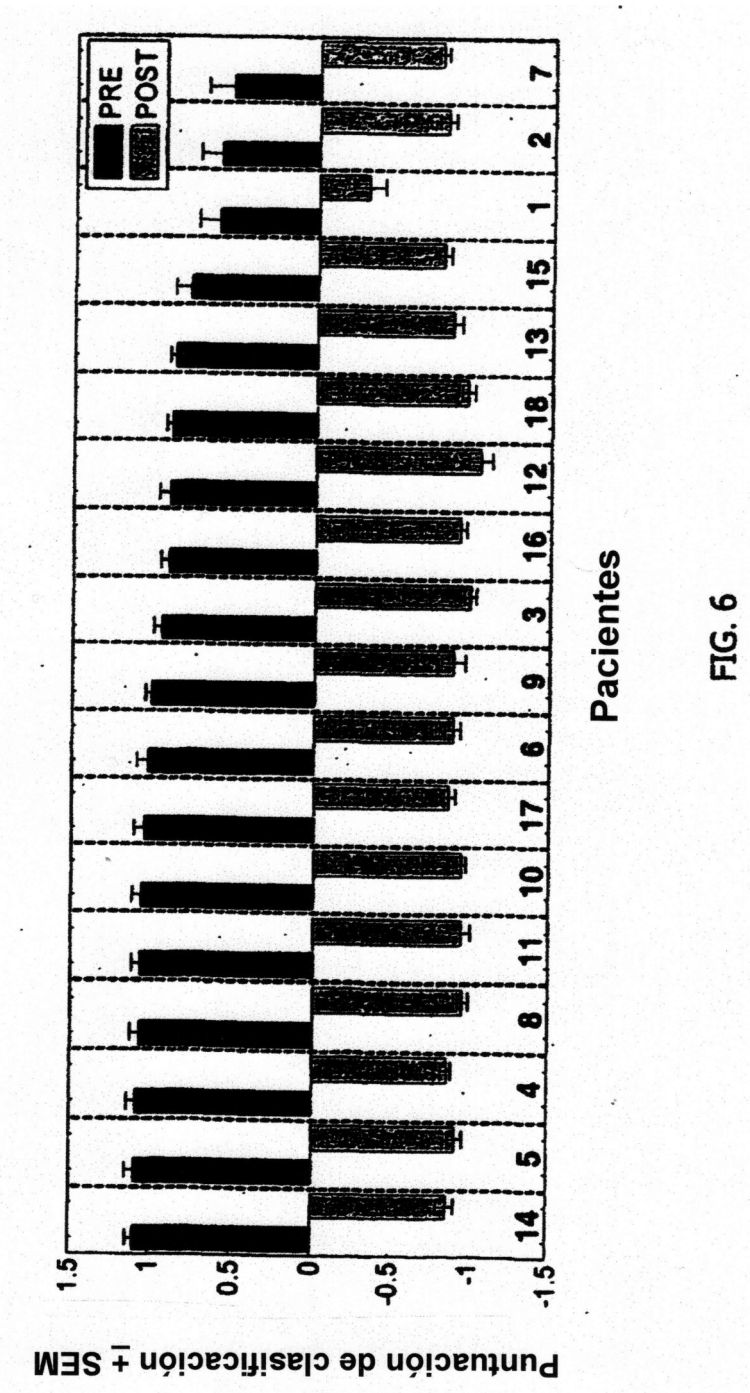
PRE
POST
1.5
1
0.5
0
-0.5
-1
-1.5
14
5
4
8
11
10
17
6
9
3
16
12
18
13
15
1
2
7
Pacientes
Puntuación de clasificación ± SEM

FIG. 6