(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 974 178**

(51) Int. Cl.:

C12Q 1/6886 (2008.01)
C12Q 1/6832 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **24.01.2020 PCT/US2020/015082**

(87) Fecha y número de publicación internacional: **30.07.2020 WO20154682**

(96) Fecha de presentación y número de la solicitud europea: **24.01.2020 E 20710332 (6)**

(97) Fecha y número de publicación de la concesión europea: **20.12.2023 EP 3914736**

(54) Título: **Detección de cáncer, tejido canceroso de origen y/o un tipo de célula cancerosa**

(30) Prioridad:

25.01.2019 US 201962797176 P
25.01.2019 US 201962797174 P
25.01.2019 US 201962797170 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.06.2024

(73) Titular/es:

GRAIL, LLC (100.0%)
1525 O'Brien Drive
Menlo Park, CA 94025, US

(72) Inventor/es:

VENN, OLIVER CLAUDE;
FIELDS, ALEXANDER P.;
GROSS, SAMUEL S.;
LIU, QINWEN;
SCHELLENBERGER, JAN;
BREDNO, JOERG;
BEAUSANG, JOHN F.;
SHOJAEE, SEYEDMEHDI;
SAKARYA, ONUR;
MAHER, M. CYRUS y
JAMSHIDI, ARASH

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

Observaciones:

Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Detección de cáncer, tejido canceroso de origen y/o un tipo de célula cancerosa

Antecedentes

La metilación del ADN juega un papel importante en la regulación de la expresión génica. La metilación del ADN aberrante se ha implicado en muchos procesos de enfermedad, incluyendo cáncer. El perfil de metilación del ADN mediante secuenciación de metilación (por ejemplo, secuenciación de bisulfito del genoma completo [WGBS]) se reconoce cada vez más como una valiosa herramienta de diagnóstico para la detección, el diagnóstico y/o el seguimiento del cáncer. Por ejemplo, los patrones específicos de regiones metiladas diferencialmente pueden ser útiles como marcadores moleculares de diversas enfermedades.

Sin embargo, WGBS no es idealmente adecuado para un ensayo de producto. La razón es que la gran mayoría del genoma no está metilado diferencialmente en el cáncer, o la densidad local de CpG es demasiado baja para proporcionar una señal robusta. Sólo un pequeño porcentaje del genoma puede ser útil para la clasificación.

Además, la identificación de regiones metiladas diferencialmente en diversas enfermedades ha planteado diversos retos. En primer lugar, la determinación de regiones metiladas diferencialmente en un grupo de enfermedad sólo tiene peso en comparación con un grupo de sujetos de control, de tal manera que si el grupo de control es pequeño en número, la determinación pierde confianza con el pequeño grupo de control. Además, entre un grupo de sujetos de control, el estado de metilación puede variar, lo que puede ser difícil de tener en cuenta a la hora de determinar las regiones metiladas diferencialmente en un grupo de enfermedad. Por otra parte, la metilación de una citosina en un sitio CpG está fuertemente correlacionada con la metilación en un sitio CpG posterior. Encapsular esta dependencia es un reto en sí mismo.

Por consiguiente, aún no se dispone de un método rentable para diagnosticar con precisión una enfermedad mediante la detección de regiones metiladas diferencialmente.

LIU, L. y col. "Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification", ANNALS OF ONCOLOGY, vol. 29, nº 6, 1 de junio de 2018 (2018-06-01), analizan el desarrollo de un ensayo de secuenciación de metilación dirigido a 9223 sitios CpG hipermetilados de forma consistente según The Cancer Genome Atlas. El estudio llevó a cabo una validación clínica del método utilizando muestras de ADNlc plasmático de 78 pacientes con cáncer colorrectal avanzado, cáncer de pulmón no microcítico (CPNM), cáncer de mama o melanoma, y comparó los resultados con los desenlaces clínicos de los pacientes.

Resumen

El alcance de la presente invención se expone en los métodos de las reivindicaciones adjuntas, donde la reivindicación 1 describe un método para analizar una muestra para detectar células de un tipo de cáncer en un sujeto. Las reivindicaciones dependientes 2-15 describen aspectos adicionales del método de la reivindicación 1.

Se proporcionan en la presente memoria (pero no se reivindican) composiciones que comprenden una pluralidad de diferentes oligonucleótidos cebo, en donde la pluralidad de diferentes oligonucleótidos cebo está configurada para hibridarse colectivamente con moléculas de ADN derivadas de al menos 200 regiones genómicas diana, en las que cada región genómica de las al menos 200 regiones genómicas diana está metilada diferencialmente en al menos un tipo de cáncer en relación con otro tipo de cáncer o en relación con un tipo distinto de cáncer, y en donde las al menos 200 regiones genómicas diana comprenden, para al menos el 80 % de todos los posibles pares de tipos de cáncer seleccionados de un conjunto que comprende al menos 10 tipos de cáncer, al menos una región genómica diana que está metilada diferencialmente entre el par de tipos de cáncer.

En algunos escenarios, los al menos 10 tipos de cáncer comprenden al menos 2, 3, 4, 5, 10, 12, 14, 16, 18 ó 20 tipos de cáncer. En algunos escenarios, los tipos de cáncer se seleccionan de cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatobiliar, hepatobiliar biliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal. En algunos escenarios, los tipos de cáncer se seleccionan de cáncer anal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago. En algunos escenarios, los tipos de cáncer se seleccionan de cáncer ADSL de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón. En algunos escenarios, las al menos 200 regiones genómicas diana se seleccionan de una

cualquiera de las listas 1-16. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 1-16. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 1-16. En algunos escenarios, las al menos 200 regiones genómicas diana se seleccionan de una cualquiera de las listas 1-3. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 1-3. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 1-3. En algunos escenarios, las al menos 200 regiones genómicas diana se seleccionan de una cualquiera de las listas 13-16. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 13-16. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 13-16. En algunos escenarios, las al menos 200 regiones genómicas diana se seleccionan de la lista 12. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en la lista 12. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en la lista 12. En algunos escenarios, las al menos 200 regiones genómicas diana se seleccionan de una cualquiera de las listas 8-11. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 8-11. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 8-11. En algunos escenarios, las al menos 200 regiones genómicas diana comprenden al menos el 40 %, el 50 %, el 60 % o el 70 % de las regiones genómicas diana indicadas en la lista 4. En algunos escenarios, en donde las al menos 200 regiones genómicas diana comprenden, para al menos el 90 % o para el 80 % de todos los pares posibles de tipos de cáncer seleccionados de un conjunto que comprende al menos 10 tipos de cáncer, al menos una región genómica diana que está diferencialmente metilada entre el par de tipos de cáncer. En algunos escenarios, la pluralidad de oligonucleótidos cebo se hibrida con al menos 15 nucleótidos o a al menos 30 nucleótidos de las moléculas de ADN derivadas de las al menos 200 regiones genómicas diana. En algunos escenarios, las moléculas de ADN derivadas de las al menos 200 regiones genómicas diana son fragmentos de ADNlc convertido. En algunos escenarios, los fragmentos de ADNlc se convierten mediante un proceso que comprende tratamiento con bisulfito. En algunos escenarios, los fragmentos de ADNlc se convierten mediante una reacción de conversión enzimática. En algunos escenarios, los fragmentos de ADNlc se convierten por una citosina desaminasa. En algunos escenarios, cada oligonucleótido cebo se conjuga con un resto de afinidad. En algunos escenarios, el resto de afinidad es biotina. En algunos escenarios, cada oligonucleótido cebo tiene entre 50 y 300 bases de longitud, entre 60 y 200 bases de longitud, entre 100 y 150 bases de longitud, entre 110 y 130 bases de longitud y/o 120 bases de longitud.

También se proporcionan en la presente memoria (pero no se reivindican) composiciones que comprenden una pluralidad de oligonucleótidos cebo diferentes configurados para hibridarse con moléculas de ADN derivadas de al menos 100 regiones genómicas diana seleccionadas de una cualquiera de las listas 1-16.

En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos 200 regiones genómicas diana. En algunos escenarios, las al menos 100 regiones genómicas diana se seleccionan de una cualquiera de las listas 1-16. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 1-16. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 1-16. En algunos escenarios, las al menos 100 regiones genómicas diana se seleccionan de una cualquiera de las listas 1-3. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 1-3. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en una cualquiera de las listas 1-3. En algunos escenarios, las al menos 100 regiones genómicas diana se seleccionan de la lista 12. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en la lista 12. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en la lista 12. En algunos escenarios, las al menos 100 regiones genómicas diana se seleccionan de la lista 8. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en la lista 8. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas diana en la lista 8. En algunos escenarios, las al menos 100 regiones genómicas diana comprenden al menos el 40 %, el 50 %, el 60 % o el 70 % de las regiones genómicas diana indicadas en la lista 4. En algunos escenarios, las moléculas

de ADN derivadas de las al menos 100 regiones genómicas diana son fragmentos de ADNlc convertido. En algunos escenarios, los fragmentos de ADNlc se convierten mediante un proceso que comprende tratamiento con bisulfito. En algunos escenarios, la composición comprende además fragmentos de ADNlc de un sujeto de prueba. En algunos escenarios, los fragmentos de ADNlc del sujeto de prueba son moléculas de ADNlc convertido. En algunos escenarios, los fragmentos de ADNlc del sujeto de prueba se convierten mediante un proceso que comprende tratamiento con bisulfito. En algunos escenarios, cada región genómica diana comprende al menos 5 dinucleótidos CpG. En algunos escenarios, cada oligonucleótido cebo tiene entre 60 y 200 bases de longitud, entre 100 y 150 bases de longitud, entre 110 y 130 bases de longitud y/o 120 bases de longitud. En algunos escenarios, los diferentes oligonucleótidos cebo comprenden una pluralidad de conjuntos de dos o más oligonucleótidos cebo, en donde cada oligonucleótido cebo dentro de un conjunto de oligonucleótidos cebo está configurado para unirse a las moléculas de ADN transformadas de la misma región genómica diana. En algunos escenarios, la razón de oligonucleótidos cebo configurados para hibridarse con regiones diana hipermetiladas para cebar oligonucleótidos configurados para hibridarse con regiones diana hipometiladas está entre 0,5 y 1,0. En algunos escenarios, cada conjunto de oligonucleótidos cebo comprende uno o más pares de un primer oligonucleótido cebo y un segundo oligonucleótido cebo, cada oligonucleótido cebo comprende un extremo 5' y un extremo 3', una secuencia de al menos X bases nucleotídicas en el extremo 3' del primer oligonucleótido cebo es idéntica a una secuencia de X bases nucleotídicas en el extremo 5' del segundo oligonucleótido cebo, y X es al menos 20, al menos 25, o al menos 30. En algunos escenarios, X es 30.

También se proporcionan en la presente memoria (pero no se reivindica en aislamiento) métodos para enriquecer una muestra de ADNlc, comprendiendo el método: poner en contacto una muestra de ADNlc convertido o sin convertir con un conjunto de cebos descrito anteriormente y enriquecer la muestra para ADNlc correspondiente a un primer conjunto de regiones genómicas mediante captura de hibridación. En algunos escenarios, la muestra de ADNlc es una muestra de ADNlc convertido.

También se proporcionan en la presente memoria (pero no se reivindica en aislamiento) métodos para obtener información de secuencia informativa de una presencia o ausencia de cáncer o un tipo de cáncer, comprendiendo el método de ADNlc convertido enriquecido en secuenciación preparado por un método que comprende poner en contacto una muestra de ADNlc convertido o sin convertir con un conjunto de cebos descrito anteriormente; y enriquecer la muestra para ADNlc correspondiente a un primer conjunto de regiones genómicas mediante captura de hibridación. En algunos escenarios, la muestra de ADNlc es una muestra de ADNlc convertido.

También se describen en la presente memoria (pero no se reivindica en tal generalidad) métodos para determinar una presencia o ausencia de cáncer en un sujeto, el método comprende capturar fragmentos de ADNlc del sujeto con una composición descrita anteriormente, secuenciar los fragmentos de ADNlc capturados y aplicar un clasificador entrenado a las secuencias de ADNlc para determinar la presencia o ausencia de cáncer. En algunos escenarios, la probabilidad de una determinación de falsos positivos de una presencia o ausencia de cáncer es inferior al 1 % y la probabilidad de una determinación precisa de una presencia o ausencia de cáncer es de al menos el 40 %. En algunos escenarios, el cáncer es un cáncer de estadio I, la probabilidad de una determinación de falsos positivos de una presencia o ausencia de cáncer es inferior al 1 %, y la probabilidad de una determinación precisa de una presencia o ausencia de cáncer es de al menos el 10 %. En algunos escenarios, los fragmentos de ADNlc son fragmentos de ADNlc convertido.

También se proporcionan en la presente memoria (pero no se reivindica en tal generalidad) métodos para detectar un tipo de cáncer que comprende capturar fragmentos de ADNlc de un sujeto con una composición que comprende una pluralidad de cebos oligonucleotídicos diferentes, secuenciar los fragmentos de ADNlc capturados y aplicar un clasificador entrenado a las secuencias de ADNlc para determinar un tipo de cáncer; en donde los cebos oligonucleotídicos se configuran para hibridarse con fragmentos de ADNlc derivados de una pluralidad de regiones genómicas diana, en donde la pluralidad de regiones genómicas diana se metila diferencialmente en uno o más tipos de cáncer en relación con un tipo de cáncer diferente o un tipo distinto de cáncer, en donde la probabilidad de una determinación de falsos positivos del cáncer es inferior al 1 %, y en donde la probabilidad de una asignación precisa de un tipo de cáncer es al menos el 75 %, al menos el 80 %, al menos el 85 % o al menos el 89 %, o al menos el 90 %. Algunos escenarios comprenden además aplicar el clasificador entrenado a las secuencias de ADNlc para determinar una presencia de cáncer antes de determinar el tipo de cáncer.

En algunos escenarios, el tipo de cáncer es un tipo de cáncer de estadio I, y la probabilidad de una asignación precisa es de al menos el 75 %. En algunos escenarios, el tipo de cáncer es un tipo de cáncer de estadio II, y la probabilidad de una asignación precisa es de al menos el 85 %. En algunos escenarios, el tipo de cáncer es cáncer de próstata y la probabilidad de una asignación precisa de cáncer de próstata es al menos el 85 % o al menos el 90 %, el tipo de cáncer es cáncer de mama y la probabilidad de una asignación precisa de cáncer de mama es de al menos el 90 % o al menos el 95 %. En algunos escenarios, el tipo de cáncer es cáncer uterino y la probabilidad de una asignación precisa del cáncer uterino es de al menos el 90 % o al menos el 95 %. En algunos escenarios, el tipo de cáncer es cáncer de ovario y la probabilidad de una asignación precisa del cáncer de ovario es de al menos el 85 % o al menos el 90 %. En algunos casos, el tipo de cáncer es de vejiga y urotelial y la probabilidad de una asignación precisa de vejiga y urotelial es de al menos el 90 % o el 95 %. En algunos escenarios, el tipo de cáncer es cáncer colorrectal y la probabilidad de una asignación precisa de cáncer colorrectal es de al menos el 65 % o al menos el 70 %. En algunos escenarios, el tipo de cáncer es cáncer de hígado y vías biliares y la probabilidad de una asignación precisa de cáncer

de hígado y vías biliares es de al menos el 90 % o al menos el 95 %. En algunos casos, el tipo de cáncer es cáncer de páncreas y vesícula biliar y la probabilidad de una asignación precisa de cáncer de páncreas y vesícula biliar es de al menos el 85 % o el 90 %. En algunos escenarios, los fragmentos de ADNlc son fragmentos de ADNlc convertido. En algunos escenarios, el de cáncer se selecciona de cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatobiliar, hepatobiliar biliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal. En algunos escenarios, el tipo de cáncer se selecciona de cáncer anal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago. En algunos escenarios, el tipo de cáncer se seleccionan de cáncer mutila de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón. En algunos escenarios, el tipo de cáncer es sarcoma y la probabilidad de un sarcoma de detección es de al menos el 35 % o al menos el 40 %. En algunos escenarios, la probabilidad de detectar un cáncer renal de estadio III o IV es de al menos el 50 % o el 70 %. En algunos escenarios, la probabilidad de detectar un cáncer de mama de estadio III o IV es de al menos el 70 % o el 85 %. En algunos escenarios, la probabilidad de detectar un cáncer de útero de estadio III o IV es de al menos el 50 %. En algunos escenarios, la probabilidad de detectar cáncer de ovario es de al menos el 60 % o al menos el 80 %. En algunos escenarios, la probabilidad de detectar cáncer de vejiga es de al menos el 35 % o al menos el 40 %. En algunos escenarios, la probabilidad de detectar cáncer anorrectal es de al menos el 60 % o el 70 %. En algunos escenarios, la probabilidad de detectar cáncer de cabeza y cuello es de al menos el 75 % o al menos el 80 %. En algunos escenarios, la probabilidad de detectar cáncer de cabeza y cuello de estado 1 es de al menos el 80 %. En algunos escenarios, la probabilidad de detectar cáncer colorrectal es de al menos el 50 % o al menos el 59 %. En algunos escenarios, la probabilidad de detectar cáncer hepático es de al menos el 75 % o el 80 %. En algunos escenarios, la probabilidad de detectar páncreas y cáncer de vesícula biliar es de al menos el 64 % o al menos el 70 %. En algunos escenarios, la probabilidad de detectar cáncer del tracto gastrointestinal superior es de al menos el 60 % o el 68 %. En algunos escenarios, la probabilidad de detectar mieloma múltiple es de al menos el 65 % o al menos el 75 %. En algunos escenarios, la probabilidad de detectar mieloma múltiple de tipo I es de al menos el 60 %. En algunos escenarios, la probabilidad de detectar neoplasia linfoide es de al menos el 65 % o al menos el 69 %. En algunos escenarios, la probabilidad de detectar cáncer de pulmón es de al menos el 50 % o al menos el 58 %. En algunos escenarios, la composición que comprende cebos oligonucleotídicos es una composición proporcionada anteriormente. En algunos escenarios, la pluralidad de regiones genómicas comprende no más de 95.000 regiones genómicas, no más de 60.000 regiones genómicas, no más de 40.000 regiones genómicas, no más de 35.000 regiones genómicas, no más de 20.000 regiones genómicas, no más de 15.000 regiones genómicas, no más de 8.000 regiones genómicas, no más de 4.000 regiones genómicas, no más de 2.000 regiones genómicas, o no más de 1.400 regiones genómicas. En algunos escenarios, el tamaño total de la pluralidad de regiones genómicas es inferior a 4 MB, inferior a 2 MB, inferior a 1 MB, inferior a 0,7 MB o inferior a 0,4 MB. En algunos escenarios, el sujeto tiene un riesgo elevado de uno o más tipos de cáncer. En algunos escenarios, el sujeto manifiesta síntomas asociados con uno o más tipos de cáncer. En algunos escenarios, el sujeto no ha sido diagnosticado con un cáncer. En algunos escenarios, el clasificador se entrenó en secuencias de ADN convertidas derivadas de al menos 100 sujetos con un primer tipo de cáncer, al menos 100 sujetos con un segundo tipo de cáncer y al menos 100 sujetos sin cáncer. En algunos escenarios, el primer tipo de cáncer es cáncer de ovario. En algunos escenarios, el primer tipo de cáncer es cáncer de hígado. En algunos escenarios, el primer tipo de cáncer se selecciona entre cáncer de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotecal, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de páncreas, cáncer de vesícula biliar, cáncer de esófago, cáncer de estómago, mieloma múltiple, neoplasia linfoide, cáncer de pulmón o leucemia. En algunos escenarios, el clasificador se entrenó en secuencias de ADN convertido derivadas de al menos 1000, al menos 2000, o al menos 4000 regiones genómicas diana seleccionadas de una cualquiera de las listas 1-16.

En algunos escenarios, el clasificador se entrena en secuencias de ADN convertido derivadas de al menos 1000, al menos 2000, o al menos 4000 regiones genómicas diana seleccionadas de una cualquiera de las listas 1-16. En algunos escenarios, el clasificador entrenado determina la presencia o ausencia de cáncer o un tipo de cáncer mediante (a) generar un conjunto de características para la muestra, en donde cada característica en el conjunto de características comprende un valor numérico; (b) introducir el conjunto de características en el clasificador, en donde el clasificador comprende un clasificador multinomio; (c) basándose en el conjunto de características, determinar, en el clasificador, un conjunto de puntuaciones de probabilidad, en donde el conjunto de puntuaciones de probabilidad comprende una puntuación de probabilidad por clase de tipo de cáncer y por clase de tipo distinto de cáncer; y (d) el umbral del conjunto de puntuaciones de probabilidad basándose en uno o más valores determinados durante el entrenamiento del clasificador para determinar una clasificación final del cáncer de la muestra. En algunos escenarios, el conjunto de características comprende un conjunto de características binarizadas. En algunos escenarios, el valor numérico comprende un único valor binario. En algunos escenarios, el clasificador multinomial comprende un conjunto de regresión logística multinomial entrenado para predecir un tejido fuente del cáncer. En algunos escenarios, el

clasificador determina una clasificación final del cáncer basándose en un diferencial de puntuación superior de dos probabilidad en relación con un valor mínimo, en donde el valor mínimo corresponde a un porcentaje predefinido de muestras de cáncer de entrenamiento que habían sido asignadas al tipo de cáncer correcto como su puntuación más alta durante el entrenamiento del clasificador. En algunos escenarios, el clasificador asigna una etiqueta de cáncer correspondiente a la puntuación de probabilidad más alta determinada por el clasificador como la clasificación final de cáncer cuando se determina que el diferencial de puntuación superior de dos probabilidad excede el valor mínimo; y asigna una etiqueta de cáncer indeterminada como la clasificación final del cáncer cuando se determina que el diferencial de puntuación superior de dos probabilidad no supera el valor mínimo.

También se proporcionan en la presente memoria (pero no se reivindican) métodos para tratar un tipo de cáncer en un sujeto que lo necesita, comprendiendo el método detectar el tipo de cáncer mediante un método descrito anteriormente, y administrar un agente terapéutico contra el cáncer al sujeto. En algunos escenarios, el agente anticanceroso es un agente quimioterapéutico seleccionado del grupo que consiste en agentes alquilantes, antimetabolitos, antraciclinas, antibióticos antitumorales, disruptores del citoesqueleto (taxanos), inhibidores de la topoisomerasa, inhibidores mitóticos, corticosteroides, inhibidores de la cinasa, análogos de nucleótidos y agentes basados en platino.

También se proporcionan en la presente memoria (pero no se reivindican) paneles de ensayo de cáncer que comprenden: al menos 500 pares de sondas, en donde cada par de los al menos 500 pares comprende dos sondas configuradas para superponerse entre sí mediante una secuencia de superposición, en donde la secuencia de superposición comprende una secuencia de 30 nucleótidos, y en donde la secuencia de 30 nucleótidos está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a, o derivada de una o más de las regiones genómicas, en donde cada una de las regiones genómicas comprende al menos cinco sitios de metilación, y en donde los al menos cinco sitios de metilación tienen un patrón de metilación anormal en muestras cancerosas.

En algunos escenarios, cada uno de los al menos 500 pares de sondas se conjuga con un resto de afinidad no nucleotídico. En algunos escenarios, el resto de afinidad sin nucleótidos es un resto de biotina. En algunos escenarios, las muestras cancerosas proceden de sujetos que padecen un cáncer seleccionado del grupo que consiste en cáncer de mama, cáncer de útero, cáncer de cuello de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial de pelvis renal, cáncer renal distinto del urotelial, cáncer de próstata, cáncer anorrectal, cáncer colorrectal, cáncer hepatobiliar derivado de hepatocitos, cáncer hepatobiliar derivado de células distintas de los hepatocitos, cáncer de páncreas, cáncer de células escamosas del tracto gastrointestinal superior, cáncer gastrointestinal superior distinto del escamoso, cáncer de cabeza y cuello, adenocarcinoma de pulmón, cáncer de pulmón de células pequeñas, cáncer de pulmón de células escamosas y cáncer distinto del adenocarcinoma o del cáncer de pulmón de células pequeñas, cáncer neuroendocrino, melanoma, cáncer de tiroides, sarcoma, mieloma múltiple, linfoma y leucemia. En algunos escenarios, el patrón de metilación anormal tiene al menos un valor de p umbral de rareza en las muestras cancerosas. En algunos escenarios, cada una de las sondas está diseñada para tener menos de 20 regiones genómicas fuera de la diana. En algunos escenarios, las menos de 20 regiones genómicas fuera de la diana se identifican usando una estrategia de siembra k-mero. En algunos casos, las menos de 20 regiones genómicas fuera de la diana se identifican mediante la estrategia de siembra de k-mero combinada con la alineación local en las ubicaciones de siembra. En algunos escenarios, el panel de ensayo de cáncer comprende al menos 10.000, 50.000, 100.000, 200.000, 300.000, 400.000, 500.000, 600.000, 700.000 u 800.000 sondas. En algunos escenarios, los al menos 500 pares de sondas juntos comprenden al menos 2 millones, 3 millones, 4 millones, 5 millones, 6 millones, 8 millones, 10 millones, 12 millones, 14 millones o 15 millones de nucleótidos. En algunos escenarios, cada una de las sondas comprende al menos 50, 75, 100 ó 120 nucleótidos. En algunos escenarios, cada una de las sondas comprende menos de 300, 250, 200 ó 150 nucleótidos. En algunos escenarios, cada una de las sondas comprende 100-150 nucleótidos. En algunos escenarios, cada una de las sondas comprende menos de 20, 15, 10, 8 ó 6 sitios de metilación. En algunos escenarios, al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los al menos cinco sitios de metilación están metilados o no metilados en las muestras cancerosas. En algunos casos, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no contienen G (guanina). En algunos escenarios, cada una de las sondas comprende múltiples sitios de unión a los sitios de metilación de la molécula de ADNlc convertido, en los que al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los múltiples sitios de unión comprenden exclusivamente CpG o CpA. En algunos escenarios, cada una de las sondas está configurada para tener menos de 15, 10 u 8 regiones genómicas fuera de la diana. En algunos escenarios, al menos el 30 % de las regiones genómicas están en exones o intrones. En algunos escenarios, al menos el 15 % de las regiones genómicas están en exones. En algunos escenarios, al menos el 20 % de las regiones genómicas están en exones. En algunos escenarios, menos del 10 % de las regiones genómicas están en regiones intergénicas. En algunos escenarios, las regiones genómicas se seleccionan de una cualquiera de las listas 1-3 o listas 4-16. En algunos escenarios, las regiones genómicas comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una cualquiera de las listas 1-3 o listas 4-16. En algunos escenarios, las regiones genómicas comprenden al menos 500, 1.000, 5.000, 10.000 o 15.000, 20.000, 30.000, 40.000, 50.000, 60.000 o 70.000 regiones genómicas en una cualquiera de las listas 1-3 o listas 4-16.

También se proporcionan en la presente memoria (pero no se reivindican) paneles de ensayo de cáncer que comprenden una pluralidad de sondas, en donde cada una de la pluralidad de sondas está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a una o más de las regiones genómicas en cualquiera de las listas 1-3 o listas 4-16.

En algunos escenarios, la pluralidad de sondas juntas está configurada para hibridarse con una pluralidad de moléculas de ADNlc convertido correspondientes a al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 % o el 90 %, el 95 % o el 100 % de las regiones genómicas de una cualquiera de las listas 1-3 o listas 4-16.

En algunos escenarios, la pluralidad de sondas juntas está configurada para hibridarse con una pluralidad de moléculas de ADNlc convertido correspondientes a al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas de una cualquiera de las listas 1-3 o listas 4-16. En algunos casos, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no contienen G (guanina). En algunos escenarios, cada una de las sondas comprende múltiples sitios de unión a los sitios de metilación de la molécula de ADNlc convertido, en los que al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los múltiples sitios de unión comprenden exclusivamente CpG o CpA. En algunos escenarios, cada una de las sondas se conjuga con un resto de afinidad no nucleotídico. En algunos escenarios, el resto de afinidad sin nucleótidos es un resto de biotina.

También se proporcionan en la presente memoria (pero no se reivindican) métodos para determinar un tejido de origen (TOO) de un cáncer, que comprende: recibir una muestra que comprende una pluralidad de moléculas de ADNlc; tratar la pluralidad de moléculas de ADNlc para convertir C (citosina) no metilada en U (uracilo), obteniendo así una pluralidad de moléculas de ADNlc convertido; aplicar un panel de ensayo de cáncer proporcionado anteriormente a la pluralidad de moléculas de ADNlc convertido, enriqueciendo así un subconjunto de las moléculas de ADNlc convertido; y secuenciar el subconjunto enriquecido de la molécula de ADNlc convertido, proporcionando de este modo un conjunto de lecturas de secuencia.

Algunos escenarios comprenden además la etapa de: determinar una condición de salud evaluando el conjunto de lecturas de secuencia, en donde la condición de salud es una presencia o ausencia de cáncer; una presencia o ausencia de cáncer de un tejido de origen (TOO); una presencia o ausencia de un tipo de célula cancerosa; o una presencia o ausencia de al menos 5, 10, 15 ó 20 tipos diferentes de cáncer. En algunos escenarios, la muestra que comprende una pluralidad de moléculas de ADNlc se obtuvo de un sujeto humano.

También se proporcionan en la presente memoria (pero no se reivindican) métodos para detectar un cáncer, que comprenden las etapas de: obtener un conjunto de lecturas de secuencia mediante secuenciación de un conjunto de fragmentos de ácido nucleico de un sujeto, en donde los fragmentos de ácido nucleico son correspondientes a, o derivados de una pluralidad de regiones genómicas seleccionadas de una cualquiera de las listas 1-3 o listas 4-16; para cada uno de los fragmentos de ácido nucleico, determinar el estado de metilación en una pluralidad de sitios de CpG, y detectar un estado de salud del sujeto evaluando el estado de metilación para las lecturas de secuencia, en donde la condición de salud es (i) una presencia o ausencia de cáncer; (ii) una presencia o ausencia de cáncer de un tejido de origen (TOO); (iii) una presencia o ausencia de un tipo de célula cancerosa; o (iv) una presencia o ausencia de al menos 5, 10, 15 ó 20 tipos diferentes de cáncer.

En algunos escenarios, la pluralidad de regiones genómicas comprende al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 %, el 95 % o el 100 % de las regiones genómicas de una cualquiera de las listas 1-3 o listas 4-16. En algunos escenarios, la pluralidad de regiones genómicas comprende 500, 1.000, 5.000, 10.000, 15.000, 20.000, 30.000, 40.000, 50.000, 60.000, 70.000 u 80.000 de las regiones genómicas de una cualquiera de las listas 1-3 o listas 4-16.

También se proporcionan en la presente memoria (pero no se reivindican) métodos para diseñar un panel de ensayo de cáncer para diagnosticar el cáncer de un tejido de origen (TOO) que comprende las etapas de: identificar una pluralidad de regiones genómicas, en donde cada una de la pluralidad de regiones genómicas (i) comprende al menos 30 nucleótidos, y (ii) comprende al menos cinco sitios de metilación, seleccionar un subconjunto de las regiones genómicas, en donde la selección se realiza cuando las moléculas de ADNlc correspondientes a, o derivadas de cada una de las regiones genómicas en muestras cancerosas tienen un patrón de metilación anormal, en donde el patrón de metilación anormal comprende al menos cinco sitios de metilación hipometilados o hipermetilados, y diseñar un panel de ensayo de cáncer que comprende una pluralidad de sondas, en donde cada una de las sondas está configurada para hibridarse con una molécula de ADNlc convertido correspondiente o derivada de uno o más del subconjunto de las regiones genómicas.

También se proporciona en la presente memoria (pero no se reivindican) conjuntos cebos para captura de hibridación, un conjunto de cebos que comprende una pluralidad de diferentes sondas que contienen oligonucleótidos, en donde cada una de las sondas que contienen oligonucleótido comprende una secuencia de al menos 30 bases de longitud que es complementaria de: (1) una secuencia de una región genómica; o (2) una secuencia que varía de la secuencia de (1) solo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica, y en donde cada sonda de las diferentes sondas que contienen oligonucleótidos es complementaria a una secuencia correspondiente a un sitio CpG que se metila diferencialmente en muestras de sujetos con un primer tipo de cáncer en relación con las muestras de sujetos con un segundo tipo de cáncer o un tipo distinto de cáncer.

En algunos casos, el primer tipo de cáncer y el segundo tipo de cáncer se seleccionan entre cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma, linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatocelular hepatobiliar, biliar hepatobiliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello de útero, cáncer de mama, cáncer de vejiga y cáncer anorrectal.

El conjunto de cebos puede comprender al menos 500, 1.000, 2.000, 2.500, 5.000, 6.000, 7.500, 10.000, 15.000, 20.000, 25.000, 50.000, 100.000, 200.000, 300.000, 500.000 u 800.000 diferentes sondas que contienen oligonucleótidos. En algunos escenarios, para cada una de las diferentes sondas que contienen oligonucleótidos, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 1-16; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 1-3; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 5 ó 7; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 4, 8 u 8-12; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 13-16; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 1-16; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en una cualquiera de las listas 4 ó 6; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 4; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 8; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 9; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 10; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 11; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la secuencia de al menos 30 bases de longitud es complementaria a (1) una secuencia dentro de una región genómica seleccionada de las regiones genómicas establecidas en la lista 12; o (2) una secuencia que varía de la secuencia de (1) sólo por una o más transiciones, en donde cada transición respectiva de una o más transiciones se produce a una citosina en la región genómica. En algunos escenarios, la pluralidad de diferentes sondas que contienen oligonucleótidos se conjuga cada una con un resto de afinidad. En algunos escenarios, el resto de afinidad es biotina. En algunos escenarios, al menos el 80 %, el 90 % o el 95 % de las sondas que contienen oligonucleótido en el conjunto de cebos no incluyen una secuencia de al menos 30, al menos 40 o al menos 45 bases que tiene 20 o más regiones fuera de diana en el genoma. En algunos escenarios, las sondas que contienen oligonucleótidos en el conjunto de cebos no incluyen una secuencia de al menos 30, al menos 40 o al menos 45 bases que tiene 20 o más regiones fuera de diana en el genoma. En algunos escenarios, la secuencia de al menos 30 bases de cada una de las sondas es de al menos 40 bases, al menos 45 bases, al menos 50 bases, al menos 60 bases, al

menos 75 o al menos 100 bases de longitud. En algunos escenarios, cada una de las sondas que contienen oligonucleótido tiene una secuencia de ácido nucleico de al menos 45, 40, 75, 100 ó 120 bases de longitud. En algunos escenarios, cada una de las sondas que contienen oligonucleótido tiene una secuencia de ácido nucleico de no más de 300, 250, 200 ó 150 bases de longitud. En algunos escenarios, cada una de la pluralidad de sondas que contienen oligonucleótidos diferentes tiene entre 60 y 200 bases de longitud, entre 100 y 150 bases de longitud, entre 110 y 130 bases de longitud y/o 120 bases de longitud. En algunos escenarios, las diferentes sondas que contienen oligonucleótidos comprenden al menos 500, al menos 1000, al menos 2.000, al menos 2.500, al menos 5.000, al menos 6.000, al menos 7.500, y al menos 10.000, al menos 15.000, al menos 20.000, o al menos 25.000 pares diferentes de sondas, en donde cada par de sondas comprende una primera sonda y la segunda sonda, en donde la segunda sonda difiere de la primera sonda y se superpone con la primera sonda mediante una secuencia superpuesta que es al menos 30, al menos 40, al menos 50, o al menos 60 nucleótidos de longitud. En algunos escenarios, el conjunto de cebos comprende sondas que contienen oligonucleótidos que están configuradas para dirigirse al menos al 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 % o el 100 % de las regiones genómicas identificadas en una cualquiera de las listas 1-16. En algunos escenarios, el conjunto de cebos comprende sondas que contienen oligonucleótidos que están configuradas para dirigirse al menos al 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 % o el 100 % de las regiones genómicas identificadas en una cualquiera de las listas 1-3. En algunos escenarios, el conjunto de cebos comprende sondas que contienen oligonucleótidos que están configuradas para dirigirse al menos al 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 % o el 100 % de las regiones genómicas identificadas en una cualquiera de las listas 4-12. En algunos escenarios, el conjunto de cebos comprende sondas que contienen oligonucleótidos que están configuradas para dirigirse al menos al 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 % o el 100 % de las regiones genómicas identificadas en una cualquiera de las listas 4, 6, u 8-12. En algunos escenarios, el conjunto de cebos comprende sondas que contienen oligonucleótidos que están configuradas para dirigirse al menos al 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 % o el 100 % de las regiones genómicas identificadas en la lista 8. En algunos escenarios, una totalidad de sondas de oligonucleótido en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una lista seleccionada de una cualquiera de las listas 1-16. En algunos escenarios, la totalidad de sondas de oligonucleótido en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una lista seleccionada de una cualquiera de las listas 1-3. En algunos escenarios, la totalidad de sondas de oligonucleótido en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una lista seleccionada de una cualquiera de las listas 4-12. En algunos escenarios, la totalidad de sondas de oligonucleótido en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una lista seleccionada de una cualquiera de las listas 4, 6, u 8-12. En algunos escenarios, la totalidad de sondas de oligonucleótido en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas en una lista seleccionada de la lista 8. En algunos escenarios, una totalidad de sondas que contienen oligonucleótidos en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos 500, 1000, 5000, 10.000, 15.000, 20.000, al menos 25.000, al menos 30.000, al menos 50.000 o al menos 80.000 regiones genómicas en una cualquiera de las listas 1-16. En algunos escenarios, la totalidad de sondas que contienen oligonucleótidos en el conjunto de cebos se configura para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos 500, 1000, 5000, 10.000, 15.000, 20.000, al menos 25.000, al menos 30.000, al menos 50.000 o al menos 80.000 regiones genómicas en una cualquiera de las listas 1-3. En algunos escenarios, la totalidad de sondas que contienen oligonucleótidos en el conjunto de cebos se configura para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos 500, 1.000, 5.000, 10.000, 15.000, 20.000, al menos 25.000, al menos 30.000, al menos 50.000 o al menos 80.000 regiones genómicas en una cualquiera de las listas 4-12. En algunos escenarios, la totalidad de sondas que contienen oligonucleótidos en el conjunto de cebos se configuran para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos 500, 1000, 5000, 10.000, 15.000, 20.000, al menos 25.000, al menos 30.000, al menos 50.000 o al menos 80.000 regiones genómicas en una cualquiera de las listas 4, 6, u 8-12. En algunos escenarios, la totalidad de sondas que contienen oligonucleótidos en el conjunto de cebos se configura para hibridarse con fragmentos obtenidos de moléculas de ADNlc correspondientes a al menos 500, 1000, 5000, 10.000, 15.000, 20.000, al menos 25.000, al menos 30.000, al menos 50.000 o al menos 80.000 regiones genómicas en una cualquiera de la lista 8. En algunos escenarios, la pluralidad de sondas que contienen oligonucleótidos comprende al menos 500, 1.000, 5.000 o 10.000 subconjuntos diferentes de sondas, en donde cada subconjunto de sondas comprende una pluralidad de sondas que se extienden colectivamente a través de una región genómica seleccionada de las regiones genómicas de una cualquiera de las listas 1-16 de una forma de dos títulos. En algunos escenarios, la pluralidad de sondas que contienen oligonucleótidos comprende al menos 500, 1.000, 5.000 o 10.000 subconjuntos diferentes de sondas, en donde cada subconjunto de sondas comprende una pluralidad de sondas que se extienden colectivamente

a través de una región genómica seleccionada de las regiones genómicas de una cualquiera de las listas 1-4, 6 u 8-12 de una forma de dos títulos. En algunos escenarios, la pluralidad de sondas que se extienden colectivamente a través de la región genómica de una forma de dos títulos comprende al menos un par de sondas que se solapan por una secuencia de al menos 30 bases, al menos 40 bases, al menos 50 bases o al menos 60 bases de longitud. En algunos escenarios, la pluralidad de sondas se extiende colectivamente a través de porciones del genoma que colectivamente son un tamaño combinado de menos de 4 MB, menos de 2 MB, menos de 1 MB, menos de 0,7 MB o menos de 0,4 MB. En algunos escenarios, la pluralidad de sondas se extiende colectivamente a través de partes del genoma que colectivamente son un tamaño combinado de entre 0,2 y 30 MB, entre 0,5 MB y 30 MB, entre 1 MB y 30 MB, entre 3 MB y 25 MB, entre 3 MB y 15, MB, entre 5 MB y 20 MB, o entre 7 MB y 12 MB. En algunos escenarios, cada una de las diferentes sondas que contienen oligonucleótidos comprende menos de 20, 15, 10, 8 ó 6 sitios de detección CpG. En algunos escenarios, al menos el 80 %, el 85 %, el 90 %, el 92 %, el 95 % o el 98 % de la pluralidad de sondas que contienen oligonucleótidos tienen exclusivamente CpG o CpA en todos los sitios de detección CpG.

También se proporciona en la presente memoria (pero no se reivindican) mezclas que comprenden: ADNlc convertido y un conjunto de cebos proporcionado anteriormente. En algunos escenarios, el ADNlc convertido comprende ADNlc convertido con bisulfito.

El ADNlc convertido puede comprender ADNlc que se ha convertido mediante una citosina desaminasa.

También se proporciona en la presente memoria (pero no se reivindican) métodos para enriquecer una muestra de ADNlc convertido, un método que comprende: poner en contacto la muestra de ADN libre de células convertida con un conjunto de cebos proporcionado anteriormente; y enriquecer la muestra para un primer conjunto de regiones genómicas mediante captura de hibridación.

También se proporciona en la presente memoria (pero no se reivindican) métodos para proporcionar información de secuencia informativa de una presencia o ausencia de un cáncer o un tipo de cáncer, el método que comprende: procesar ADNlc a partir de una muestra biológica con un agente de desaminación para generar una muestra de ADN libre de células que comprende nucleótidos desaminados; enriquecer la muestra de ADNlc para moléculas de ADN libre de células informativas; y secuenciar las moléculas de ADNlc enriquecidas, obteniendo así un conjunto de lecturas de secuencia informativa de una presencia o ausencia de un cáncer o un tipo de cáncer.

En algunos escenarios, enriquecer el ADNlc comprende amplificar, mediante PCR, porciones de los fragmentos de ADN libres de células usando cebadores configurados para hibridarse con una pluralidad de regiones genómicas seleccionadas de una cualquiera de las listas 1-16. En algunos escenarios, enriquecer la muestra de ADNlc comprende poner en contacto el ADN libre de células con una pluralidad de sondas configuradas para hibridarse con fragmentos convertidos obtenidos de las moléculas de ADNlc correspondientes o derivadas de las regiones genómicas en una cualquiera de las listas 1-16. En algunos escenarios, la muestra de ADNlc comprende poner en contacto el ADN libre de células con una pluralidad de sondas configuradas para hibridarse con fragmentos convertidos obtenidos de las moléculas de ADNlc correspondientes o derivadas de al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 %, el 95 % de las regiones genómicas en una cualquiera de las listas 1-16. En algunos escenarios, las regiones genómicas se seleccionan de una cualquiera de las listas 1-3. En algunos escenarios, las regiones genómicas se seleccionan de una cualquiera de las listas 4-12. En algunos escenarios, las regiones genómicas se seleccionan de una cualquiera de las listas 4, 6 u 8-12. En algunos escenarios, las regiones genómicas se seleccionan de la lista 8. En algunos escenarios, la muestra de ADNlc se enriquece mediante un método proporcionado anteriormente. En algunos escenarios, el método comprende además determinar una clasificación de cáncer evaluando el conjunto de lecturas de secuencia, en donde la clasificación de cáncer es una presencia o ausencia de cáncer; o una presencia o ausencia de un tipo de cáncer. En algunos escenarios, la etapa de determinar una clasificación de cáncer comprende: generar un vector de característica de prueba basándose en el conjunto de lecturas de secuencia; y aplicar el vector de características de prueba a un clasificador. En algunos escenarios, el clasificador comprende un modelo que se entrena mediante un proceso de entrenamiento con un primer conjunto de fragmentos de fragmentos de uno o más sujetos de entrenamiento con un primer tipo de cáncer y un segundo conjunto de cáncer de fragmentos de uno o más sujetos de entrenamiento con un segundo tipo de cáncer, en donde tanto el primer conjunto de fragmentos de fragmentos como el segundo conjunto de cáncer de fragmentos comprenden una pluralidad de fragmentos de entrenamiento. En algunos escenarios, la clasificación del cáncer es una presencia o ausencia de cáncer. En algunos escenarios, tiene un área bajo una curva característica operativa del receptor de al menos 0,8. En algunos escenarios, la clasificación del cáncer es un tipo de cáncer. En algunos escenarios, el tipo de cáncer se selecciona entre al menos 12, 14, 16, 18 ó 20 tipos de cáncer. En algunos escenarios, los tipos de cáncer se seleccionan de cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatocelular hepatobiliar, hepatobiliar biliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal. En algunos escenarios, los tipos de cáncer se seleccionan de cáncer anal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago. En algunas realizaciones, los tipos de cáncer se seleccionan de cáncer hipo metilados de tiroides,

melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón. En algunos escenarios, al 99 % de especificidad, la sensibilidad del método para cáncer de cabeza y cuello es de al menos el 79 % o al menos el 84 % ; al 99 % de especificidad, la sensibilidad del método para cáncer de hígado de es al menos el 82 % o al menos el 85 %; al 99 % de especificidad, la sensibilidad del método para cáncer de tracto gastrointestinal superior es de al menos el 62 % o al menos el 68 %; en donde al 99 % de especificidad, la sensibilidad del método para cáncer de vejiga o de vesícula biliar es de al menos el 62 % o al menos el 68 %; al 99 % de especificidad, la sensibilidad del método para cáncer colorrectal de es al menos el 60 % o al menos el 65 %; al 99 % de especificidad, la sensibilidad del método para cáncer de ovario de es al menos el 75 % o al menos el 80 %; al 99 % de especificidad, la sensibilidad del método para cáncer de hígado de es al menos el 60 % o al menos el 65 %; al 99 % de especificidad, la sensibilidad del método para mieloma múltiple es de al menos el 68 % o al menos el 75 % ; al 99 % de especificidad, la sensibilidad del método para neoplasia linfoide de es al menos el 65 % o al menos el 70 %; al 99 % de especificidad, la sensibilidad del método para cáncer anorrectal de es al menos el 60 % o al menos el 65 %; y al 99 % de especificidad, la sensibilidad del método para cáncer de vejiga de es al menos el 40 % o al menos el 44 %. En algunos escenarios, la clasificación del cáncer es una presencia o ausencia de un tipo de cáncer. En algunos escenarios, la etapa de determinar una clasificación de cáncer comprende: generar un vector de característica de prueba basándose en el conjunto de lecturas de secuencia; y aplicar el vector de características de prueba a un clasificador. En algunos escenarios, el clasificador comprende un modelo que es entrenado por un proceso de entrenamiento con un primer conjunto de tipos de cáncer de secuencias de ADN convertidos de uno o más sujetos de entrenamiento con un primer tipo de cáncer y un segundo conjunto de tipos de cáncer de secuencias de ADN convertidas de uno o más sujetos de entrenamiento con un segundo tipo de cáncer, en donde tanto el primer conjunto de tipos de cáncer de secuencias de ADN convertidos como el segundo conjunto de tipos de cáncer de secuencias de ADN convertidas comprenden una pluralidad de secuencias de ADN convertidas en entrenamiento. En algunos casos, el tipo de cáncer se selecciona del grupo formado por cáncer de cabeza y cuello, cáncer de hígado/vía biliar, cáncer del tracto gastrointestinal superior, cáncer de páncreas/vesícula biliar; cáncer colorrectal, cáncer de ovario, cáncer de pulmón, mieloma múltiple, neoplasias linfoides, melanoma, sarcoma, cáncer de mama y cáncer de útero. En algunos escenarios, el tipo de cáncer es cáncer de cabeza y cuello, y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 79 % o al menos el 84 %, el tipo de cáncer es cáncer de hígado y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 82 % o al menos el 85 %. En algunos casos, el tipo de cáncer es un cáncer del tracto gastrointestinal superior, y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 62 % o el 68 %. En algunos escenarios, el tipo de cáncer es un cáncer de páncreas o de vesícula biliar, y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 62 % o al menos el 68 %. En algunos escenarios, el tipo de cáncer es cáncer colorrectal, y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 60 % o al menos el 65 %. En algunos escenarios, el tipo de cáncer es cáncer de ovario, y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 75 % o al menos el 80 %. En algunos escenarios, el tipo de cáncer es cáncer de pulmón y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 60 % o al menos el 65 %. En algunos escenarios, el tipo de cáncer es mieloma múltiple y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 68 % o al menos el 75 %. En algunos escenarios, el tipo de cáncer es una neoplasia linfoide y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 65 % o al menos el 70 %. En algunos escenarios, el tipo de cáncer es cáncer anorrectal y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 60 % o al menos el 65 %. En algunos escenarios, el tipo de cáncer es cáncer de vejiga y el método, al 99,0 % de especificidad, tiene una sensibilidad de al menos el 40 % o al menos el 44 %. En algunos escenarios, el tamaño total de las regiones genómicas diana es inferior a 4 Mb, inferior a 2 Mb, inferior a 1 Mb, inferior a 0,7 Mb o inferior a 0,4 Mb. En algunos escenarios, la etapa de determinar una clasificación de cáncer comprende:

generar un vector de característica de prueba basándose en el conjunto de lecturas de secuencia; y aplicar el vector de características de prueba a un modelo obtenido mediante un proceso de entrenamiento con un conjunto de fragmentos de fragmentos de uno o más sujetos de entrenamiento con un cáncer y un conjunto de fragmentos no cancerosas de uno o más sujetos de entrenamiento sin cáncer, en donde tanto el conjunto de fragmentos de fragmentos como el conjunto de fragmentos no cancerosas comprenden una pluralidad de fragmentos de entrenamiento. En algunos escenarios, el proceso de entrenamiento comprende: obtener información de secuencia de fragmentos de entrenamiento de una pluralidad de sujetos de entrenamiento; para cada fragmento de entrenamiento, determinar si ese fragmento de entrenamiento está hipometilado o hipermetilado, donde cada uno de los fragmentos de entrenamiento hipometilados e hipermetilados comprende al menos un número umbral de sitios CpG con al menos un porcentaje umbral de los sitios CpG no metilados o metilados, respectivamente, para cada sujeto de entrenamiento, generar un vector de características de entrenamiento basándose en los fragmentos de entrenamiento hipometilados y los fragmentos de entrenamiento hipermetilados, y entrenar el modelo con los vectores de características de entrenamiento de uno o más sujetos de entrenamiento sin cáncer y los vectores de características de entrenamiento de uno o más sujetos de entrenamiento con cáncer. En algunos escenarios, el proceso de entrenamiento comprende: obtener información de secuencia de fragmentos de entrenamiento de una pluralidad de sujetos de entrenamiento; para cada fragmento de entrenamiento, determinar si ese fragmento de entrenamiento está hipometilado o hipermetilado, en el que cada uno de los fragmentos de entrenamiento hipometilados e hipermetilados comprende al menos un número umbral de sitios CpG con al menos un porcentaje umbral de los sitios CpG no metilados o metilados, respectivamente, para cada uno de una pluralidad de sitios CpG en un genoma de referencia:

cuantificar un recuento de fragmentos de entrenamiento hipometilados que se solapan con el sitio CpG y un recuento de fragmentos de entrenamiento hipermetilados que se solapan con el sitio CpG; y generar una puntuación de hipometilación y una puntuación de hipermetilación basadas en el recuento de fragmentos de entrenamiento hipometilados y fragmentos de entrenamiento hipermetilados; para cada fragmento de entrenamiento, generar una puntuación de hipometilación agregada basándose en la puntuación de hipometilación de los sitios CpG en el fragmento de entrenamiento y una puntuación de hipermetilación agregada basándose en la puntuación de hipermetilación de los sitios CpG en el fragmento de entrenamiento; para cada sujeto de entrenamiento: clasificación de la pluralidad de fragmentos de entrenamiento en función de la puntuación de hipometilación agregada y clasificación de la pluralidad de fragmentos de entrenamiento en función de la puntuación de hipermetilación agregada; y generar un vector de características basándose en la clasificación de los fragmentos de entrenamiento; obtener vectores de características de entrenamiento para uno o más sujetos de entrenamiento sin vectores de formación de cáncer y de entrenamiento para el uno o más sujetos de entrenamiento con cáncer; y entrenar el modelo con los vectores de características para el uno o más sujetos de entrenamiento sin cáncer y los vectores de características para el uno o más sujetos de entrenamiento con cáncer. En algunos escenarios, el modelo comprende uno de un clasificador de regresión logística de núcleo, un clasificador de bosque aleatorio, un modelo de mezcla, una red neuronal convolucional y un modelo de autocodificador. En algunos escenarios, el método comprende además las etapas de: obtener una probabilidad de cáncer para la muestra de prueba basándose en el modelo; y comparar la probabilidad de cáncer con una probabilidad umbral para determinar si la muestra de prueba es de un sujeto con cáncer o sin cáncer. En algunos escenarios, el método comprende además las etapas de: obtener una probabilidad de tipo de cáncer para la muestra de prueba en base al modelo; y comparar la probabilidad de tipo de cáncer con una probabilidad umbral para determinar si la muestra de prueba es de un sujeto con el tipo de cáncer u otro tipo de cáncer o sin cáncer. En algunos escenarios, el método comprende además administrar un agente anticanceroso al sujeto.

También se proporciona en la presente memoria (pero no se reivindican) métodos para tratar un paciente con cáncer, el método que comprende:

administrar un agente anticanceroso a un sujeto que se ha identificado como un cáncer sujeto por un método proporcionado anteriormente. En algunos escenarios, el agente anticanceroso es un agente quimioterapéutico seleccionado del grupo que consiste en agentes alquilantes, antimetabolitos, antraciclinas, antibióticos antitumorales, disruptores del citoesqueleto (taxanos), inhibidores de la topoisomerasa, inhibidores mitóticos, corticosteroides, inhibidores de la cinasa, análogos de nucleótidos y agentes basados en platino.

También se proporciona en la presente memoria (pero no se reivindican) métodos para tratar a un paciente con cáncer, el método comprende administrar un agente anticancerígeno a un sujeto que se ha identificado como un cáncer sujeto por un método proporcionado en la presente memoria. En algunos escenarios, el agente anticanceroso es un agente quimioterapéutico seleccionado del grupo que consiste en agentes alquilantes, antimetabolitos, antraciclinas, antibióticos antitumorales, disruptores del citoesqueleto (taxanos), inhibidores de la topoisomerasa, inhibidores mitóticos, corticosteroides, inhibidores de la cinasa, análogos de nucleótidos y agentes basados en platino.

También se proporciona en la presente memoria (pero no se reivindican) métodos para evaluar si un sujeto tiene un cáncer, el método comprende: obtener ADNlc del sujeto; aislar una porción del ADNlc del sujeto mediante captura de hibridación; obtener lecturas de secuencia derivadas del ADNlc capturado para determinar los estados de metilación de ADN; aplicar un clasificador a las lecturas de secuencia; y determinar si el sujeto tiene cáncer basándose en la aplicación del clasificador; en donde el clasificador tiene un área bajo la curva característica del operador receptor de al menos 0,80. En algunos escenarios, el método comprende además determinar un tipo de cáncer, en donde la sensibilidad del método para cáncer de cabeza y cuello es de al menos el 79 % o al menos el 84 % ; en donde la sensibilidad del método para cáncer de hígado de es al menos el 82 % o al menos el 85 %; en el que la sensibilidad del método para el cáncer del tracto gastrointestinal superior es de al menos el 62 % o el 68 %; en donde la sensibilidad del método para el cáncer de páncreas o de vesícula biliar es de al menos el 62 % o al menos el 68 %%; en donde la sensibilidad del método para cáncer colorrectal es de al menos el 60 % o al menos el 65 %; en donde la sensibilidad del método para cáncer de ovario es de al menos el 75 % o al menos el 80 % ; en donde la sensibilidad del método para cáncer de pulmón de es al menos el 60 % o al menos el 65 %; en donde la sensibilidad del método para mieloma múltiple es de al menos el 68 % o al menos el 75 % ; en donde la sensibilidad del método para neoplasia linfoide de es al menos el 65 % o al menos el 70 %; en donde la sensibilidad del método para cáncer de pulmón de es al menos el 60 % o al menos el 65 %; y en donde la sensibilidad del método de cáncer de vejiga es de al menos el 40 % o al menos el 44 %. En algunos escenarios, el tamaño total de las regiones genómicas diana es inferior a 4 Mb, inferior a 2 Mb, inferior a 1 Mb, inferior a 0,7 Mb o inferior a 0,4 Mb. En algunos escenarios, el método comprende además convertir citosinas no metiladas en el ADNlc en uracilo antes de aislar la porción del ADNlc del sujeto mediante captura de hibridación. En algunos escenarios, el método comprende además citosinas no metiladas en el ADNlc en uracilo después de aislar la porción del ADNlc del sujeto mediante captura de hibridación. En algunos escenarios, el clasificador es un clasificador binario. En algunos escenarios, el clasificador es un clasificador modelo de mezcla. En algunos escenarios, aislar una porción del ADNlc del sujeto por captura de hibridación comprende poner en contacto el ADN libre de células con un conjunto de cebos que comprende una pluralidad de sondas que contienen oligonucleótidos diferentes. En algunos escenarios, el conjunto de cebos es un conjunto de cebos proporcionado en la presente memoria.

También se proporciona en la presente memoria (pero no se reivindican) métodos que comprenden las etapas de: obtener un conjunto de lecturas de secuencia de fragmentos de prueba modificados, en donde los fragmentos de prueba modificados se han obtenido mediante el procesamiento de un conjunto de fragmentos de ácido nucleico de un sujeto de prueba, en donde cada uno de los fragmentos de ácido nucleico corresponde o se deriva de una pluralidad de regiones genómicas seleccionadas de cualquiera de las listas 1-16; y aplicar el conjunto de lecturas de secuencia o un vector de características de prueba obtenido en base al conjunto de lecturas de secuencia con un modelo obtenido mediante un proceso de entrenamiento con un primer conjunto de fragmentos de una pluralidad de sujetos de entrenamiento con un primer tipo de cáncer y un segundo conjunto de fragmentos de una pluralidad de sujetos de entrenamiento con un segundo tipo de cáncer, en donde tanto el primer conjunto de fragmentos como el segundo conjunto de fragmentos comprenden una pluralidad de fragmentos de entrenamiento.

En algunos escenarios, el modelo comprende uno de un clasificador de regresión logística de núcleo, un clasificador de bosque aleatorio, un modelo de mezcla, una red neuronal convolucional y un modelo de autocodificador. En algunos escenarios, el conjunto de lecturas de secuencia se obtiene usando un panel de ensayo proporcionado anteriormente.

Breve descripción de los dibujos

Las características de la descripción se exponen con particularidad en las reivindicaciones adjuntas. Se obtendrá una mejor comprensión de las características y ventajas de la presente descripción mediante referencia a la siguiente descripción detallada que establece escenarios ilustrativos, en los que se utilizan los principios de la descripción, y los dibujos adjuntos de los cuales:

la Figura 1A ilustra un diseño de sonda en forma de dos títulos, con tres sondas dirigidas a una región diana pequeña, donde cada base en una región diana (recuadrada en el rectángulo punteado) está cubierta por al menos dos sondas, según un escenario.

La Figura 1B ilustra un diseño de sonda en forma de dos títulos, con más de tres sondas dirigidas a una región diana más grande, donde cada base en una región diana (recuadrada en el rectángulo punteado) está cubierta por al menos dos sondas, según un escenario.

La Figura 1C ilustra el diseño de sondas dirigidas a fragmentos hipometilados y/o hipermetilados en regiones genómicas, según un escenario.

La Figura 2 ilustra un proceso para generar un panel de ensayo de cáncer, según un escenario.

La Figura 3A es un diagrama de flujo que describe un proceso para crear una estructura de datos para un grupo de control, según un escenario.

La Figura 3B es un diagrama de flujo que describe una etapa adicional de validar la estructura de datos para el grupo de control de **la Figura 3A,** según un escenario.

La Figura 4 es un diagrama de flujo que describe un proceso para seleccionar regiones genómicas para diseñar sondas para un panel de ensayo de cáncer, según un escenario.

La Figura 5 es una ilustración de un cálculo de puntuación de valor de p de ejemplo, según un escenario.

La Figura 6A es un diagrama de flujo que describe un proceso de entrenamiento de un clasificador basándose en fragmentos hipometilados e hipermetilados indicativos de cáncer, según un escenario.

La Figura 6B es un diagrama de flujo que describe un proceso de identificación de fragmentos indicativos del cáncer determinado por modelos probabilísticos, según un escenario.

La Figura 7A es un diagrama de flujo que describe un proceso de secuenciación de un fragmento de ADN (cf) libre de células, según un escenario.

La Figura 7B es una ilustración del proceso de **la Figura 7A** de secuenciación de un fragmento de ADN libre de células (lc) para obtener un vector de estado de metilación, según un escenario.

La Figura 8 ilustra el grado de conversión con bisulfito (panel superior) y la profundidad media de cobertura/secuenciación (panel inferior) en diferentes estadios de cáncer.

La Figura 9 ilustra la concentración de ADNlc por muestra en diferentes estadios de cáncer.

La Figura 10 es un gráfico de las cantidades de fragmentos de ADN que se unen a sondas dependiendo de los tamaños de solapamiento entre los fragmentos de ADN y las sondas.

La Figura 11A resume las frecuencias de las anotaciones genómicas de las regiones genómicas dirigidas de la lista 1 (negro) y las regiones genómicas seleccionadas aleatoriamente (gris). **La Figura 11B** resume las frecuencias de las anotaciones genómicas de las regiones genómicas dirigidas de la lista 2 (negro) y las regiones genómicas seleccionadas aleatoriamente (gris). **La Figura 11C** resume las frecuencias de las anotaciones genómicas de las regiones genómicas dirigidas de la lista 3 (negro) y las regiones genómicas seleccionadas aleatoriamente (gris).

La Figura 12A ilustra un diagrama de flujo de dispositivos para secuenciar muestras de ácido nucleico según un escenario. **La Figura 12B** ilustra un sistema analítico que analiza el estado de metilación del ADNlc según un escenario.

La Figura 13 es una matriz sombreada que presenta números de regiones genómicas seleccionadas para diferenciar cada TOO objetivo (eje x) de un TOO de contraste (eje y).

La Figura 14 proporciona datos para verificar regiones genómicas seleccionadas mediante el uso de ADNlc y ADNg de WBC. Las fracciones (eje y) clasifican correctamente cada TOO (eje x).

La Figura 15A representa una curva del operador receptor (ROC) que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 4. **La Figura 15B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 4.

La Figura 16A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 5. **La Figura 16B** ilustra el tipo de cáncer real frente al tipo de cáncer previsto usando un clasificador generado con las regiones genómicas de la lista 5.

La Figura 17A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 6. **La Figura 17B** ilustra el tipo de cáncer real frente al tipo de cáncer previsto usando un clasificador generado con las regiones genómicas de la lista 6.

La Figura 18A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 7. **La Figura 18B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 7.

La Figura 19A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 8. **La Figura 19B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 8.

La Figura 20A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 9. **La Figura 20B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 9.

La Figura 21A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 10. **La Figura 21B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 10.

La Figura 22A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 11. **La Figura 22B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 11.

La Figura 23A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 12. **La Figura 23B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 12.

La Figura 24A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 13. **La Figura 24B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 13.

La Figura 25A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 14. **La Figura 25B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 14.

La Figura 26A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 15. **La Figura 26B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 15.

La Figura 27A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para las regiones genómicas diana de la lista 16. **La Figura 27B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para el objetivo de la lista 16.

La Figura 28A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para un subconjunto seleccionado aleatoriamente del 10 % de las regiones genómicas diana de la lista 12. **La Figura 28B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para un subconjunto seleccionado aleatoriamente del 10 % de las regiones genómicas diana de la lista 12.

La Figura 29A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer mediante el uso de datos de metilación para un subconjunto seleccionado aleatoriamente del 25 % de las regiones genómicas diana de la lista 12. **La Figura 29B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para un subconjunto seleccionado aleatoriamente del 25 % de las regiones genómicas diana de la lista 12.

La Figura 30A representa una ROC que muestra la sensibilidad y especificidad de la detección del cáncer usando datos de metilación para un subconjunto seleccionado aleatoriamente del 50 % de las regiones genómicas diana de la lista 4. **La Figura 30B** es una matriz de confusión que representa la precisión de las clasificaciones del tipo de cáncer para sujetos que se determina que tienen cáncer mediante el uso de datos de metilación para un subconjunto seleccionado aleatoriamente del 50 % de las regiones genómicas diana de la lista 4.

Descripción detallada

Definiciones

A menos que se defina lo contrario, todos los términos técnicos y/o científicos usados en la presente memoria tienen el mismo significado que entiende comúnmente un experto en la técnica a la que pertenece la invención. Como se usa en la presente memoria, los siguientes términos tienen los significados atribuidos a continuación.

Como se usa en la presente memoria, cualquier referencia a "un escenario" significa que un elemento, característica, estructura o característica particular descrita en relación con el escenario se incluye en al menos un escenario. Las apariciones de la frase "en un escenario" en varios lugares de la memoria descriptiva no se refieren necesariamente al mismo escenario, proporcionando de este modo un marco para diversas posibilidades de las situaciones descritas en el escenario.

Como se usa en la presente memoria, los términos "comprende", "que comprende", "incluye", "que incluye", "tiene", "que tiene" o cualquier otra variación de los mismos, pretenden cubrir una inclusión no exclusiva. Por ejemplo, un proceso, método, artículo o aparato que comprende una lista de elementos no se limita necesariamente a solo esos elementos, sino que puede incluir otros elementos no expresamente enumerados o inherentes a dicho proceso, método, artículo o aparato. Además, a menos que se indique expresamente lo contrario, "o" se refiere a un o inclusivo y no a un o exclusivo. Por ejemplo, una condición A o B se satisface por uno cualquiera de los siguientes: A es verdadero (o presente) y B es falso (o no está presente), A es falso (o no está presente) y B es verdadero (o presente), y tanto A como B son verdaderos (o presentes).

Además, el uso de "un" o "una" se emplea para describir elementos y componentes de los escenarios en la presente memoria. Esto se hace simplemente por conveniencia y para dar un sentido general de la descripción. Esta descripción debe leerse para incluir uno o al menos uno y el singular también incluye el plural a menos que sea obvio que se pretende de otro modo.

Como se usa en la presente memoria, los intervalos y cantidades pueden expresarse como "aproximadamente" un valor o intervalo particular. También incluye la cantidad exacta. Por tanto, "aproximadamente 5 µg" significa "aproximadamente 5 µg" y también "5 µg." Generalmente, el término "aproximadamente" incluye una cantidad que se esperaría que esté dentro del error experimental. En algunas realizaciones, "aproximadamente" se refiere al número o valor mencionado, "+" o "-" 20 %, 10 % o 5 % del número o valor. Además, se entiende que los intervalos

citados en la presente memoria son abreviados para todos los valores dentro del intervalo, incluidos los puntos finales enumerados. Por ejemplo, se entiende que un intervalo de 1 a 50 incluye cualquier número, combinación de números o subintervalo del grupo que consiste en 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 40, 41, 42, 43, 44, 40, 41, 42, 43, 49 y 50.

El término "metilación", como se usa en la presente memoria, se refiere a un proceso mediante el cual se añade un grupo metilo a una molécula de ADN. Por ejemplo, un átomo de hidrógeno en el anillo de pirimidina de una base de citosina se puede convertir en un grupo metilo, formando 5-metilcitosina. El término también se refiere a un proceso por el que se añade un grupo hidroximetilo a una molécula de ADN, por ejemplo mediante oxidación de un grupo metilo en el anillo de pirimidina de una base de citosina. La metilación y la hidrometilación tienden a producirse a dinucleótidos de citosina y guanina a los que se hace referencia en la presente memoria como "sitios CpG"

El término "metilación" también puede referirse al estado de metilación de un sitio CpG. Un sitio CpG con un resto 5-metilcitosina está metilado. Un sitio CpG con un átomo de hidrógeno en el anillo de pirimidina de la base de citosina no está metilado.

Si también se cubre el estado de metilación en un sitio, es decir, presencia o ausencia de un grupo metilo. Cuando la presencia de un grupo metilo es un sitio metilado/ausencia de un grupo metilo es un sitio no metilado o un sitio no metilado.

En tales escenarios, el ensayo de laboratorio húmedo utilizado para detectar la metilación puede variar de los descritos en la presente memoria como es bien conocido en la técnica.

El término "sitio de metilación" como se usa en la presente memoria se refiere a una región de una molécula de ADN donde se puede añadir un grupo metilo. Los sitios "CpG" son el sitio de metilación más común, pero los sitios de metilación no se limitan a sitios CpG. Por ejemplo, la metilación del ADN puede producirse en citosinas en CHG y CHH, donde H es adenina, citosina o timina. La metilación de citosina en forma de 5-hidroximetilcitosina también puede evaluarse (véase, por ejemplo, los documentos WO 2010/037001 y WO 2011/127136), y sus características, mediante el uso de los métodos y procedimientos descritos en la presente memoria.

El término "sitio CpG" como se usa en la presente memoria se refiere a una región de una molécula de ADN donde un nucleótido de citosina es seguido por un nucleótido de guanina en la secuencia lineal de bases a lo largo de su dirección 5' a 3'. "CpG" es una abreviatura de 5'-C-fosfato-G-3', es decir, citosina y guanina separadas por un solo grupo fosfato. Las citosinas en los dinucleótidos CpG se pueden metilar para formar 5-metilcitosina.

El término "sitio de detección CpG" como se usa en la presente memoria se refiere a una región en una sonda que está configurada para hibridarse con un sitio CpG de una molécula de ADN diana. El sitio CpG en la molécula de ADN diana puede comprender citosina y guanina separadas por un grupo fosfato, donde citosina está metilada o no metilada. El sitio CpG en la molécula de ADN diana puede comprender uracilo y guanina separados por un grupo fosfato, donde el uracilo se genera por la conversión de citosina no metilada.

El término "UpG" es una abreviatura de 5'-U-fosfato-G-3', es decir, uracilo y guanina separados por un solo grupo fosfato. UpG puede generarse mediante un tratamiento con bisulfito de un ADN que convierte las citosinas no metiladas en uracilos. Las citosinas pueden convertirse en uracilos por otros métodos conocidos en la técnica, como la modificación química, la síntesis o la conversión enzimática.

Los términos "hipometilado" o "hipermetilado" utilizados en la presente memoria se refieren al estado de metilación de una molécula de ADN que contiene múltiples sitios CpG (por ejemplo, más de 3, 4, 5, 6, 7, 8, 9, 10, etc.) en los que un alto porcentaje de los sitios CpG (por ejemplo, más del 80 %, 85 %, 90 % o 95 %, o cualquier otro porcentaje dentro del intervalo del 50 %-100 %) no están metilados o están metilados, respectivamente.

Los términos "vector de estado de metilación" o "vector de estado de metilación" como se usan en la presente memoria se refieren a un vector que comprende múltiples elementos, donde cada elemento indica el estado de metilación de un sitio de metilación en una molécula de ADN que comprende múltiples sitios de metilación, en el orden en que aparecen de 5' a 3' en la molécula de ADN. Por ejemplo, $< M_x, M_{x+1}, M_{x+2} >$, $< M_x, M_{x+1}, U_{x+2} >$, . . ., $< U_x, U_{x+1}, U_{x+2} >$ pueden ser vectores de metilación para moléculas de ADN que comprenden tres sitios de metilación, donde M representa un sitio de metilación metilado y U representa un sitio de metilación no metilado.

El término "patrón de metilación anormal" o "patrón de metilación anómalo" como se usa en la presente memoria se refiere al patrón de metilación de una molécula de ADN o un vector de estado de metilación que se espera que se encuentre en una muestra con menos frecuencia que un valor umbral en una muestra no cancerosa o sana. En un escenario particular proporcionado en la presente memoria, la expectativa de encontrar un vector de estado de metilación específico en un grupo de control sano que comprende individuos sanos está representado por un valor de p. Una puntuación de valor de p baja corresponde generalmente a un vector de estado de metilación que es relativamente inesperado en comparación con otros vectores de estado de metilación dentro de muestras de individuos sanos. Una puntuación de valor de p alta corresponde generalmente a un vector de estado de metilación que es

relativamente más esperado en comparación con otros vectores de estado de metilación que se encuentran en muestras de individuos sanos en el grupo de control sano. Un vector de estado de metilación que tiene un valor de p inferior a un valor umbral (por ejemplo, 0,1, 0,01, 0,001, 0,0001, etc.) puede definirse como un patrón de metilación anormal/anómalo. Pueden usarse diversos métodos conocidos en la técnica para calcular un valor de p o expectativa de un patrón de metilación o un vector de estado de metilación. Los métodos ilustrativos proporcionados en la presente memoria implican el uso de una probabilidad de cadena de Markov que asume que los estados de metilación de los sitios CpG dependen de los estados de metilación de los sitios CpG vecinos. Los métodos alternativos proporcionados en la presente memoria calculan la expectativa de observar un vector de estado de metilación específico en individuos sanos utilizando un modelo de mezcla que incluye múltiples componentes de mezcla, siendo cada uno un modelo de sitios independientes donde se supone que la metilación en cada sitio CpG es independiente de los estados de metilación en otros sitios CpG.

El término "muestra cancerosa" como se usa en la presente memoria se refiere a una muestra que comprende ADN genómicos de un individuo diagnosticado con cáncer. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico de un sujeto con cáncer. Los ADN genómicos pueden secuenciarse y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, secuenciación con bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (por ejemplo, The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando el genoma de un individuo diagnosticado de cáncer, la muestra cancerosa puede referirse a ADN genómico o fragmentos de ANDlc que tienen las secuencias genómicas. El término "muestras cancerosas" como un plural se refiere a muestras que comprenden ADN genómicos de múltiples individuos, cada individuo diagnosticado con cáncer. En diversos escenarios, se usan muestras cancerosas de más de 100, 300, 500, 1.000, 2.000, 5.000, 10.000, 20.000, 40.000, 50.000, o más individuos diagnosticados con cáncer.

El término "muestra no cancerosa", como se usa en la presente memoria, se refiere a una muestra que comprende ADN genómicos de un individuo no diagnosticado con cáncer. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico de un sujeto sin cáncer. Los ADN genómicos pueden secuenciarse y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, la secuenciación por bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (por ejemplo, The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando el genoma de un individuo sin cáncer, la muestra no cancerosa puede referirse a ADN genómico o fragmentos de ANDlc que tienen las secuencias genómicas. El término "muestras no cancerosas" como un plural se refiere a muestras que comprenden ADN genómicos de múltiples individuos, cada individuo está sin cáncer. En diversos escenarios, se usan muestras cancerosas de más de 100, 300, 500, 1.000, 2.000, 5.000, 10.000, 20.000, 40.000, 50.000, o más individuos sin cáncer.

El término "muestra de entrenamiento" como se usa en la presente memoria se refiere a una muestra usada para entrenar un clasificador descrito en la presente memoria y/o para seleccionar una o más regiones genómicas para la detección del cáncer o detectar un tejido canceroso de origen o tipo de célula cancerosa. Las muestras de entrenamiento pueden comprender ADN genómicos o una modificación de, de uno o más sujetos sanos y de uno o más sujetos que tienen una afección de enfermedad (por ejemplo, cáncer, un tipo específico de cáncer, una etapa específica de cáncer, etc.). Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico. Los ADN genómicos pueden secuenciarse y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, la secuenciación por bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (por ejemplo, The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando el genoma de un individuo, una muestra de entrenamiento puede referirse a ADN genómico o fragmentos de ADNlc que tengan las secuencias genómicas.

El término "muestra de prueba" como se usa en la presente memoria se refiere a una muestra de un sujeto, cuya condición de salud fue, o se prueba mediante el uso de un clasificador y/o un panel de ensayo descrito en la presente memoria. La muestra de prueba puede comprender ADN genómicos o una modificación de la misma. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico.

El término "región genómica diana", como se usa en la presente memoria, se refiere a una región en un genoma seleccionado para el análisis en muestras de prueba. Se genera un panel de ensayo con sondas diseñadas para hibridarse con fragmentos de ácido nucleico (y opcionalmente eliminar) derivados de la región genómica diana o un fragmento de la misma. Un fragmento de ácido nucleico derivado de la región genómica diana se refiere a un fragmento de ácido nucleico generado por degradación, escisión, conversión con bisulfito u otro procesamiento del ADN de la región genómica diana.

Se describen varias regiones genómicas diana según su ubicación cromosómica en el listado de secuencias presentado aquí. El ADN cromosómico es bicatenario, por lo que una región genómica diana incluye dos cadenas de ADN: uno con la secuencia proporcionada en el listado y un segundo que es un complemento inverso a la secuencia en el listado. Las sondas pueden diseñarse para hibridarse con una o ambas secuencias. Opcionalmente, las sondas se hibridan con secuencias convertidas resultantes de, por ejemplo, tratamiento con bisulfito de sodio.

El término "región genómica fuera de diana", como se usa en la presente memoria, se refiere a una región en un genoma que no se ha seleccionado para su análisis en muestras de prueba pero tiene una homología suficiente con una región genómica diana para potencialmente unirse y extraerse por una sonda diseñada para dirigirse a la región genómica diana. En un escenario, una región genómica fuera de la diana es una región genómica que se alinea con una sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 %.

Los términos "moléculas de ADN convertidas", "moléculas de ADNlc convertido" y "fragmento modificado obtenido a partir del procesamiento de las moléculas de ADNlc" se refieren a moléculas de ADN obtenidas procesando moléculas de ADN o ADNlc en una muestra con el fin de diferenciar un nucleótido metilado y un nucleótido no metilado en las moléculas de ADN o ADNlc. Por ejemplo, en un escenario, la muestra puede tratarse con ion bisulfito (por ejemplo, usando bisulfito de sodio), como es bien conocido en la técnica, para convertir citosinas no metiladas ("C") en uracilos ("U"). En otro escenario, la conversión de citosinas no metiladas en uracilos se logra mediante el uso de una reacción de conversión enzimática, por ejemplo, mediante el uso de una citidina desaminasa (tal como APOBEC). Después del tratamiento, las moléculas de ADN transformadas o las moléculas de ADNlc incluyen uracilos adicionales que no están presentes en la muestra de ADNlc original. Replicación por ADN polimerasa de una cadena de ADN que comprende un uracilo resulta en la adición de una adenina a la cadena complementaria naciente en lugar de la guanina añadida normalmente como complemento a una citosina o metilcitosina.

Los términos "ácido nucleico libre de células", "ADN libre de células" o "ADNlc" se refieren a fragmentos de ácido nucleico que circulan en el cuerpo de un individuo (por ejemplo, torrente sanguíneo) y se originan a partir de una o más células sanas y/o de una o más células cancerosas. Además, el ADNlc puede provenir de otras fuentes tales como virus, fetos, etc.

El término "ADN tumoral circulante" o "ADNtc" se refiere a fragmentos de ácido nucleico que se originan a partir de células tumorales, que pueden liberarse en el torrente sanguíneo de un individuo como resultado de procesos biológicos tales como apoptosis o necrosis de células que mueren o liberan activamente por células tumorales viables.

El término "fragmento" como se usa en la presente memoria puede referirse a un fragmento de una molécula de ácido nucleico. Por ejemplo, en un escenario, un fragmento puede referirse a una molécula de ADNlc en una muestra de sangre o plasma, o una molécula de ADNlc que se ha extraído de una muestra de sangre o plasma. Un producto de amplificación de una molécula de ADNlc también puede denominarse "fragmento". En otro escenario, el término "fragmento" se refiere a una lectura de secuencia o conjunto de lecturas de secuencia, que se han procesado para el análisis posterior (por ejemplo, para la clasificación basándose en aprendizaje automático), como se describe en la presente memoria. Por ejemplo, como se conoce bien en la técnica, las lecturas de secuencia sin procesar pueden alinearse con un genoma de referencia y las lecturas de secuencia de extremos emparejados coincidentes ensambladas en un fragmento más largo para el análisis posterior.

El término "individuo" se refiere a un individuo humano. El término "individuo sano" se refiere a un individuo que se supone que no tiene un cáncer o enfermedad.

El término "sujeto" se refiere a un individuo cuyo ADN se está analizando. Un sujeto puede ser un sujeto de prueba cuyo ADN se evalúa mediante el uso de un panel diana como se describe en la presente memoria para evaluar si la persona tiene cáncer u otra enfermedad. Un sujeto también puede ser parte de un grupo de control que se sabe que no tiene cáncer u otra enfermedad. Un sujeto también puede formar parte de un grupo con cáncer u otra enfermedad conocida. Los grupos de control y de cáncer/enfermedad pueden usarse para ayudar a diseñar o validar el panel específico.

El término "lecturas de secuencia" como se usa en la presente memoria se refiere a las secuencias de nucleótidos lecturas de una muestra. Las lecturas de secuencia se pueden obtener a través de diversos métodos proporcionados en la presente memoria o como se conoce en la técnica.

El término "profundidad de secuenciación", como se usa en la presente memoria, se refiere al recuento del número de veces que se ha secuenciado un ácido nucleico diana determinado dentro de una muestra (por ejemplo, el recuento de lecturas de secuencia en una región diana dada). El aumento de la profundidad de secuenciación puede reducir las cantidades requeridas de ácidos nucleicos necesarios para evaluar un estado de enfermedad (por ejemplo, tejido de origen cáncer o cáncer).

El término "tejido de origen" o "TOO" como se usa en la presente memoria se refiere al órgano, grupo de órganos, región de cuerpo o tipo de célula que surge un cáncer o se origina. La identificación de un tejido de origen o de células cancerosas típicamente permite la identificación de las siguientes etapas más apropiadas en el continuo de cuidado del cáncer para diagnosticar aún más, etapa y decidir el tratamiento.

El término "transición" generalmente se refiere a cambios en la composición base de una purina a otra purina, o de una pirimidina a otra pirimidina. Por ejemplo, los siguientes cambios son transiciones C→U, U→C, G→A, A→G, C→T y T→C.

"Una totalidad de las sondas" de un panel o conjunto de cebos o "una totalidad de sondas que contienen polinucleótidos" de un panel o conjunto de cebos generalmente se refiere a todas las sondas administradas con un panel específico o conjunto de cebos. Por ejemplo, en algunos escenarios, un panel o conjunto de cebos puede incluir (1) sondas que tienen características especificadas en la presente memoria (por ejemplo, sondas para unirse a fragmentos de ADN libres de células correspondientes o derivadas de regiones genómicas expuestas en la presente memoria en una o más listas) y (2) sondas adicionales que no contienen dicha(s) característica(s). La totalidad de las sondas de un panel generalmente se refiere a todas las sondas suministradas con el panel o conjunto de cebos, incluyendo tales sondas que no contienen la(s) característica(s) especificada(s).

Panel de ensayo del cáncer

En un primer escenario, la presente descripción proporciona un panel de ensayo de cáncer que comprende una pluralidad de sondas o una pluralidad de pares de sondas. Los paneles de ensayo descritos en la presente memoria pueden denominarse alternativamente conjuntos cebos o como composiciones que comprenden oligonucleótidos cebo. Las sondas pueden ser sondas que contienen polinucleótidos que se diseñan específicamente para dirigirse a una o más regiones genómicas metiladas diferencialmente entre muestras cancerosas y no cancerosas, entre diferentes tipos de tejido de cáncer de origen (TOO), entre diferentes tipos de células cancerosas, entre muestras de diferentes estadios del cáncer, como se identifica por los métodos proporcionados en la presente memoria. En algunos escenarios, las regiones genómicas diana se seleccionan para maximizar la precisión de la clasificación, sujeto a un presupuesto de tamaño (que se determina por presupuesto de secuenciación y profundidad de secuenciación deseada).

Para diseñar el panel de ensayo de cáncer, un sistema de análisis puede recoger muestras correspondientes a diversos resultados en consideración, por ejemplo, muestras que se sabe que tienen cáncer, muestras consideradas para ser sanas, muestras de un tejido de origen conocido, etc. Las fuentes del ADNlc y/o ADNct usados para seleccionar regiones genómicas diana pueden variar dependiendo del propósito del ensayo. Por ejemplo, pueden ser deseables diferentes fuentes para un ensayo destinado a detectar el cáncer generalmente, un tipo específico de cáncer, una etapa de cáncer o un tejido de origen. Estas muestras pueden procesarse mediante secuenciación con bisulfito de genoma completo (WGBS) u obtenerse de una base de datos pública (por ejemplo, TCGA). El sistema de análisis puede ser cualquier sistema informático genérico con un procesador de ordenador y un medio de almacenamiento legible por ordenador con instrucciones para ejecutar el procesador de ordenador para realizar cualquiera o todas las operaciones descritas en esta presente descripción.

El sistema de análisis puede seleccionar regiones genómicas diana basadas en patrones de metilación de fragmentos de ácido nucleico. Un enfoque considera la capacidad de distinción por pares entre pares de resultados para regiones (o más específicamente para sitios CpG dentro de regiones). Otro enfoque considera la capacidad de distinción de las regiones (o más específicamente para los sitios CpG dentro de las regiones) cuando se considera cada resultado contra los resultados restantes. A partir de las regiones genómicas diana seleccionadas con alta potencia de distinción, el sistema de análisis puede diseñar sondas para dirigirse a fragmentos de las regiones genómicas seleccionadas. El sistema de análisis puede generar tamaños variables del panel de ensayo de cáncer, por ejemplo, donde un panel de ensayo de cáncer de pequeño tamaño incluye sondas dirigidas a las regiones genómicas más informativas, un panel de ensayo de cáncer de tamaño medio incluye sondas del panel de ensayo de cáncer de pequeño tamaño y sondas adicionales dirigidas a un segundo nivel de regiones genómicas informativas, y un panel de ensayo de cáncer de gran tamaño incluye sondas de los paneles de ensayo de cáncer de tamaño pequeño y de tamaño medio junto con incluso más sondas dirigidas a un tercer nivel de regiones genómicas informativas. Con los datos obtenidos tales paneles de ensayo de cáncer (por ejemplo, el estado de metilación en ácidos nucleicos derivados de los paneles de ensayo de cáncer), el sistema de análisis puede entrenar clasificadores con diversas técnicas de clasificación para predecir la probabilidad de una muestra de tener un resultado o estado particular, por ejemplo, cáncer, tipo específico de cáncer, otro trastorno, otra enfermedad, etc.

La metodología ilustrativa para diseñar un panel de ensayo de cáncer se describe generalmente en la **Figura 2.** Por ejemplo, para diseñar un panel de ensayo de cáncer, un sistema de análisis puede recoger información sobre el estado de metilación de los sitios CpG de fragmentos de ácido nucleico de muestras correspondientes a diversos resultados en consideración, por ejemplo, muestras que se sabe que tienen cáncer, muestras consideradas para ser sanas, muestras de un TOO conocido, etc. Estas muestras pueden procesarse (por ejemplo, con secuenciación con bisulfito de genoma completo (WGBS) para determinar el estado de metilación de los sitios CpG, o la información puede obtenerse de TCGA. El sistema de análisis puede ser cualquier sistema informático genérico con un procesador de ordenador y un medio de almacenamiento legible por ordenador con instrucciones para ejecutar el procesador de ordenador para realizar cualquiera o todas las operaciones descritas en esta presente descripción.

El sistema de análisis puede seleccionar regiones genómicas diana basadas en patrones de metilación de fragmentos de ácido nucleico. Un enfoque considera la capacidad de distinción por pares entre pares de resultados para regiones (o más específicamente sitios CpG). Otro enfoque considera la capacidad de distinción de las regiones (o más específicamente sitios CpG) cuando se considera cada resultado contra los resultados restantes. A partir de las regiones genómicas diana seleccionadas con alta potencia de distinción, el sistema de análisis puede diseñar sondas para dirigirse a fragmentos de las regiones genómicas seleccionadas. El sistema de análisis puede generar tamaños

variables del panel de ensayo de cáncer, por ejemplo, donde un panel de ensayo de cáncer de pequeño tamaño incluye sondas dirigidas a las regiones genómicas más informativas, un panel de ensayo de cáncer de tamaño medio incluye sondas del panel de ensayo de cáncer de pequeño tamaño y sondas adicionales dirigidas a un segundo nivel de regiones genómicas informativas, y un panel de ensayo de cáncer de gran tamaño incluye sondas de los paneles de ensayo de cáncer de tamaño pequeño y de tamaño medio junto con incluso más sondas dirigidas a un tercer nivel de regiones genómicas informativas. Con tales paneles de ensayo de cáncer, el sistema de análisis puede entrenar clasificadores con diversas técnicas de clasificación para predecir la probabilidad de una muestra de tener un resultado o estado particular, por ejemplo, cáncer, tipo específico de cáncer, otro trastorno, otra enfermedad, etc.

En algunos escenarios, el panel de ensayo del cáncer comprende al menos 500 pares de sondas, en donde cada par de al menos 500 pares comprende dos sondas configuradas para superponerse entre sí mediante una secuencia superpuesta, en donde la secuencia superpuesta comprende al menos 30 nucleótidos, y en donde cada sonda está configurada para hibridarse con la misma cadena de una molécula de ADN (opcionalmente convertida) (por ejemplo, una molécula de ADNlc) correspondiente a una o más regiones genómicas. En algunos escenarios, cada una de las regiones genómicas comprende al menos cinco sitios de metilación, y en donde los al menos cinco sitios de metilación tienen un patrón de metilación anormal en muestras cancerosas o un patrón de metilación diferente entre muestras de un TOO diferente. Por ejemplo, en un escenario, los al menos cinco sitios de metilación se metilan diferencialmente entre muestras cancerosas y no cancerosas o entre uno o más pares de muestras de cánceres con diferentes tejidos de origen. En algunos escenarios, cada par de sondas comprende una primera sonda y una segunda sonda, en donde la segunda sonda difiere de la primera sonda. La segunda sonda puede solaparse con la primera sonda mediante una secuencia superpuesta que es al menos 30, al menos 40, al menos 50 o al menos 60 nucleótidos de longitud.

Las regiones genómicas diana pueden seleccionarse de una cualquiera de las listas 1-16 (**tabla 1**). En algunos escenarios, el panel de ensayo de cáncer comprende una pluralidad de sondas, en donde cada una de la pluralidad de sondas está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a una o más de las regiones genómicas en una cualquiera de las listas 1-16. En algunos escenarios, la pluralidad de oligonucleótidos cebo diferentes se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 1-16. En algunos escenarios, la pluralidad de oligonucleótidos cebo diferentes se configura para hibridarse con moléculas de ADN derivadas de al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 % o el 80 % de las regiones genómicas diana de una cualquiera de las listas 1-16.

Las regiones genómicas diana pueden seleccionarse de la lista 1. Las regiones genómicas diana pueden seleccionarse de la lista 2. Las regiones genómicas diana pueden seleccionarse de la lista 3. Las regiones genómicas diana pueden seleccionarse de la lista 4. Las regiones genómicas diana pueden seleccionarse de la lista 5. Las regiones genómicas diana pueden seleccionarse de la lista 6. Las regiones genómicas diana pueden seleccionarse de la lista 7. Las regiones genómicas diana pueden seleccionarse de la lista 8. Las regiones genómicas diana pueden seleccionarse de la lista 9. Las regiones genómicas diana pueden seleccionarse de la lista 10. Las regiones genómicas diana pueden seleccionarse de la lista 11. Las regiones genómicas diana pueden seleccionarse de la lista 12. Las regiones genómicas diana pueden seleccionarse de la lista 13. Las regiones genómicas diana pueden seleccionarse de la lista 14. Las regiones genómicas diana pueden seleccionarse de la lista 15. Las regiones genómicas diana pueden seleccionarse de la lista 16.

Dado que las sondas están configuradas para hibridarse con una molécula de ADN o ADNlc convertido correspondiente a, o derivada de, una o más regiones genómicas, las sondas pueden tener una secuencia diferente de la región genómica diana. Por ejemplo, un ADN que contiene un sitio CpG no metilado se convertirá para incluir UpG en lugar de CpG porque las citosinas no metiladas se convierten en uracilos mediante una reacción de conversión (por ejemplo, tratamiento con bisulfito). Como resultado, una sonda se configura para hibridarse con una secuencia que incluye UpG en lugar de un CpG no metilado existente natural. Por consiguiente, un sitio complementario en la sonda al sitio no metilado puede comprender CpA en lugar de CpG, y algunas sondas dirigidas a un sitio hipometilado donde todos los sitios de metilación no son metilados pueden no tener bases de guanina (G). En algunos escenarios, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no comprenden secuencias CpG.

El panel de ensayo de cáncer puede usarse para detectar la presencia o ausencia de cáncer generalmente y/o proporcionar una clasificación de cáncer tal como el tipo de cáncer, la etapa de cáncer tal como I, II, III o IV, o proporcionar el TOO donde se cree que el cáncer se origina. El panel puede incluir sondas dirigidas a regiones genómicas que se metilan diferencialmente entre muestras generales cancerosas (pan-cancerosas) y muestras no cancerosas, o solo en muestras cancerosas con un tipo específico de cáncer (por ejemplo, dianas específicas de cáncer de pulmón). Por ejemplo, en algunos escenarios, un panel de ensayo de cáncer está diseñado para incluir regiones genómicas metiladas diferencialmente basadas en datos de secuenciación convertidos (por ejemplo, bisulfito) generados a partir del ADNlc de individuos con cáncer y sin cáncer.

Cada una de las sondas (o pares de sondas) puede diseñarse para dirigirse a una o más regiones genómicas diana. Las regiones genómicas diana se pueden seleccionar basándose en varios criterios diseñados para aumentar el enriquecimiento selectivo de fragmentos de ADNlc informativos mientras se reducen el ruido y las uniones no específicas.

En un ejemplo, un panel puede incluir sondas que pueden unirse selectivamente y enriquecer fragmentos de ADNlc que están diferencialmente metilados en muestras cancerosas. En este caso, la secuenciación de los fragmentos enriquecidos puede proporcionar información relevante para la detección del cáncer. Además, en algunos escenarios, las sondas (o una porción de la misma) están diseñadas para dirigirse a regiones genómicas que se determina que tienen un patrón de metilación anormal en muestras de cáncer, o en muestras de ciertos tipos de cáncer, tipos de tejido o tipos de células. En un escenario, las sondas se diseñan para dirigirse a regiones genómicas determinadas para ser hipermetiladas o hipometiladas en ciertos cánceres o tipos de cáncer para proporcionar selectividad y especificidad adicionales de la detección. En algunos escenarios, un panel comprende sondas dirigidas a fragmentos hipometilados. En algunos escenarios, un panel comprende sondas dirigidas a fragmentos hipermetilados. En algunos escenarios, un panel comprende tanto un primer conjunto de sondas que se dirigen a fragmentos hipermetilados como un segundo conjunto de sondas dirigidas a fragmentos hipometilados. En algunos escenarios, un panel de ensayo de cáncer incluye no solo sondas diseñadas para dirigirse a una región que tiene un primer estado de metilación (por ejemplo, hipometilación), pero también incluye sondas que están diseñadas para hibridarse con la misma región diana con el estado de metilación opuesto (por ejemplo, hipermetilación). El direccionamiento de las sondas a fragmentos hipo- y hipermetilados de las mismas regiones puede denominarse direccionamiento "binario" (véase información en la lista de secuencias) (**Figura 1C**). En algunos escenarios, la razón entre el primer conjunto de sondas dirigidas a fragmentos hipermetilados y el segundo conjunto de sondas dirigidas a fragmentos hipometilados (razón HyperHypo) oscila entre 0,4 y 2, entre 0,5 y 1,8, entre 0,5 y 1,6, entre 0,5 y 1,0, entre 1,4 y 1,6, entre 1,2 y 1,4, entre 1 y 1,2, entre 0,8 y 1, entre 0,6 y 0,8 o entre 0,4 y 0,6. Los métodos para identificar regiones genómicas (es decir, regiones genómicas que dan lugar a moléculas de ADN metiladas diferencialmente (o moléculas de ADN débilmente metiladas) entre muestras cancerosas y no cancerosas, entre diferentes tipos de tejido de cáncer de origen (TOO), entre diferentes tipos de células cancerosas, o entre muestras de diferentes estadios del cáncer se proporcionan en detalle en la presente memoria y los métodos para identificar moléculas de ADN débilmente metiladas o fragmentos que se identifican como indicativos de cáncer también se proporcionan en detalle en la presente memoria.

En un segundo ejemplo, las regiones genómicas pueden seleccionarse cuando las regiones genómicas dan lugar a moléculas de ADN débilmente metiladas en muestras de cáncer o muestras con tipos conocidos de tejido canceroso (TOO). Por ejemplo, como se describe en la presente memoria, se puede usar un modelo de Markov entrenado en un conjunto de muestras no cancerosas para identificar regiones genómicas que dan lugar a moléculas de ADN anormalmente metiladas (es decir, moléculas de ADN que tienen un patrón de metilación por debajo de un umbral de valor de p).

Cada una de las sondas puede dirigirse a una región genómica que comprende al menos 30 pb, 35 pb, 40 pb, 45 pb, 50 pb, 60 pb, 70 pb, 80 pb, 90 pb, 100 pb o más. En algunos escenarios, las regiones genómicas pueden seleccionarse para tener menos de 30, 25, 20, 15, 12, 10, 8 o 6 sitios de metilación.

En algunos casos, las regiones genómicas pueden seleccionarse cuando al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los al menos cinco sitios de metilación (por ejemplo, CpG) dentro de la región están metilados o no metilados en muestras no cancerosas o cancerosas, o en muestras de cáncer de un tejido de origen (TOO).

Las regiones genómicas pueden filtrarse adicionalmente para seleccionar solo aquellas que probablemente sean informativas basándose en sus patrones de metilación, por ejemplo, sitios CpG que están metilados diferencialmente entre muestras cancerosas y no cancerosas (por ejemplo, anormalmente metilados o no metilados en cáncer frente a no cáncer), entre muestras cancerosas de una TOO y muestras cancerosas de un TOO diferente, sitios CpG que se metilan diferencialmente solo en muestras cancerosas de un TOO. Para la selección, el cálculo se puede realizar con respecto a cada CpG o una pluralidad de sitios CpG. Por ejemplo, se puede determinar un primer recuento que es el número de muestras que contienen cáncer (recuento_cáncer) que incluyen un fragmento que se solapa con ese CpG, y se determina un segundo recuento que es el número de muestras totales que contienen fragmentos que se solapan con ese sitio CpG (total). Las regiones genómicas pueden seleccionarse basándose en criterios correlacionados positivamente con el número de muestras que contienen cáncer (recuento_cáncer) que incluyen un fragmento indicativo de cáncer superpuesto a ese sitio CpG, e inversamente correlacionados con el número total de muestras que contienen fragmentos indicativos de cáncer superpuestos a ese sitio CpG (total). En un escenario, se cuenta el número de muestras no cancerosas (no-cáncer) y el número de muestras cancerosas (cáncer) que tienen un fragmento superpuesto a un sitio CpG. A continuación, se calcula la probabilidad de que una muestra sea cancerosa, por ejemplo como (cáncer + 1) / (cáncer + no-cáncer + 2). Este principio podría aplicarse de manera similar a otros resultados.

Los sitios CpG puntuados por esta métrica se clasifican y se añaden suavemente a un panel hasta que el presupuesto del tamaño del panel se agota. El proceso de selección de regiones genómicas indicativas del cáncer se detalla adicionalmente en la presente memoria. En algunos escenarios, pueden seleccionarse diferentes regiones diana dependiendo de si el ensayo está destinado a ser un ensayo pan-cáncer o un ensayo de cáncer único, o dependiendo de qué tipo de flexibilidad se desea a la hora de elegir qué sitios CpG contribuyen al panel. Un panel para detectar un tipo específico de cáncer puede diseñarse usando un proceso similar. En este escenario, para cada tipo de cáncer, y para cada sitio CpG, la ganancia de información se calcula para determinar si incluir una sonda que se dirige a ese sitio CpG. La ganancia de información se puede calcular para muestras con un tipo de cáncer dado de un TOO en comparación con todas las demás muestras. Por ejemplo, considerar dos variables aleatorias, "AF" y "CT". "AF" es

una variable binaria que indica si hay un fragmento anormal que se superpone a un sitio CpG particular en una muestra particular (sí o no). "CT" es una variable aleatoria binaria que indica si el cáncer es de un tipo particular (por ejemplo, cáncer de pulmón o cáncer distinto de pulmón). Se puede calcular la información mutua con respecto a "CT" dada "AF". Es decir, cuántos bits de información sobre el tipo de cáncer (pulmonar frente a no pulmonar en el ejemplo) se obtienen si se sabe si hay un fragmento anómalo que se superpone a un sitio CpG particular. Esto puede usarse para clasificar CpG en base a cómo son específicos de pulmón. Este procedimiento se repite para una pluralidad de tipos de cáncer. Si una región particular está comúnmente metilada diferencialmente solo en el cáncer de pulmón (y no otros tipos de cáncer o no cáncer), CpG en esa región tenderían a tener altas ganancias de información para el cáncer de pulmón. Para cada tipo de cáncer, los sitios CpG se clasifican por esta métrica de ganancia de información y luego se añaden en gris a un panel hasta que el presupuesto de tamaño para ese tipo de cáncer se agota.

Se puede realizar una filtración adicional para seleccionar sondas con alta especificidad para el enriquecimiento (es decir, alta eficiencia de unión) de ácidos nucleicos derivados de regiones genómicas específicas. Las sondas pueden filtrarse para reducir la unión inespecífica (o unión fuera de la diana) a ácidos nucleicos derivados de regiones genómicas no dirigidas. Por ejemplo, las sondas pueden filtrarse para seleccionar solo aquellas sondas que tengan menos de un umbral establecido de acontecimientos de unión fuera de la diana. En un escenario, las sondas pueden alinearse con un genoma de referencia (por ejemplo, un genoma de referencia humana) para seleccionar sondas que se alinean a menos de un umbral establecido de regiones a través del genoma. Por ejemplo, se pueden seleccionar sondas que se alinean a menos de 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9 u 8 regiones fuera de la diana a través del genoma de referencia. En otros casos, la filtración se realiza para eliminar las regiones genómicas cuando la secuencia de las regiones genómicas diana aparece más de 5, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 o 35 veces en un genoma. Puede realizarse un filtrado adicional para seleccionar regiones genómicas diana cuando una secuencia de sonda, o un conjunto de secuencias de sonda que son homólogas en el 90 %, el 91 %, el 92 %, el 93 %, el 94 %, el 95 %, el 96 %, el 97 %, el 98 % o el 99 % a las regiones genómicas diana, aparecen menos de 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9 u 8 veces en un genoma de referencia, o para eliminar regiones genómicas diana cuando la secuencia de la sonda, o un conjunto de secuencias de la sonda diseñadas para enriquecer la región genómica diana son homólogas en el 90 %, el 91 %, el 92 %, el 93 %, el 94 %, el 95 %, el 96 %, el 97 %, el 98 % o el 99 % a las regiones genómicas diana, aparecen más de 5, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 o 35 veces en un genoma de referencia. Esto es para excluir sondas repetitivas que pueden tirar de fragmentos fuera de diana, que no son deseables y pueden afectar la eficiencia del ensayo.

En algunos escenarios, se demostró que un solapamiento de sonda-sonda de al menos 45 pb es eficaz para lograr una cantidad no despreciable de extracción (aunque un experto en la técnica apreciaría este número puede muy) como se proporciona en el ejemplo 1. En algunos escenarios, más de un 10 % de tasa de emparejamiento erróneo entre la sonda y las secuencias de fragmentos en la región de solapamiento es suficiente para interrumpir en gran medida la unión y, por tanto, la eficiencia de la extracción. Por tanto, las secuencias que pueden alinearse con la sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 % pueden ser candidatas para la extracción fuera de la diana. Por tanto, en un escenario, se puntúa el número de tales regiones. Las mejores sondas tienen una puntuación de 1, lo que significa que coinciden en un solo lugar (la región diana prevista). Las sondas con una puntuación intermedia (digamos, menos de 5 o 10) pueden aceptarse en algunos casos y, en algunos casos, se descartan cualquier sonda por encima de una puntuación particular. Pueden usarse otros valores de corte para muestras específicas.

Una vez que las sondas hibridan y capturan fragmentos de ADN correspondientes a, o derivados de, una región genómica diana, los productos intermedios de fragmento de ADN-sonda hibridado se extraen (o aíslan), y el ADN diana se amplifica y se secuencia. La lectura de secuencia proporciona información relevante para la detección del cáncer. Para este fin, un panel puede diseñarse para incluir una pluralidad de sondas que pueden capturar fragmentos que juntos pueden proporcionar información relevante para la detección del cáncer. En algunos escenarios, un panel incluye al menos 500, 1.000, 2.000, 2.500, 5.000, 6.000, 7.500, 10.000, 15.000, 20.000, 25.000, 30.000, 35.000, 40.000, 50.000, 60.000, 70.000 u 80.000 pares de sondas. En otros escenarios, un panel incluye al menos 1.000, 2.000, 5.000, 10.000, 12.000, 15.000, 20.000, 30.000, 40.000, 50.000, 100.000, 200.000, 250.000, 300.000, 400.000, 500.000, 550.000, 600.000, 700.000 u 800.000 sondas. La pluralidad de sondas juntas puede comprender al menos 0,2 millones, 0,4 millones, 0,6 millones, 0,8 millones, 1 millón, 2 millones, 3 millones, 4 millones, 5 millones, 6 millones, 7 millones, 8 millones, 9 millones, 10 millones, 12 millones, 14 millones, 15 millones, 20 millones o 25 millones de nucleótidos.

Las regiones genómicas diana seleccionadas pueden ubicarse en varias posiciones en un genoma, que incluyen, pero no se limitan a, exones, intrones, regiones intergénicas y otras partes. **FIG. 11** En algunos escenarios, pueden añadirse sondas dirigidas a regiones genómicas no humanas, tales como las que dirigen las regiones genómicas virales.

En algunos casos, pueden usarse cebadores para amplificar específicamente diana/biomarcadores de interés (por ejemplo, por PCR), enriqueciendo así la muestra para dianas/biomarcadores deseados (opcionalmente sin captura de hibridación). Por ejemplo, pueden prepararse cebadores directos e inversos para cada región genómica de interés y usarse para amplificar fragmentos que corresponden a o se derivan de la región genómica deseada. Por tanto, aunque

la presente descripción paga particular atención a los paneles de ensayo de cáncer y conjuntos cebos para captura de hibridación, la descripción es lo suficientemente amplia como para abarcar otros métodos para el enriquecimiento del ADN libre de células. Por consiguiente, un experto en la materia, con el beneficio de esta descripción, reconocerá que los métodos análogos a los descritos en la presente memoria en relación con la captura de hibridación pueden lograrse alternativamente reemplazando la captura de hibridación con alguna otra estrategia de enriquecimiento, tal como la amplificación por PCR de fragmentos de ADN libres de células que se corresponden con regiones genómicas de interés. En algunos escenarios, la captura de sonda de candado con bisulfito se usa para enriquecer regiones de interés, tal como se describe en Zhang y col. (documento US 2016/0340740). En algunos escenarios, se usan métodos adicionales o alternativos para el enriquecimiento (por ejemplo, enriquecimiento no dirigido) tal como secuenciación de bisulfito de representación reducida, secuenciación de enzimas de restricción de metilación, secuenciación de inmunoprecipitación de ADN de metilación, secuenciación de proteínas de dominio de unión a metil-CpG, secuenciación de captura de ADN de metilo o PCR de microgotas.

Sondas

Los paneles de ensayo de cáncer (alternativamente denominados "conjuntos cebos") proporcionados en la presente memoria pueden ser un panel que incluye un conjunto de sondas de hibridación (también denominadas en la presente memoria "sondas") diseñadas para, durante el enriquecimiento, diana y eliminar fragmentos de ácido nucleico de interés para el ensayo. En algunos escenarios, las sondas están diseñadas para hibridarse y enriquecer moléculas de ADN o ADNlc de muestras cancerosas que han sido tratadas para convertir citosinas (C) no metiladas en uracilos (U). En otros escenarios, las sondas están diseñadas para hibridarse y enriquecer moléculas de ADN o ADNlc de muestras cancerosas de un TOO (o una pluralidad de TOO) que han sido tratadas para convertir citosinas (C) no metiladas en uracilos (U). Las sondas pueden diseñarse para hibridarse (o hibridarse) con una cadena diana (complementaria) de ADN o ARN. La cadena diana puede ser la cadena "positiva" (por ejemplo, la cadena transcrita en el ARNm y posteriormente traducida en una proteína) o la cadena "negativa" complementaria. En un escenario particular, un panel de ensayo de cáncer puede incluir conjuntos de dos sondas, una sonda dirigida a la cadena positiva y la otra sonda dirigida a la cadena negativa de una región genómica diana.

Para cada región genómica diana, se pueden diseñar al menos cuatro posibles secuencias de sonda. Cada región diana es bicatenaria, y como tal, una sonda o conjunto de sonda puede dirigirse a la cadena "positiva" o hacia adelante o su complemento inverso (la cadena "negativa"). Además, en algunos escenarios, las sondas o conjuntos de sondas se diseñan para enriquecer moléculas de ADN o fragmentos que se han tratado para convertir citosinas (C) no metiladas en uracilos (U). Dado que las sondas o conjuntos de sondas están diseñados para enriquecer moléculas de ADN correspondientes a las regiones diana o derivadas de ellas tras la conversión, la secuencia de la sonda puede diseñarse para enriquecer moléculas de ADN de fragmentos en los que las C no metiladas se han convertido en U (usando A en lugar de G en sitios que son citosinas no metiladas en moléculas o fragmentos de ADN correspondientes a la región diana o derivadas de ella). En un escenario, las sondas están diseñadas para unirse, o hibridarse con, moléculas de ADN o fragmentos de regiones genómicas que se sabe que contienen patrones de metilación específicos del cáncer (por ejemplo, moléculas de ADN hipermetiladas o hipometiladas), enriqueciendo así las moléculas o fragmentos de ADN específicos del cáncer. Las regiones genómicas específicas, o los patrones de metilación específicos del cáncer, pueden ser ventajosas, lo que permite enriquecer específicamente moléculas de ADN o fragmentos identificados como informativos para el cáncer o el cáncer TOO, y por lo tanto, disminuir las necesidades de secuenciación y los costes de secuenciación. En otras realizaciones, se pueden diseñar dos secuencias sonda por una región genómica diana (una para cada cadena de ADN). En todavía otros casos, las sondas están diseñadas para enriquecer todas las moléculas de ADN o fragmentos correspondientes a, o derivadas de, una región dirigida (es decir, independientemente del estado de cadena o de metilación). Esto podría deberse a que el estado de metilación del cáncer no está altamente metilado o no metilado, o porque las sondas están diseñadas para dirigirse a mutaciones pequeñas u otras variaciones en lugar de cambios de metilación, con estas otras variaciones de manera similar indicativas de la presencia o ausencia de un cáncer o la presencia o ausencia de un cáncer de uno o más TOO. En ese caso, las cuatro secuencias de sonda posibles pueden incluirse por una región genómica diana.

Las sondas pueden variar en longitud de 10 s, 100 s, 200 s o 300 s de pares de bases. Las sondas pueden comprender al menos 50, 75, 100 o 120 nucleótidos. Las sondas pueden comprender menos de 300, 250, 200 o 150 nucleótidos. En un escenario, las sondas comprenden 100-150 nucleótidos. En un escenario particular, las sondas comprenden 120 nucleótidos.

En algunos escenarios, las sondas están diseñadas en una forma "de dos títulos" para cubrir porciones superpuestas de una región diana. Cada sonda se solapa opcionalmente en cobertura al menos parcialmente con otra sonda en la biblioteca. En tales escenarios, el panel contiene múltiples pares de sondas, con cada sonda en un par que se superpone al otro en al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 o 100 nucleótidos. En algunos escenarios, la secuencia de superposición puede diseñarse para ser complementaria a una región genómica diana (o a partir de ADNlc derivado de la misma) o para ser complementaria a una secuencia con homología con una región diana o ADNlc. Por tanto, en algunos escenarios, al menos dos sondas son complementarias a la misma secuencia dentro de una región genómica diana, y un fragmento de nucleótido correspondiente o derivado de la región genómica diana puede unirse y extraerse por al menos una de las sondas. Son posibles otros niveles de titulación, tales como tres títulos, cuatro títulos, etc., en donde cada nucleótido en una región diana puede unirse a más de dos sondas.

En un escenario, cada base en una región genómica diana se superpone con exactamente dos sondas, como se ilustra en la **Figura 1B.** Las sondas que se extienden en ambas direcciones más allá de una región genómica diana son útiles para eliminar fragmentos de ADNlc que comprenden una porción de la región genómica diana y secuencias de ADN adyacentes a la región genómica diana. En algunos casos, incluso las regiones diana relativamente pequeñas pueden dirigirse con tres sondas (véase la **Figura 1A**). Un conjunto de sondas que comprende tres o más sondas se usa opcionalmente para capturar una región genómica más grande (véase la **Figura 1B**). En algunos escenarios, los subconjuntos de sondas se extenderán colectivamente a través de una región genómica completa (por ejemplo, puede ser complementaria a fragmentos no convertidos o convertidos de la región genómica). Un conjunto de sonda de mosaico comprende opcionalmente sondas que incluyen colectivamente al menos dos sondas que se solapan cada nucleótido en la región genómica. Esto se hace para asegurar que los ADNlc que comprenden una pequeña porción de una región genómica diana en un extremo tendrán una superposición sustancial que se extiende en la región genómica adyacente no dirigida con al menos una sonda, para proporcionar una captura eficiente.

Por ejemplo, puede garantizarse que un fragmento de ADNlc de 100 pb que comprende una región genómica diana de 30 nt tenga al menos 65 pb de superposición con al menos una de las sondas solapantes. Son posibles otros niveles de titulación. Por ejemplo, para aumentar el tamaño objetivo y añadir más sondas en un panel, las sondas pueden diseñarse para expandir una región diana de 30 pb en al menos 70 pb, 65 pb, 60 pb, 55 pb o 50 pb. Para capturar cualquier fragmento que se superponga a la región diana (aunque sólo sea 1 pb), las sondas pueden diseñarse para que se extiendan más allá de los extremos de la región diana a ambos lados.

Las sondas están diseñadas para analizar el estado de metilación de regiones genómicas diana (por ejemplo, del ser humano o de otro organismo) que se sospecha que están correlacionadas con la presencia o ausencia de cáncer en general, la presencia o ausencia de determinados tipos de cáncer, el estadio del cáncer, o la presencia o ausencia de otros tipos de enfermedades.

Además, las sondas están diseñadas para unirse eficazmente a fragmentos de ANDlc que contengan una región genómica diana. En algunos casos, las sondas están diseñadas para cubrir partes solapadas de una región diana, de modo que cada sonda tiene una cobertura "titulada" tal que cada sonda se solapa en cobertura al menos parcialmente con otra sonda de la biblioteca. En tales escenarios, el panel contiene múltiples pares de sondas, donde cada par comprende al menos dos sondas superpuestas entre sí por una secuencia superpuesta de al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 ó 100 nucleótidos. En algunos escenarios, la secuencia de superposición puede diseñarse para ser complementaria a una región genómica diana (o una versión convertida de la misma), por tanto, un fragmento de nucleótido derivado de o que contiene la región genómica diana puede unirse y extraerse por al menos una de las sondas. Además, las sondas pueden diseñarse para cubrir ambas cadenas de una secuencia de ADNlc bicatenaria.

En un escenario, la región genómica diana más pequeña es de 30 ó 31 pb. Cuando se añade una nueva región diana al panel (basándose en la selección de voraz como se describió anteriormente), la nueva región diana de 30 pb se puede centrar en un sitio CpG específico de interés. Luego, se comprueba si cada borde de esta nueva diana está lo suficientemente cerca como para otros objetivos de manera que puedan fusionarse. Esto se basa en un parámetro de "distancia de fusión" que puede ser de 200 pb por defecto pero puede ajustarse. Esto permite que las regiones objetivo cercanas pero distintas se enriquezcan con sondas superpuestas. Dependiendo de si existen dianas suficientemente cercanas a la izquierda o a la derecha de la nueva diana, la nueva diana puede fusionarse sin nada (aumentando el número de dianas de panel en uno), fusionado con solo una diana a la izquierda o a la derecha (no cambiando el número de dianas de panel), o fusionado con dianas existentes tanto a la izquierda como a la derecha (reduciendo el número de dianas de panel en uno).

Métodos para seleccionar regiones genómicas diana

En otro escenario, los métodos para seleccionar regiones genómicas diana para detectar cáncer y/o un TOO. Las regiones genómicas dirigidas pueden usarse para diseñar y fabricar sondas para un panel de ensayo de cáncer. El estado de metilación de las moléculas de ADN o ADNlc correspondientes a, o derivados de, las regiones genómicas diana pueden seleccionarse usando el panel de ensayo de cáncer. Los métodos alternativos, por ejemplo, por WGBS u otros métodos conocidos en la técnica, también pueden implementarse para detectar el estado de metilación de moléculas de ADN o fragmentos correspondientes a, o derivados de, las regiones genómicas diana.

Procesamiento de muestra

La **Figura 7A** es un diagrama de flujo de un proceso 100 para procesar una muestra de ácido nucleico y generar vectores de estado de metilación para fragmentos de ADN, según un escenario. El método incluye, pero no se limita a, las siguientes etapas. Por ejemplo, cualquier etapa del método puede comprender una subetapa de cuantificación para el control de calidad u otros procedimientos de ensayo de laboratorio conocidos por un experto en la técnica.

En la etapa 105, una muestra de ácido nucleico (ADN o ARN) se extrae de un sujeto. En la presente descripción, el ADN y el ARN pueden usarse indistintamente a menos que se indique lo contrario. Es decir, los escenarios descritos en la presente memoria pueden ser aplicables tanto a tipos de ADN como a ARN de secuencias de ácido nucleico.

Sin embargo, los ejemplos descritos en la presente memoria pueden centrarse en ADN para fines de claridad y explicación. La muestra puede ser cualquier subconjunto del genoma humano, incluyendo el genoma completo. La muestra puede incluir sangre, plasma, suero, orina, heces, saliva, otros tipos de fluidos corporales o cualquier combinación de los mismos. En algunos escenarios, los métodos para extraer una muestra de sangre (por ejemplo, jeringa o punción de dedo) pueden ser menos invasivos que los procedimientos para obtener una biopsia de tejido, que puede requerir cirugía. La muestra extraída puede comprender ADNlc y/o ADNct. Para individuos sanos, el cuerpo humano puede eliminar naturalmente el ADNlc y otros restos celulares. Si un sujeto tiene un cáncer o enfermedad, el ADNlc y/o ADNct en una muestra extraída pueden estar presentes a un nivel detectable para detectar el cáncer o enfermedad.

En la etapa 110, los fragmentos de ADNlc se tratan para convertir citosinas no metiladas en uracilos. En un escenario, el método usa un tratamiento con bisulfito del ADN que convierte las citosinas no metiladas en uracilos sin convertir las citosinas metiladas. Por ejemplo, se usa un kit comercial tal como el kit EZ DNA Methylation™-Gold, EZ DNA Methylation™-Direct o un DNA Metylaation™-Lightning (comercializado por Zymo Research Corp (Irvine, CA) para la conversión con bisulfito. En otro escenario, la conversión de citosinas no metiladas en uracilos se logra mediante el uso de una reacción enzimática. Por ejemplo, la conversión puede usar un kit disponible comercialmente para la conversión de citosinas no metiladas en uracilos, tales como APOBEC-Seq (NEBiolabs, Ipswich, MA).

En la etapa 115, se prepara una biblioteca de secuenciación. En una primera etapa, se añade un adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito usando una reacción de ligamiento de ADNmc. En un escenario, la reacción de ligamiento de ADNmc usa CircLigase II (Epicentre) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida en bisulfito, en donde el extremo 5' del adaptador está fosforilado y el ADNmc convertido en bisulfito ha sido desfosforilado (es decir, el extremo 3' tiene un grupo hidroxilo). En otro escenario, la reacción de ligamiento de ADNmc usa la ligasa termoestable 5' AppDNA/RNA (disponible de New England BioLabs (Ipswich, MA)) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito. En este ejemplo, el primer adaptador UMI se adenila en el extremo 5' y se bloquea en el extremo 3'. En otro escenario, la reacción de ligamiento de ADNmc usa una ligasa de ARN T4 (disponible de New England BioLabs) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito. En una segunda etapa, se sintetiza un ADN de segunda cadena en una reacción de extensión. Por ejemplo, un cebador de extensión, que se hibrida con una secuencia de cebador incluida en el adaptador de ADNmc, se usa en una reacción de extensión de cebador para formar una molécula de ADN bicatenario convertida con bisulfito. Opcionalmente, en un escenario, la reacción de extensión usa una enzima que es capaz de leer a través de residuos de uracilo en la cadena de molde convertida con bisulfito. Opcionalmente, en una tercera etapa, se añade un adaptador de ADNbc a la molécula de ADN bicatenario convertida con bisulfito. Finalmente, el ADN bicatenario convertido con bisulfito se amplifica para añadir adaptadores de secuenciación. Por ejemplo, la amplificación por PCR usando un cebador directo que incluye una secuencia P5 y un cebador inverso que incluye una secuencia P7 se usa para añadir las secuencias P5 y P7 al ADN convertido con bisulfito. Opcionalmente, durante la preparación de la biblioteca, pueden añadirse identificadores moleculares únicos (UMI) a las moléculas de ácido nucleico (por ejemplo, moléculas de ADN) mediante ligamiento de adaptador. Las UMI son secuencias cortas de ácido nucleico (por ejemplo, 4-10 pares de bases) que se añaden a los extremos de los fragmentos de ADN durante el ligamiento del adaptador. En algunos escenarios, las UMI son pares de bases degenerados que sirven como una etiqueta única que puede usarse para identificar lecturas de secuencia que se originan en un fragmento de ADN específico. Durante la amplificación por PCR después de la ligamiento del adaptador, las UMI se replican junto con el fragmento de ADN unido, lo que proporciona una forma de identificar lecturas de secuencia que provienen del mismo fragmento original en el análisis posterior.

En la etapa 120, las secuencias de ADN diana pueden enriquecerse de la biblioteca. Esto se usa, por ejemplo, donde se realiza un ensayo de panel objetivo en las muestras. Durante el enriquecimiento, las sondas de hibridación (también denominadas en la presente memoria "sondas") se usan para dirigirse, y eliminar, fragmentos de ácido nucleico informativos para la presencia o ausencia de cáncer (o enfermedad), estado del cáncer o una clasificación de cáncer (por ejemplo, tipo de cáncer o tejido de origen). Para un flujo de trabajo determinado, las sondas pueden diseñarse para hibridar (o hibridarse) con una cadena diana (complementaria) de ADN o ARN. La cadena diana puede ser la cadena "positiva" (por ejemplo, la cadena transcrita en el ARNm y posteriormente traducida en una proteína) o la cadena "negativa" complementaria. Las sondas pueden variar en longitud de 10 s, 100 s o 1000 s de pares de bases. Además, las sondas pueden cubrir porciones superpuestas de una región diana.

Después de una etapa de hibridación 120, los fragmentos de ácido nucleico hibridados se capturan y también se pueden amplificar usando PCR (enriquecimiento 125). Por ejemplo, las secuencias diana pueden enriquecerse para obtener secuencias enriquecidas que se pueden secuenciar posteriormente. En general, cualquier método conocido en la técnica puede usarse para aislar, y enriquecer para, ácidos nucleicos diana hibridados por sonda. Por ejemplo, como es bien conocido en la técnica, puede añadirse un resto de biotina al extremo 5' de las sondas (es decir, biotiniladas) para facilitar el aislamiento de ácidos nucleicos diana hibridados con sondas usando una superficie recubierta con estreptavidina (por ejemplo, perlas recubiertas con estreptavidina).

En la etapa 130, se generan lecturas de secuencia a partir de las secuencias de ADN enriquecidas, por ejemplo, secuencias enriquecidas. Los datos de secuenciación pueden adquirirse a partir de las secuencias de ADN

enriquecidas por medios conocidos en la técnica. Por ejemplo, el método puede incluir técnicas de secuenciación de próxima generación (NGS) que incluyen tecnología de síntesis (Illumina), pirosecuenciación (454 Life Sciences), tecnología de semiconductores de iones (secuenciación Ion Torrent), secuenciación en tiempo real de una sola molécula (Pacífico Biosciences), secuenciación por ligamiento (secuenciación SOLiD), secuenciación de nanoporos (Oxford Nanopore Technologies) o secuenciación de extremos emparejados. En algunas realizaciones, la secuenciación masivamente paralela se realiza mediante secuenciación por síntesis con terminadores de colorante reversibles.

En la etapa 140, se generan vectores de estado de metilación a partir de las lecturas de secuencia. Para hacerlo, una lectura de secuencia se alinea con un genoma de referencia. El genoma de referencia ayuda a proporcionar el contexto en cuanto a qué posición en un genoma humano se origina el fragmento ADNlc En un ejemplo simplificado, la lectura de secuencia se alinea de manera que los tres sitios CpG se correlacionan con los sitios CpG 23, 24 y 25 (identificadores de referencia arbitrarios usados por conveniencia de la descripción). Después de la alineación, hay información tanto en el estado de metilación de todos los sitios CpG en el fragmento de ADNlc como en qué posición en el genoma humano se correlaciona con los sitios CpG. Con el estado y ubicación de metilación, se puede generar un vector de estado de metilación para el fragmento ADNlc.

Generación de la estructura de datos

La **Figura 3A** es un diagrama de flujo que describe un proceso 300 para generar una estructura de datos para un grupo de control sano, según un escenario. Para crear una estructura de datos del grupo de control saludable, el sistema de análisis obtiene información relacionada con el estado de metilación de una pluralidad de sitios CpG en lecturas de secuencia derivadas de una pluralidad de moléculas de ADN o fragmentos de una pluralidad de sujetos sanos. El método proporcionado en la presente memoria para crear una estructura de datos del grupo de control saludable puede realizarse de manera similar para sujetos con cáncer, sujetos con cáncer de un TOO, sujetos con un tipo de cáncer conocido o sujetos con otro estado de enfermedad conocido. Se genera un vector de estado de metilación para cada molécula o fragmento de ADN, por ejemplo, a través del proceso 100.

El sistema de análisis subdivide 310 el vector de estado de metilación de cada fragmento de ADN en cadenas de sitios CpG. En un escenario, el sistema de análisis subdivide 310 el vector de estado de metilación de manera que las cadenas resultantes son todas menos de una longitud dada. Por ejemplo, un vector de estado de metilación de longitud 11 puede subdividirse en cadenas de longitud menor o igual a 3 daría como resultado 9 cadenas de longitud 3, 10 cadenas de longitud 2 y 11 cadenas de longitud 1. En otro ejemplo, un vector de estado de metilación de longitud 7 se subdivide en cadenas de longitud menor o igual a 4 daría como resultado 4 cadenas de longitud 4, 5 cadenas de longitud 3, 6 cadenas de longitud 2 y 7 cadenas de longitud 1. Si el vector de estado de metilación resultante de un fragmento de ADN es más corto que o la misma longitud que la longitud de cadena especificada, entonces el vector de estado de metilación puede convertirse en una sola cadena que contiene todos los sitios CpG del vector.

El sistema de análisis consiste en 320 las cadenas contando, para cada sitio CpG posible y la posibilidad de estados de metilación en el vector, el número de cadenas presentes en el grupo de control que tiene el sitio CpG especificado como el primer sitio CpG en la cadena y que tiene esa posibilidad de estados de metilación. Para una longitud de cadena de tres en un sitio CpG dado, hay 2^3 u 8 configuraciones de cuerda posibles. Para cada sitio CpG, el sistema de análisis se detiene 320 cuántas apariciones de cada vector de estado de metilación posible aparecen en el grupo de control. Esto puede implicar cortar las siguientes cantidades: $< M_x, M_{x+1}, M_{x+2} >$, $< M_x, M_{x+1}, U_{x+2} >$, ..., $< U_x, U_{x+1}, U_{x+2} >$ para cada sitio CpG inicial en el genoma de referencia. El sistema de análisis crea 330 una estructura de datos que almacena los recuentos talados para cada sitio de CpG inicial y la posibilidad de cadena en cada CpG inicial.

Existen varias ventajas para establecer un límite superior en la longitud de la cuerda. En primer lugar, dependiendo de la longitud máxima para una cadena, el tamaño de la estructura de datos creada por el sistema de análisis puede aumentar drásticamente el tamaño. Por ejemplo, una longitud de cuerda máxima de 4 significa que hay como máximo 2A4 números para tazar en cada CpG. El aumento de la longitud máxima de la cuerda a 5 duplica el posible número de estados de metilación. La reducción del tamaño de la cadena ayuda a reducir la carga de almacenamiento computacional y de datos de la estructura de datos. En algunos escenarios, el tamaño de la cadena es 3. En algunos escenarios, el tamaño de la cadena es 4. Una segunda razón para limitar la longitud máxima de la cuerda es evitar el sobreajuste de modelos aguas abajo. Las probabilidades de cálculo basadas en cadenas largas de sitios CpG pueden ser problemáticas si las cadenas de CpG largas no tienen un fuerte efecto biológico en el resultado (por ejemplo, predicciones de anomalía que predicen la presencia de cáncer), ya que requiere una cantidad significativa de datos que pueden no estar disponibles y, por lo tanto, serían demasiado dispersos para que un modelo funcione adecuadamente. Por ejemplo, calcular una probabilidad de anomalía/cáncer condicionado en los 100 sitios CpG anteriores requeriría recuentos de cadenas en la estructura de datos de longitud 100, idealmente alguna coincidencia exactamente con los 100 estados de metilación anteriores. Si solo hay recuentos dispersos de cadenas de longitud 100, habrá datos insuficientes para determinar si una cadena dada de longitud de 100 en una muestra de prueba es anómalo o no.

Validación de la estructura de datos

Una vez que se ha creado la estructura de datos, el sistema de análisis puede buscar validar 340 la estructura de datos y/o cualquier modelo aguas abajo que haga uso de la estructura de datos.

Este primer tipo de validación asegura que las posibles muestras cancerosas se eliminen del grupo de control sano para no afectar la pureza del grupo de control. Este tipo de validación verifica la coherencia dentro de la estructura de datos del grupo de control. Por ejemplo, el grupo de control sano puede contener una muestra de un individuo con un cáncer no diagnosticado que contiene una pluralidad de fragmentos anormalmente metilados. El sistema de análisis puede realizar diversos cálculos para determinar si excluir los datos de un sujeto con cáncer aparentemente no diagnosticado.

Un segundo tipo de validación verifica el modelo probabilístico usado para calcular los valores de p con los recuentos de la propia estructura de datos (es decir, del grupo de control sano). Un proceso para el cálculo de valor de p se describe a continuación junto con la **Figura 5.** Una vez que el sistema de análisis genera un valor de p para los vectores de estado de metilación en el grupo de validación, el sistema de análisis construye una función de densidad acumulativa (CDF) con los valores de p. Con el CDF, el sistema de análisis puede realizar diversos cálculos en el CDF para validar la estructura de datos del grupo de control. Una prueba utiliza el hecho de que la CDF debe estar idealmente en o por debajo de una función de identidad, de manera que $CDF(x) \leq x$. En el contrario, por encima de la función de identidad revela alguna deficiencia dentro del modelo probabilístico utilizado para la estructura de datos del grupo de control. Por ejemplo, si 1/100 de fragmentos tienen una puntuación de valor de p de 1/1000 que significa CDF (1/1000) = 1/100>1/1000, entonces el segundo tipo de validación falla indicando un problema con el modelo probabilístico. *Véase* por ejemplo, la solicitud estadounidense n.º 16/352.602, publicada como publicación estadounidense n.º 2019/0287652.

Un tercer tipo de validación usa un conjunto saludable de muestras de validación separadas de las utilizadas para construir la estructura de datos. Estas pruebas si la estructura de datos se construye correctamente y el modelo funciona. Un proceso ilustrativo para llevar a cabo este tipo de validación se describe a continuación junto con la **Figura 3B**. El tercer tipo de validación puede cuantificar cómo el grupo de control sano generaliza la distribución de muestras sanas. Si falla el tercer tipo de validación, entonces el grupo de control sano no generaliza bien en la distribución saludable.

Un cuarto tipo de pruebas de validación con muestras de un grupo de validación no saludable. El sistema de análisis calcula valores de p y construye el CDF para el grupo de validación no saludable. Con un grupo de validación no saludable, el sistema de análisis espera ver el $CDF(x)>x$ para al menos algunas muestras o, indicadas de manera diferente, la inversa de lo que se esperaba en el segundo tipo de validación y el tercer tipo de validación con el grupo de control sano y el grupo de validación saludable. Si falla el cuarto tipo de validación, esto es indicativo de que el modelo no identifica adecuadamente la anomalía que estaba diseñado para identificar.

La **Figura 3B** es un diagrama de flujo que describe la etapa adicional 340 de validar la estructura de datos para el grupo de control de la **Figura 3A**, según un escenario. En este escenario de la etapa 340 de validar la estructura de datos, el sistema de análisis realiza el cuarto tipo de prueba de validación como se describió anteriormente que utiliza un grupo de validación con una composición supuestamente similar de sujetos, muestras y/o fragmentos como el grupo de control. Por ejemplo, si el sistema de análisis seleccionó sujetos sanos sin cáncer para el grupo de control, entonces el sistema de análisis también usa sujetos sanos sin cáncer en el grupo de validación.

El sistema de análisis toma el grupo de validación y genera 100 un conjunto de vectores de estado de metilación como se describe en la **Figura 3A**. El sistema de análisis realiza un cálculo del valor de p para cada vector de estado de metilación del grupo de validación. El proceso de cálculo de valor de p se describirá adicionalmente junto con las **Figuras 4-5**. Para cada vector de estado de metilación posible, el sistema de análisis calcula una probabilidad de la estructura de datos del grupo de control. Una vez que se calculan las probabilidades para las posibilidades de los vectores de estado de metilación, el sistema de análisis calcula 350 una puntuación de valor de p para ese vector de estado de metilación basándose en las probabilidades calculadas. La puntuación de valor de p representa una expectativa de encontrar ese vector de estado de metilación específico y otros vectores de estado de metilación posibles que tienen probabilidades incluso menores en el grupo de control. Una puntuación de valor de p baja, por lo tanto, corresponde generalmente a un vector de estado de metilación que es relativamente inesperado en comparación con otros vectores de estado de metilación dentro del grupo de control, mientras que una puntuación de valor de p alta corresponde generalmente a un vector de estado de metilación que es relativamente más esperado en comparación con otros vectores de estado de metilación encontrados en el grupo de control. Una vez que el sistema de análisis genera una puntuación de valor de p para los vectores de estado de metilación en el grupo de validación, el sistema de análisis construye 360 una función de densidad acumulativa (CDF) con las puntuaciones de valor de p del grupo de validación. El sistema de análisis valida la coherencia 370 del CDF como se ha descrito anteriormente en el cuarto tipo de pruebas de validación.

Fragmentos finamente metilados

Los fragmentos anomalamente metilados que tienen patrones de metilación anormales en muestras de pacientes con cáncer, sujetos con cáncer de un TOO, sujetos con un tipo de cáncer conocido o sujetos con otro estado de

enfermedad conocido, se seleccionan como regiones genómicas diana, según un escenario como se describe en las regiones genómicas diana en la **Figura 4**. Los procesos ilustrativos de fragmentos aleatorios seleccionados de forma anómala 440 se ilustran visualmente en la **Figura 5** y se describe adicionalmente a continuación la descripción de la **Figura 4**. En el proceso 400, el sistema de análisis genera 100 Vectores de estado de metilación a partir de fragmentos de ADNlc de la muestra. El sistema de análisis maneja cada vector de estado de metilación de la siguiente manera.

Para un vector de estado de metilación dado, el sistema de análisis enumera 410 todas las posibilidades de los vectores de estado de metilación que tienen el mismo sitio CpG inicial y la misma longitud (es decir, conjunto de sitios CpG) en el vector de estado de metilación. Como cada estado de metilación puede estar metilado o no metilado, solo hay dos estados posibles en cada sitio CpG y, por tanto, el recuento de posibilidades distintas de vectores de estado de metilación depende de una potencia de 2, de manera que un vector de estado de metilación de longitud n estaría asociado con 2^n posibilidades de los vectores de estado de metilación.

El sistema de análisis calcula 420 la probabilidad de observar cada posibilidad del vector de estado de metilación para el sitio CpG inicial identificado/longitud del vector de estado de metilación accediendo a la estructura de datos del grupo de control saludable. En un escenario, calcular la probabilidad de observar una posibilidad dada usa una probabilidad de cadena de Markov para modelar el cálculo de probabilidad de articulación que se describirá con mayor detalle con respecto a la **Figura 5** a continuación. En otros escenarios, se usan métodos de cálculo distintos de las probabilidades de cadena de Markov para determinar la probabilidad de observar cada posibilidad del vector de estado de metilación.

El sistema de análisis calcula 430 una puntuación de valor de p para el vector de estado de metilación usando las probabilidades calculadas para cada posibilidad. En un escenario, esto incluye identificar la probabilidad calculada correspondiente a la posibilidad que coincida con el vector de estado de metilación en cuestión. Específicamente, esto es la posibilidad que tiene el mismo conjunto de sitios CpG, o similar al mismo sitio de CpG inicial y longitud como vector de estado de metilación. El sistema de análisis suma las probabilidades calculadas de cualquier posibilidades que tenga probabilidades inferiores o iguales a la probabilidad identificada para generar la puntuación de valor de p.

Este valor de p representa la probabilidad de observar el vector de estado de metilación del fragmento u otros vectores de estado de metilación incluso menos probables en el grupo de control sano. Una puntuación de valor de p baja, por lo tanto, corresponde generalmente a un vector de estado de metilación que es raro en un sujeto sano, y que hace que el fragmento se marque anormalmente metilado, en relación con el grupo de control sano. Se espera que una puntuación de valor de p alta generalmente se refiera a un vector de estado de metilación, en un sentido relativo, en un sujeto sano. Si el grupo de control sano es un grupo no canceroso, por ejemplo, un valor de p bajo indica que el fragmento está anormalmente metilado en relación con el grupo no cáncer y, por tanto, posiblemente indicativo de la presencia de cáncer en el sujeto de prueba.

Como anteriormente, el sistema de análisis calcula las puntuaciones de valor de p para cada uno de una pluralidad de vectores de estado de metilación, cada uno de los cuales representa un fragmento de ADNlc en la muestra de prueba. Para identificar cuál de los fragmentos están anormalmente metilados, el sistema de análisis puede filtrar 440 el conjunto de vectores de estado de metilación basándose en sus puntuaciones de valor de p. En un escenario, el filtrado se realiza comparando las puntuaciones de los valores de p con respecto a un umbral y manteniendo solo aquellos fragmentos por debajo del umbral. Esta puntuación de valor de p umbral podría ser del orden de 0,1, 0,01, 0,001, 0,0001 o similar.

Cálculo de la puntuación de valor de p

La **Figura 5** es una ilustración 500 de un cálculo de puntuación de valor de p de ejemplo, según un escenario. Para calcular una puntuación de valor de p dada un vector 505 de estado de metilación de prueba, el sistema de analítica toma ese vector 505 de estado de metilación de prueba y enumera 410 posibilidades de vectores de estado de metilación. En este ejemplo ilustrativo, el vector de estado de metilación de prueba 505 es < M23, M24, M25, U26 >. Como la longitud del vector de estado de metilación de prueba 505 es 4, hay 2^4 posibilidades de vectores de estado de metilación que abarcan los sitios CpG 23 - 26. En un ejemplo genérico, el número de posibilidades de vectores de estado de metilación es 2^n, donde n es la longitud del vector de estado de metilación de prueba o, alternativamente, la longitud de la ventana deslizante (descrita más adelante).

El sistema de análisis calcula 420 probabilidades 515 para las posibilidades enumeradas de vectores de estado de metilación. Como la metilación depende condicionalmente del estado de metilación de los sitios CpG cercanos, una forma de calcular la probabilidad de observar una posibilidad de vector de estado de metilación dada es usar el modelo de cadena de Markov. Generalmente, un vector de estado de metilación tal como $<S_1, S_2, \ldots, S_n>$, donde S indica el estado de metilación ya sea metilado (indicado como M), no metilado (indicado como U) o indeterminado (indicado como I), tiene una probabilidad conjunta que puede expandirse usando la regla de cadena de probabilidades como:

$$P(< S_1, S_2, \ldots, S_n >) = P(S_n | S_1, \ldots, S_{n-1} >) * P(S_{n-1} | S_1, \ldots, S_{n-2} >) * \ldots * P(S_2 | S_1) * P(S_1) \quad (1)$$

El modelo de cadena de Markov se puede usar para hacer el cálculo de las probabilidades condicionales de cada posibilidad más eficiente. En un escenario, el sistema de análisis selecciona un orden de cadena de Markov k que corresponde a cuántos sitios de CpG anteriores en el vector (o ventana) considerar en el cálculo de probabilidad condicional, de modo que la probabilidad condicional se modela como $P(S_n | S_1, \ldots, S_{n-1}) \sim P(S_n | S_{n-k-2}, \ldots, S_{n-1})$.

Para calcular cada probabilidad modelada por Markov para una posibilidad de vector de estado de metilación, el sistema de análisis accede a la estructura de datos del grupo de control, específicamente los recuentos de varias cadenas de sitios y estados CpG. Para calcular $P(M_n | S_{n-k-2}, \ldots, S_{n-1})$, el sistema de análisis toma una razón del recuento almacenado del número de cadenas de la estructura de datos $< S_{n-k-2}, \ldots, S_{n-1}, M_n >$ dividido entre la suma del recuento almacenado del número de cadenas de la estructura de datos que coincide $< S_{n-k-2}, \ldots, S_{n-1}, M_n >$ and $< S_{n-k-2}, \ldots, S_{n-1}, U_n >$. Por tanto, $P(M_n | S_{n-k-2}, \ldots, S_{n-1})$, se calcula una razón que tiene la forma:

$$\frac{\#\text{ de } < S_{n-k-2}, \ldots, S_{n-1}, M_n >}{\text{N.º de } < S_{n-k-2}, \ldots, S_{n-1}, M_n > + \#\text{ de } < S_{n-k-2}, \ldots, S_{n-1}, U_n >} \quad (2)$$

El cálculo puede implementar adicionalmente un suavizado de los recuentos aplicando una distribución previa. En un escenario, la distribución previa es un valor uniforme antes del suavizado de Laplace. Como ejemplo de esto, se añade una constante al numerador y otra constante (por ejemplo, dos veces la constante en el numerador) se añade al denominador de la ecuación anterior. En otros escenarios, se usa una técnica algorítmica tal como suavizado de Knesser-Ney.

En la ilustración, las fórmulas indicadas anteriormente se aplican al vector de estado de metilación de prueba 505 que cubre los sitios 23-26. Una vez que se completan las probabilidades calculadas 515, el sistema de análisis calcula 430 una puntuación de valor de p 525 que suma las probabilidades que son menores o iguales a la probabilidad de posibilidad de que el vector de estado de metilación coincida con el vector de estado de metilación de prueba 505.

En un escenario, la carga computacional de las probabilidades de cálculo y/o las puntuaciones de valor de p puede reducirse aún más al almacenar en caché al menos algunos cálculos. Por ejemplo, el sistema analítico puede almacenar en caché cálculos de probabilidades de memoria transitoria o persistente para posibilidades de vectores de estado de metilación (o ventanas de los mismos). Si otros fragmentos tienen los mismos sitios CpG, el almacenamiento en caché de las probabilidades de posibilidad permite un cálculo eficiente de las puntuaciones de valor de p sin necesidad de volver a calcular las probabilidades de posibilidad subyacente. De manera equivalente, el sistema de análisis puede calcular puntuaciones de valor de p para cada una de las posibilidades de los vectores de estado de metilación asociados con un conjunto de sitios CpG del vector (o ventana del mismo). El sistema de análisis puede almacenar en caché las puntuaciones de valor de p para su uso en la determinación de las puntuaciones de valor de p de otros fragmentos que incluyen los mismos sitios CpG. Generalmente, las puntuaciones de valor de p de las posibilidades de los vectores de estado de metilación que tienen los mismos sitios CpG pueden usarse para determinar la puntuación de valor de p de una diferente de las posibilidades del mismo conjunto de sitios CpG.

Ventana deslizante

En un escenario, el sistema de análisis utiliza 435 una ventana deslizante para determinar las posibilidades de los vectores de estado de metilación y calcular los valores de p. En lugar de enumerar las posibilidades y calcular los valores de p para todos los vectores de estado de metilación, el sistema de análisis enumera las posibilidades y calcula los valores de p solo para una ventana de sitios CpG secuenciales, donde la ventana es más corta en longitud (de sitios CpG) que al menos algunos fragmentos (de lo contrario, la ventana no serviría). La longitud de la ventana puede ser estática, determinada por el usuario, dinámica o seleccionada de otro modo.

Al calcular los valores de p para un vector de estado de metilación mayor que la ventana, la ventana identifica el conjunto secuencial de sitios CpG del vector dentro de la ventana comenzando desde el primer sitio CpG en el vector. El sistema analítico calcula una puntuación de valor de p para la ventana que incluye el primer sitio CpG. El sistema de análisis "desliza" la ventana al segundo sitio CpG en el vector, y calcula otra puntuación de valor de p para la segunda ventana. Por lo tanto, para un tamaño de ventana *l* y longitud del vector de metilación *m,* cada vector de estado de metilación generará *m-l+1* puntuaciones de valor de p. Después de completar los cálculos de valor de p para cada parte del vector, la puntuación de valor de p más baja de todas las ventanas deslizantes se toma como la puntuación de valor de p global para el vector de estado de metilación. En otro escenario, el sistema de análisis agrega las puntuaciones de valor de p para los vectores de estado de metilación para generar una puntuación de valor de p general.

Usando la ventana deslizante ayuda a reducir el número de posibilidades enumeradas de vectores de estado de metilación y sus cálculos de probabilidad correspondientes que de otro modo necesitaría realizarse. Los cálculos de probabilidad de ejemplo se muestran en la **Figura 5**, pero generalmente el número de posibilidades de vectores de estado de metilación aumenta exponencialmente en un factor de 2 con el tamaño del vector de estado de metilación. Para dar un ejemplo realista, es posible que los fragmentos tengan hacia arriba de 54 sitios CpG. En lugar de probabilidades informáticas para 2A54 (~1,8x10^16) posibilidades para generar un solo valor de p, el sistema de análisis puede usar en cambio una ventana de tamaño 5 (por ejemplo) que da como resultado los cálculos de 50 p para cada una de las 50 Ventanas del vector de estado de metilación para ese fragmento. Cada uno de los 50 cálculos enumera 2A5 (32) posibilidades de vectores de estado de metilación, cuyos resultados totales dan como resultado 50 x 2^5 (1,6 x 10^3) cálculos de probabilidad. Esto da como resultado que se realice una gran reducción de los cálculos, sin impacto significativo para la identificación precisa de fragmentos anómalos. Esta etapa adicional también se puede aplicar cuando se valida 340 el grupo de control con los vectores de estado de metilación del grupo de validación.

Identificar fragmentos indicativos del cáncer

El sistema de análisis identifica 450 fragmentos de ADN indicativos del cáncer del conjunto filtrado de fragmentos anormalmente metilados.

Fragmentos hipometilados e hipermetilados

Según un primer método, el sistema de análisis puede identificar fragmentos de ADN que se consideran hipometilados o hipermetilados como fragmentos indicativos del cáncer del conjunto filtrado de fragmentos anormalmente metilados. Los fragmentos hipometilados y hipermetilados pueden definirse como fragmentos de una cierta longitud de sitios CpG (por ejemplo, más de 3, 4, 5, 6, 7, 8, 9, 10, etc.) con un alto porcentaje de sitios CpG metilados (por ejemplo, más del 80 %, el 85 %, el 90 % o el 95 %, o cualquier otro porcentaje dentro del intervalo del 50 % -100 %) o un alto porcentaje de sitios CpG no metilados (por ejemplo, más del 80 %, 85 %, 90 % o 95 %, o cualquier otro porcentaje dentro del intervalo del 50 % -100 %).

Modelos probabilísticos

Según un método descrito en la presente memoria, el sistema de análisis identifica fragmentos indicativos de cáncer que utilizan modelos probabilísticos de patrones de metilación ajustados a cada tipo de cáncer y tipo distinto de cáncer. El sistema de análisis calcula relaciones de probabilidad logarítmica para una muestra usando fragmentos de ADN en las regiones genómicas considerando los diversos tipos de cáncer con los modelos probabilísticos ajustados para cada tipo de cáncer y tipo distinto de cáncer. El sistema de análisis puede determinar que un fragmento de ADN es indicativo de cáncer basándose en si al menos una de las relaciones de probabilidad logarítmica consideradas contra los diversos tipos de cáncer está por encima de un valor umbral.

En un escenario de partición del genoma, el sistema de análisis divide el genoma en regiones por múltiples etapas. En una primera etapa, el sistema de análisis separa el genoma en bloques de sitios CpG. Cada bloque se define cuando hay una separación entre dos sitios CpG adyacentes que excede algún umbral, por ejemplo, más de 200 pb, 300 pb, 400 pb, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb o 1.000 pb. A partir de cada bloque, el sistema de análisis subdivide en una segunda etapa cada bloque en regiones de una cierta longitud, por ejemplo, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1.000 pb, 1.100 pb, 1.200 pb, 1.300 pb, 1.400 pb o 1.500 pb. El sistema de análisis puede superponerse adicionalmente a regiones adyacentes por un porcentaje de la longitud, por ejemplo, el 10 %, el 20 %, el 30 %, el 40 %, el 50 % o el 60 %.

El sistema de análisis analiza lecturas de secuencia derivadas de fragmentos de ADN para cada región. El sistema de análisis puede procesar muestras de tejido y/o ADNlc de alta señal. Las muestras de ADNlc de alta señal pueden determinarse mediante un modelo de clasificación binaria, por estadio de cáncer o por otra métrica.

Para cada tipo de cáncer y no cáncer, el sistema de análisis se ajusta a un modelo probabilístico separado para fragmentos. En un ejemplo, cada modelo probabilístico es un modelo de mezcla que comprende una combinación de una pluralidad de componentes de la mezcla con cada componente de la mezcla que es un modelo de sitios independientes donde se supone que la metilación en cada sitio CpG es independiente de los estados de metilación en otros sitios CpG.

En escenarios alternativos, el cálculo se realiza con respecto a cada sitio CpG. Específicamente, se determina un primer recuento que es el número de muestras cancerosas (recuento de cáncer) que incluyen un fragmento de ADN mal metilado que se superpone que CpG, y se determina un segundo recuento que es el número total de muestras que contienen fragmentos que se superponen que CpG (total) en el conjunto. Las regiones genómicas pueden seleccionarse en base a los números, por ejemplo, en base a criterios correlacionados positivamente con el número de muestras cancerosas (recuento de cáncer) que incluyen un fragmento de ADN que se superpone que CpG, e inversamente se correlaciona con el número total de muestras que contienen fragmentos que se superponen que CpG (total) en el conjunto.

Los cánceres de diversos tipos que tienen diferentes TOO pueden seleccionarse del grupo que consiste en cáncer de mama, cáncer de útero, cáncer de cuello uterino, cáncer de ovario, cáncer de vejiga, cáncer urotelial de pelvis renal, cáncer renal distinto del urotelial, cáncer de próstata, cáncer anorrectal, cáncer anal, cáncer colorrectal, cáncer hepatobiliar derivado de hepatocitos, cáncer hepatobiliar derivado de células distintas de los hepatocitos, cáncer de hígado/conducto biliar, cáncer de esófago, cáncer de páncreas, cáncer de estómago, cáncer de células escamosas del tracto gastrointestinal superior, cáncer gastrointestinal superior distinto del escamoso, cáncer de cabeza y cuello, cáncer de pulmón, adenocarcinoma de pulmón, cáncer de pulmón de células pequeñas, cáncer de pulmón de células escamosas y cáncer distinto del adenocarcinoma o del cáncer de pulmón de células pequeñas, cáncer neuroendocrino, melanoma, cáncer de tiroides, sarcoma, neoplasia de células plasmáticas, mieloma múltiple, neoplasia mieloide, linfoma y leucemia.

En algunos escenarios, varios tipos de cáncer pueden clasificarse y etiquetarse utilizando métodos de clasificación disponibles en la técnica, como la Clasificación Internacional de Enfermedades Oncológicas (CIE-O-3) (codes.iarc.fr) o el Programa de Vigilancia, Epidemiología y Resultados Finales (SEER) (seer.cancer.gov). En otros escenarios, los tipos de cáncer se clasifican en tres códigos ortogonales, (i) códigos topográficos, (ii) códigos morfológicos, o (iii) códigos de comportamiento. En los códigos conductuales, el tumor benigno es 0, el comportamiento incierto es 1, el carcinoma in situ es 2, maligno, el sitio primario es 3 y el sitio metastásico y maligno es 6.

En algunos escenarios, un TOO de cáncer puede seleccionarse de un grupo definido por la guía que se usará para la etapa de un cáncer detectado. Por ejemplo, la referencia, Amin, M.B., Edge, S., Greene, F., Byrd, D.R., Brookland, R.K., Washington, M.K., Gershenwald, J.E., Compton, C.C., Hess, K.R., Sullivan, D.C., Jessup, J.M., Brierley, J.D., Gaspar, L.E., Schilsky, R.L., Balch, C.M., Winchester, D.P., Asare, E.A., Madera, M., Gress, D.M., Meyer, L.R. (Eds.), AJCC Cancer Staging Manual, 8ª edición, Springer, 2017, identifica grupos de diferentes cánceres que se escalonan juntos siguiendo las pautas convencionales. El envejecimiento es típicamente una siguiente etapa en la gestión del cáncer después de su detección y diagnóstico.

El sistema de análisis puede calcular además las relaciones de probabilidad logarítmica ("R") para un fragmento que indica una probabilidad de que el fragmento sea indicativo de cáncer considerando los diversos tipos de cáncer con los modelos probabilísticos ajustados para cada tipo de cáncer y tipo distinto de cáncer, o para un TOO de cáncer. Las dos probabilidades pueden tomarse de modelos probabilísticos ajustados para cada uno de los tipos de cáncer y el tipo de no cáncer, los modelos probabilísticos definidos para calcular la probabilidad de observar un patrón de metilación en un fragmento dado cada uno de los tipos de cáncer y el tipo de no cáncer. Por ejemplo, los modelos probabilísticos pueden definirse ajustados para cada uno de los tipos de cáncer y el tipo de no cáncer.

Selección de regiones genómicas indicativas del cáncer

En algunos escenarios, el sistema de análisis puede identificar 460 regiones genómicas indicativas de cáncer. Para identificar estas regiones informativas, el sistema de análisis calcula una ganancia de información para cada región genómica o más específicamente cada sitio CpG que describe una capacidad para distinguir entre diversos resultados.

Un método para identificar regiones genómicas capaces de distinguir entre el tipo de cáncer y el tipo de no cáncer utiliza un modelo de clasificación entrenado que puede aplicarse en el conjunto de moléculas de ADN débilmente metiladas o fragmentos correspondientes o derivados de un grupo canceroso o no canceroso. El modelo de clasificación entrenado puede entrenarse para identificar cualquier condición de interés que pueda identificarse a partir de los vectores de estado de metilación.

En un escenario, el modelo de clasificación entrenado es un clasificador binario entrenado basándose en estados de metilación para fragmentos de ADNlc o secuencias genómicas obtenidas de una cohorte de sujeto con cáncer o un TOO de cáncer, y una cohorte de sujeto sano sin cáncer, y luego se usa para clasificar una probabilidad de sujeto de prueba de tener cáncer, un TOO de cáncer o no tener cáncer, basándose en vectores de estado de metilación de forma anómala. En otros escenarios, se pueden entrenar diferentes clasificadores utilizando cohortes de sujetos que se sabe que padecen un cáncer concreto (por ejemplo, de mama, pulmón, próstata, etc.); conocido por tener cáncer de TOO particular donde se cree que el cáncer original; o se sabe que padecen distintos estadios de un cáncer concreto (por ejemplo, de mama, pulmón, próstata, etc.). En estos escenarios, se pueden entrenar diferentes clasificadores utilizando lecturas de secuencias obtenidas de muestras enriquecidas para células tumorales de cohortes de sujetos que se sabe que padecen un cáncer concreto (por ejemplo, de mama, pulmón, próstata, etc.). Cada capacidad de la región genómica para distinguir entre el tipo de cáncer y el tipo de no cáncer en el modelo de clasificación se usa para clasificar las regiones genómicas de la mayoría informativa al menos informativa en el rendimiento de clasificación. El sistema de análisis puede identificar regiones genómicas de la clasificación según la ganancia de información en la clasificación entre el tipo de cáncer y el tipo de cáncer.

Calcular información de información de fragmentos hipometilados y hipermetilados indicativos del cáncer

Con fragmentos indicativos de cáncer, el sistema de análisis puede entrenar un clasificador según un proceso 600 ilustrado en la **Figura 6A,** según un escenario. El proceso 600 accede a dos grupos de entrenamiento de muestras-

un grupo no canceroso y un grupo de cáncer-y obtiene 605 un conjunto no canceroso de vectores de estado de metilación y un conjunto de cáncer de vectores de estado de metilación que comprenden fragmentos débilmente metilados, por ejemplo, a través de la etapa 440 del proceso 400.

El sistema de análisis determina 610, para cada vector de estado de metilación, si el vector de estado de metilación es indicativo de cáncer. Aquí, los fragmentos indicativos del cáncer pueden definirse como fragmentos hipermetilados o hipometilados determinados si al menos algún número de sitios CpG tiene un estado particular (metilado o no metilado, respectivamente) y/o tener un porcentaje umbral de sitios que son el estado particular (nuevamente, metilado o no metilado, respectivamente). En un ejemplo, los fragmentos de ADNlc se identifican como hipometilados o hipermetilados, respectivamente, si el fragmento se superpone al menos 5 sitios CpG, y al menos el 80 %, el 90 % o el 100 % de sus sitios CpG están metilados o al menos el 80 %, el 90 % o el 100 % no están metilados.

En un escenario alternativo, el sistema de análisis considera porciones del vector de estado de metilación y determina si la porción está hipometilada o hipermetilada, y puede distinguir esa porción que va a hipometilarse o hipermetilarse. Esta alternativa resuelve los vectores de estado de metilación ausentes que son grandes de tamaño pero contienen al menos una región de hipometilación densa o hipermetilación. Este proceso de definir hipometilación e hipermetilación puede aplicarse en la etapa 450 de la **Figura 4.** En otro escenario, los fragmentos indicativos del cáncer pueden definirse según las probabilidades emitidas desde modelos probabilísticos entrenados.

En un escenario, el sistema de análisis genera 620 una puntuación de hipometilación (P_{hipo}) y una puntuación de hipermetilación (P_{hiper}) por sitio CpG en el genoma. Para generar una puntuación en un sitio CpG dado, el clasificador toma cuatro recuentos en ese recuento de sitio CpG- (1) de los vectores (estado de metilación) del conjunto de cáncer marcado hipometilado que se superpone al sitio de CpG; (2) recuento de vectores del conjunto de cáncer marcado hipermetilado que se solapan con el sitio de CpG; (3) recuento de vectores del conjunto de cáncer sin cáncer marcado hipometilado que se superpone al sitio de CpG; y (4) recuento de vectores del conjunto de cáncer sin cáncer marcado hipermetilado que se solapan con el sitio CpG. Además, el proceso puede normalizar estos recuentos para cada grupo para tener en cuenta la varianza en el tamaño del grupo entre el grupo no cáncer y el grupo de cáncer. En realizaciones alternativas en donde se usan más generalmente fragmentos indicativos de cáncer, las puntuaciones pueden definirse más ampliamente como recuentos de fragmentos indicativos del cáncer en cada región genómica y/o sitio CpG.

En un escenario, para generar 620 la puntuación de hipometilación en un sitio CpG dado, el proceso toma una razón de (1) sobre (1) sumado con (3). De manera similar, la puntuación de hipermetilación se calcula tomando una relación de (2) en (2) y (4). Además, estas razones pueden calcularse con una técnica de suavizado adicional como se discutió anteriormente. La puntuación de hipometilación y la puntuación de hipermetilación se relacionan con una estimación de la probabilidad del cáncer dada la presencia de hipometilación o hipermetilación de fragmentos del conjunto de cáncer.

El sistema de análisis genera 630 una puntuación de hipometilación agregada y una puntuación de hipermetilación agregada para cada vector de estado de metilación anómalo. Las puntuaciones agregadas de hiper e hipometilación se determinan basándose en las puntuaciones de hipermetilación de los sitios CpG en el vector de estado de metilación. En un escenario, las puntuaciones agregadas de hiper e hipometilación se asignan como las puntuaciones más grandes de hipermetilación de la hipermetilación de los sitios en cada vector de estado, respectivamente. Sin embargo, en escenarios alternativos, las puntuaciones agregadas podrían basarse en medios, medianos u otros cálculos que usan las puntuaciones de hipermetilación de los sitios en cada vector.

El sistema de análisis ordena 640 todos los vectores del estado de metilación del sujeto por su puntuación de hipometilación agregada y por su puntuación de hipermetilación agregada, lo que da como resultado dos vías por sujeto. El proceso selecciona puntuaciones de hipometilación agregadas a partir de la clasificación de hipometilación y las puntuaciones de hipermetilación agregada de la clasificación de hipermetilación. Con las puntuaciones seleccionadas, el clasificador genera 650 un único vector de características para cada sujeto. En un escenario, las puntuaciones seleccionadas de cualquier clasificación se seleccionan con un orden fijo que es el mismo para cada vector de característica generado para cada sujeto en cada uno de los grupos de entrenamiento. Como ejemplo, en un escenario, el clasificador selecciona el primer, el segundo, el cuarto y el octavo agregado de hipermetilación agregado, y de manera similar para cada puntuación de hipometilación agregada, de cada clasificación y escribe esas puntuaciones en el vector de características para ese sujeto.

Los trenes del sistema de análisis 660 un clasificador binario para distinguir vectores de características entre los grupos de entrenamiento cáncer y no cáncer. Generalmente, puede usarse una cualquiera de una serie de técnicas de clasificación. En un escenario, el clasificador es un clasificador no lineal. En un escenario específico, el clasificador es un clasificador no lineal que utiliza una regresión logística de núcleo regularizado de L2 con un núcleo de función base radial gaussiana (RBF).

Específicamente, en un escenario, el número de muestras no cancerosas o tipos de cáncer diferentes (n_{otro}) y el número de muestras de cáncer o tipo(s) de cáncer ($n_{cáncer}$) que tiene un fragmento anormalmente metilado que se superpone a un sitio CpG. A continuación, la probabilidad de que una muestra sea cancerosa se estima mediante una puntuación ("S") que se correlaciona positivamente con $n_{cáncer}$ e inversamente con n_{otro}. La puntuación se puede

calcular usando la ecuación: $(n_{cáncer} + 1) / (n_{cáncer} + n_{otro} + 2)$ o $(n_{cáncer}) / (n_{cáncer} + n_{otro})$. El sistema de análisis calcula 670 una ganancia de información para cada tipo de cáncer y para cada región genómica o sitio CpG para determinar si la región genómica o el sitio CpG es indicativo de cáncer. La ganancia de información se calcula para las muestras de entrenamiento con un determinado tipo de cáncer en comparación con todas las demás muestras. Por ejemplo, se usan dos variables aleatorias "fragmentos anómalos ('AF') y tipo de cáncer ('CT'). En un escenario, AF es una variable binaria que indica si hay un fragmento anómalo que se superpone a un sitio CpG dado en una muestra dada según se determina para la puntuación de anomalía/vector de características anterior. La CT es una variable aleatoria que indica si el cáncer es de un tipo particular. El sistema de análisis calcula la información mutua con respecto a CT dada AF. Es decir, cuántos bits de información sobre el tipo de cáncer se obtienen si se sabe si hay un fragmento anómalo que se superpone a un sitio CpG particular.

Para un tipo de cáncer dado, el sistema de análisis usa esta información para clasificar sitios CpG en función de cómo son específicos del cáncer. Este procedimiento se repite para todos los tipos de cáncer en consideración. Si una región particular se ha metilado comúnmente de forma anómala en muestras de entrenamiento de un cáncer dado, pero no en muestras de entrenamiento de otros tipos de cáncer o en muestras de entrenamiento saludable, entonces los sitios CpG superpuestos por esos fragmentos anómalos tenderán a tener altas ganancias de información para el tipo de cáncer dado. Los sitios CpG clasificados para cada tipo de cáncer se añaden (seleccionan) con avidez a un conjunto seleccionado de sitios CpG en función de su rango para su uso en el clasificador de cáncer.

Calcular ganancia de información por pares a partir de fragmentos indicativos de cáncer identificado a partir de modelos probabilísticos

Con fragmentos indicativos del cáncer identificado según el segundo método descrito en la presente memoria, la analítica puede identificar regiones genómicas según el proceso 680 en la **Figura 6B.** El sistema de análisis define 690 un vector de característica para cada muestra, para cada región, para cada tipo de cáncer por un recuento de fragmentos de ADN que tienen una relación de probabilidad logarítmica calculada que el fragmento es indicativo de cáncer por encima de una pluralidad de umbrales, en donde cada recuento es un valor en el vector de características. En un escenario, el sistema de análisis cuenta el número de fragmentos presentes en una muestra en una región para cada tipo de cáncer con relaciones de probabilidad logarítmica por encima de uno o una pluralidad de posibles valores umbral. El sistema de análisis define un vector de características para cada muestra, por un recuento de fragmentos de ADN para cada región genómica para cada tipo de cáncer que proporciona una relación logarítmica de probabilidad calculada para el fragmento por encima de una pluralidad de umbrales, en donde cada recuento es un valor en el vector de características. El sistema de análisis utiliza los vectores de características definidos para calcular una puntuación informativa para cada región genómica que describe la capacidad de la región genómica para distinguir entre cada par de tipos de cáncer. Para cada par de tipos de cáncer, el sistema de análisis clasifica las regiones en base a las puntuaciones informativas. El sistema de análisis puede seleccionar regiones en base a la clasificación según las puntuaciones informativas.

El sistema de análisis calcula 695 una puntuación informativa para cada región que describe la capacidad de esa región para distinguir entre cada par de tipos de cáncer. Para cada par de tipos de cáncer distintos, el sistema de análisis puede especificar un tipo como un tipo positivo y el otro como un tipo negativo. En un escenario, la capacidad de una región para distinguir entre el tipo positivo y el tipo negativo se basa en información mutua, calculada usando la fracción estimada de muestras de ADNlc del tipo positivo y del tipo negativo para el que se esperaría que la característica sea distinta de cero en el ensayo final, es decir, al menos un fragmento de ese nivel que se secuenciaría en un ensayo de metilación dirigido. Esas fracciones se estiman usando las tasas observadas a las que se produce la característica en ADNlc sano y en muestras de ADNlc y/o tumorales de alta señal de cada tipo de cáncer. Por ejemplo, si una característica se produce con frecuencia en ADNlc sano, entonces también se estima que se produce con frecuencia en ADNlc de cualquier tipo de cáncer y probablemente daría como resultado una puntuación informativa baja. El sistema de análisis puede elegir un cierto número de regiones para cada par de tipos de cáncer de la clasificación, por ejemplo, 1024.

En escenarios adicionales, el sistema de análisis identifica además regiones predominantemente hipermetiladas o hipometiladas a partir de la clasificación de regiones. El sistema de análisis puede cargar el conjunto de fragmentos en el tipo o tipos positivos para una región que se identificó como informativa. El sistema de análisis, de los fragmentos cargados, evalúa si los fragmentos cargados están predominantemente hipermetilados o hipometilados. Si los fragmentos cargados están predominantemente hipermetilados o hipometilados, el sistema de análisis puede seleccionar sondas correspondientes al patrón de metilación predominante. Si los fragmentos cargados no están predominantemente hipermetilados o hipometilados, el sistema de análisis puede usar una mezcla de sondas para dirigir tanto la hipermetilación como la hipometilación. El sistema de análisis puede identificar además un conjunto mínimo de sitios CpG que se superponen más que algún porcentaje de los fragmentos.

En otros escenarios, el sistema de análisis, después de clasificar las regiones basadas en puntuaciones informativas, marca cada región con la clasificación informativa más baja en todos los pares de tipos de cáncer. Por ejemplo, si una región fuera la décima más informativa para distinguir entre mama y pulmón, y la quinta más informativa para distinguir entre mama y colorrectal, se le daría una etiqueta global de "5". El sistema de análisis puede diseñar sondas que

comienzan con las regiones con etiqueta más baja mientras se añaden regiones al panel, por ejemplo, hasta que se ha agotado el presupuesto del tamaño del panel.

Regiones genómicas fuera de la diana

En algunos escenarios, las sondas que se dirigen a regiones genómicas seleccionadas se filtran adicionalmente 475 en base al número de regiones fuera de diana. Esto es para sondas de cribado que extraen demasiados fragmentos de ADNlc correspondientes a, o derivados de, regiones genómicas fuera de diana. La exclusión de sondas que tienen muchas regiones fuera de diana puede ser valiosa disminuyendo las tasas fuera de diana y aumentando la cobertura objetivo para una cantidad dada de secuenciación.

Una región genómica fuera de diana es una región genómica que tiene una homología suficiente con respecto a una región genómica diana, de manera que las moléculas de ADN o fragmentos derivados de regiones genómicas fuera de diana hibridan y arrastran por una sonda diseñada para hibridarse con una región genómica diana. Una región genómica fuera de diana puede ser una región genómica (o una secuencia convertida de esa misma región) que se alinea con una sonda a lo largo de al menos 35 pb, 40 pb, 45 pb, 50 pb, 60 pb, 70 pb o 80 pb con al menos el 80 %, el 85 %, el 90 %, el 95 % o el 97 % de tasa de coincidencia. En un escenario, una región genómica fuera de la diana es una región genómica (o una secuencia convertida de esa misma región) que se alinea con una sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 %. Pueden adoptarse diversos métodos conocidos en la técnica para cribar regiones genómicas fuera de diana.

La búsqueda rápida del genoma para encontrar todas las regiones genómicas fuera de la diana puede ser computacionalmente desafiante. En un escenario, una estrategia de siembra de k-mero (que puede permitir una o más faltas de coincidencia) se combina con la alineación local en las ubicaciones de las semillas. En este caso, se puede garantizar una búsqueda exhaustiva de buenas alineaciones en base a la longitud de k-mero, se permite el número de faltas de coincidencia y el número de aciertos de semilla de k-mero en una ubicación particular. Esto requiere la alineación local de programación dinámica en un gran número de ubicaciones, por lo que este enfoque está altamente optimizado para usar las instrucciones de la CPU de vectores (por ejemplo, AVX2, AVX512) y también puede paralelizarse en muchos núcleos dentro de una máquina y también en muchas máquinas conectadas por una red. Un experto en la técnica reconocerá que las modificaciones y variaciones de este enfoque pueden implementarse con el fin de identificar regiones genómicas fuera de diana.

En algunos escenarios, se excluyen (o filtran) las sondas que tienen homología de secuencia con regiones genómicas fuera de diana, o moléculas de ADN correspondientes a, o derivadas de regiones genómicas fuera de diana que comprenden más de un número umbral. Por ejemplo, las sondas que tienen homología de secuencia con regiones genómicas fuera de diana, o moléculas de ADN correspondientes a, o derivadas de regiones genómicas fuera de diana de más de 30, más de 25, más de 20, más de 18, más de 15, más de 12, más de 10 o más de 5 regiones fuera de diana se excluyen.

En algunos escenarios, las sondas se dividen en 2, 3, 4, 5, 6 o más grupos separados dependiendo de los números de regiones fuera de la diana. Por ejemplo, las sondas que tienen homología de secuencia sin regiones fuera de diana o moléculas de ADN correspondientes a, o derivadas de regiones fuera de diana se asignan al grupo de alta calidad, las sondas que tienen homología de secuencia con las regiones fuera de diana de 1-18 o las moléculas de ADN correspondientes a, o derivadas de, las regiones fuera de diana 1-18, se asignan al grupo de baja calidad, y las sondas que tienen homología de secuencia con más de 19 regiones fuera de diana o moléculas de ADN correspondientes, o derivadas de 19 regiones fuera de diana, se asignan a un grupo de calidad deficiente. Pueden usarse otros valores de corte para el agrupamiento.

En algunos escenarios, se excluyen las sondas en el grupo de calidad más baja. En algunos escenarios, se excluyen las sondas en grupos distintos del grupo de mayor calidad. En algunos escenarios, se hacen paneles separados para las sondas en cada grupo. En algunos escenarios, todas las sondas se colocan en el mismo panel, pero el análisis separado se realiza en base a los grupos asignados.

En algunos escenarios, un panel comprende un mayor número de sondas de alta calidad que el número de sondas en grupos inferiores. En algunos escenarios, un panel comprende un número menor de sondas de baja calidad que el número de sondas en otro grupo. En algunos escenarios, más del 95 %, el 90 %, el 85 %, el 80 %, el 75 % o el 70 % de las sondas en un panel son sondas de alta calidad. En algunos escenarios, menos del 35 %, el 30 %, el 20 %, el 10 %, el 5 %, el 4 %, el 3 %, el 2 % o el 1 % de las sondas en un panel son sondas de baja calidad. En algunos escenarios, menos del 5 %, el 4 %, el 3 %, el 2 % o el 1 % de las sondas en un panel son sondas de baja calidad. En algunos escenarios, no se incluyen sondas de baja calidad en un panel.

En algún escenario, se excluyen las sondas que tienen por debajo del 50 %, por debajo del 40 %, por debajo del 30 %, por debajo del 20 %, por debajo del 10 % o por debajo del 5 %. En algún escenario, las sondas que tienen por encima del 30 %, por encima del 40 %, por encima del 50 %, por encima del 60 %, por encima del 70 %, por encima del 80 %, o por encima del 90 % se incluyen selectivamente en un panel.

Métodos de uso del panel de ensayo de cáncer

En otro aspecto más, se proporcionan métodos para usar un panel de ensayo de cáncer (alternativamente denominado "conjunto de cebos"). Los métodos pueden comprender las etapas de tratar moléculas o fragmentos de ADN para convertir citosinas no metiladas en uracilos (por ejemplo, usando tratamiento con bisulfito), aplicar un panel de cáncer (como se describe en la presente memoria) a las moléculas o fragmentos de ADN convertidos, enriquecer un subconjunto de moléculas de ADN transformadas o fragmentos que se unen a las sondas en el panel y secuenciar los fragmentos de ADNlc enriquecidos. En algunos escenarios, las lecturas de secuencia pueden compararse con un genoma de referencia (por ejemplo, un genoma de referencia humana), lo que permite la identificación de estados de metilación en una pluralidad de sitios CpG dentro de las moléculas de ADN o fragmentos y, por lo tanto, proporciona información relevante para la detección del cáncer.

Análisis de lecturas de secuencia

En algunos escenarios, las lecturas de secuencia pueden alinearse con un genoma de referencia usando métodos conocidos en la técnica para determinar la información de posición de alineación. La información de la posición de alineación puede indicar una posición inicial y una posición final de una región en el genoma de referencia que corresponde a una base de nucleótidos inicial y una base de nucleótidos final de una secuencia determinada leída. La información de la posición de alineación también puede incluir una longitud de lectura de secuencia, que puede determinarse desde la posición inicial y la posición final. Una región en el genoma de referencia puede asociarse con un gen o un segmento de un gen.

En diversos escenarios, una lectura de secuencia comprende un par de lectura indicado como R_1 y R_2. Por ejemplo, la primera lectura R_1 puede secuenciarse desde un primer extremo de un fragmento de ácido nucleico mientras que la segunda lectura R_2 puede secuenciarse desde el segundo extremo del fragmento de ácido nucleico. Por tanto, pares de bases de nucleótidos de la primera lectura R_1 y segunda lectura R_2 puede alinearse de manera consistente (por ejemplo, en orientaciones opuestas) con bases de nucleótidos del genoma de referencia. Información de posición de alineación derivada del par de lectura R_1 y R_2 puede incluir una posición inicial en el genoma de referencia que corresponde a un extremo de una primera lectura (por ejemplo, R_1) y una posición final en el genoma de referencia que corresponde a un extremo de una segunda lectura (por ejemplo, R_2). En otras palabras, la posición inicial y la posición final en el genoma de referencia representan la ubicación probable dentro del genoma de referencia al que corresponde el fragmento de ácido nucleico. Se puede generar un archivo de salida que tiene un formato SAM (mapa de alineación de secuencia) o un formato BAM (mapa de alineación binaria) y se emite para un análisis adicional.

A partir de las lecturas de secuencia, la ubicación y el estado de metilación para cada sitio CpG se pueden determinar en función de la alineación con un genoma de referencia. Además, puede generarse un vector de estado de metilación para cada fragmento que especifica una ubicación del fragmento en el genoma de referencia (por ejemplo, como se especifica por la posición del primer sitio CpG en cada fragmento, U otra métrica similar), un número de sitios CpG en el fragmento y el estado de metilación de cada sitio CpG en el fragmento ya sea metilado (por ejemplo, indicado como M), no metilado (por ejemplo, indicado como U), o indeterminado (por ejemplo, indicado como I). Los vectores de estado de metilación pueden almacenarse en memoria informática temporal o persistente para su uso y procesamiento posterior. Además, pueden eliminarse lecturas duplicadas o vectores de estado de metilación duplicados de un solo sujeto. En un escenario adicional, se puede determinar que un cierto fragmento tiene uno o más sitios CpG que tienen un estado de metilación indeterminado. Dichos fragmentos pueden excluirse del procesamiento posterior o incluirse selectivamente donde el modelo de datos aguas abajo representa dichos estados de metilación indeterminados.

La **Figura 7B** es una ilustración del proceso 100 de la **Figura 7A** de secuenciación de un fragmento de ADNlc para obtener un vector de estado de metilación, según un escenario. Como ejemplo, el sistema de análisis toma un fragmento de ADNlc 112. En este ejemplo, el fragmento de ADNlc 112 contiene tres sitios CpG. Como se muestra, el primer y tercer sitios de CpG del fragmento de ADNlc 112 están metilados 114. Durante la etapa de tratamiento 120, el fragmento de ADNlc 112 se convierte para generar un fragmento de ADNlc convertido 122. Durante el tratamiento 120, el segundo sitio CpG que no estaba metilado tenía su citosina convertida en uracilo. Sin embargo, el primer y tercer sitios de CpG no se convierten.

Después de la conversión, se prepara y secuenció una biblioteca 130 de secuenciación generando una lectura de secuencia 142. El sistema de análisis alinea 150 la lectura de secuencia 142 con un genoma de referencia 144. El genoma de referencia 144 proporciona el contexto en cuanto a qué posición en un genoma humano se origina el fragmento ADNlc En este ejemplo simplificado, el sistema de análisis alinea 150 la secuencia leída de manera que los tres sitios CpG se correlacionan con los sitios CpG 23, 24 y 25 (identificadores de referencia arbitrarios usados por conveniencia de la descripción). Por tanto, el sistema de análisis genera información tanto en el estado de metilación de todos los sitios CpG en el fragmento de ADNlc 112 como en la posición en el genoma humano, el mapa de sitios CpG. Como se muestra, los sitios CpG en la secuencia de lectura 142 que estaban metilados se leen como citosinas. En este ejemplo, las citosina aparecen en la lectura de secuencia 142 solo en el primer y tercer sitio CpG que permite inferir que el primer y tercer sitios CpG en el fragmento de ADNlc original estaban metilados. El segundo sitio CpG se lee como una timina (U se convierte en T durante el proceso de secuenciación) y, por lo tanto, se puede inferir que el segundo sitio CpG no estaba metilado en el fragmento de ADNlc original. Con estos dos fragmentos de información,

el estado y ubicación de metilación, el sistema de análisis genera 160 un vector de estado de metilación 152 para el fragmento ADNlc 112. En este ejemplo, el vector de estado de metilación 152 resultante es < M_{23}, U_{24}, M_{25} >, en donde M corresponde a un sitio CpG metilado, U corresponde a un sitio CpG no metilado, y los números de subíndice corresponden a las posiciones de cada sitio CpG en el genoma de referencia.

Las **Figuras 8A-8B** muestran tres gráficos de datos que validan la coherencia de la secuenciación a partir de un grupo de control. El primer gráfico 170 muestra la exactitud de la conversión de citosinas no metiladas a uracilo (paso 120) en el fragmento de ADNlc obtenido de una muestra de prueba a través de sujetos en diferentes estadios de cáncer - estadio 0, estadio I, estadio II, estadio III, estadio IV y sin cáncer. Como se muestra, hubo una consistencia uniforme en la conversión de citosinas no metiladas en fragmentos de ADNlc en uracilos. Hubo una precisión general de conversión del 99,47 % con una precisión de ±0,024 %. El segundo gráfico 180 compara la cobertura (profundidad de secuenciación) en diferentes etapas del cáncer. Recuento solo de lecturas de secuencia que se mapearon de forma segura con un genoma de referencia, la cobertura media sobre todos los grupos fue ~34. El tercer gráfico 190 muestra la concentración de ADNlc por muestra en diferentes estadios de cáncer.

Detección del cáncer

Las lecturas de secuencia obtenidas por los métodos proporcionados en la presente memoria se procesan adicionalmente mediante algoritmos automatizados. Por ejemplo, el sistema de análisis se utiliza para recibir datos de secuenciación de un secuenciador y realizar diversos aspectos del procesamiento como se describe en la presente memoria. El sistema de análisis puede ser uno de un ordenador personal (PC), un ordenador de sobremesa, un ordenador portátil, un cuaderno, una tablet, un dispositivo móvil. Un dispositivo informático puede acoplarse comunicativamente al secuenciador a través de una combinación inalámbrica, cableada o de comunicación por cable. Generalmente, el dispositivo informático está configurado con un procesador y memoria que almacena instrucciones informáticas que, cuando son ejecutadas por el procesador, hacen que el procesador realice las etapas como se describe en el resto de este documento. Generalmente, la cantidad de datos genéticos y datos derivados de los mismos es lo suficientemente grande, y la cantidad de potencia computacional requerida es tan grande, por lo que debe ser imposible realizarse en papel o por la mente humana.

La interpretación clínica del estado de metilación de las regiones genómicas específicas es un proceso que incluye clasificar el efecto clínico de cada uno o una combinación del estado de metilación e informar los resultados de formas que son significativas para un profesional médico. La interpretación clínica puede basarse en la comparación de las lecturas de secuencia con base de datos específica para sujetos con cáncer o no cáncer, y/o basarse en números y tipos de los fragmentos de ADNlc que tienen patrones de metilación específicos del cáncer identificados a partir de una muestra. En algunos escenarios, las regiones genómicas dirigidas se clasifican o clasifican en función de su similitud para metilarse diferencialmente en muestras de cáncer, y los rangos o clasificaciones se usan en el proceso de interpretación. Los montones y clasificaciones pueden incluir (1) el tipo de efecto clínico, (2) la resistencia de evidencia del efecto y (3) el tamaño del efecto. Pueden adoptarse diversos métodos para el análisis clínico e interpretación de los datos del genoma para el análisis de las lecturas de secuencia. En algunos otros escenarios, la interpretación clínica de los estados de metilación de dichas regiones metiladas diferencialmente puede basarse en enfoques de aprendizaje automático que interpretan una muestra actual basándose en un método de clasificación o regresión que se entrenó mediante el uso de los estados de metilación de tales regiones metiladas diferencialmente a partir de muestras de pacientes con cáncer y no cáncer con estado de cáncer conocido, tipo de cáncer, estadio de cáncer, TOO, etc.

La información con significado clínico puede incluir la presencia o ausencia de cáncer en general, la presencia o ausencia de determinados tipos de cáncer, el estadio del cáncer, o la presencia o ausencia de otros tipos de enfermedades. En algunos escenarios, la información se refiere a la presencia o ausencia de uno o más tipos de cáncer, seleccionados del grupo que consiste en cáncer de mama, cáncer de endometrio, cáncer de cuello uterino, cáncer de ovario, cáncer de vejiga, cáncer urotelial de pelvis renal, carcinoma de células renales, cáncer de próstata, cáncer anorrectal, cáncer anal, cáncer colorrectal, cáncer hepatocelular, cáncer de hígado/conducto biliar, colangiocarcinoma y cáncer hepatobiliar, cáncer de páncreas, adenocarcinoma del tracto gastrointestinal superior, cáncer esofágico de células escamosas, cáncer de cabeza y cuello, cáncer de pulmón, cáncer de pulmón de células escamosas, adenocarcinoma de pulmón, cáncer de pulmón de células pequeñas, cáncer neuroendocrino, melanoma, cáncer de tiroides, sarcoma, neoplasia de células plasmáticas, mieloma múltiple, neoplasia mieloide, linfoma y leucemia. En algunos escenarios, la información se refiere a la presencia o ausencia de uno o más tipos de cáncer, seleccionados del grupo que consiste en cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma, linfoma, cáncer de pulmón de células pequeñas, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatobiliar (hcc), biliar hepatobiliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal. En algunos escenarios, la información se refiere a la presencia o ausencia de uno o más tipos de cáncer, seleccionados del grupo que consiste en cáncer anal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago. En algunos escenarios, la información se refiere a la presencia o ausencia de uno

o más tipos de cáncer, seleccionados del grupo que consiste en cáncer de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello de útero, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón. En algunos escenarios, las muestras no son cancerosas y son de sujetos que tienen expansión clonal de glóbulos blancos o ningún cáncer.

Clasificador de cáncer

En algunos ejemplos, el panel de ensayo descrito en la presente memoria puede usarse con un clasificador de tipo cáncer que predice un estado de enfermedad para una muestra, tal como una predicción de cáncer o no cáncer, una predicción de tejido de origen y/o una predicción indeterminada. En algunos ejemplos, el clasificador de tipo cáncer puede generar características basadas en lecturas de secuencia teniendo en cuenta fragmentos de ADN metilados o no metilados en ciertas áreas genómicas de interés. Por ejemplo, si el clasificador de tipo cáncer determina que un patrón de metilación en un fragmento se asemeja al de un cierto tipo de cáncer, entonces el clasificador de tipo cáncer puede establecer una característica para ese fragmento como 1, y de lo contrario si no hay dicho fragmento presente, entonces la característica puede establecerse como 0. De esta manera, el clasificador de tipo cáncer puede producir un conjunto de características binarias (simplemente a modo de ejemplo, 30.000 características) para cada muestra. Además, en algunos ejemplos, toda o una parte del conjunto de características binarias para una muestra puede introducirse en el clasificador de tipo cáncer para proporcionar un conjunto de puntuaciones de probabilidad, tal como una puntuación de probabilidad por clase de tipo cáncer y para una clase de tipo no cancerosa. Además, en algunos ejemplos, el clasificador de tipo cáncer puede incorporar o usarse de otra manera junto con el umbral para determinar si una muestra debe denominarse cáncer o no cáncer, y/o un umbral indeterminado para reflejar la confianza en una llamada TOO específica. Dichos métodos se describen adicionalmente a continuación.

Para entrenar el clasificador de tipo cáncer, el sistema de análisis (por ejemplo, el sistema de análisis 800, la Figura 12B) puede obtener un conjunto de muestras de entrenamiento. En algunos ejemplos, cada muestra de entrenamiento incluye archivo(s) de fragmento (por ejemplo, datos de lectura de secuencia que contienen archivo), una etiqueta correspondiente a un tipo de cáncer (TOO) o estado no canceroso de la muestra y/o sexo del individuo de la muestra. El sistema de análisis puede utilizar el conjunto de entrenamiento para entrenar el clasificador de tipo cáncer para predecir el estado de enfermedad de la muestra.

En algunos ejemplos, para el entrenamiento, el sistema de análisis divide el genoma (por ejemplo, el genoma completo) o un subconjunto del genoma (por ejemplo, regiones de metilación específicas) en regiones. Simplemente a modo de ejemplo, las porciones del genoma pueden separarse en "bloques" de CpG, por lo que un nuevo bloque comienza siempre que haya una separación entre los CpG más cercanos a los vecinos es al menos una distancia de separación mínima (por ejemplo, al menos 500 pb). Además, en algunos ejemplos, cada bloque puede dividirse en regiones de 1000 pb y colocarse de manera que las regiones vecinas tengan una cierta cantidad (por ejemplo, 50 % o 500 pb) de superposición.

Además, en algunos ejemplos, el sistema de análisis puede dividir el conjunto de entrenamiento en K subconjuntos o pliegues que se utilizarán en una validación cruzada de K pliegues. En algunos ejemplos, los pliegues pueden equilibrarse en función del estado oncológico/no oncológico, el tejido de origen, el estadio del cáncer, la edad (por ejemplo, agrupados en grupos de 10 años) y/o el hábito de fumar. En algunos ejemplos, el conjunto de entrenamiento se divide en 5 pliegues, por lo que 5 clasificadores separados son entrenados, en cada caso entrenamiento en 4/5 de las muestras de entrenamiento y usando el 1/5 restante para la validación.

Durante el entrenamiento con el conjunto de entrenamiento, el sistema de análisis puede, para cada tipo de cáncer (y para el ADNlc sano), ajustar un modelo probabilístico a los fragmentos derivados de las muestras de ese tipo. Como se usa en la presente memoria, un "modelo probabilístico" es cualquier modelo matemático capaz de asignar una probabilidad a una secuencia leída en función del estado de metilación en uno o más sitios en la lectura. Durante el entrenamiento, el sistema de análisis se ajusta a lecturas de secuencia derivadas de una o más muestras de sujetos que tienen una enfermedad conocida y pueden usarse para determinar las probabilidades de lecturas de secuencia indicativas de un estado de enfermedad que utiliza información de metilación o vectores de estado de metilación. En particular, en algunos casos, el sistema de análisis determina las tasas observadas de metilación para cada sitio CpG dentro de una secuencia leída. La tasa de metilación representa una fracción o porcentaje de pares de bases que están metilados dentro de un sitio CpG. El modelo probabilístico entrenado puede parametrizarse mediante productos de las tasas de metilación. En general, puede usarse cualquier modelo probabilístico conocido para asignar probabilidades a lecturas de secuencia de una muestra. Por ejemplo, el modelo probabilístico puede ser un modelo binomial, en donde cada sitio (por ejemplo, sitio CpG) en un fragmento de ácido nucleico se asigna una probabilidad de metilación, o un modelo de sitios independientes, en donde cada metilación de CpG se especifica por una probabilidad de metilación distinta con metilación en un sitio que se supone que es independiente de la metilación en uno o más sitios diferentes en el fragmento de ácido nucleico.

En algunos ejemplos, el modelo probabilístico es un modelo de Markov, en donde la probabilidad de metilación en cada sitio CpG depende del estado de metilación en algún número de sitios CpG anteriores en la secuencia leída, o

la molécula de ácido nucleico de la que se deriva la lectura de secuencia. Véase por ejemplo, la solicitud de patente estadounidense n.º 16/352.602, titulada "Anomalous Fragment Detection and Classification," y presentada el 13 de marzo de 2019.

En algunos ejemplos, el modelo probabilístico es un "modelo de mezcla" equipado con una mezcla de componentes de modelos subyacentes. Por ejemplo, en algunos escenarios, los componentes de la mezcla pueden determinarse mediante el uso de múltiples modelos de sitios independientes, donde se supone que la metilación (por ejemplo, las tasas de metilación) en cada sitio CpG es independiente de la metilación en otros sitios CpG. Utilizando un modelo de sitios independientes, la probabilidad asignada a una lectura de secuencia, o la molécula de ácido nucleico de la que deriva, es el producto de la probabilidad de metilación en cada sitio CpG donde la lectura de secuencia está metilada y una menos la probabilidad de metilación en cada sitio CpG donde la lectura de secuencia no está metilada. Según este ejemplo, el sistema de análisis determina las tasas de metilación de cada uno de los componentes de la mezcla. El modelo de mezcla se parametriza mediante una suma de los componentes de la mezcla, cada uno asociado con un producto de las tasas de metilación. Un modelo probabilístico *Pr* de *n* componentes de mezcla pueden representarse como:

$$Pr(\textit{fragmento}\,|\{\beta_{ki}, f_k\}) = \sum_{k=1}^{n} f_k \prod_i \beta_{ki}^{m_i}(1-\beta_{ki})^{1-m_i}$$

Para un fragmento de entrada, $m_i \in \{0,1\}$ representa el estado de metilación observado del fragmento en la posición *i* de un genoma de referencia, indicando 0 no metilación e indicando 1 metilación. Una asignación fraccional A cada componente de mezcla *k* es f_k, donde $f_k \geq 0$ and $\sum_{k=1}^{n} f_k = 1$. La probabilidad de metilación en posición *i* en un sitio CpG del componente de mezcla *k* es β_{ki}. Por tanto, la probabilidad de no metilación es 1 - β_{ki}. El número de componentes de mezcla n puede ser 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, etc.

En algunos ejemplos, el sistema de análisis se ajusta al modelo probabilístico usando una estimación máxima de la probabilidad para identificar un conjunto de parámetros $\{\beta_{ki}, f_k\}$ que maximiza la probabilidad logarítmica de todos los fragmentos derivados de un estado de enfermedad, sujeto a una penalización de regularización aplicada a cada probabilidad de metilación con resistencia de regularización r. La cantidad maximizada para N fragmentos totales puede representarse como:

$$\sum_{j}^{N} ln\left(Pr\left(\textit{fragmento}_j\,|\{\beta_{ki}, f_k\}\right)\right) + r \cdot ln\left(\beta_{ki}(1-\beta_{ki})\right)$$

En algunos ejemplos, el sistema de análisis realiza ajustes por separado para cada tipo de cáncer y para ADNlc sano. Como apreciaría un experto en la técnica, pueden usarse otros medios para ajustar los modelos probabilísticos o para identificar parámetros que maximizan la probabilidad logarítmica de todas las lecturas de secuencia derivadas de las muestras de referencia. Por ejemplo, en algunos ejemplos, se usa el ajuste bayesiano (usando, por ejemplo, la cadena Markov Monte Carlo), en donde cada parámetro no se asigna un valor único, sino que se asocia a una distribución. En algunos ejemplos, se usa la optimización basándose en gradiente, en la que el gradiente de la probabilidad (o probabilidad logarítmica) con respecto a los valores de parámetro se usa para paso a través del espacio de parámetro hacia un óptimo. En todavía algunos ejemplos, la maximización de expectativa, en la que se deriva un conjunto de parámetros latentes (tales como identidades del componente de mezcla de las cuales se deriva cada fragmento) en sus valores esperados bajo los parámetros del modelo anteriores, y luego los parámetros del modelo se asignan para maximizar la probabilidad condicional de los valores supuestos de esas variables latentes. El proceso de dos etapas se repite hasta la convergencia.

Además, en algunos ejemplos, el sistema de análisis puede generar características para cada muestra en el conjunto de entrenamiento. Por ejemplo, para cada muestra (independientemente de la etiqueta), en cada región, para cada tipo de cáncer, para cada fragmento, el sistema de análisis puede evaluar la relación log-verosimilitud *R* con los modelos probabilísticos ajustados según:

$$R_{\textit{cáncer tipo A}}(\textit{fragmento}) \equiv En\left(\frac{Pr\,(\textit{fragmento}|\textit{cáncer tipo A})}{Pr\,(\textit{fragmento}|\textit{ADNlc sano})}\right)$$

A continuación, para cada muestra, para cada región, para cada tipo de cáncer, para cada uno de un conjunto de valores de "nivel", el sistema de análisis puede contar el número de fragmentos con tipo de cáncer>asignar los recuentos como funciones no negativas de valor entero. Por ejemplo, los títulos incluyen valores umbral de 1, 2, 3, 4, 5, 6, 7, 8 y 9, lo que da como resultado cada región que aloja 9 características por tipo de cáncer.

En algunos ejemplos, el sistema de análisis puede seleccionar ciertas características para su inclusión en un vector de características para cada muestra. Por ejemplo, para cada par de tipos de cáncer distintos, el sistema de análisis puede especificar un tipo como el "tipo positivo" y el otro como el "tipo negativo" y clasificar las características por su capacidad para distinguir esos tipos. En algunos casos, la clasificación se basa en información mutua calculada por el sistema de análisis. Por ejemplo, la información mutua puede calcularse usando la fracción estimada de muestras del tipo positivo y tipo negativo (por ejemplo, tipos de cáncer a y B) para los cuales se espera que la característica sea distinta de cero en un ensayo resultante. Por ejemplo, si una característica se produce con frecuencia en ADNlc sano, el sistema de análisis determina que la característica es poco probable que ocurra con frecuencia en ADNlc asociado con diversos tipos de cáncer. Por consiguiente, la característica puede ser una medida débil para distinguir entre estados de enfermedad. Al calcular la información mutua *I* , la variable *X* es una determinada característica (por ejemplo, binaria) y variable *Y* representa un estado de enfermedad, por ejemplo, cáncer tipo A o B:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y)\ log\ log\ \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

$$I \approx \frac{1}{2}\left(p(1|A) \cdot log\left(\frac{p(1|A)}{\frac{1}{2}\left(p(1|A) + p(1|B)\right)} \right) + p(1|B) \cdot log\left(\frac{p(1|B)}{\frac{1}{2}\left(p(1|A)p(1|B)\right)} \right) \right)$$

$$p(1|A) = f_A + f_H - f_H f_A$$

La función de masa de probabilidad conjunta de *x* y *y* es p (x, y) y las funciones de masa de probabilidad marginal son p(x) y p(y). El sistema de análisis puede asumir que la ausencia de características no es informativa y el estado de la enfermedad es igualmente probable *a priori,* por ejemplo, $p(Y = A) = p(Y = B) = 0{,}5$. La probabilidad de observar (por ejemplo, en ADNlc) una característica binaria dada del cáncer tipo a está representada por *p* (1aspárticoa), donde f_a es la probabilidad de observar la característica en muestras de ADNc de muestras de tumor (o ADNlc de alta señal) asociadas con el cáncer tipo a, y f_H es la probabilidad de observar la característica en una muestra de ADNlc sana o no cancerosa.

En algunos ejemplos, solo las características correspondientes al tipo positivo se incluyen en la clasificación, y solo cuando esas características de la tasa de ocurrencia prevista es mayor en el tipo positivo que en el tipo negativo. Por ejemplo, si "hígado" es el tipo positivo y "mama" es el tipo negativo, entonces solo se consideran características "hígado_x", y solo si su aparición estimada en el ADNlc hepático es mayor que su aparición estimada en ADNlc de mama. Además, en algunos ejemplos, para cada región, para cada par de tipos de cáncer (que incluye no cáncer como tipo negativo), el sistema de análisis mantiene solo el nivel de mejor rendimiento. Además, en algunos ejemplos, el sistema de análisis transforma los valores característicos mediante binarización, por lo que cualquier valor característico mayor que 0 se establece en 1, de modo que todas las características sean 0 ó 1.

En algunos ejemplos, el sistema de análisis trenes de un clasificador de regresión logística multinomial en los datos de entrenamiento para un pliegue, y genera predicciones para los datos de salida. Por ejemplo, para cada uno de los K pliegues, se puede entrenar una regresión logística para cada combinación de hiperparámetros. Dichos hiperparámetros pueden incluir penalización de L2 y/o topoK (por ejemplo, el número de regiones de alto rango para mantenerse por par de tipos de tejido (incluido no cáncer), como se clasifica por el procedimiento de información mutua descrito anteriormente). Para cada conjunto de hiperparámetros, el rendimiento se evalúa en las predicciones validadas cruzadas del conjunto de entrenamiento completo, y el conjunto de hiperparámetros con el mejor rendimiento se selecciona para reentrenar en el conjunto de entrenamiento completo. En algunos ejemplos, el sistema de análisis usa la pérdida de registro como una métrica de rendimiento, por lo que la pérdida de registro se calcula tomando el logaritmo negativo de la predicción para la etiqueta correcta para cada muestra, y luego sumando sobre muestras (es decir, una predicción perfecta de 1,0 para la etiqueta correcta daría una pérdida logarítmica de 0).

Para generar predicciones para una nueva muestra, los valores de las características se calculan utilizando el mismo método descrito anteriormente, pero restringido a las características (combinaciones región/clase positiva) seleccionadas bajo el valor topK elegido. Las características generadas se usan entonces para crear una predicción usando el modelo de regresión logística entrenado anteriormente.

En algunos ejemplos, los trenes de analíticas un clasificador de dos etapas. Por ejemplo, el sistema de análisis trenes de un clasificador de cáncer binario para distinguir entre las etiquetas, el cáncer y el no cáncer, basándose en los vectores de características de las muestras de entrenamiento. En este caso, el clasificador binario emite una puntuación de predicción que indica la probabilidad de la presencia o ausencia de cáncer. En otro ejemplo, el sistema

de análisis entrena un clasificador multiclase de cáncer para distinguir entre muchos tipos de cáncer. En este clasificador de cáncer de múltiples clases, el clasificador de cáncer se entrena para determinar una predicción de cáncer que comprende un valor de predicción para cada uno de los tipos de cáncer que se clasifican. Los valores de predicción pueden corresponder a una probabilidad de que una muestra dada tenga cada uno de los tipos de cáncer. Por ejemplo, el clasificador de cáncer devuelve una predicción del cáncer que incluye un valor de predicción para el cáncer de mama, cáncer de pulmón y no cáncer. Por ejemplo, el clasificador de cáncer puede devolver una predicción de cáncer para una muestra de prueba que incluya una puntuación de predicción de cáncer de mama, cáncer de pulmón y/o ningún cáncer.

El sistema de análisis puede entrenar el clasificador de cáncer según uno cualquiera de una serie de métodos. Por ejemplo, el clasificador binario de cáncer puede ser un clasificador de regresión logística L2-regularizado que se entrena utilizando una función de pérdida logarítmica. Como otro ejemplo, el clasificador multicáncer (TOO) puede ser una regresión logística multinomial. En la práctica, cualquier tipo de clasificador de cáncer puede entrenarse utilizando otras técnicas. Estas técnicas son numerosas incluyendo uso potencial de métodos de núcleo, algoritmos de aprendizaje automático tales como redes neuronales multicapa, etc. En particular, métodos como se describe en el documento PCT/US 2019/022122 y la solicitud de patente estadounidense n.º 16/352.602 pueden usarse para diversos escenarios. Aún más, en algunos ejemplos, el clasificador TOO se entrena solo en muestras de cáncer que se denominan con éxito cáncer por el clasificador binario, asegurando así una señal de cáncer suficiente en la muestra de cáncer. Por otro lado, en algunos ejemplos, el clasificador binario se entrena en las muestras de entrenamiento independientemente del TOO.

Secuenciador y sistema de análisis ilustrativos

La **Figura 12A** es un diagrama de flujo de sistemas y dispositivos para secuenciar muestras de ácido nucleico según un escenario. Este diagrama de flujo ilustrativo incluye dispositivos tales como un secuenciador 820 y un sistema 800 de análisis. El secuenciador 820 y el sistema 800 de analítica pueden funcionar en tándem para realizar una o más etapas en los procesos descritos en la presente memoria.

En diversos escenarios, el secuenciador 820 recibe una muestra de ácido nucleico enriquecida 810. Como se muestra en la **Figura 12A,** el secuenciador 820 puede incluir una interfaz gráfica de usuario 825 que permite interacciones de usuario con tareas particulares (por ejemplo, iniciar secuenciación o terminar secuenciación) así como una más estaciones de carga 830 para cargar un cartucho de secuenciación que incluye las muestras de fragmentos enriquecidas y/o para cargar los tampones necesarios para realizar los ensayos de secuenciación. Por tanto, una vez que un usuario del secuenciador 820 ha proporcionado los reactivos necesarios y el cartucho de secuenciación a la estación 830 de carga del secuenciador 820, el usuario puede iniciar la secuenciación interactuando con la interfaz gráfica de usuario 825 del secuenciador 820. Una vez iniciada, el secuenciador 820 realiza la secuenciación y emite las lecturas de secuencia de los fragmentos enriquecidos de la muestra de ácido nucleico 810.

En algunos escenarios, el secuenciador 820 está acoplado comunicativamente con el sistema de análisis 800. El sistema 800 de análisis incluye algún número de dispositivos informáticos utilizados para procesar las lecturas de secuencia para diversas aplicaciones, tales como evaluar el estado de metilación en uno o más sitios CpG, llamada variante o control de calidad. El secuenciador 820 puede proporcionar las lecturas de secuencia en un formato de archivo BAM al sistema 800 de analítica. El sistema 800 de analítica se puede acoplar comunicativamente al secuenciador 820 a través de una red inalámbrica, cableada o de comunicación inalámbrica. Generalmente, el sistema 800 de analítica está configurado con un procesador y un medio de almacenamiento legible por ordenador no transitorio que almacena instrucciones informáticas que, cuando son ejecutadas por el procesador, hacen que el procesador procese las lecturas de secuencia o realice una o más etapas de cualquiera de los métodos o procesos descritos en la presente memoria.

En algunos escenarios, las lecturas de secuencia pueden alinearse con un genoma de referencia usando métodos conocidos en la técnica para determinar la información de posición de alineación. La posición de alineación generalmente puede describir una posición inicial y una posición final de una región en el genoma de referencia que corresponde a una base de nucleótidos inicial y una base de nucleótidos final de una secuencia de lectura dada. Correspondientes a la secuenciación de metilación, la información de la posición de alineación puede generalizarse para indicar un primer sitio CpG y un último sitio CpG incluido en la secuencia leída según la alineación con el genoma de referencia. La información de la posición de alineación puede indicar además estados de metilación y ubicaciones de todos los sitios CpG en una secuencia dada. Una región en el genoma de referencia puede asociarse con un gen o un segmento de un gen; como tal, el sistema de análisis 800 puede etiquetar una secuencia leída con uno o más genes que se alinean con la secuencia leída. En un escenario, la longitud del fragmento (o tamaño) se determina desde las posiciones inicial y final.

En diversos escenarios, por ejemplo cuando se usa un proceso de secuenciación de extremos emparejados, una lectura de secuencia comprende un par de lectura indicado como R_1 y R_2. Por ejemplo, la primera lectura R_1 se puede secuenciar desde un primer extremo de una molécula de ADN bicatenario (ADNbc) mientras que la segunda lectura R_2 se puede secuenciar desde el segundo extremo del ADN bicatenario (ADNbc). Por tanto, los pares de bases de nucleótidos de la primera lectura R_1 y la segunda lectura R_2 pueden alinearse de manera consistente (por

ejemplo, en orientaciones opuestas) con bases de nucleótidos del genoma de referencia. La información de posición de alineación derivada del par de lectura R_1 y R_2 puede incluir una posición inicial en el genoma de referencia que corresponde a un extremo de una primera lectura (por ejemplo, R_1) y una posición final en el genoma de referencia que corresponde a un extremo de una segunda lectura (por ejemplo, R_2). En otras palabras, la posición inicial y la posición final en el genoma de referencia representan la ubicación probable dentro del genoma de referencia al que corresponde el fragmento de ácido nucleico. En un escenario, el par de lectura R_1 y R_2 se pueden ensamblar en un fragmento, y el fragmento usado para el análisis y/o clasificación posterior. Se puede generar un archivo de salida que tenga formato SAM (mapa de alineación de secuencia) o formato BAM (binario) y se emita para su posterior análisis.

Con referencia ahora a la **Figura 12B, Figura 12B** es un diagrama de bloques de un sistema de análisis 800 para procesar muestras de ADN según un escenario. El sistema de análisis implementa uno o más dispositivos informáticos para su uso en el análisis de muestras de ADN. El sistema de análisis 800 incluye un procesador de secuencia 840, base de datos de secuencias 845, base de datos de modelos 855, modelos 850, base de datos de parámetros 865 y motor de puntuación 860. En algunos escenarios, el sistema de análisis 800 realiza una o más etapas en los procesos 300 de la **Figura 3A,** 340 de la **Figura 3B,** 400 de la **Figura 4,** 500 de la **Figura 5,** 600 de la **Figura 6A,** o 680 de la **Figura 6B** y otro proceso descrito en la presente memoria.

El procesador de secuencia 840 genera vectores de estado de metilación para fragmentos de una muestra. En cada sitio CpG en un fragmento, el procesador de secuencia 840 genera un vector de estado de metilación para cada fragmento que especifica una ubicación del fragmento en el genoma de referencia, un número de sitios CpG en el fragmento y el estado de metilación de cada sitio CpG en el fragmento ya sea metilado, no metilado o indeterminado a través del proceso 300 de la **Figura 3A.** El procesador de secuencia 840 puede almacenar vectores de estado de metilación para fragmentos en la base de datos de secuencias 845. Los datos en la base de datos de secuencias 845 pueden organizarse de manera que los vectores de estado de metilación de una muestra están asociados entre sí.

Además, pueden almacenarse múltiples modelos 850 diferentes en la base de datos 855 de modelo o recuperarse para su uso con muestras de prueba. En un ejemplo, un modelo es un clasificador de cáncer entrenado para determinar una predicción de cáncer para una muestra de prueba mediante el uso de un vector de características derivado de fragmentos anómalos. El entrenamiento y uso del clasificador de cáncer se describe en otra parte de la presente memoria. El sistema de análisis 800 puede entrenar uno o más modelos 850 y almacenar diversos parámetros entrenados en la base de datos de parámetros 865. El sistema de análisis 800 almacena los modelos 850 junto con funciones en la base de datos de modelos 855.

Durante la inferencia, el motor 860 de puntuación usa los uno o más modelos 850 para devolver salidas. El motor de puntuación 860 accede a los modelos 850 en la base de datos de modelos 855 junto con parámetros entrenados de la base de datos de parámetros 865. Según cada modelo, el motor de puntuación recibe una entrada adecuada para el modelo y calcula una salida basándose en la entrada recibida, los parámetros y una función de cada modelo relacionan la entrada y la salida. En algunos casos de uso, el motor 860 de puntuación calcula además métricas que se correlacionan con una confianza en las salidas calculadas del modelo. En otros casos de uso, el motor 860 de puntuación calcula otros valores intermediarios para su uso en el modelo.

Cáncer y seguimiento del tratamiento

En ciertos escenarios, el primer punto de tiempo es antes de un tratamiento contra el cáncer (por ejemplo, antes de una cirugía de resección o una intervención terapéutica), y el segundo punto de tiempo es después de un tratamiento contra el cáncer (por ejemplo, después de una cirugía de resección o intervención terapéutica), y el método utilizado para controlar la efectividad del tratamiento. Por ejemplo, si la segunda probabilidad o puntuación de probabilidad disminuye en comparación con la primera probabilidad o puntuación de probabilidad, entonces se considera que el tratamiento ha sido exitoso. Sin embargo, si la segunda probabilidad o puntuación de probabilidad aumenta en comparación con la primera probabilidad o puntuación de probabilidad, entonces se considera que el tratamiento no ha sido exitoso. En otras realizaciones, tanto el primer como el segundo punto de tiempo son antes de un tratamiento contra el cáncer (por ejemplo, antes de una cirugía de resección o una intervención terapéutica). En otros escenarios, tanto el primer como el segundo punto de tiempo son después de un tratamiento contra el cáncer (por ejemplo, antes de una cirugía de resección o una intervención terapéutica) y el método se usa para controlar la efectividad del tratamiento o la pérdida de efectividad del tratamiento. En otros escenarios, se pueden obtener muestras de ADNlc de un paciente con cáncer en un primer y segundo punto de tiempo y analizarse, por ejemplo, para controlar la progresión del cáncer, para determinar si un cáncer está en remisión (por ejemplo, después del tratamiento), para controlar o detectar la enfermedad residual o recurrencia de la enfermedad, o para controlar la eficacia del tratamiento (por ejemplo, terapéutica).

Los expertos en la técnica apreciarán fácilmente que las muestras de prueba pueden obtenerse de un paciente con cáncer durante cualquier conjunto de puntos de tiempo deseados y analizarse según los métodos de la invención para controlar un estado de cáncer en el paciente. En algunos escenarios, el primer y segundo puntos temporales están separados por una cantidad de tiempo que oscila entre aproximadamente 15 minutos hasta aproximadamente 30 años, tal como aproximadamente 30 minutos, tal como aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,

15, 16, 17, 18, 19, 20, 21, 22, 23, o aproximadamente 24 horas, como aproximadamente 1, 2, 3, 4, 5, 10, 15, 20, 25 o aproximadamente 30 días, o como aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, o 12 meses, o como aproximadamente 1, 1. 5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12, 12.5, 13, 13.5, 14, 14.5, 15, 15.5, 16, 16.5, 17, 17. 5, 18, 18,5, 19, 19,5, 20, 20,5, 21, 21,5, 22, 22,5, 23, 23,5, 24, 24,5, 25, 25,5, 26, 26,5, 27, 27,5, 28, 28,5, 29, 29,5 o aproximadamente 30 años. En otros escenarios, las muestras de prueba se pueden obtener del paciente al menos una vez cada 3 meses, al menos una vez cada 6 meses, al menos una vez al año, al menos una vez cada 2 años, al menos una vez cada 3 años, al menos una vez cada 4 años, o al menos una vez cada 5 años.

Tratamiento (no reivindicado)

En otro escenario más, la información obtenida de cualquier método descrito en la presente memoria (por ejemplo, la probabilidad o puntuación de probabilidad) puede usarse para hacer o influir en una decisión clínica (por ejemplo, diagnóstico de cáncer, selección de tratamiento, evaluación de la efectividad del tratamiento, etc.). Por ejemplo, en un escenario, si la probabilidad o puntuación de probabilidad excede un umbral, un médico puede prescribir un tratamiento apropiado (por ejemplo, una cirugía de resección, radioterapia, quimioterapia y/o inmunoterapia). En algunos escenarios, la información tal como una puntuación de probabilidad o probabilidad puede proporcionarse como una lectura a un médico o sujeto.

Un clasificador (como se describe en la presente memoria) puede usarse para determinar una probabilidad o puntuación de probabilidad de que un vector de características de muestra sea de un sujeto que tenga cáncer. En un escenario, se prescribe un tratamiento apropiado (por ejemplo, cirugía o terapéutica de resección) cuando la probabilidad o probabilidad excede un umbral. Por ejemplo, en un escenario, si la probabilidad o puntuación de probabilidad es mayor o igual a 60, se prescriben uno o más tratamientos apropiados. En otros escenarios, si la puntuación de probabilidad es mayor o igual a 65, mayor o igual a 70, mayor o igual a 75, mayor o igual a 80, mayor o igual a 85, mayor o igual a 90, o mayor o igual a 95, se prescriben uno o más tratamientos apropiados. En otros escenarios, una relación de probabilidad logarítmica de cáncer puede indicar la efectividad de un tratamiento contra el cáncer. Por ejemplo, un aumento en la relación de probabilidad logarítmica del cáncer con el tiempo (por ejemplo, en un segundo, después del tratamiento) puede indicar que el tratamiento no fue eficaz. De manera similar, una disminución en la relación de probabilidad logarítmica del cáncer con el tiempo (por ejemplo, en un segundo, después del tratamiento) puede indicar un tratamiento exitoso. En otro escenario, si la relación de probabilidad logarítmica del cáncer es mayor que 1, mayor que 1,5, mayor que 2, mayor que 2,5, mayor que 3, mayor que 3,5 o mayor que 4, se prescribe uno o más tratamientos apropiados.

En algunos escenarios, el tratamiento es uno o más agentes terapéuticos contra el cáncer seleccionados del grupo que consiste en un agente de quimioterapia, un agente de terapia dirigida contra el cáncer, un agente de terapia de diferenciación, un agente de terapia hormonal y un agente de inmunoterapia. Por ejemplo, el tratamiento puede consistir en uno o más agentes quimioterapéuticos seleccionados del grupo formado por agentes alquilantes, antimetabolitos, antraciclinas, antibióticos antitumorales, disruptores del citoesqueleto (taxanos), inhibidores de la topoisomerasa, inhibidores mitóticos, corticosteroides, inhibidores de la cinasa, análogos de nucleótidos, agentes basados en platino y cualquier combinación de los mismos. En algunas realizaciones, el tratamiento es uno o más agentes de terapia dirigida contra el cáncer seleccionados del grupo que consiste en inhibidores de la transducción de señales (por ejemplo, inhibidores de la tirosina quinasa y del receptor del factor de crecimiento), inhibidores de la histona deacetilasa (HDAC), agonistas del receptor retinoico, inhibidores del proteosoma, inhibidores de la angiogénesis y conjugados de anticuerpos monoclonales. En algunos casos, el tratamiento consiste en uno o más agentes de terapia diferenciadora que incluyen retinoides, como la tretinoína, la alitretinoína y el bexaroteno. En algunos escenarios, el tratamiento es uno o más agentes de terapia hormonal seleccionados del grupo que consiste en estrógenos, inhibidores de aromatasa, progestinas, estrógenos, antiandrógenos y agonistas o análogos de GnRH. En un escenario, el tratamiento es uno o más agentes inmunoterapéuticos seleccionados del grupo que comprende terapias de anticuerpos monoclonales, como rituximab (RITUXAN) y alemtuzumab (CAMPATH), inmunoterapias no específicas y adyuvantes, como BCG, interleucina-2 (IL-2) e interferón alfa, fármacos inmunomoduladores, por ejemplo, talidomida y lenalidomida (REVLIMID). Un médico u oncólogo experto puede seleccionar un agente terapéutico contra el cáncer adecuado en función de características como el tipo de tumor, el estadio del cáncer, la exposición previa a un tratamiento o agente terapéutico contra el cáncer y otras características del cáncer.

Ejemplos

Los siguientes ejemplos se presentan para proporcionar a los expertos en la técnica una descripción completa y descripción de cómo hacer y usar la presente descripción, y no pretenden limitar el alcance de lo que los inventores consideran su descripción ni pretenden representar que los experimentos siguientes son todos o los únicos experimentos realizados. Se han realizado esfuerzos para garantizar la precisión con respecto a los números utilizados (por ejemplo, cantidades, temperatura, etc.) pero se deben tener en cuenta algunos errores experimentales y desviaciones.

EJEMPLO 1 - Análisis de las cualidades de la sonda

Para comprobar cuánto solapamiento es necesario entre un fragmento de ANDlc y una sonda para lograr una cantidad no despreciable de extracción, se probaron varias longitudes de solapamientos utilizando paneles diseñados para incluir tres tipos diferentes de sondas (V1D3, V1D4, V1E2) con varios solapamientos con fragmentos de ADN diana de 175 pb específicos para cada sonda. Los solapamientos probados oscilaban entre 0 pb y 120 pb. Las muestras que contenían fragmentos de ADN diana de 175 pb se aplicaron al panel y se lavaron, y a continuación se recogieron los fragmentos de ADN unidos a las sondas. Se midieron las cantidades de los fragmentos de ADN recogidos y las cantidades se trazaron como densidades sobre los tamaños de los solapamientos, tal como se indica en la **FIG. 10.**

No hubo una unión significativa y la extracción de fragmentos de ADN diana cuando había menos de 45 pb de superposición. Estos resultados sugieren que generalmente se requiere un solapamiento de sonda-sonda de al menos 45 pb para lograr una cantidad no despreciable de extracción, aunque este número puede variar dependiendo de las condiciones del ensayo.

Además, se ha sugerido que más de una tasa de emparejamiento erróneo del 10 % entre la sonda y las secuencias de fragmentos en la región de solapamiento es suficiente para interrumpir en gran medida la unión y, por tanto, la eficiencia de la extracción. Por tanto, las secuencias que pueden alinearse con la sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 % son candidatas para la extracción fuera de la diana.

Por tanto, hemos realizado una búsqueda exhaustiva de todas las regiones genómicas que tienen alineaciones de 45 pb con una tasa de coincidencia de 90 % + (es decir, regiones fuera de diana) para cada sonda. Específicamente, se combinó una estrategia de siembra de k-mero (que puede permitir una o más faltas de coincidencia) con alineación local en las ubicaciones de las semillas. Esto garantizó que no falte ningún buen alineamiento basándose en la longitud de k-mero, se permitió el número de faltas de coincidencia y el número de aciertos de semilla de k-mero en una ubicación particular. Esto implica realizar una alineación local de programación dinámica en un gran número de ubicaciones, por lo que la implementación se optimizó para usar instrucciones de CPU de vectores (por ejemplo, AVX2, AVX512) y paralelizó a través de muchos núcleos dentro de una máquina y también a través de muchas máquinas conectadas por una red. Esto permite una búsqueda exhaustiva que es valiosa para diseñar un panel de alto rendimiento (es decir, velocidad baja fuera de la diana y alta cobertura objetivo para una cantidad dada de secuenciación).

Después de la búsqueda exhaustiva, cada sonda se calificó en base al número de regiones fuera de la diana. La mayoría de las sondas tienen una puntuación de 1, lo que significa que coinciden en un solo lugar. Las sondas con puntuaciones entre 2-19 se aceptaron pero las sondas con puntuaciones de más de 20 se descartaron. Pueden usarse otros valores de corte para muestras específicas. Las sondas dirigidas a regiones hipermetiladas tienden a tener regiones significativamente menos fuera de la diana que las sondas dirigidas a otras regiones.

EJEMPLO 2 - Anotación de regiones genómicas diana

Regiones genómicas diana identificadas por el proceso descrito en la **Figura 4** se analizaron para comprender características de las regiones diana. Específicamente, las regiones genómicas diana seleccionadas se alinearon con un genoma de referencia para determinar las posiciones de alineación. La información de la posición de alineación se recogió para cada región genómica diana seleccionada, incluyendo el número de cromosoma, la base de nucleótidos inicial, la base de nucleótidos final y las anotaciones genómicas de la región genómica dada. Las regiones genómicas diana se situaron en intrones, exones, regiones intergénicas, 5'UTR, 3'UTR o regiones de control como promotores o potenciadores. El número de regiones genómicas diana que se encuentran dentro de cada anotación genómica se contó y se representó gráficamente en el gráfico proporcionado en la **Figura 11.** La **Figura 11** también compara números de las regiones genómicas diana seleccionadas (barras negras) o números de regiones genómicas seleccionadas aleatoriamente (barras grises) que se sitúan dentro de cada anotación genómica.

El análisis muestra que las regiones genómicas diana seleccionadas no son aleatorias en su distribución genómica y presentaban un mayor enriquecimiento en elementos reguladores y funcionales como promotores y 5'UTR y una menor representación de secuencias intergénicas en comparación con dianas del mismo tamaño seleccionadas al azar. Por ejemplo, se observó que las regiones genómicas diana se situaban en promotores, 5'UTR, exones, límites intrón/exón, intrones, 3'UTR o potenciadores, en lugar de en regiones intergénicas.

EJEMPLO 3 - Paneles de ensayo para detectar cáncer y tipos de cáncer

Muestras usadas para la selección de región genómica: Las muestras de ADN para este trabajo provienen de varias fuentes.

El genoma circulante libre de células Atlas ("CCGA"; Clinical Trial.gov identificador NCT02889978) es un estudio prospectivo, multicéntrico, de casos y controles, observacional y con seguimiento longitudinal. Los bioespecímenes identificados se recogieron de aproximadamente 15.000 participantes a partir de 142 sitios. Se seleccionaron muestras para garantizar una distribución preespecificada de tipos de cáncer y no cánceres a través de sitios en cada cohorte, y las muestras cancerosas y no cancerosas tenían una frecuencia de edad equivalente por el género.

The Cancer Genome Atlas ("TCGA"; Clinical Trial.gov identificador NCT02889978) es un recurso público desarrollado a través de una colaboración entre el Instituto Nacional del Cáncer (NCI) y el Instituto Nacional de Investigación del Genoma Humano (NHGRI).

Se adquirieron células tumorales disociadas (DTC) de Conversant.

Las células no cancerosas fueron proporcionadas por Yuval Dor y Ben Glaser (Universidad Hebrea) y procedían de tejido humano obtenido mediante procedimientos clínicos convencionales. Por ejemplo, las células epiteliales luminales y basales de mama eran de cirugía de reducción de mama; las células epiteliales del colon procedían de tejido cercano al lugar de reimplantación tras una resección segmentaria por patología localizada del colon; las células de médula ósea procedían de una operación de prótesis articular; las células endoteliales vasculares y arteriales procedían de cirugía vascular; y el epitelio de cabeza y cuello procedía de una amigdalectomía.

La WGBS se llevó a cabo en más de 1.000 muestras de ADN genómico recogidas de individuos sanos y de individuos diagnosticados de cáncer en diferentes estadios y tejidos de origen. Las muestras incluían bloques de tejido fijados con formaldehído e incluidos en parafina (FFPE), células tumorales diseminadas (DTC) de cánceres de diferentes TOO, células mononucleares de médula ósea (BMMC), glóbulos blancos (WBC) y células mononucleares de sangre periférica (PBMC). Las DTC se sometieron a selección negativa para eliminar leucocitos, fibroblastos y células endoteliales usando un kit de selección negativa antes del aislamiento de ADNg. La selección negativa produjo células tumorales purificadas que permitieron que las regiones metiladas diferencialmente se identificaron más claramente.

Los datos de TCGA se recogieron mediante hibridación de fragmentos de ADN convertidos con bisulfito de 8809 muestras a matrices de oligonucleótidos sensibles a la metilación. Los valores β de este estudio representan la abundancia relativa de metilación a 480.000 sitios CpG individuales. Se analizaron 75.000 de estos sitios CpG después de excluir los CpG de las regiones genómicas ruidosas (360.000) y los sitios CpG con sondas de hibridación cruzada (45.000). Los datos TCGA se analizaron mediante el uso de diferentes algoritmos porque describe la metilación de sitios CpG individuales, mientras que los datos de WGBS revelan el patrón de metilación de cadenas de sitios CpG adyacentes en fragmentos de ADN.

Tejido de clases de origen: Cada muestra se categorizó en una de veinticinco (25) clases diferentes de tejido de origen (TOO) (es decir, tipos de cáncer): cáncer de mama, cáncer de útero, cáncer de cuello de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial de pelvis renal, cáncer renal distinto del urotelial, cáncer de próstata, cáncer anorrectal, cáncer colorrectal, cáncer hepatobiliar derivado de hepatocitos, cáncer hepatobiliar derivado de células distintas de los hepatocitos, cáncer de páncreas, cáncer de células escamosas del tracto gastrointestinal superior, cáncer gastrointestinal superior distinto del escamoso, cáncer de cabeza y cuello, adenocarcinoma de pulmón, cáncer de pulmón de células pequeñas, cáncer de pulmón de células escamosas y cáncer distinto del adenocarcinoma o del cáncer de pulmón de células pequeñas, cáncer neuroendocrino, melanoma, cáncer de tiroides, sarcoma, mieloma múltiple, linfoma y leucemia. Estas clases de TOO abarcan el 97 % de la incidencia de cáncer notificada por el programa de Vigilancia, Epidemiología y Resultados Finales (SEER; seer.cancer.gov), tras filtrar el líquido, el cerebro, el intestino delgado, la vagina y la vulva y el pene y los testículos. Los cánceres de incidencia poco frecuente, como el sarcoma, y los neuroendocrinos se agregaron para evitar clasificaciones erróneas. Se utilizaron los códigos topográficos, morfológicos y de comportamiento de la Clasificación Internacional de Enfermedades Oncológicas (CIE-O-3) y las designaciones topográficas de la Organización Mundial de la Salud (OMS) para clasificar las muestras individuales en las clases TOO. Por ejemplo, los 34 estudios de TCGA se mapearon a 25 clases TOO como se muestra en la **tabla 1** . La clasificación de TOO se refinó iterativamente contra el rendimiento de clasificación observado.

TABLA 1 - Clasificación del tejido de origen (TOO) de los tipos de TCGA

Clase de TOO	**Tipo de TCGA**	**N**
Mama	BRCA	779
Renal	KIRC, KIRP, KICH	657
Cerebro	LGG, GBM	654
Tracto gastrointestinal superior	ESCA, STDA	580
Melanoma	SKCM, UVM	550
Cabeza y cuello	HNSC	528
Tiroides	THCA	507
Próstata	PRAD	498
De útero	UCEC, UCS	484
Adenocarcinoma de pulmón	LUAD	444

Clase de TOO	Tipo de TCGA	N
Vejiga	BLCA	409
Colorrectal	COAD, READ	382
Carcinoma hepatocelular	LIHC	377
Pulmón escamoso	LUSC	370
De cuello uterino	CESC	307
Sarcoma	SARC	261
Suprarrenal	ACC, PCPG	259
Páncreas	PAAD	184
Leucemia	LAML, LCML	140
Testicular	TGCT	134
Timo	THYM	124
Mesotelioma	MESO	87
Linfoma	DLBC	48
Hepatobiliar biliar	CHOL	36
De ovario	OV	10

Selección de región Para la selección de la diana, se seleccionaron fragmentos que tienen patrones de metilación anormales en muestras de cáncer mediante el uso de uno o más métodos como se describe en la presente memoria. El uso de estos métodos permitió la identificación de regiones de bajo ruido como dianas putativas. Entre las regiones de bajo ruido, se clasificaron y seleccionaron fragmentos más informativos en los tipos de cáncer discriminatorios.

Específicamente, en algunos escenarios, cuando se utilizaron los datos WGBS, las secuencias de fragmentos en la base de datos se filtraron en función del valor de p utilizando una distribución no cancerígena, y sólo se retuvieron los fragmentos con $p < 0,001$, como se describe en la presente memoria. En algunos casos, los ADNlc seleccionados se filtraron adicionalmente para retener solo aquellos que estaban al menos 90 % metilados o 90 % no metilados. A continuación, para cada sitio CpG en los fragmentos seleccionados, se contaron los números de muestras de cáncer o muestras no cancerosas que incluyen fragmentos que se superponen en el sitio CpG. En concreto, se calculó P (cáncer | fragmento solapado) para cada CpG y se seleccionaron los sitios genómicos con altos valores de P como dianas generales del cáncer. Por diseño, los fragmentos seleccionados tenían un ruido muy bajo (es decir, se superponen pocos fragmentos sin cáncer).

Para encontrar dianas específicas del tipo de cáncer, se realizaron procesos de selección similares. Los sitios CpG se clasificaron en base a su ganancia de información, comparando (i) el número de muestras de una TOO específica u otras muestras, incluidas las muestras no cancerosas y las muestras de un TOO diferente, (ii) los números de muestras de una muestra TOO o no cancerosa específica, y/o (iii) los números de muestras de un TOO específico o un TOO diferente que incluyen fragmentos que se superponen en el sitio CpG. El proceso se aplicó a cada uno de los 25 TOO y la comparación se realizó para todas las combinaciones por pares para 25 TOO. Por ejemplo, se calculó P (cáncer de un fragmento superpuesto de TOO) y luego se comparó con P (cáncer de un fragmento superpuesto de TOO diferente). Un fragmento atípicos en cada TOO que tiene una probabilidad mucho mayor en el cáncer de un TOO que bajo el cáncer de un TOO diferente se seleccionó como un objetivo para el TOO. Por consiguiente, las regiones genómicas seleccionadas por las comparaciones por pares incluyeron regiones genómicas diferencialmente metiladas para separar un TOO objetivo y un TOO de contraste.

Se seleccionaron regiones genómicas diana adicionales según los métodos descritos en la sección anterior titulada "Calcular la ganancia de información por pares a partir de fragmentos indicativos de cáncer identificados a partir de modelos probabilísticos." Los números de regiones genómicas para diferenciar cada TOO objetivo (eje x) de un TOO de contraste (eje y) se proporcionan en la Figura 13.

Cuando se usaron datos TCGA, se usaron los valores beta CpG que indican la intensidad de metilación para identificar las regiones genómicas diana. Esto se debe a que los datos de matriz no están en los niveles del sitio CpG y, por tanto, son propensos a resultar en falsos positivos. Para evitar falsos positivos, los sitios CpG se convirtieron en intervalos de 350 pb a través del genoma. Los valores beta de cada intervalo se calcularon como la media de los valores beta de CpG en ese intervalo. Los intervalos con menos de 2 CpG se excluyeron del análisis. A continuación, se seleccionaron los intervalos con una diferencia beta > 0,95 entre (i) muestras de un TOO específico y otras muestras, incluyendo muestras no cancerosas y muestras de un TOO diferente, (ii) muestras de un TOO específico y

muestras no cancerosas, y/o (iii) muestras de un TOO específico y un TOO diferente que incluyan fragmentos que se solapen con ese sitio CpG.

Las regiones genómicas seleccionadas como se describió anteriormente se filtraron en base a los números de sus regiones genómicas fuera de diana como se especifica en 4,4,7. Específicamente, el número de localizaciones genómicas con alineaciones >=45 pb con >=90 % de identidad se calculó como el número de regiones genómicas fuera del objetivo. Se descartaron las regiones genómicas que tienen regiones genómicas fuera de diana más de 20.

Diversas listas de regiones genómicas diana seleccionadas como se describe en esta sección se identifican en la **tabla 2**. Estas listas tienen conjuntos diferentes pero superpuestos de regiones genómicas diana. Se diferencian en sus números totales de regiones genómicas diana, el total de las longitudes de sus regiones genómicas diana y las ubicaciones cromosómicas de sus regiones genómicas diana. Las listas 1-3 son paneles pequeños, medianos y grandes. Las regiones genómicas diana de las listas 4-16 tienen subconjuntos de los sitios de metilación CpG encontrados en las regiones genómicas diana de la lista 3. Las listas 4, 6 y 8-16 se filtraron para excluir las regiones genómicas diana previamente conocidas.

TABLA 2 - Id. de sec. n.° correspondientes a las listas 1-16. Para cada lista, la tabla identifica el número total de regiones genómicas diana de la lista, un intervalo de la Id. de sec. n.° correspondientes a todas las regiones genómicas diana de la lista que se encuentran en el listado de secuencias presentado con esta solicitud, y el total de las longitudes de todas las regiones genómicas diana de la lista. El listado de secuencias identifica la ubicación cromosómica de cada región genómica diana, si los fragmentos de ADNlc se enriquecen de la región están hipermetilados o hipometilados, y la secuencia de una cadena de ADN de la región genómica diana. Los números de cromosomas y las posiciones de inicio y parada se proporcionan en relación con un genoma de referencia humana conocido, hg19. La secuencia del genoma humano de referencia, hg19, está disponible en Genome Reference Consortium con un número de referencia, GRCh37/hg19, y también está disponible en Genome Browser proporcionado por Santa Cruz Genomics Institute.

		Id. de sec. n.°		
Lista	Regiones genómicas diana	Primera	Última	Tamaño del panel (Mb)
1	34844	1	34844	6,43
2	67431	34845	102275	12,14
3	94955	102276	197230	17,72
4	23941	197231	221171	4,63
5	56624	221172	277795	16,42
6	52850	277796	330645	10,45
7	14284	330646	344929	8,48
8	1370	344930	346299	0,39
9	2842	346300	349141	0,79
10	7483	349142	356624	1,94
11	12328	356625	368952	3,08
12	14725	368953	383677	3,65
13	3814	383678	387491	0,62
14	7730	387492	395221	1,26
15	19424	395222	414645	3,23
16	38061	414646	452706	6,58

Las Id. de sec. n.° 452,706-483,478 proporcionan información adicional sobre ciertas regiones genómicas diana hipermetiladas o hipometiladas. Estos registros de la Id. de sec. n.° identifican regiones genómicas diana que pueden metilarse diferencialmente en muestras de pares específicos de tipos de cáncer. Las regiones genómicas diana de las Id. de sec. n.° 452,706-483,478 se dibujan de la lista 6. Muchas de las mismas regiones genómicas diana también se encuentran en las listas 1-5 y 7-16. La entrada para cada Id. de sec. indica la ubicación cromosómica de la región genómica diana en relación con hg19, si los fragmentos de ADNlc a enriquecer de la región están hipermetilados o hipometilados, la secuencia de una cadena de ADN de la región genómica diana, y el par o pares de tipos de cáncer que están diferencialmente metilados en esa región genómica. Como el estado de metilación de algunas regiones genómicas diana distingue más de un par de tipos de cáncer, cada entrada identifica un primer tipo de cáncer como se indica en la **tabla 3** y uno o más segundos tipos de cáncer.

Tabla 3 - Id. de sec. n.° que identifican regiones genómicas diana que están metiladas diferencialmente entre pares de tipos de cáncer

Primer tipo de cáncer	Regiones genómicas diana	Id. de sec. n.°	
		Primera	Última
Anorrectal	1377	452707	454083
Vejiga y urotelio	1411	454084	455494
Mama	1748	455495	457242
De cuello uterino	2011	457243	459253
Colorrectal	1321	459254	460574
Cabeza y cuello	1624	460575	462198
Hígado y vías biliares	1810	462199	464008
Pulmón	1863	464009	465871
Neoplasia linfoide	2660	465872	468531
Melanoma	1378	468532	469909
Mieloma múltiple	986	469910	470895
Neoplasia mieloide	1595	470896	472490
Ovario	1041	472491	473531
Páncreas y vesícula biliar	1682	473532	475213
Próstata	1395	475214	476608
Renal	1236	476609	477844
Sarcoma	1418	477845	479262
Tiroides	895	479263	480157
Tracto gastrointestinal superior	1606	480158	481763
De útero	1715	481764	483478

Verificación de regiones genómicas seleccionadas:

Algunas de las regiones genómicas seleccionadas se verificaron (1) sin referencia (usando ADNlc a partir de la base de datos CCA1 30 x WGBS limitada a ADNlc de muestras con una relación log-verosimilitud indicativa de cáncer mayor de 0,9) o (2) con referencia (usando muestras de tejido y WBC). La **Figura 14** proporciona los resultados de verificación basados en fracciones correctamente clasificadas. Los resultados son de (1) verificación realizada con ADNlc sobre regiones genómicas entrenadas en ADNlc; (2) la verificación realizada con ADNlc sobre regiones genómicas entrenadas en todos los tipos diferentes de muestras usadas en la presente memoria; y (3) verificación realizada con muestra de ADNg de tejido y WBC sobre las regiones genómicas seleccionadas. Los datos de verificación se resumen en la **tabla 4,** incluyendo adicionalmente datos de verificación realizados con todas las muestras. Los resultados de la verificación demuestran que las regiones genómicas seleccionadas por el método descrito en la presente memoria pueden proporcionar información para la detección del cáncer y diversos tipos de cáncer.

Tabla 4 - Datos de verificación

Tipo de cáncer	**ADNlc entrenado en ADNlc**	**ADNlc entrenado en tejido**	**Tejido + otro no ADNlc**	**Todas las muestras**
Leucemia	1/5 [20 %]	4/5 [80 %]	49/66 [74 %]	53/71 [75 %]
Linfoma	22/23 [96 %]	22/23 [96 %]	39/47 [83 %]	61/70 [87 %]
Mieloma múltiple	13/14 [93 %]	14/14 [100 %]	17/23 [74 %]	31/37 [84 %]
Sarcoma		0/1 [0 %]	5/6 [83 %]	5/7 [71 %]
Tiroides			11/11 [100 %]	11/11 [100 %]
Melanoma		2/2 [100 %]	13/13 [100 %]	15/15 [100 %]

Tipo de cáncer	ADNlc entrenado en ADNlc	ADNlc entrenado en tejido	Tejido + otro no ADNlc	Todas las muestras
Neuroendocrina	1/7 [14 %]	1/7 [14 %]	0/2 [0 %]	1/9 [11 %]
Pulmón	86/95 [91 %]	84/95 [88 %]	51/55 [93 %]	135/150 [90 %]
Cabeza y cuello	10/16 [62 %]	13/16 [81 %]	36/43 [84 %]	49/59 [83 %]
Tracto gastrointestinal superior	17/25 [68 %]	15/25 [60 %]	43/49 [88 %]	58/74 [78 %]
Páncreas	23/30 [77 %]	26/30 [87 %]	15/15 [100 %]	41/45 [91 %]
Colangio y biliar	4/9 [44 %]	5/9 [56 %]	2/5 [40 %]	7/14 [50 %]
Hepatocelular	9/11 [82 %]	11/11 [100 %]	5/5 [100 %]	16/16 [100 %]
Colorrectal	50/58 [86 %]	49/58 [84 %]	70/72 [97 %]	119/130 [92 %]
Anorrectal	6/7 [86 %]	6/7 [86 %]	0/1 [0 %]	6/8 [75 %]
Próstata	3/3 [100 %]	3/3 [100 %]	58/58 [100 %]	61/61 [100 %]
Renal	2/5 [40 %]	4/7 [57 %]	50/56 [89 %]	54/63 [86 %]
Vejiga		0/2 [0 %]	28/32 [88 %]	28/34 [82 %]
De ovario	11/14 [79 %]	13/14 [93 %]	43/50 [86 %]	56/64 [88 %]
De cuello uterino	0/6 [0 %]	1/6 [17 %]	21/23 [91 %]	22/29 [76 %]
De endometrio	1/5 [20 %]	3/5 [60 %]	47/49 [96 %]	50/54 [93 %]
Mama	69/83 [83 %]	71/83 [86 %]	117/118 [99 %]	188/201 [94 %]
Total	**328/416 [79 %]**	**3457/423 [82 %]**	**720/799 [90 %]**	**1067/1222 [87 %]**

Ejemplo 4 - Generación de un clasificador de modelo de mezcla

Para maximizar el rendimiento, los modelos predictivos de cáncer descritos en este Ejemplo se entrenaron utilizando datos de secuencia obtenidos de una pluralidad de muestras de tipos de cáncer conocidos y no cancerosos de ambos subestudios CCGA (CCGA1 y CCGA22), una pluralidad de muestras de tejido para cánceres conocidos obtenidas de CCGA1, y una pluralidad de muestras no cancerosas del estudio STRIVE (véase Clinical Trail.gov Identifier: NCT03085888 (//clinicaltrials.gov/ct2/show/NCT03085888)). El estudio STRIVE es un estudio de cohortes prospectivo, multicéntrico y observacional para validar un ensayo para la detección precoz del cáncer de mama y otros cánceres invasivos, del que se obtuvieron muestras de entrenamiento no cancerosas adicionales para entrenar el clasificador descrito en la presente memoria. Los tipos de cáncer conocidos incluidos en el conjunto de muestras CCGA fueron los siguientes: mama, pulmón, próstata, colorrectal, renal, uterino, páncreas, esofágico, linfoma, cabeza y cuello, ovario, hepatobiliar, melanoma, cervical, mieloma múltiple, leucemia, tiroides, vejiga, gástrico y anorrectal. Como tal, un modelo puede ser un modelo multicáncer (o un clasificador multicáncer) para detectar uno o más, dos o más, tres o más, cuatro o más, cinco o más, diez o más, o 20 o más tipos diferentes de cáncer.

Los datos de rendimiento del clasificador que se muestran a continuación se obtuvieron para un clasificador bloqueado entrenado con muestras de cáncer y no cáncer obtenidas de CCGA2, un subestudio de CCGA, y con muestras no cancerosas de STRIVE. Los individuos del subestudio CCGA2 eran diferentes de los individuos del subestudio CCGA1 cuyo ADNlc se usó para seleccionar los genomas diana. En el estudio CCGA2 se recogieron muestras de sangre de individuos diagnosticados de cáncer sin tratar (incluidos 20 tipos de tumores y todos los estadios del cáncer) y de individuos sanos sin diagnóstico de cáncer (controles). Para STRIVE, se recogieron muestras de sangre de mujeres en los 28 días siguientes a su mamografía de cribado. El ADN libre de células (ADNlc) se extrajo de cada muestra y se trató con bisulfito para convertir citosinas no metiladas en uracilos. El ADNlc tratado con bisulfito se enriqueció para dar moléculas de ADNlc informativas usando sondas de hibridación diseñadas para enriquecer ácidos nucleicos convertidos por bisulfito derivados de cada una de una pluralidad de regiones genómicas dirigidas en un panel de ensayo que comprende todas las regiones genómicas de listas 1-16. Las moléculas de ácido nucleico enriquecidas con bisulfito enriquecidas se secuenciaron mediante el uso de secuenciación de extremos emparejados en una plataforma Illumina (San Diego, CA) para obtener un conjunto de lecturas de secuencia para cada una de las muestras de entrenamiento, y los pares de lecturas resultantes se alinearon con el genoma de referencia, se ensamblaron en fragmentos y se identificaron sitios CpG metilados y no metilados.

Caracterización basándose en el modelo de mezcla

Para cada tipo de cáncer (incluido el no canceroso) se entrenó y utilizó un modelo de mezcla probabilística para asignar una probabilidad a cada fragmento de cada muestra cancerosa y no cancerosa en función de la probabilidad de que el fragmento se observara en un tipo de muestra determinado.

Análisis a nivel de fragmento

Brevemente, para cada tipo de muestra (muestras de cáncer y de no cáncer), para cada región (donde cada región se utilizó tal cual si era menor de 1 kb, o bien se subdividió en regiones de 1 kb de longitud con un solapamiento del 50 % (por ejemplo, solapamiento de 500 pares de bases) entre regiones adyacentes), se ajustó un modelo probabilístico a los fragmentos derivados de las muestras de entrenamiento para cada tipo de cáncer y de no cáncer. El modelo probabilístico entrenado para cada tipo de muestra fue un modelo de mezcla, en el que cada uno de los tres componentes de la mezcla era un modelo de sitios independientes en el que se supone que la metilación en cada CpG es independiente de la metilación en otros CpG. Los fragmentos se excluyeron del modelo si: tenían un valor de p (de un modelo de Markov no canceroso) superior a 0,01; se marcaron como fragmentos duplicados; los fragmentos tenían un tamaño de bolsa superior a 1 (sólo para las muestras de metilación dirigida); no cubrían al menos un sitio CpG; o si el fragmento tenía una longitud superior a 1000 bases. Los fragmentos de entrenamiento retenidos se asignaron a una región si se solaparon al menos un CpG de esa región. Si un fragmento solapaba CpG en varias regiones, se asignaba a todas ellas.

Modelos de fuentes locales

Cada modelo probabilístico se ajustó utilizando la estimación de máxima verosimilitud para identificar un conjunto de parámetros que maximizaran la probabilidad logarítmica de todos los fragmentos derivados de cada tipo de muestra, sujeto a una penalización de regularización.

Específicamente, en cada región de clasificación se entrenó un conjunto de modelos probabilísticos, uno para cada etiqueta de entrenamiento (es decir, uno para cada tipo de cáncer y otro para los no cancerosos). Cada modelo tomó la forma de un modelo de mezcla Bernoulli con tres componentes. Matemáticamente,

$$(1)\ Pr\left(\text{fragmento}_j \mid \{\beta_{ki}, f_k\}\right) = \sum_{k=1}^{n} f_k \prod_i \beta_{ki}^{m_i}(1-\beta_{ki})^{1-m_i}$$

cuando n es el número de componentes de mezcla, establecidos en 3; $m_i \in \{0, 1\}$ es la metilación observada del fragmento en la posición i; f_k es la asignación fraccionaria al componente k (con $f_k \geq 0$ y $\Sigma f_k = 1$); y β_{ki} es la fracción de metilación en el componente k en CpG i. El producto sobre i incluyó solo aquellas posiciones para las que podría identificarse un estado de metilación a partir de la secuenciación. Se estimaron valores de probabilidad máxima de los parámetros $\{f_k, \beta_{ki}\}$ de cada modelo usando el algoritmo rprop (por ejemplo, el algoritmo rprop tal como se describe en Riedmiller M, Braun H. RPROP-A Fast Adaptive Learning Algorithm. Proceedings of the International Symposium on Computer and Information Science VII, 1992) para maximizar la probabilidad logarítmica total de los fragmentos de un marcador de entrenamiento, sujeto a una penalización de regularización en β_{ki} que tomó la forma de un antes distribuido beta. Matemáticamente, la cantidad maximizada fue

$$(2)\ \sum_j \ln\left(Pr\left(\text{fragmento}_j \mid \{\beta_{ki}, f_k\}\right)\right) + \sum_{k,i} r \ln\left(\beta_{ki}(1-\beta_{ki})\right)$$

donde r es la resistencia de regularización, que se estableció en 1.

Caracterización

Una vez entrenados los modelos probabilísticos, se calculó un conjunto de características numéricas para cada muestra. Específicamente, se extrajeron características para cada fragmento de cada muestra de entrenamiento, para cada muestra de cáncer y muestra no cancerosa, en cada región. Las características extraídas fueron los recuentos de fragmentos atípicos (es decir, fragmentos anómalamente metilados), que se definieron como aquellos cuya probabilidad logarítmica bajo un primer modelo de cáncer superaba la probabilidad logarítmica bajo un segundo modelo de cáncer o modelo de no cáncer en al menos un valor de nivel umbral. Los fragmentos atípicos se contaron por separado para cada región genómica, modelo de muestra (es decir, tipo de cáncer) y nivel (para los niveles 1, 2, 3, 4, 5, 6, 7, 8 y 9), obteniéndose 9 características por región para cada tipo de muestra. De esta manera, cada característica se definió por tres propiedades: una región genómica; un marcador de tipo cáncer "positivo" (que excluye el no cáncer); y el valor del nivel seleccionado del conjunto {1, 2, 3, 4, 5, 6, 7, 8, 9}. El valor numérico de cada característica se definió como el número de fragmentos en esa región de manera que

$$(3)\ En\left(\frac{Pr(fragmento|tipo\ de\ cáncer\ positivo)}{Pr(fragmento|no\ cáncer)}\right) > nivel$$

donde las probabilidades se definieron mediante la ecuación (1) usando los valores de parámetros estimados de probabilidad máxima correspondientes al tipo de cáncer "positivo" (en el numerador del logaritmo) o al no cáncer (en el denominador).

Clasificación de características

Para cada conjunto de características por pares, las características se clasificaron mediante el uso de información mutua basándose en su capacidad para distinguir el primer tipo de cáncer (que definió el modelo de probabilidad logarítmica a partir del cual se obtuvo la característica) del segundo tipo de cáncer o no cáncer. Específicamente, se compilaron dos listas clasificadas de características para cada par único de marcadores de clase: uno con el primer marcador asignado como "positivo" y el segundo como el "negativo", y el otro con la asignación positiva/negativa intercambiada (con la excepción del marcador "no canceroso", que solo se permitió como marcador negativo). Para cada una de estas listas clasificadas, solo las características cuyo marcador de tipo cáncer positivo (como en la ecuación (3)) coincidió con el marcador positivo en consideración se incluyeron en la clasificación. Para cada una de dichas características, la fracción de muestras de entrenamiento con valor de característica distinto de cero se calculó por separado para los marcadores positivos y negativos. Las características para las cuales esta fracción fue mayor en el marcador positivo se clasificaron por su información mutua con respecto a esa pareja de marcadores de clase.

Las 256 características superiores clasificadas de cada comparación por pares se identificaron y se añadieron al conjunto final de características para cada tipo de cáncer y no cáncer. Para evitar la redundancia, si se seleccionó más de una característica del mismo tipo positivo y región genómica (es decir, para múltiples tipos negativos), solo se retuvo el rango más bajo (más informativo) para su par de tipos de cáncer, rompiendo el valor del nivel más alto. Las características del conjunto final de características de cada muestra (tipo de cáncer y no cáncer) se binarizaron (cualquier valor de característica superior a 0 se fijó en 1, de modo que todas las características fueran 0 o 1).

Entrenamiento del clasificador

A continuación, las muestras de entrenamiento se dividieron en distintos conjuntos de entrenamiento de validación cruzada de 5 pliegues, y se entrenó un clasificador de dos etapas para cada pliegue, en cada caso entrenando en 4/5 de las muestras de entrenamiento y usando las 1/5 restantes para la validación.

En la primera etapa del entrenamiento, se entrenó un modelo de regresión logística binaria (de dos clases) para detectar la presencia de cáncer para discriminar las muestras de cáncer (independientemente del TOO) de no cáncer. Al entrenar este clasificador binario, se asignó un peso de muestra a las muestras macho sin cáncer para contrarrestar el desequilibrio del sexo en el conjunto de entrenamiento. Para cada muestra, el clasificador binario emite una puntuación de predicción que indica la probabilidad de una presencia o ausencia de cáncer.

En la segunda etapa del entrenamiento, un modelo paralelo de regresión logística de múltiples clases para determinar el tejido canceroso de origen fue entrenado con TOO como marcador de destino. Solo se incluyeron las muestras de cáncer que recibieron una puntuación por encima del percentil 95 de las muestras no cancerosas en el clasificador de primera etapa en el entrenamiento de este clasificador de múltiples clases. Para cada muestra de cáncer usada en el entrenamiento del clasificador de múltiples clases, el clasificador de múltiples clases emite valores de predicción para los tipos de cáncer que se clasifican, donde cada valor de predicción es una probabilidad de que la muestra dada tenga un cierto tipo de cáncer. Por ejemplo, el clasificador de cáncer puede devolver una predicción de cáncer para una muestra de prueba que incluye una puntuación de predicción para el cáncer de mama, una puntuación de predicción para el cáncer de pulmón y/o una puntuación de predicción para ningún cáncer.

Tanto los clasificadores binarios como los multiclase se entrenaron mediante descenso de gradiente estocástico con minilotes y, en cada caso, el entrenamiento se detuvo antes de tiempo cuando el rendimiento en el pliegue de validación (evaluado mediante la pérdida de entropía cruzada) empezó a degradarse. Para predecir en muestras fuera del conjunto de entrenamiento, en cada etapa se promediaron las puntuaciones asignadas por los cinco clasificadores de validación cruzada. Las puntuaciones asignadas a los tipos de cáncer inapropiados para el sexo se fijaron en cero, y los valores restantes se renormalizaron para sumar uno.

Las puntuaciones asignadas a los pliegues de validación dentro del conjunto de entrenamiento se retuvieron para su uso en la asignación de valores de corte (umbrales) para apuntar a determinadas métricas de rendimiento. En particular, las puntuaciones de probabilidad asignadas a las muestras de conjuntos de entrenamiento no cancerosas se usaron para definir umbrales correspondientes a niveles de especificidad particulares. Por ejemplo, para una diana de especificidad deseada del 99,4 %, el umbral se estableció en el percentil 99,4º de las puntuaciones de probabilidad de detección del cáncer de validación cruzada asignadas a las muestras no cancerosas en el conjunto de

entrenamiento. Las muestras de entrenamiento con una puntuación de probabilidad que excedió un umbral se denominaron positiva para el cáncer.

Posteriormente, para cada muestra de entrenamiento determinada para ser positiva para el cáncer, se realizó una evaluación de tipo TOO o cáncer a partir del clasificador de múltiples clases. Primero, el clasificador de regresión logística de múltiples clases asignó un conjunto de puntuaciones de probabilidad, uno para cada tipo de cáncer prospectivo, a cada muestra. A continuación, la confianza de estas puntuaciones se evaluó como la diferencia entre las puntuaciones más altas y segundas más altas asignadas por el clasificador de múltiples clases para cada muestra. Luego, las puntuaciones del conjunto de entrenamiento validado en cruz se usaron para identificar el valor umbral más bajo de manera que de las muestras de cáncer en el conjunto de entrenamiento con un diferencial de puntuación superior-dos que exceda el umbral, se le ha asignado el 90 % del marcador TOO correcto como su puntuación más alta. De esta manera, las puntuaciones asignadas a los pliegues de validación durante el entrenamiento se usaron además para determinar un segundo umbral para distinguir entre las llamadas TOO de confianza e indeterminada.

En el tiempo de predicción, se asignaron muestras que reciben una puntuación del clasificador binario (primera etapa) por debajo del umbral de especificidad predefinido un marcador "no canceroso". Para las muestras restantes, aquellas cuyo diferencial de puntuación de TOO superior del clasificador de segunda etapa estaba por debajo del segundo umbral predefinido se asignaron el marcador de "cáncer indeterminado". Se asignaron las muestras restantes a la etiqueta de cáncer a la que el clasificador TOO asignó la puntuación más alta.

EJEMPLO 5 - Clasificador con las regiones genómicas diana de las listas 4-16

El valor discriminatorio de las regiones genómicas diana de las listas 4-16 se evaluó probando la capacidad de un clasificador de cáncer para detectar cáncer y cualquiera de los 20 tipos de cáncer diferentes según el estado de metilación de estas regiones genómicas diana. El rendimiento se evaluó sobre un conjunto de muestras de cáncer de 1.532 y 1.521 muestras no cancerosas que no se usaron para entrenar el clasificador, como se muestra en la **tabla 5.** Para cada muestra, el ADNlc metilado diferencialmente se enriqueció usando un conjunto de cebos que comprendía todas las regiones genómicas diana de las listas 1-16. El clasificador se restringió entonces para proporcionar determinaciones de cáncer basándose solo en el estado de metilación de las regiones genómicas diana de la lista que se está evaluando.

Tabla 5 - Diagnósticos de cáncer de los individuos cuyo ADNlc se usó para entrenar el clasificador

Tipo de cáncer	**Total**	**Estadio**				
		I	**II**	**III**	**IV**	**No informado**
Sin cáncer	1521	-	-	-	-	-
Pulmón	261	60	23	72	106	0
Mama	247	102	110	27	8	0
Próstata	188	39	113	19	17	0
Neoplasia linfoide	147	15	27	27	39	39
Colorrectal	121	13	22	41	45	0
Páncreas y vesícula biliar	95	15	15	19	46	0
De útero	84	73	3	5	3	0
Tracto gastrointestinal superior	67	9	12	19	27	0
Cabeza y cuello	62	7	13	16	26	0
Renal	56	37	4	4	11	0
Ovario	37	4	2	25	6	0
Mieloma múltiple	34	10	13	11	0	0
No notificado	29	8	5	7	6	3
Conducto biliar hepático	29	5	7	7	10	0
Sarcoma	17	2	4	5	6	0
Vejiga y urotelial	16	6	7	3	1	0
Anorrectal	14	4	5	5	0	0
De cuello uterino	11	8	1	2	0	0

Tipo de cáncer	Total	Estadio				
		I	II	III	IV	No informado
Melanoma	7	3	1	0	3	0
Neoplasia mieloide	4	2	1	0	1	0
Tiroides	4	0	0	0	0	4
Sólo predicción	2	0	0	0	2	0

Los resultados del análisis de rendimiento del clasificador para listas 4-16 se presentan en las **Figuras 15-27.** En cada figura, la parte a es una curva del operador receptor (ROC) que muestra resultados positivos verdaderos y resultados falsos positivos para una determinación del cáncer o no cáncer. La forma asimétrica de estas curvas ROC ilustra que el clasificador estaba diseñado para minimizar resultados falsos positivos. Las áreas bajo la curva están estrechamente agrupadas entre 0,78 y 0,83, como se muestra en la **tabla 6.** Estos resultados indican que una determinación del cáncer no se ve comprometida enormemente mediante el uso de paneles más pequeños de menos de 1 MB, tales como las listas 8, 9 y 13, en comparación con paneles más grandes de más de 10 MB, tales como las listas 6 y 6.

Tabla 6.

Regiones diana	AUC
Lista 4	0,81
Lista 5	0,83
Lista 6	0,81
Lista 7	0,83
Lista 8	0,80
Lista 9	0,81
Lista 10	0,81
Lista 11	0,81
Lista 12	0,81
Lista 13	0,78
Lista 14	0,79
Lista 15	0,80
Lista 16	0,80

También se evaluó el rendimiento del clasificador para subconjuntos seleccionados aleatoriamente de las regiones genómicas diana de la lista 4 y la lista 12, como se muestra en las **Figuras 28-30** y la **tabla 7.** Nuevamente, el panel más pequeño (el 10 % aleatorio de la lista 12, 0,36 MB) tuvo resultados similares al panel más grande (lista 4, 4,63 MB), lo que indica que los resultados del estado de metilación para al menos una mayoría sustancial de las regiones diana en todas las listas son informativos de la presencia o ausencia de cáncer.

Tabla 7

Regiones diana	AUC
Lista 4	0,81
El 50 % aleatorio de la lista 4	0,81
Lista 12	0,81
El 10 % aleatorio de la lista 12	0,78
El 25 % aleatorio de la lista 12	0,79

Se intentó una determinación de tipo de cáncer (es decir, TOO) para todas las muestras con una determinación del cáncer. EL panel B en las **Figuras 15-30** muestra la precisión de estas determinaciones. Por ejemplo, el valor en la esquina superior derecha de la **Figura 15B** indica que 151 muestras clasificadas como cáncer de pulmón basándose en el estado de metilación de las regiones genómicas diana de la lista 4 se habían obtenido de sujetos que se sabe

que tienen cáncer de pulmón. El valor "3" situado 3 posiciones a la izquierda en la misma matriz de confusión indica que tres muestras en las que se predijo cáncer de pulmón procedían de sujetos que en realidad tenían un cáncer del tracto gastrointestinal superior. En general, la gran mayoría de las determinaciones del tipo de cáncer realizadas usando las regiones genómicas diana de cualquiera de las listas 4-16 caen en las diagonales de las matrices de confusión, lo que indica que el clasificador determinó el tipo de cáncer correcto. Se obtuvieron resultados similares usando regiones genómicas diana seleccionadas al azar de las listas 4 y 12.

Estos resultados del clasificador se resumen además en las **TABLAS 8 - 23,** que indican la precisión de las detecciones de cáncer y las determinaciones del tipo de cáncer realizadas con una especificidad de 0,990, lo que indica una tasa de falsos positivos del 1 %. Estos resultados están delimitados por el estadio del cáncer. Muestran una mejor detección del cáncer y determinación del tipo de cáncer en muestras de individuos con cánceres en estadios avanzados (por ejemplo, estadio IV) en comparación con muestras de individuos con cánceres en estadios más tempranos (por ejemplo, estadio I). Para todos los estadios del cáncer (sin segregación por estadios), la determinación del tipo de cáncer fue precisa aproximadamente el 90 % de las veces para todas las listas de regiones genómicas diana y para subconjuntos aleatorios de la lista 4 y la lista 12. En el caso de los cánceres en estadio I, se determinó con precisión el tipo de cáncer en aproximadamente el 75 % de las ocasiones. En particular, el 75,6 % de las determinaciones del tipo de cáncer fueron precisas para el panel de ensayo más pequeño, la lista 8, con sólo 1.370 regiones genómicas diana de un tamaño total de 395 kb.

Los mismos resultados de precisión se desglosan según el tipo de cáncer en la **TABLA 24**, que demuestra determinaciones altamente precisas del tipo de cáncer con las regiones genómicas diana de todas las listas para cánceres comunes como el cáncer de hígado y de vías biliares, cánceres raros como el sarcoma y cánceres difíciles de detectar como el cáncer de mama.

La sensibilidad para detectar 20 tipos diferentes de cáncer utilizando las regiones genómicas diana de las listas 4-16 o porciones seleccionadas aleatoriamente de las listas 4 y 12 se presenta en las **TABLAS 25-40**. Los resultados de sensibilidad se presentan para una especificidad de 0,990 (una tasa de falsos positivos del 1 %). La sensibilidad se presenta para todos los cánceres del tipo especificado y para los cánceres en los estadios I a IV. En general, la sensibilidad fue mayor en los cánceres en estadios más avanzados. Para los cánceres en estadio IV, la sensibilidad fue superior al 60 % para todos los cánceres con más de una muestra y fue superior al 90 % para el cáncer de mama, el cáncer de ovario, el cáncer de vejiga y urotelectal, el cáncer de cabeza y cuello, el cáncer colorrectal, el cáncer de hígado, el cáncer de páncreas y vesícula biliar, el cáncer del tracto gastrointestinal superior, la neoplasia linfoide y el cáncer de pulmón. En el estadio II, la sensibilidad fue mejor para el cáncer de cabeza y cuello, el cáncer de hígado, el cáncer de páncreas y vesícula biliar, el cáncer del tracto gastrointestinal superior, la neoplasia linfoide y el cáncer de pulmón. La lista 8, el grupo más pequeño de regiones genómicas diana, proporciona una sensibilidad de al menos el 50 % para estos cánceres en estadio II.

Tabla 8. La precisión de clasificación usa las regiones genómicas de la lista 4. Los datos para el tipo de presencia y cáncer del cáncer en una especificidad de 0,990 muestran una precisión porcentual, un intervalo de confianza del 95 % en paréntesis y el número asignado correctamente sobre el total entre paréntesis.

Estadio	Presencia del cáncer	Tipo de cáncer
I	13 % [10-16,6] (55/422)	78,6 % [63,2-89,7] (33/42)
II	34,5 % [29,8-39,5] (134/388)	88,1 % [81,1-93,2] (111/126)
III	72,2 % [66,9-77,1] (226/313)	91 % [86,1-94,6] (182/200)
IV	85,1 % [81-88,6] (309/363)	91,9 % [88,3-94,8] (274/298)
Todos	49,2 % [46,6-51,7] (753/1532)	90,5 % [88-92,6] (627/693)

TABLA 9. La precisión de clasificación usa las regiones genómicas de la lista 5.

Estadio	Presencia del cáncer	Tipo de cáncer
I	20,9 % [17,1-25] (88/422)	77,3 % [66,2-86,2] (58/75)
II	45,9 % [40,8-51] (178/388)	88,3 % [82,4-92,8] (144/163)
III	82,7 % [78,1-86,8] (259/313)	89,9 % [85,5-93,4] (223/248)
IV	90,6 % [87,2-93,4] (329/363)	92,1 % [88,5-94,8] (291/316)
Todos	57,4 % [54,9-59,9] (880/1532)	89,5 % [87,2-91,5] (740/827)

TABLA 10. La precisión de clasificación usa las regiones genómicas de la lista 6.

Estadio	Presencia del cáncer	Tipo de cáncer
I	13,3 % [10,2-16,9] (56/422)	76 % [61,8-86,9] (38/50)
II	36,3 % [31,5-41,3] (141/388)	87,7 % [80,8-92,8] (114/130)
III	72,5 % [67,2-77,4] (227/313)	91,1 % [86,3-94,7] (185/203)
IV	85,1 % [81-88,6] (309/363)	91,6 % [87,8-94,5] (271/296)
Todos	49,6 % [47,1-52,1] (760/1532)	89,9 % [87,4-92] (633/704)

TABLA 11. La precisión de clasificación usa las regiones genómicas de la lista 7.

Estadio	Presencia del cáncer	Tipo de cáncer
I	21,3 % [17,5-25,5] (90/422)	77 % [65,8-86] (57/74)
II	45,9 % [40,8-51] (178/388)	88,5 % [82,6-92,9] (146/165)
III	82,1 % [77,4-86,2] (257/313)	89,8 % [85,4-93,3] (221/246)
IV	90,4 % [86,8-93,2] (328/363)	92,7 % [89,2-95,3] (292/315)
Todos	57,4 % [54,9-59,9] (879/1532)	89,6 % [87,3-91,6] (740/826)

TABLA 12. La precisión de clasificación usa las regiones genómicas de la lista 8.

Estadio	Presencia del cáncer	Tipo de cáncer
I	13 % [10-16,6] (55/422)	75,6 % [59,7-87,6] (31/41)
II	33 % [28,3-37,9] (128/388)	89,2 % [81,9-94,3] (99/111)
III	67,7 % [62,2-72,9] (212/313)	89,9 % [84,7-93,8] (169/188)
IV	84,6 % [80,4-88,1] (307/363)	91 % [87,1-94] (262/288)
Todos	47,5 % [45-50,1] (728/1532)	89,5 % [86,9-91,8] (582/650)

TABLA 13. La precisión de clasificación usa las regiones genómicas de la lista 9.

Estadio	Presencia del cáncer	Tipo de cáncer
I	12,1 % [9,1-15,6] (51/422)	76,3 % [59,8-88,6] (29/38)
II	35,1 % [30,3-40] (136/388)	88,1 % [81,1-93,2] (111/126)
III	68,4 % [62,9-73,5] (214/313)	92,1 % [87,2-95,5] (174/189)
IV	85,1 % [81-88,6] (309/363)	90,7 % [86,8-93,8] (264/291)
Todos	48,1 % [45,6-50,6] (737/1532)	89,9 % [87,3-92] (602/670)

TABLA 14. La precisión de clasificación usa las regiones genómicas de la lista 10.

Estadio	Presencia del cáncer	Tipo de cáncer
I	14,2 % [11-17,9] (60/422)	72,3 % [57,4-84,4] (34/47)
II	36,9 % [32-41,9] (143/388)	87,2 % [80,3-92,4] (116/133)
III	71,9 % [66,6-76,8] (225/313)	92,6 % [88-95,8] (187/202)
IV	85,1 % [81-88,6] (309/363)	90,9 % [87-93,9] (269/296)
Todos	49,7 % [47,2-52,3] (762/1532)	89,6 % [87,1-91,7] (627/700)

TABLA 15. La precisión de clasificación usa las regiones genómicas de la lista 11.

Estadio	Presencia del cáncer	Tipo de cáncer
I	13 % [10-16,6] (55/422)	78,3 % [63,6-89,1] (36/46)
II	35,3 % [30,6-40,3] (137/388)	90,7 % [84,3-95,1] (117/129)
III	72,5 % [67,2-77,4] (227/313)	87,7 % [82,5-91,8] (185/211)
IV	85,1 % [81-88,6] (309/363)	91,1 % [87,3-94,1] (277/304)
Todos	49,3 % [46,8-51,9] (756/1532)	89,4 % [86,9-91,6] (641/717)

TABLA 16. La precisión de clasificación usa las regiones genómicas de la lista 12.

Estadio	Presencia del cáncer	Tipo de cáncer
I	13,5 % [10,4-17,1] (57/422)	73,5 % [58,9-85,1] (36/49)
II	36,9 % [32-41,9] (143/388)	88,5 % [81,7-93,4] (115/130)
III	72,2 % [66,9-77,1] (226/313)	92,5 % [87,9-95,7] (185/200)
IV	84,8 % [80,7-88,4] (308/363)	91,5 % [87,7-94,4] (269/294)
Todos	49,6 % [47,1-52,1] (760/1532)	90,1 % [87,6-92,2] (628/697)

TABLA 17. La precisión de clasificación usa las regiones genómicas de la lista 13.

Estadio	Presencia del cáncer	Tipo de cáncer
I	9 % [6,5-12,2] (38/422)	78,9 % [54,4-93,9] (15/19)
II	29,9 % [25,4-34,7] (116/388)	86 % [76,9-92,6] (74/86)
III	57,5 % [51,8-63,1] (180/313)	92,1 % [86,3-96] (128/139)
IV	80,7 % [76,3-84,6] (293/363)	90,7 % [86,3-94,1] (215/237)
Todos	42 % [39,5-44,5] (643/1532)	90,1 % [87,1-92,6] (445/494)

TABLA 18. La precisión de clasificación usa las regiones genómicas de la lista 14.

Estadio	Presencia del cáncer	Tipo de cáncer
I	8,5 % [6-11,6] (36/422)	75 % [50,9-91,3] (15/20)
II	30,2 % [25,6-35] (117/388)	85,9 % [77-92,3] (79/92)
III	61,3 % [55,7-66,8] (192/313)	91,4 % [85,7-95,3] (138/151)
IV	81 % [76,6-84,9] (294/363)	90,2 % [85,8-93,6] (222/246)
Todos	43,4 % [40,9-45,9] (665/1532)	89,6 % [86,7-92,1] (474/529)

TABLA 19. La precisión de clasificación usa las regiones genómicas de la lista 15.

Estadio	Presencia del cáncer	Tipo de cáncer
I	10,2 % [7,5-13,5] (43/422)	70,4 % [49,8-86,2] (19/27)
II	31,7 % [27,1-36,6] (123/388)	87,4 % [79,4-93,1] (90/103)
III	62 % [56,4-67,4] (194/313)	91,7 % [86,3-95,5] (144/157)
IV	82,1 % [77,8-85,9] (298/363)	90,5 % [86,2-93,7] (237/262)
Todos	44,7 % [42,2-47,2] (685/1532)	89,7 % [86,9-92,1] (514/573)

TABLA 20. La precisión de clasificación usa las regiones genómicas de la lista 16.

Estadio	Presencia del cáncer	Tipo de cáncer
I	10,2 % [7,5-13,5] (43/422)	65,4 % [44,3-82,8] (17/26)
II	33 % [28,3-37,9] (128/388)	88,5 % [81,1-93,7] (100/113)
III	65,5 % [59,9-70,8] (205/313)	90,9 % [85,4-94,8] (150/165)
IV	83,2 % [78,9-86,9] (302/363)	91,3 % [87,3-94,4] (242/265)
Todos	46 % [43,5-48,6] (705/1532)	89,7 % [87-92] (532/593)

TABLA 21. Precisión de la clasificación usando un subconjunto seleccionado al azar del 10 % de las regiones genómicas de la lista 12.

Estadio	Presencia del cáncer	Tipo de cáncer
I	10 % [7,3-13,2] (42/422)	78,1 % [60-90,7] (25/32)
II	32 % [27,3-36,9] (124/388)	87,5 % [79,9-93] (98/112)
III	61 % [55,4-66,5] (191/313)	89,6 % [84,1-93,7] (155/173)
IV	82,9 % [78,6-86,6] (301/363)	90,5 % [86,4-93,7] (247/273)
Todos	44,1 % [41,6-46,7] (676/1532)	89,3 % [86,6-91,6] (542/607)

TABLA 22. Precisión de la clasificación usando un subconjunto seleccionado al azar del 25 % de las regiones genómicas de la lista 12.

Estadio	Presencia del cáncer	Tipo de cáncer
I	11,8 % [8,9-15,3] (50/422)	71,4 % [55,4-84,3] (30/42)
II	33,2 % [28,6-38,2] (129/388)	90,8 % [84,2-95,3] (109/120)
III	65,5 % [59,9-70,8] (205/313)	89,9 % [84,7-93,8] (169/188)
IV	84,6 % [80,4-88,1] (307/363)	91,5 % [87,7-94,5] (260/284)
Todos	46,5 % [44-49,1] (713/1532)	89,9 % [87,4-92,1] (589/655)

TABLA 23. Precisión de la clasificación usando un subconjunto seleccionado al azar del 50 % de las regiones genómicas de la lista 4.

Estadio	Presencia del cáncer	Tipo de cáncer
I	11,4 % [8,5-14,8] (48/422)	73,8 % [58-86,1] (31/42)
II	33,2 % [28,6-38,2] (129/388)	88,5 % [81,5-93,6] (108/122)
III	64,9 % [59,3-70,1] (203/313)	92,9 % [88,2-96,2] (171/184)
IV	83,2 % [78,9-86,9] (302/363)	90,4 % [86,4-93,5] (263/291)
Todos	46,3 % [43,8-48,9] (710/1532)	89,8 % [87,2-92] (598/666)

TABLA 24. Precisión de clasificación del tipo de cáncer con diversas regiones de destino genómico.

Tipo de cáncer	Lista 4		Lista 5		Lista 6		Lista 7		Lista 8		Lista 9		Lista 10	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos los tipos	92	613/666	91	740/815	91	622/686	90	747/827	91	576/633	90	593/656	91	627/688
Tiroides	n/a	0/0	0	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Melanoma	n/a	0/0	100	3/3	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Sarcoma	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3
Neoplasia mieloide	100	3/3	0	0/0	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3
Renal	n/a	0/0	92	12/13	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Próstata	91	10/11	90	17/19	91	10/11	92	12/13	90	9/10	90	9/10	91	10/11
Mama	100	13/13	95	80/84	100	13/13	90	18/20	100	15/15	100	14/14	100	13/13
De útero	97	60/62	86	19/22	95	62/65	93	82/88	97	57/59	97	60/62	96	63/66
Ovario	100	8/8	96	26/27	100	10/10	91	20/22	89	8/9	100	8/8	92	12/13
Vejiga y urotelio	100	28/28	100	2/2	100	27/27	96	26/27	96	22/23	92	23/25	96	26/27
De cuello uterino	n/a	0/0	50	2/4	100	1/1	100	2/2	100	2/2	100	2/2	67	2/3
Anorrectal	0	0/1	100	1/1	n/a	0/0	40	2/5	0	0/1	0	0/1	50	1-2;
Cabeza y cuello	n/a	0/0	75	46/61	n/a	0/0	100	1/1	n/a	0/0	100	1/1	100	1/1
Colorrectal	73	37/51	99	93/94	69	38/55	73	46/63	76	37/49	71	37/52	73	37/51
Hígado y vías biliares	100	74/74	95	18/19	99	73/74	99	94/95	100	63/63	99	64/65	100	71/71
Páncreas y vesícula biliar	95	18/19	90	68/76	90	18/20	95	18/19	90	17/19	90	17/19	95	18/19

(continuación)

Tipo de cáncer	Lista 11		Lista 12		Lista 13		Lista 14		Lista 15		Lista 16	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos los tipos	92	627/684	91	620/681	91	416/456	91	450/497	92	488/533	91	513/565
Tiroides	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Melanoma	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Sarcoma	100	3/3	100	3/3	100	2/2	100	1/1	100	2/2	100	3/3
Neoplasia mieloide	100	3/3	100	3/3	100	1/1	100	1/1	100	2/2	100	2/2
Renal	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Próstata	91	10/11	91	10/11	100	3/3	100	3/3	100	6/6	100	6/6
Mama	100	14/14	100	13/13	100	5/5	100	6/6	100	10/10	100	9/9
De útero	97	62/64	94	63/67	98	39/40	98	42/43	98	47/48	98	50/51
Ovario	100	11/11	92	11/12	100	1/1	100	2/2	67	2/3	100	2/2
Vejiga y urotelio	100	26/26	100	27/27	100	16/16	94	17/18	95	18/19	96	21/22
De cuello uterino	100	2/2	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Anorrectal	0	0/1	0	0/1	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Cabeza y cuello	100	1/1	100	1/1	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Colorrectal	69	38/55	70	37/53	71	25/35	68	28/41	64	27/42	65	30/46
Hígado y vías biliares	99	75/76	100	71/71	100	54/54	100	56/56	98	63/64	100	64/64
Páncreas y vesícula biliar	95	19/20	95	19/20	93	13/14	82	14/17	94	15/16	88	15/17

(continuación)

Tipo de cáncer	Lista 4		El 50 % aleatorio de la lista 4		Lista 12		El 10 % aleatorio de la lista 12		El 25 % aleatorio de la lista 12	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos los tipos	92	613/666	91	578/635	91	620/681	91	531/586	91	567/622
Tiroides	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Melanoma	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Sarcoma	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3
Neoplasia mieloide	100	3/3	100	3/3	100	3/3	100	3/3	100	3/3
Renal	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0	n/a	0/0
Próstata	91	10/11	89	8/9	91	10/11	100	7/7	90	9/10
Mama	100	13/13	100	13/13	100	13/13	93	13/14	100	14/14
De útero	97	60/62	93	57/61	94	63/67	96	54/56	95	58/61
Ovario	100	8/8	88	7/8	92	11/12	83	5/6	100	7/7
Vejiga y urotelio	100	28/28	96	26/27	100	27/27	100	19/19	100	23/23
De cuello uterino	n/a	0/0	n/a	0/0	n/a	0/0	100	1/1	100	1/1
Anorrectal	0	0/1	0	0/1	0	0/1	0	0/1	n/a	0/0
Cabeza y cuello	n/a	0/0	n/a	0/0	100	1/1	n/a	0/0	n/a	0/0
Colorrectal	73	37/51	71	37/52	70	37/53	72	33/46	69	34/49
Hígado y vías biliares	100	74/74	99	69/70	100	71/71	100	65/65	99	66/67
Páncreas y vesícula biliar	95	18/19	94	17/18	95	19/20	95	18/19	95	18/19
Ovario	100	8/8	88	7/8	92	11/12	83	5/6	100	7/7

TABLA 25. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 4.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	25	1/4	20	11/56	9	16/188	27	67/247	18	15/84	87	32/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	3	1/39	2	2/102	14	10/73	25	1/4	33	2/6
I	0	0/1	0	0/1	0	0/4			0	0/4	1	1/113	33	36/110	33	1/3	0	0/2	43	3/7
III					40	2/5			50	2/4	11	2/19	85	23/27	60	3/5	100	25/25	50	1/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	71	12/17	75	6/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	84	52/62	65	79/121	86	25/29	68	65/95	73	49/67	77	26/34	73	107/147	64	168/261
I	13	1/8	25	1/4	86	6/7	8	1/13	60	3/5	33	5/15	11	1/9	60	6/10	27	4/15	12	7/60
II	100	1/1	60	3/5	77	10/13	36	8/22	86	6/7	60	9/15	58	7/12	69	9/13	85	23/27	65	15/23
III	50	1/2	100	5/5	75	12/16	68	28/41	86	6/7	58	11/19	79	15/19	100	11/11	78	21/27	75	54/72
IV					92	24/26	93	42/45	100	10/10	87	40/46	96	26/27			80	31/39	87	92/106

TABLA 26. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 5

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	53	9/17	0	0/4	23	13/56	13	25/188	35	86/247	30	25/84	92	34/37	63	10/16
I	0	0/2	0	0/3	100	2/2			0	0/37	5	2/39	5	5/102	25	18/73	75	3/4	67	4/6
II	0	0/1	0	0/1	25	1/4			50	2/4	5	6/113	45	49/110	67	2/3	0	0/2	57	4/7
III					40	2/5			50	2/4	11	2/19	89	24/27	60	3/5	100	25/25	50	1/2
IV	0	0/1	100	3/3	67	4/6			82	9/11	88	15/17	100	8/8	67	2/3	100	6/6	100	1/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	46	5/11	71	10/14	92	57/62	83	100/121	86	25/29	85	81/95	85	57/67	74	25/34	71	105/147	74	194/261
I	25	2/8	50	2/4	86	6/7	46	6/13	60	3/5	60	9/15	22	2/9	50	5/10	27	4/15	20	12/60
II	100	1/1	60	3/5	92	12/13	77	17/22	86	6/7	80	12/15	83	10/12	69	9/13	82	22/27	87	20/23
III	100	2/2	100	5/5	94	15/16	85	35/41	86	6/7	84	16/19	95	18/19	100	11/11	85	23/27	89	64/72
IV					92	24/26	93	42/45	100	10/10	96	44/46	100	27/27			80	31/39	93	98/106

TABLA 27. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 6.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	20	11/56	9	17/188	29	71/247	18	15/84	87	32/37	44	7/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	3	3/102	14	10/73	25	1/4	67	4/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	34	37/110	33	1/3	0	0/2	43	3/7
III					60	3/5			50	2/4	11	2/19	85	23/27	60	3/5	100	25/25	0	0/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	71	12/17	100	8/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	81	50/62	68	82/121	86	25/29	70	66/95	69	46/67	79	27/34	72	106/147	65	169/261
I	13	1/8	25	1/4	86	6/7	15	2/13	60	3/5	27	4/15	11	1/9	60	6/10	20	3/15	12	7/60
II	100	1/1	60	3/5	77	10/13	41	9/22	86	6/7	67	10/15	58	7/12	77	10/13	89	24/27	65	15/23
III	50	1/2	100	5/5	69	11/16	71	29/41	86	6/7	63	12/19	68	13/19	100	11/11	82	22/27	76	55/72
IV					89	23/26	93	42/45	100	10/10	87	40/46	93	25-27;			80	31/39	87	92/106

TABLA 28. Sensibilidad de clasificación al 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 7.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	53	9/17	0	0/4	23	13/56	12	23/188	35	86/247	30	25/84	89	33/37	63	10/16
I	0	0/2	0	0/3	100	2/2			0	0/37	5	2/39	5	5/102	25	18/73	75	3/4	67	4/6
II	0	0/1	0	0/1	25	1/4			50	2/4	4	5/113	45	49/110	67	2/3	0	0/2	57	4/7
III					40	2/5			50	2/4	11	2/19	89	24/27	60	3/5	96	24/25	50	1/2
IV	0	0/1	100	3/3	67	4/6			82	9/11	82	14/17	100	8/8	67	2/3	100	6/6	100	1/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	36	4/11	71	10/14	94	58/62	81	98/121	86	25/29	86	82/95	85	57/67	79	27/34	72	106/147	74	194/261
I	13	1/8	50	2/4	100	7/7	46	6/13	60	3/5	60	9/15	22	2/9	60	6/10	33	5/15	20	12/60
II	100	1/1	60	3/5	92	12/13	73	16/22	86	6/7	87	13/15	83	10/12	77	10/13	82	22/27	87	20/23
III	100	2/2	100	5/5	94	15/16	83	34/41	86	6/7	84	16/19	95	18/19	100	11/11	85	23/27	89	64/72
IV					92	24/26	93	42/45	100	10/10	96	44/46	100	27/27			80	31/39	93	98/106

TABLA 29. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 8.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	25	1/4	43	3/7	41	7/17	25	1/4	20	11/56	11	20/188	26	64/247	16	13/84	84	31/37	44	7/16
I	0	0/2	0	0/3	100	2/2			0	0/37	5	2/39	2	2/102	11	8/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	29	32/110	33	1/3	0	0/2	57	4/7
III					40	2/5			50	2/4	11	2/19	85	23/27	60	3/5	96	24/25	50	1/2
IV	100	1/1	100	3/3	50	3/6			82	9/11	77	13/17	88	7/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	71	10/14	81	50/62	59	71/121	83	24/29	64	61/95	69	46/67	79	27/34	69	102/147	62	161/261
I	13	1/8	50	2/4	100	7/7	15	2/13	40	2/5	27	4/15	11	1/9	60	6/10	27	4/15	12	7/60
II	100	1/1	60	3/5	77	10/13	23	5/22	86	6/7	53	8/15	50	6/12	77	10/13	85	23/27	57	13/23
III	50	1/2	100	5/5	63	10/16	56	23/41	86	6/7	47	9/19	68	13/19	100	11/11	74	20/27	74	53/72
IV					89	23/26	91	41/45	100	10/10	87	40/46	96	26/27			80	31/39	83	88/106

TABLA 30. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 9.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	21	12/56	10	19/188	27	66/247	17	14/84	81	30/37	44	7/16
I	0	0/2	0	0/3	50	1/2			3	1/37	0	0/39	3	3/102	12	9/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	4	4/113	31	34/110	33	1/3	0	0/2	57	4/7
III					40	2/5			50	2/4	11	2/19	82	22/27	60	3/5	92	23/25	50	1/2
IV	0	0/1	100	3/3	50	3/6			82	9/11	77	13/17	88	7/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	81	50/62	61	74/121	83	24/29	68	65/95	69	46/67	77	26/34	69	102/147	63	165/261
I	13	1/8	25	1/4	86	6/7	15	2/13	40	2/5	27	4/15	11	1/9	60	6/10	20	3/15	10	6/60
II	100	1/1	60	3/5	77	10/13	32	7/22	86	6/7	67	10/15	58	7/12	69	9/13	85	23/27	65	15/23
III	50	1/2	100	5/5	69	11/16	61	25/41	86	6/7	58	11/19	68	13/19	100	11/11	70	19/27	74	53/72
IV					89	23/26	89	40/45	100	10/10	87	40/46	93	25/27			82	32/39	86	91/106

TABLA 31. Sensibilidad de clasificación al 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 10.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	25	1/4	20	11/56	10	18/188	28	69/247	19	16/84	84	31/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	3	1/39	3	3/102	15	11/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	34	37/110	33	1/3	0	0/2	43	3/7
III					40	2/5			50	2/4	11	2/19	85	23/27	60	3/5	96	24/25	50	1/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	71	12/17	75	6/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	82	51/62	68	82/121	86	25/29	72	68/95	70	47/67	79	27/34	69	102/147	67	174/261
I	13	1/8	25	1/4	86	6/7	15	2/13	60	3/5	33	5/15	11	1/9	60	6/10	27	4/15	17	10/60
II	100	1/1	60	3/5	77	10/13	41	9/22	86	6/7	67	10/15	67	8/12	77	10/13	89	24/27	70	16/23
III	50	1/2	100	5/5	75	12/16	71	29/41	86	6/7	68	13/19	63	44184	100	11/11	74	20/27	76	55/72
IV					89	23/26	93	42/45	100	10/10	87	40/46	96	26/27			80	31/39	88	93/106

TABLA 32. Sensibilidad de clasificación al 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 11.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	20	11/56	10	18/188	28	69/247	18	15/84	84	31/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	4	4/102	14	10/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	33	36/110	33	1/3	0	0/2	43	3/7
III					60	3/5			50	2/4	11	2/19	85	23/27	60	3/5	96	24/25	50	1/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	77	13/17	75	6/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	84	52/62	65	78/121	86	25/29	71	67/95	70	47/67	79	27/34	70	103/147	66	171/261
I	13	1/8	25	1/4	86	6/7	8	1/13	60	3/5	33	5/15	11	1/9	60	6/10	20	3/15	13	8/60
II	100	1/1	60	3/5	77	10/13	27	6/22	86	6/7	67	10/15	67	8/12	77	10/13	85	23/27	65	15/23
III	50	1/2	100	5/5	75	12/16	71	29/41	86	6/7	63	12/19	68	13/19	100	11/11	74	20/27	78	56/72
IV					92	24/26	93	42/45	100	10/10	87	40/46	93	25/27			80	31/39	87	92/106

TABLA 33. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 12.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	20	11/56	10	18/188	28	69/247	18	15/84	87	32/37	44	7/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	2	2/102	14	10/73	25	1/4	50	3/6
II	0	0/1	0	0/1	0	0/4			0	0/4	4	4/113	35	38/110	33	1/3	0	0/2	43	3/7
III					60	3/5			50	2/4	11	2/19	85	23/27	60	3/5	100	25/25	50	1/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	71	12/17	75	6/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	81	50/62	66	80/121	86	25/29	76	72/95	70	47/67	79	27/34	69	101/147	65	169/261
I	13	1/8	25	1/4	86	6/7	15	2/13	60	3/5	40	6/15	11	1/9	60	6/10	20	3/15	13	8/60
II	100	1/1	60	3/5	77	10/13	36	8/22	86	6/7	73	11/15	67	8/12	77	10/13	85	23/27	65	15/23
III	50	1/2	100	5/5	69	11/16	68	28/41	86	6/7	74	14/19	68	13/19	100	11/11	74	20/27	75	54/72
IV					89	23/26	93	42/45	100	10/10	89	41-46;	93	25/27			80	31/39	87	92/106

TABLA 34. Sensibilidad de clasificación al 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 13.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	14	8/56	9	16/188	21	52/247	11	9/84	62	23/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	1	1/102	8	6/73	25	1/4	50	3/6
II	0	0/1	0	0/1	0	0/4			0	0/4	2	2/113	25	27/110	33	1/3	0	0/2	43	3/7
III					40	2/5			25	1/4	11	2/19	67	18/27	20	1/5	68	17/25	0	0/2
IV	0	0/1	100	3/3	50	3/6			64	7/11	71	12/17	75	6/8	33	1/3	83	5/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	71	44/62	55	66/121	83	24/29	60	57/95	54	36/67	65	22/34	62	91/147	59	154/261
I	13	1/8	25	1/4	86	6/7	8	1/13	40	2/5	7	1/15	0	0/9	40	4/10	20	3/15	8	5/60
II	100	1/1	60	3/5	77	10/13	18	4/22	86	6/7	60	9/15	33	4/12	54	7/13	89	24/27	57	13/23
III	50	1/2	100	5/5	44	7/16	56	23/41	86	6/7	37	7/19	37	7-19;	100	11/11	70	19/27	68	49/72
IV					81	21/26	84	38/45	100	10/10	87	40/46	93	25/27			77	30/39	82	87/106

TABLA 35. Sensibilidad de clasificación al 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 14.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	50	2/4	16	9/56	9	16/188	22	55/247	12	10/84	65	24/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	1	1/102	8	6/73	25	1/4	50	3/6
II	0	0/1	0	0/1	0	0/4			0	0/4	2	2/113	26	28/110	33	1/3	0	0/2	43	3/7
III					40	2/5			25	1/4	11	2/19	74	20/27	40	2/5	72	18/25	0	0/2
IV	0	0/1	100	3/3	33	2/6			73	8/11	71	12/17	75	6/8	33	1/3	83	5/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	79	49/62	58	70/121	83	24/29	57	54/95	60	40/67	65	22/34	67	99/147	58	151/261
I	13	1/8	25	1/4	86	6/7	8	1/13	40	2/5	7	1/15	0	0/9	40	4/10	20	3/15	5	3/60
II	100	1/1	60	3/5	77	10/13	18	4/22	86	6/7	53	8/15	42	5/12	54	7/13	85	23/27	61	14/23
III	50	1/2	100	5/5	63	10/16	61	25/41	86	6/7	42	8/19	53	10/19	100	11/11	70	19/27	67	48/72
IV					89	23/26	89	40/45	100	10/10	80	37/46	93	25/27			80	31/39	81	86/106

TABLA 36. Sensibilidad de clasificación con el 99,0 % de especificidad usando las regiones genómicas diana de la lista 15.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	35	6/17	25	1/4	16	9/56	9	17/188	23	56/247	13	11/84	65	24/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	1	1/102	8	6/73	25	1/4	50	3/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	27	30/110	33	1/3	0	0/2	43	3/7
III					40	2/5			25	1/4	11	2/19	70	19/27	60	3/5	72	18/25	0	0/2
IV	0	0/1	100	3/3	50	3/6			73	8/11	71	12/17	75	6/8	33	1/3	83	5/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	81	50/62	60	73/121	86	25/29	60	57/95	63	42/67	68	23/34	70	103/147	59	154/261
I	13	1/8	25	1/4	86	6/7	15	2/13	60	3/5	20	3/15	11	1/9	50	5/10	27	4/15	5	3/60
II	100	1/1	60	3/5	77	10/13	27	6/22	86	6/7	53	8/15	50	6/12	54	7/13	85	23/27	61	14/23
III	50	1/2	100	5/5	69	11/16	61	25/41	86	6/7	37	7/19	53	10/19	100	11/11	70	19/27	69	50/72
IV					89	23/26	89	40/45	100	10/10	85	39/46	93	25/27			80	31/39	82	87/106

TABLA 37. Sensibilidad de clasificación con el 99,0 % de especificidad mediante el uso de las regiones genómicas diana de la lista 16.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	41	7/17	25	1/4	16	9/56	9	16/188	25	62/247	13	11/84	73	27/37	44	7/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	1	1/102	8	6/73	25	1/4	67	4/6
II	0	0/1	0	0/1	0	0/4			0	0/4	2	2/113	30	33/110	33	1/3	0	0/2	43	3/7
III					60	3/5			25	1/4	11	2/19	74	20/27	60	3/5	80	20/25	0	0/2
IV	0	0/1	100	3/3	50	3/6			73	8/11	71	12/17	100	8/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	81	50/62	62	75/121	86	25/29	64	61/95	64	43/67	68	23/34	71	104/147	60	156/261
I	13	1/8	25	1/4	86	6/7	15	2/13	60	3/5	20	3/15	0	0/9	40	4/10	27	4/15	7	4/60
II	100	1/1	60	3/5	77	10/13	27	6/22	86	6/7	60	9/15	58	7/12	62	8/13	85	23/27	61	14/23
III	50	1/2	100	5/5	69	11/16	63	26/41	86	6/7	53	10/19	58	11/19	100	11/11	74	20/27	71	51/72
IV					89	23/26	91	41/45	100	10/10	85	39/46	93	25/27			80	31/39	82	87/106

TABLA 38. Sensibilidad de clasificación con una especificidad del 99,0 % utilizando un subconjunto seleccionado al azar del 10 % de las regiones genómicas diana de la lista 12.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	0	0/4	16	9/56	9	16/188	25	62/247	11	9/84	65	24/37	38	6/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	2	2/102	7	5/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	2	2/113	29	32/110	33	1/3	0	0/2	43	3/7
III					40	2/5			25	1/4	11	2/19	78	21/27	40	2/5	68	17/25	0	0/2
IV	0	0/1	100	3/3	33	2/6			73	8/11	71	12/17	88	7/8	33	1/3	100	6/6	100	1/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	77	48/62	59	71/121	86	25/29	62	59/95	60	40/67	74	25/34	65	95/147	59	154/261
I	13	1/8	25	1/4	86	6/7	8	1/13	60	3/5	33	5/15	0	0/9	50	5/10	20	3/15	7	4/60
II	100	1/1	60	3/5	77	10/13	32	7/22	86	6/7	53	8/15	42	5/12	69	9/13	89	24/27	48	11/23
III	50	1/2	100	5/5	56	9/16	56	23/41	86	6/7	42	8/19	47	9/19	100	11/11	70	19/27	71	51/72
IV					89	23/26	89	40/45	100	10/10	83	38/46	96	26/27			80	31/39	83	88/106

TABLA 39. Sensibilidad de clasificación con una especificidad del 99,0 % utilizando un subconjunto seleccionado al azar del 25 % de las regiones genómicas diana de la lista 12.

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	25	1/4	20	11/56	9	17/188	27	66/247	14	12/84	78	29/37	44	7/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	2	2/102	10	7/73	25	1/4	50	3/6
II	0	0/1	0	0/1	0	0/4			0	0/4	3	3/113	32	35/110	33	1/3	0	0/2	43	3/7
III					40	2/5			50	2/4	11	2/19	82	22/27	60	3/5	88	22/25	0	0/2
IV	0	0/1	100	3/3	33	2/6			82	9/11	71	12/17	88	7/8	33	1/3	100	6/6	100	1/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	71	10/14	79	49/62	61	74/121	86	25/29	65	62/95	64	43/67	68	23/34	67	98/147	62	161/261
I	13	1/8	50	2/4	86	6/7	15	2/13	60	3/5	33	5/15	11	1/9	50	5/10	27	4/15	7	4/60
II	100	1/1	60	3/5	77	10/13	27	6/22	86	6/7	73	11/15	33	4/12	54	7/13	85	23/27	61	14/23
III	50	1/2	100	5/5	63	10/16	59	24/41	86	6/7	42	8/19	63	12/19	100	11/11	70	19/27	72	52/72
IV					89	23/26	93	42/45	100	10/10	83	38/46	96	26/27			80	31/39	86	91/106

TABLA 40. Sensibilidad de clasificación con una especificidad del 99,0 % utilizando un subconjunto seleccionado al azar del 50 % de las regiones genómicas diana de la lista 4

Estadio	Tiroides		Melanoma		Sarcoma		Neoplasia mieloide		Renal		Próstata		Mama		De útero		Ovario		Vejiga y urotelio	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	0	0/4	43	3/7	29	5/17	0	0/4	18	10/56	8	15/188	25	62/247	17	14/84	81	30/37	31	5/16
I	0	0/2	0	0/3	50	1/2			0	0/37	0	0/39	1	1/102	12	9/73	25	1/4	33	2/6
II	0	0/1	0	0/1	0	0/4			0	0/4	1	1/113	29	32/110	33	1/3	0	0/2	43	3/7
III					40	2/5			50	2/4	11	2/19	82	22/27	60	3/5	92	23/25	0	0/2
IV	0	0/1	100	3/3	33	2/6			73	8/11	71	12/17	88	7/8	33	1/3	100	6/6	0	0/1

Estadio	De cuello uterino		Anorrectal		Cabeza y cuello		Colorrectal		Hígado		Páncreas y vesícula biliar		Tracto gastrointestinal superior		Mieloma múltiple		Neoplasia linfoide		Pulmón	
	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	27	3/11	64	9/14	76	47/62	63	76/121	86	25/29	60	57/95	66	44/67	74	25/34	69	101/147	63	165/261
I	13	1/8	25	1/4	86	6/7	8	1/13	60	3/5	20	3/15	11	1/9	50	5/10	20	3/15	13	8/60
II	100	1/1	60	3/5	77	10/13	41	9/22	86	6/7	60	9/15	50	6/12	69	9/13	85	23/27	61	14/23
III	50	1/2	100	5/5	56	9/16	63	26/41	86	6/7	37	7/19	58	11/19	100	11/11	67	18/27	71	51/72
IV					85	22/26	89	40/45	100	10/10	83	38/46	96	26/27			77	30/39	87	92/106

EJEMPLO 6 - Detección de cáncer usando un panel de ensayos de cáncer

Se recogen muestras de sangre de un grupo de individuos previamente diagnosticados de cáncer de un TOO ("grupo de prueba"), y de otros grupos de individuos sin cáncer o diagnosticados de un tipo diferente de cáncer ("otro grupo"). Se extraen fragmentos de ADNlc de las muestras de sangre y se tratan con bisulfito para convertir las citosinas no metiladas en uracilos. A las muestras tratadas con bisulfito se les aplicó el panel de ensayos de cáncer descrito en la presente memoria. Los fragmentos de ADNlc no unidos se lavan y se recogen fragmentos de ADNlc unidos a las sondas. Los fragmentos de ADNlc recogidos se amplifican y secuencian. Las lecturas de secuencia confirman que las sondas enriquecen específicamente fragmentos de ADNlc que tienen patrones de metilación indicativos de cáncer de un TOO y muestras del grupo de prueba incluyen significativamente más de los fragmentos de ADNlc metilados diferencialmente en comparación con otros grupos.

REIVINDICACIONES

1. Un método describe un método para analizar una muestra para detectar células de un tipo de cáncer en un sujeto, comprendiendo el método:

(a) capturar fragmentos de ADN libre de células (ADNlc) convertidos o no convertidos a partir de la muestra o productos de amplificación de los mismos con una composición que comprende una pluralidad de oligonucleótidos cebo diferentes, en donde el ADNlc convertido se refiere a ADNlc que ha sido tratado para convertir citosinas no metiladas en uracilos, en donde:

(i) cada oligonucleótido cebo en la pluralidad de oligonucleótidos cebo diferentes tiene al menos 45 nucleótidos de longitud;

(ii) la pluralidad de oligonucleótidos cebo diferentes se hibridan colectivamente con al menos 200 regiones genómicas diana;

(iii) las al menos 200 regiones genómicas diana se metilan diferencialmente en ADNlc de sujetos de referencia con al menos un tipo de cáncer en relación con un tipo de cáncer diferente o en relación con el no cáncer; y

(iv) las al menos 200 regiones genómicas diana comprenden, para al menos el 80 % de todos los pares posibles de tipos de cáncer seleccionados de un conjunto que comprende al menos 10 tipos de cáncer, al menos una región genómica diana que está diferencialmente metilada entre el par de tipos de cáncer;

(b) secuenciar los fragmentos de ADNlc capturados o los productos de amplificación de los mismos para producir lecturas de secuenciación, en donde, si el ADNlc capturado en la etapa a) no se ha convertido, entonces es ADNlc se convierte antes de la secuenciación tratando el ADNlc para convertir las citosinas no metiladas en uracilos; y

(c) aplicar un clasificador entrenado a las lecturas de secuenciación para determinar la presencia o ausencia de cáncer, en donde:

(i) el clasificador entrenado es un clasificador entrenado mediante el uso de muestras de ADNlc convertido, en donde el ADNlc convertido se refiere a ADNlc que se ha tratado para convertir citosinas no metiladas en uracilos;

(ii) el clasificador entrenado compara un número de lecturas de secuenciación para una pluralidad de las al menos 200 regiones genómicas diana que se identifican como hipermetiladas o hipometiladas en los fragmentos de ADNlc con un umbral para cada uno de los al menos 10 tipos de cáncer; y

(iii) detección por encima del umbral para un tipo de cáncer respectivo identifica la presencia de las células del tipo de cáncer en el sujeto.

2. El método de la reivindicación 1, en donde los al menos 10 tipos de cáncer se seleccionan de:

(a) cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatobiliar, hepatobiliar biliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal;

(b) cáncer anorrectal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago; o

(c) cáncer de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón.

3. El método de la reivindicación 1, en donde las al menos 200 regiones genómicas diana:

(a) comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas diana en una cualquiera de las listas 1-16, o sus complementos;

(b) comprenden al menos 500 regiones genómicas diana en una cualquiera de las listas 1-16, o sus complementos;

(c) comprenden al menos el 20 % de las regiones genómicas diana en una cualquiera de las listas 1-3, o sus complementos;

(d) comprenden al menos el 10 % de las regiones genómicas diana en una cualquiera de las listas 13-16, o sus complementos;

(e) comprenden al menos el 10 % de las regiones genómicas diana en la lista 12, o sus complementos;

(f) comprenden al menos el 20 % de las regiones genómicas diana en una cualquiera de las listas 8-11, o sus complementos;

(g) comprenden al menos el 40 % de las regiones genómicas diana en la lista 4, o sus complementos;

(h) comprenden, para al menos el 90 % de todos los pares posibles de tipos de cáncer seleccionados del conjunto que comprende los al menos 10 tipos de cáncer, al menos una región genómica diana que está diferencialmente metilada entre el par de tipos de cáncer; o

(i) son fragmentos de ADNlc convertido; opcionalmente en donde los fragmentos de ADNlc se convierten mediante tratamiento con bisulfito, una reacción de conversión enzimática o una citosina desaminasa.

4. El método de la reivindicación 1, en donde cada oligonucleótido cebo de la pluralidad de oligonucleótidos cebo:

(a) se hibrida con al menos 45 nucleótidos de una región genómica diana seleccionada de las al menos 200 regiones genómicas diana;

(b) se conjuga con un resto de afinidad; opcionalmente en donde el resto de afinidad es biotina; o

(c) tiene una longitud de entre 50 y 300 bases.

5. El método de la reivindicación 1, en donde las al menos 200 regiones genómicas diana se seleccionan de una cualquiera de las listas 1-16, o sus complementos.

6. El método de la reivindicación 1, en donde el tamaño total de las al menos 200 regiones genómicas diana comprende de 0,2 MB a 30 MB.

7. El método de la reivindicación 1, en donde el clasificador es un clasificador entrenado en secuencias de ADNlc convertido derivadas de al menos 100 sujetos con un primer tipo de cáncer, al menos 100 sujetos con un segundo tipo de cáncer, y al menos 100 sujetos sin cáncer.

8. El método de la reivindicación 1, en donde los fragmentos de ADNlc de la muestra son moléculas de ADNlc convertido; opcionalmente en donde los fragmentos de ADNlc se convierten mediante un proceso que comprende tratamiento con bisulfito.

9. El método de la reivindicación 1, en donde cada una de las al menos 200 regiones genómicas diana comprende al menos 5 dinucleótidos CpG.

10. El método de la reivindicación 1, en donde la pluralidad de oligonucleótidos cebo diferentes comprende una pluralidad de conjuntos de dos o más oligonucleótidos cebo, en donde cada oligonucleótido cebo dentro de un conjunto de oligonucleótidos cebo está configurado para unirse a moléculas de ADN transformadas de la misma región genómica diana con diferentes estados de metilación.

11. El método de la reivindicación 1, en donde la razón de oligonucleótidos cebo configurados para hibridarse con regiones diana hipermetiladas para cebar oligonucleótidos configurados para hibridarse con regiones diana hipometiladas está entre 0,5 y 1,0.

12. El método de la reivindicación 1, en donde:

(i) la pluralidad de oligonucleótidos cebo diferentes comprende uno o más pares de un primer oligonucleótido cebo y un segundo oligonucleótido cebo,

(ii) cada oligonucleótido cebo comprende un extremo 5' y un extremo 3',

(iii) una secuencia de al menos X bases de nucleótido en el extremo 3' del primer oligonucleótido cebo es idéntica a una secuencia de X bases de nucleótido en el extremo 5' del segundo oligonucleótido cebo, y

(iv) X es al menos 20, al menos 25, o al menos 30.

13. El método de la reivindicación 1, en donde:

(i) la probabilidad de una determinación de falsos positivos de una presencia o ausencia de cáncer es inferior al 1 % y la probabilidad de una determinación precisa de una presencia o ausencia de cáncer es de al menos el 40 % ; o

(ii) el cáncer es un cáncer de estadio I, la probabilidad de una determinación de falsos positivos de una presencia o ausencia de cáncer es inferior al 1 %, y la probabilidad de una determinación precisa de una presencia o ausencia de cáncer es de al menos el 10 %.

14. El método de la reivindicación 1, en donde al menos el 3 % de la pluralidad de oligonucleótidos cebo diferentes no comprenden G (guanina).

15. El método de la reivindicación 1, en donde cada oligonucleótido cebo de la pluralidad de oligonucleótidos cebo diferentes comprende múltiples sitios de unión a sitios de metilación de moléculas de ADNIc convertido derivadas de las al menos 200 regiones genómicas diana, en donde al menos el 80 % de los múltiples sitios de unión comprenden exclusivamente CpG o CpA.

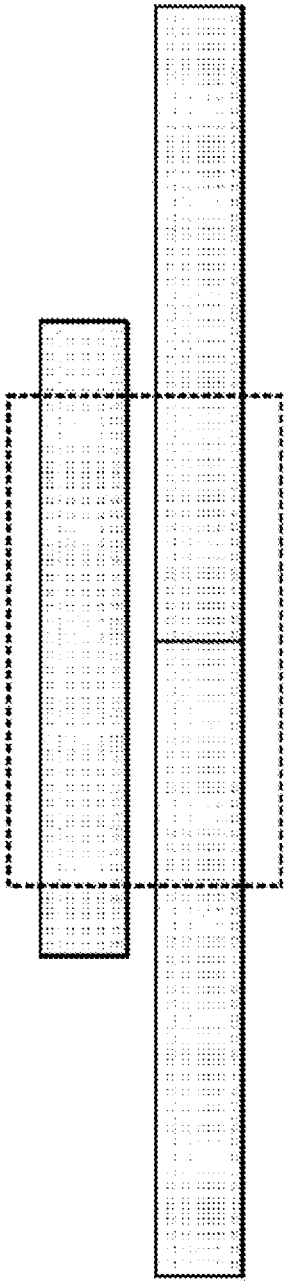

Fig. 1A

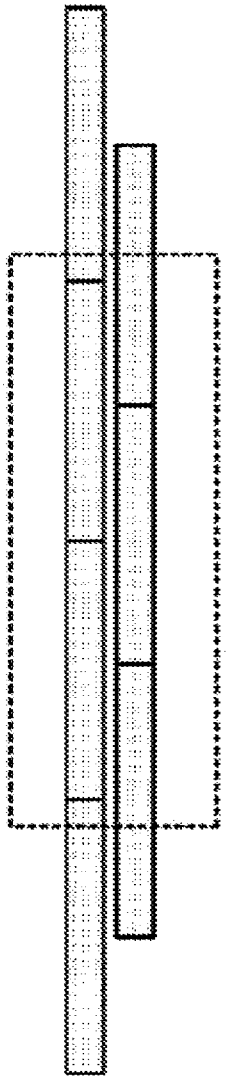

Fig. 1B

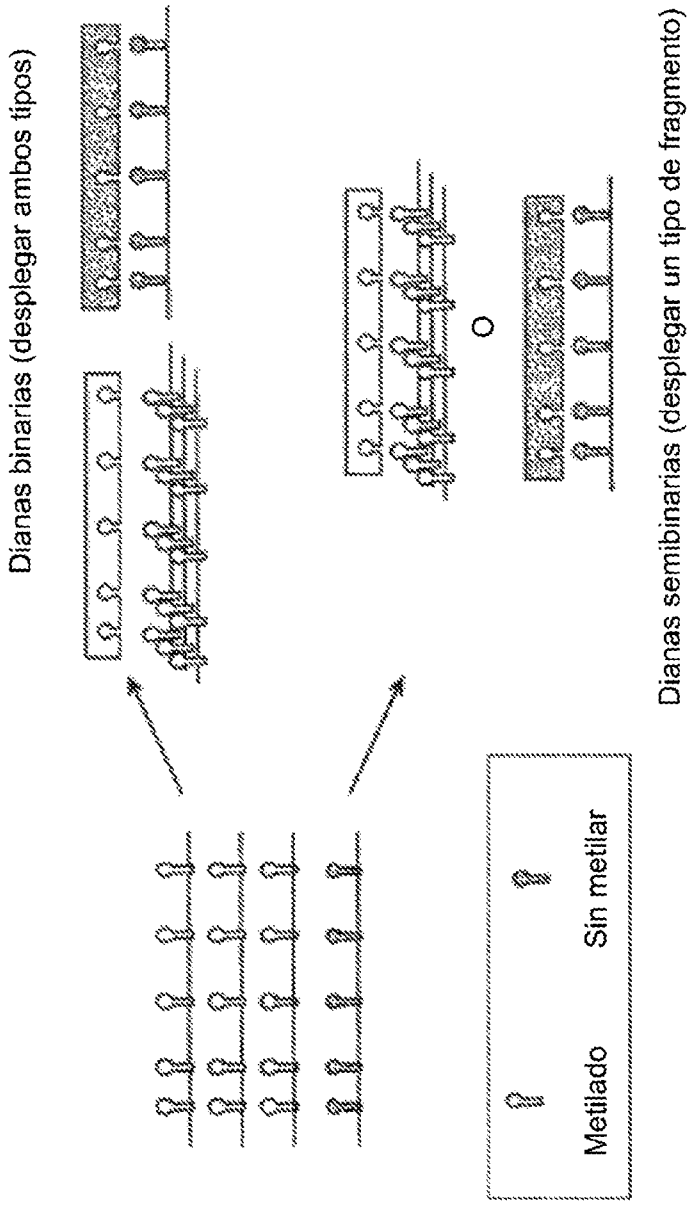
Dianas binarias (desplegar ambos tipos)
O
Dianas semibinarias (desplegar un tipo de fragmento)
Metilado
Sin metilar

Fig. 1C

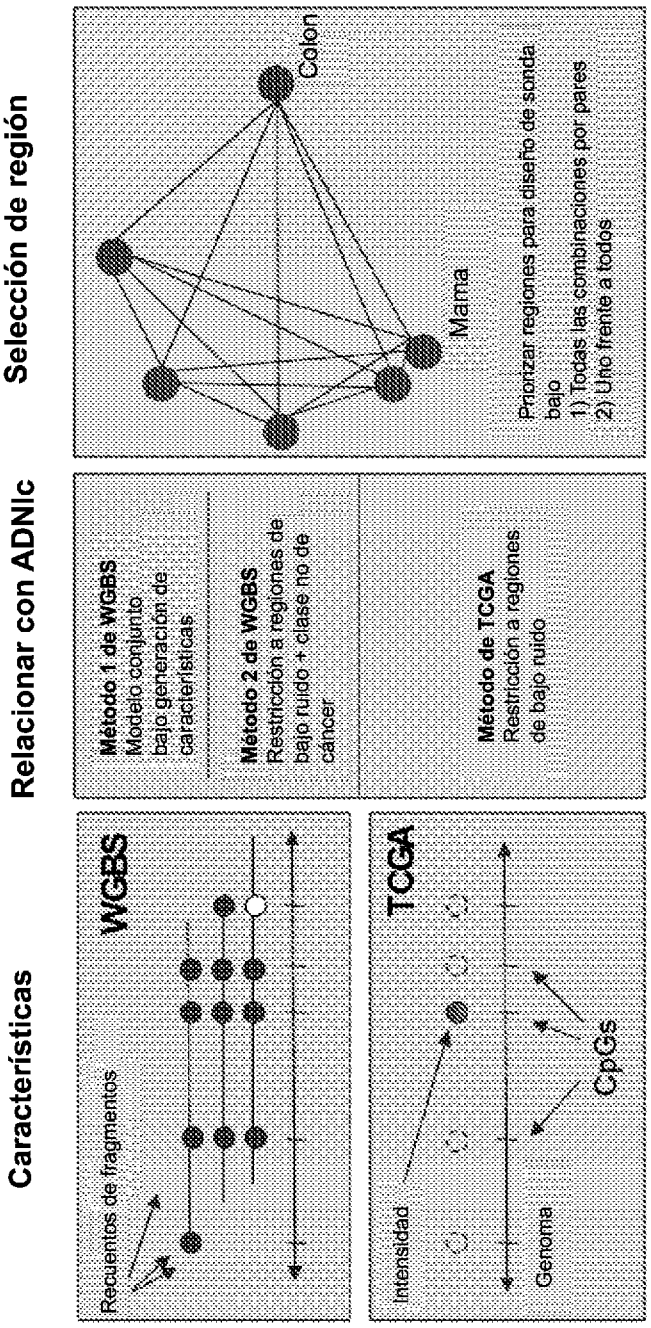
Características
Relacionar con ADNlc
Selección de región
WGBS
Recuentos de fragmentos
TCGA
Intensidad
Genoma
CpGs
Método 1 de WGBS
Modelo conjunto
bajo generación de
características
Método 2 de WGBS
Restricción a regiones de
bajo ruido + clase no de
cáncer
Método de TCGA
Restricción a regiones
de bajo ruido
Colon
Mama
Priorizar regiones para diseño de sonda
bajo
1) Todas las combinaciones por pares
2) Uno frente a todos

Fig. 2

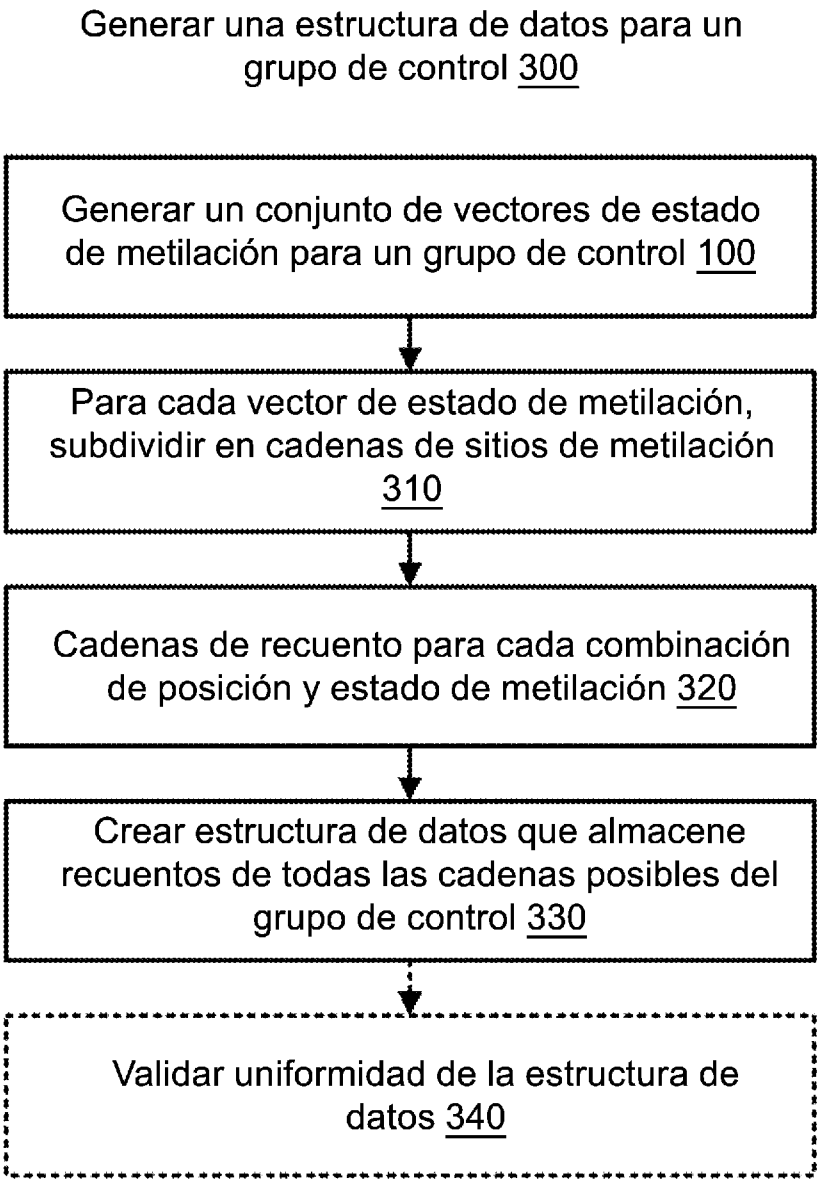
Generar una estructura de datos para un grupo de control 300
Generar un conjunto de vectores de estado de metilación para un grupo de control 100
Para cada vector de estado de metilación, subdividir en cadenas de sitios de metilación 310
Cadenas de recuento para cada combinación de posición y estado de metilación 320
Crear estructura de datos que almacene recuentos de todas las cadenas posibles del grupo de control 330
Validar uniformidad de la estructura de datos 340

Fig. 3A

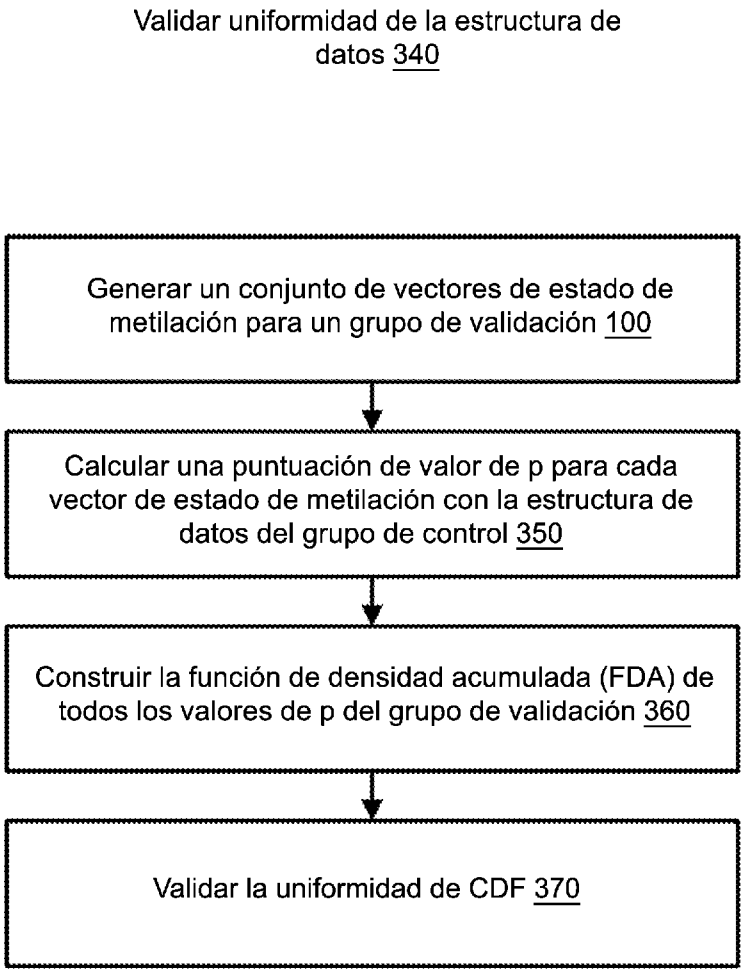
Validar uniformidad de la estructura de datos 340
Generar un conjunto de vectores de estado de metilación para un grupo de validación 100
Calcular una puntuación de valor de p para cada vector de estado de metilación con la estructura de datos del grupo de control 350
Construir la función de densidad acumulada (FDA) de todos los valores de p del grupo de validación 360
Validar la uniformidad de CDF 370

Fig. 3B

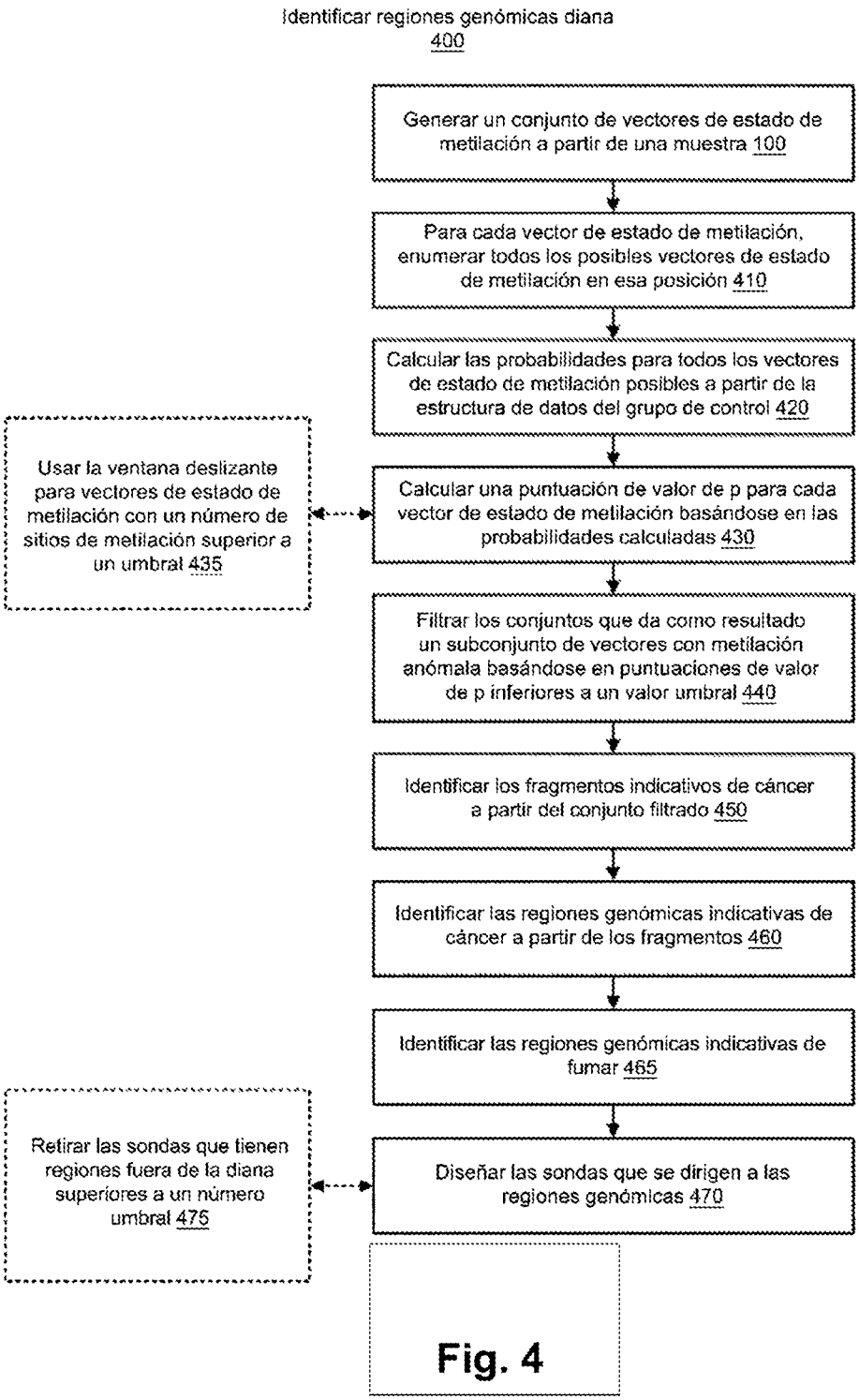
Identificar regiones genómicas diana
400
Generar un conjunto de vectores de estado de metilación a partir de una muestra 100
Para cada vector de estado de metilación, enumerar todos los posibles vectores de estado de metilación en esa posición 410
Calcular las probabilidades para todos los vectores de estado de metilación posibles a partir de la estructura de datos del grupo de control 420
Usar la ventana deslizante para vectores de estado de metilación con un número de sitios de metilación superior a un umbral 435
Calcular una puntuación de valor de p para cada vector de estado de metilación basándose en las probabilidades calculadas 430
Filtrar los conjuntos que da como resultado un subconjunto de vectores con metilación anómala basándose en puntuaciones de valor de p inferiores a un valor umbral 440
Identificar los fragmentos indicativos de cáncer a partir del conjunto filtrado 450
Identificar las regiones genómicas indicativas de cáncer a partir de los fragmentos 460
Identificar las regiones genómicas indicativas de fumar 465
Retirar las sondas que tienen regiones fuera de la diana superiores a un número umbral 475
Diseñar las sondas que se dirigen a las regiones genómicas 470

Fig. 4

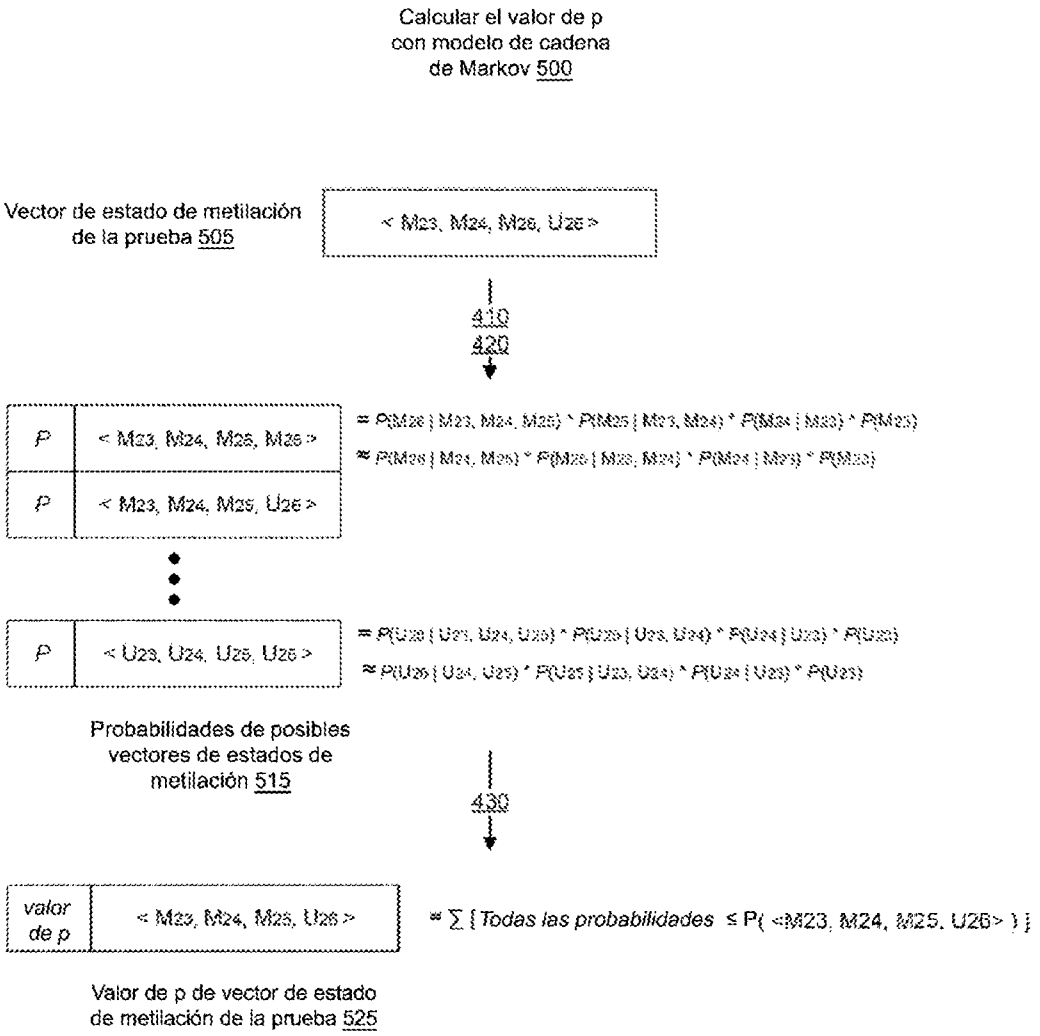
Calcular el valor de p con modelo de cadena de Markov 500
Vector de estado de metilación de la prueba 505
< M23, M24, M25, U26 >
410
420
P
< M23, M24, M25, M26 >
= P(M26 | M23, M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
≈ P(M26 | M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
P
< M23, M24, M25, U26 >
P
< U23, U24, U25, U26 >
= P(U26 | U23, U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
≈ P(U26 | U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
Probabilidades de posibles vectores de estados de metilación 515
430
valor de p
< M23, M24, M25, U26 >
= ∑ { Todas las probabilidades ≤ P(<M23, M24, M25, U26>) }
Valor de p de vector de estado de metilación de la prueba 525

Fig. 5

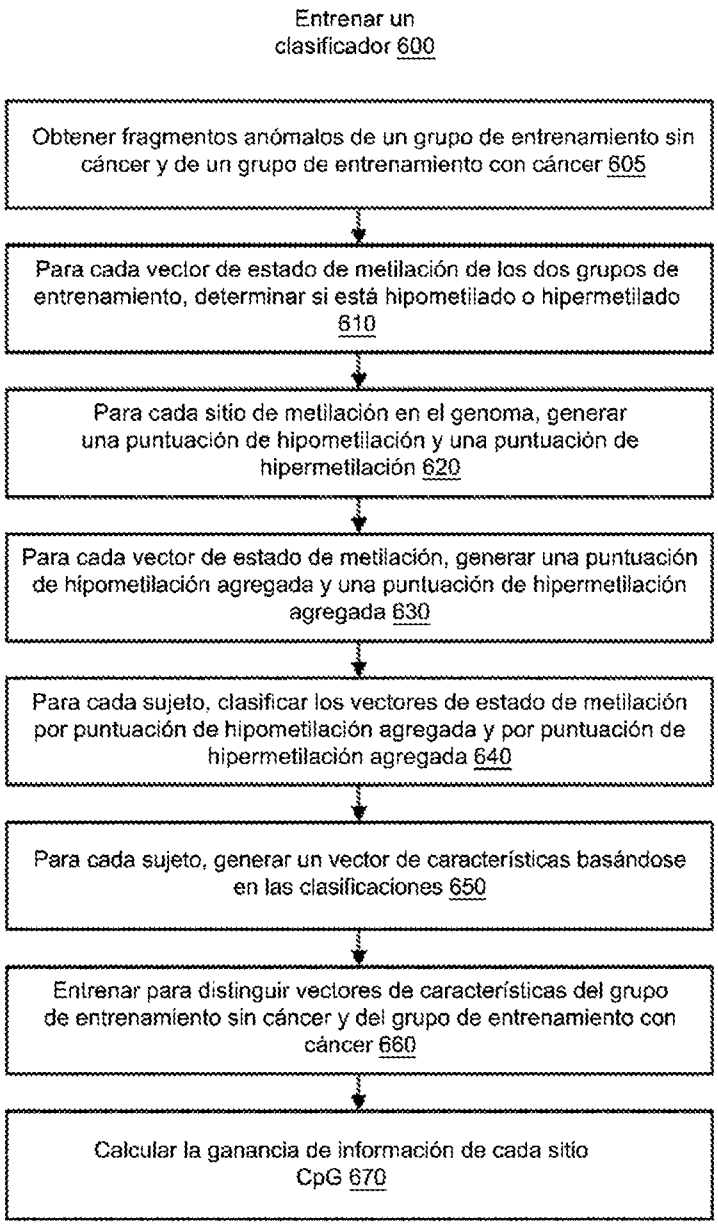
Entrenar un clasificador 600
Obtener fragmentos anómalos de un grupo de entrenamiento sin cáncer y de un grupo de entrenamiento con cáncer 605
Para cada vector de estado de metilación de los dos grupos de entrenamiento, determinar si está hipometilado o hipermetilado 610
Para cada sitio de metilación en el genoma, generar una puntuación de hipometilación y una puntuación de hipermetilación 620
Para cada vector de estado de metilación, generar una puntuación de hipometilación agregada y una puntuación de hipermetilación agregada 630
Para cada sujeto, clasificar los vectores de estado de metilación por puntuación de hipometilación agregada y por puntuación de hipermetilación agregada 640
Para cada sujeto, generar un vector de características basándose en las clasificaciones 650
Entrenar para distinguir vectores de características del grupo de entrenamiento sin cáncer y del grupo de entrenamiento con cáncer 660
Calcular la ganancia de información de cada sitio CpG 670

Fig. 6A

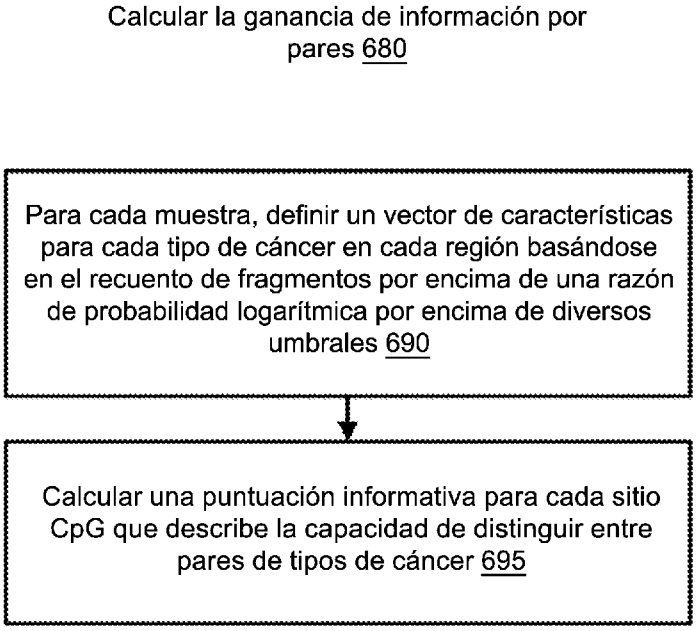
Calcular la ganancia de información por pares 680
Para cada muestra, definir un vector de características para cada tipo de cáncer en cada región basándose en el recuento de fragmentos por encima de una razón de probabilidad logarítmica por encima de diversos umbrales 690
Calcular una puntuación informativa para cada sitio CpG que describe la capacidad de distinguir entre pares de tipos de cáncer 695

Fig. 6B

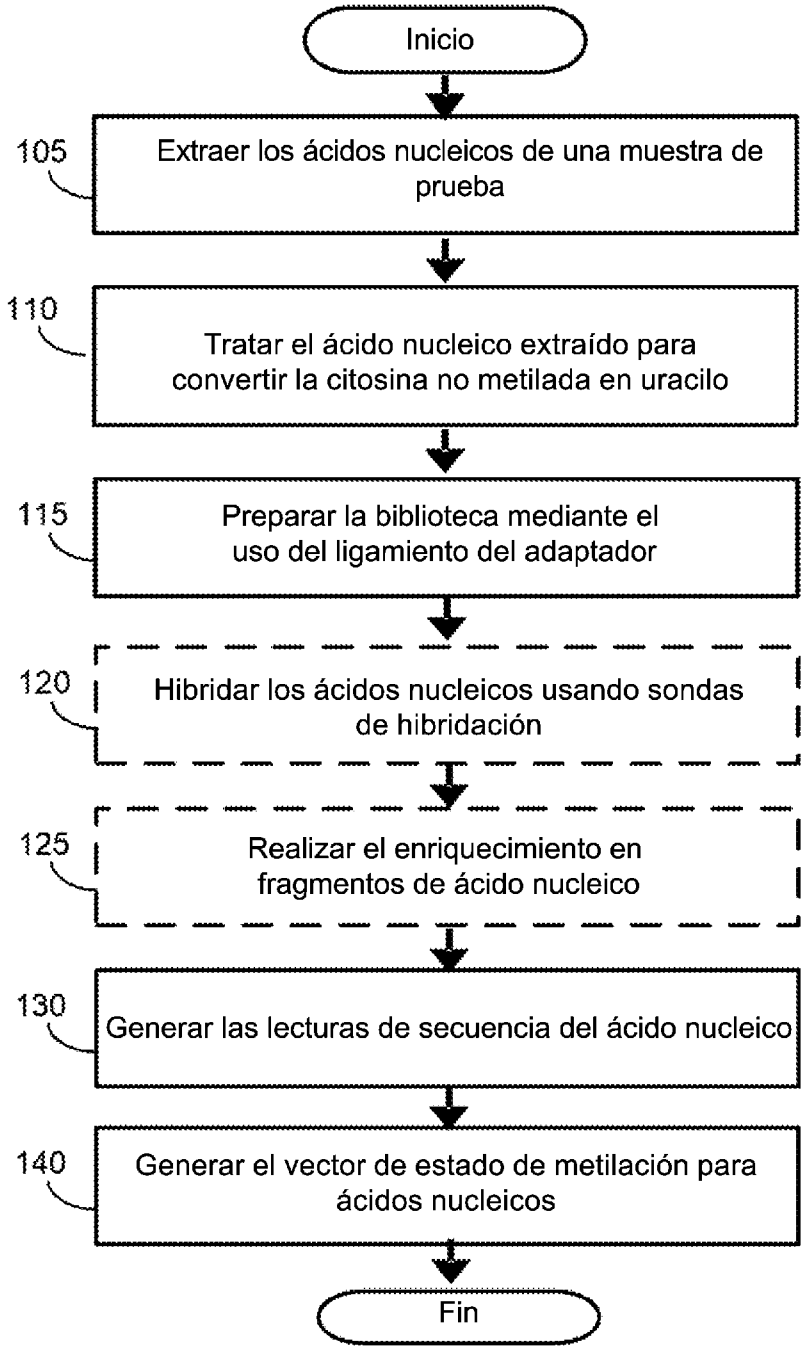
Inicio
105
Extraer los ácidos nucleicos de una muestra de prueba
110
Tratar el ácido nucleico extraído para convertir la citosina no metilada en uracilo
115
Preparar la biblioteca mediante el uso del ligamiento del adaptador
120
Hibridar los ácidos nucleicos usando sondas de hibridación
125
Realizar el enriquecimiento en fragmentos de ácido nucleico
130
Generar las lecturas de secuencia del ácido nucleico
140
Generar el vector de estado de metilación para ácidos nucleicos
Fin

Fig. 7A

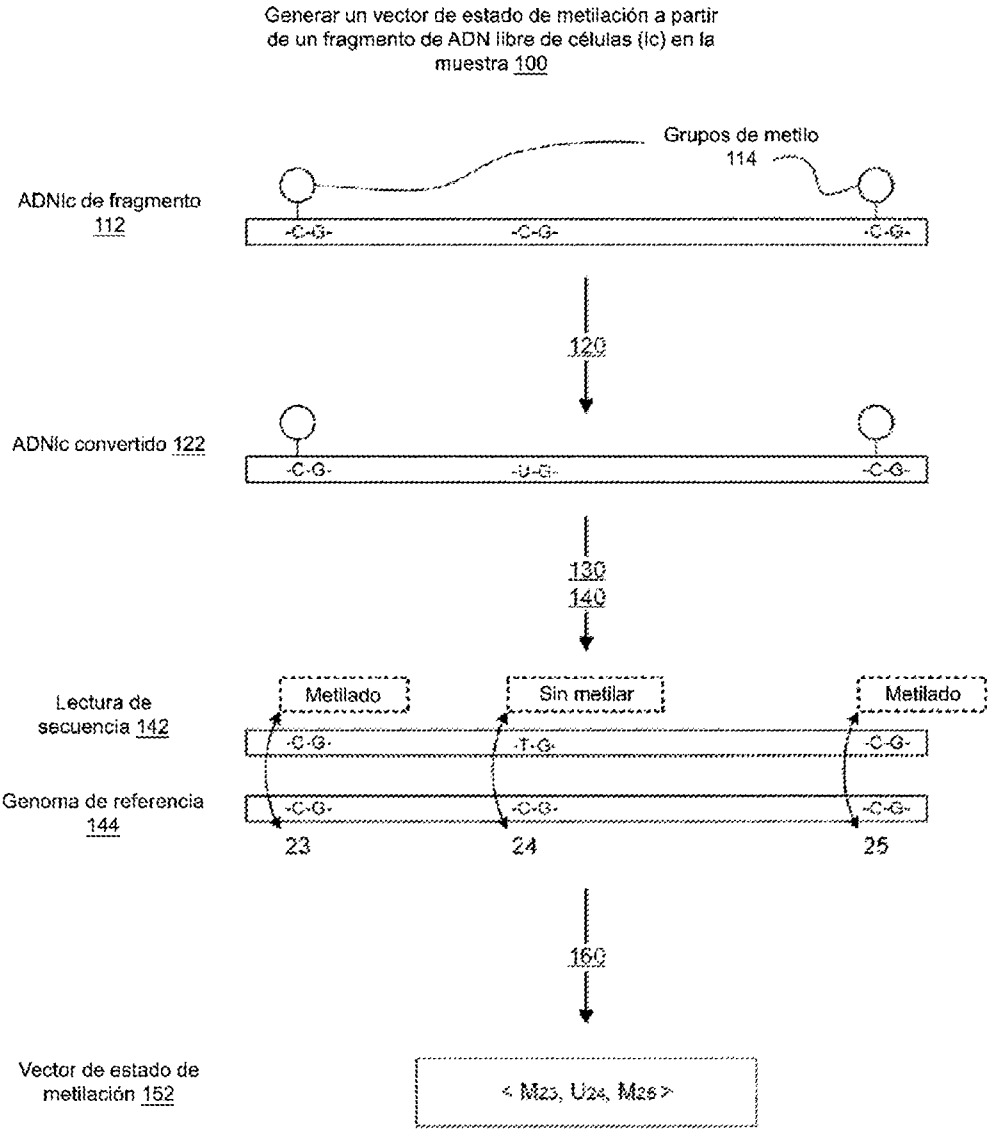
Generar un vector de estado de metilación a partir de un fragmento de ADN libre de células (lc) en la muestra 100
Grupos de metilo 114
ADNlc de fragmento 112
-C-G-
-C-G-
-C-G-
120
ADNlc convertido 122
-C-G-
-U-G-
-C-G-
130
140
Lectura de secuencia 142
Metilado
Sin metilar
Metilado
-C-G-
-T-G-
-C-G-
Genoma de referencia 144
-C-G-
-C-G-
-C-G-
23
24
25
160
Vector de estado de metilación 152
< M23, U24, M25 >

Fig. 7B

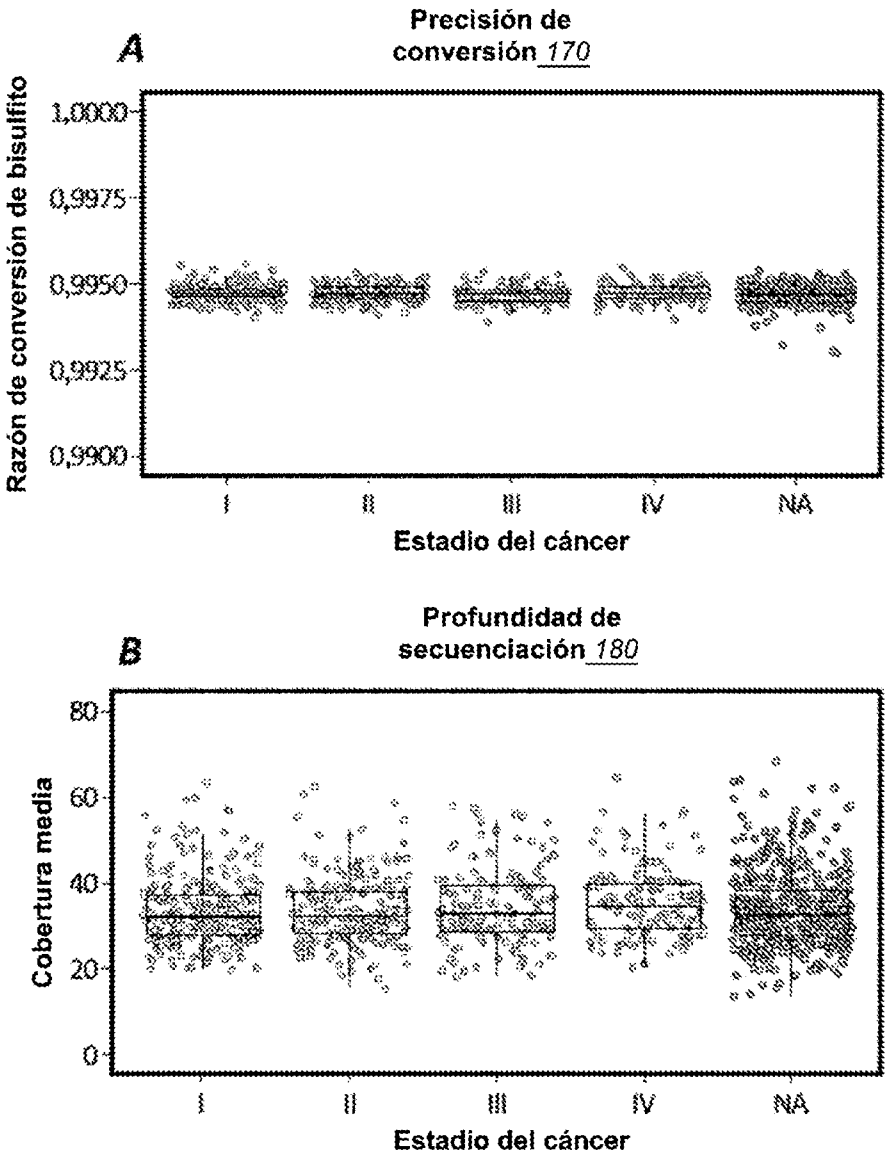
A
Precisión de conversión 170
Razón de conversión de bisulfito
1,0000
0,9975
0,9950
0,9925
0,9900
I
II
III
IV
NA
Estadio del cáncer
B
Profundidad de secuenciación 180
Cobertura media
80
60
40
20
0
I
II
III
IV
NA
Estadio del cáncer

Fig. 8

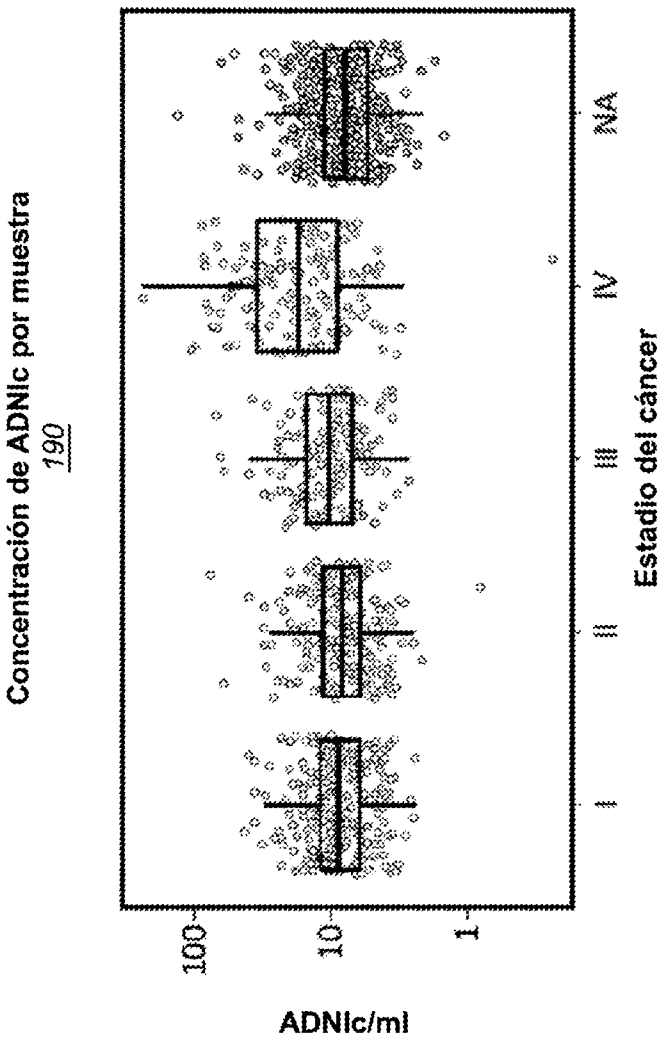
Concentración de ADNlc por muestra
190
100
10
1
ADNlc/ml
I
II
III
IV
NA
Estadio del cáncer

Fig. 9

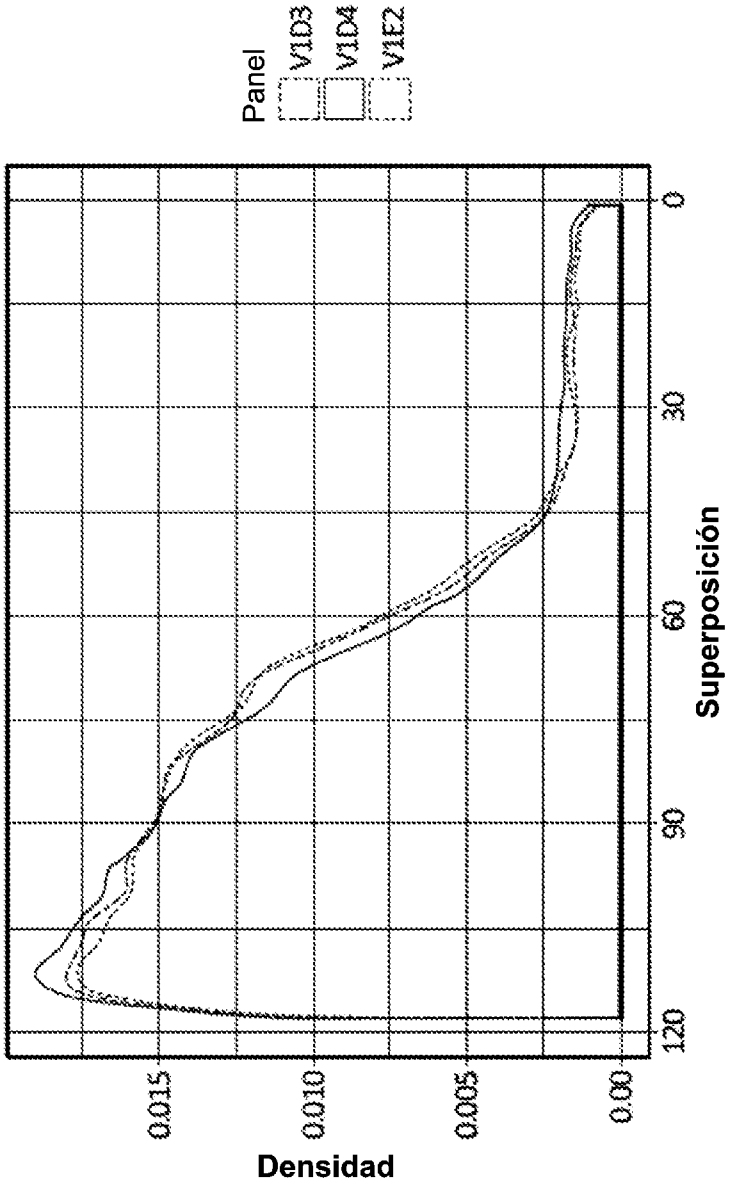
Panel
V1D3
V1D4
V1E2
0.015
0.010
0.005
0.00
Densidad
0
30
60
90
120
Superposición

Fig. 10

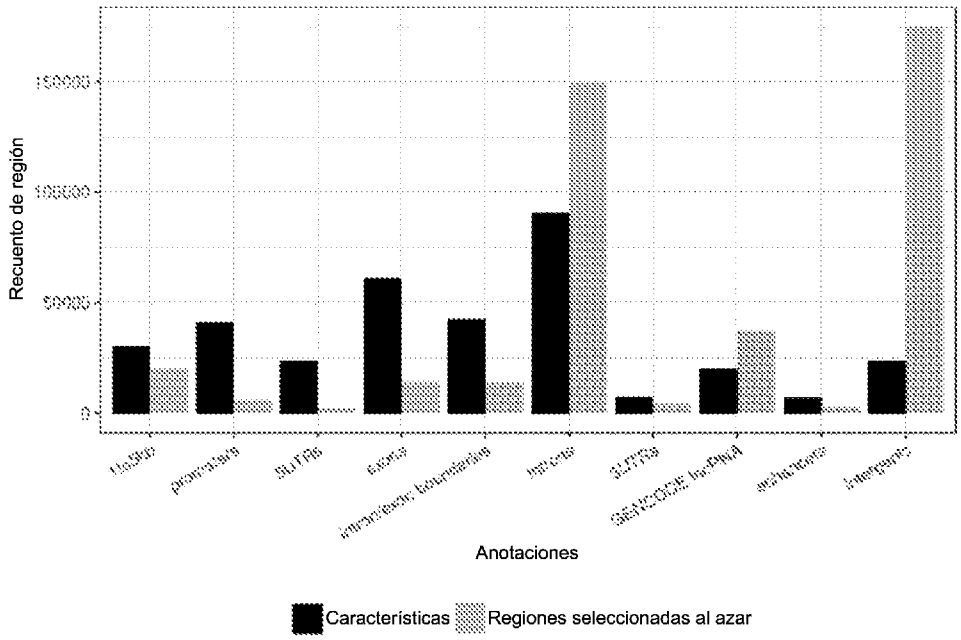

Fig. 11A

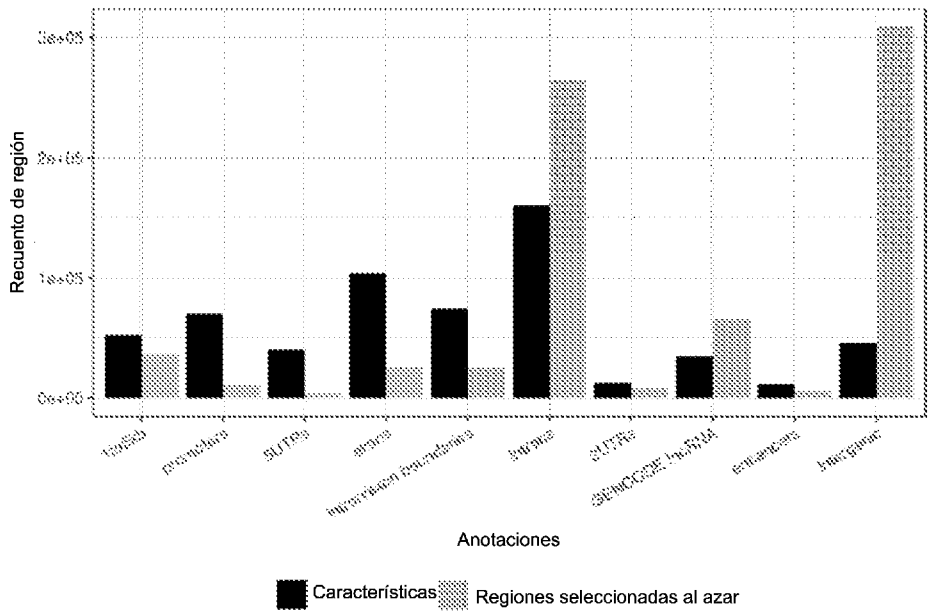

Fig. 11B

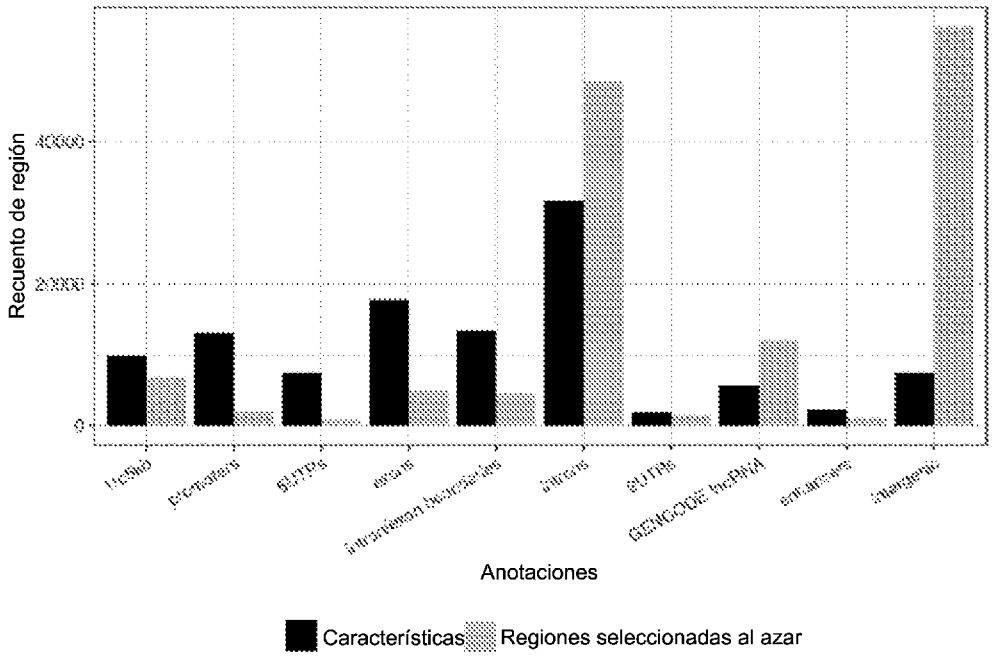
Recuento de región
4000
2000
0
Anotaciones
Características
Regiones seleccionadas al azar

Fig. 11C

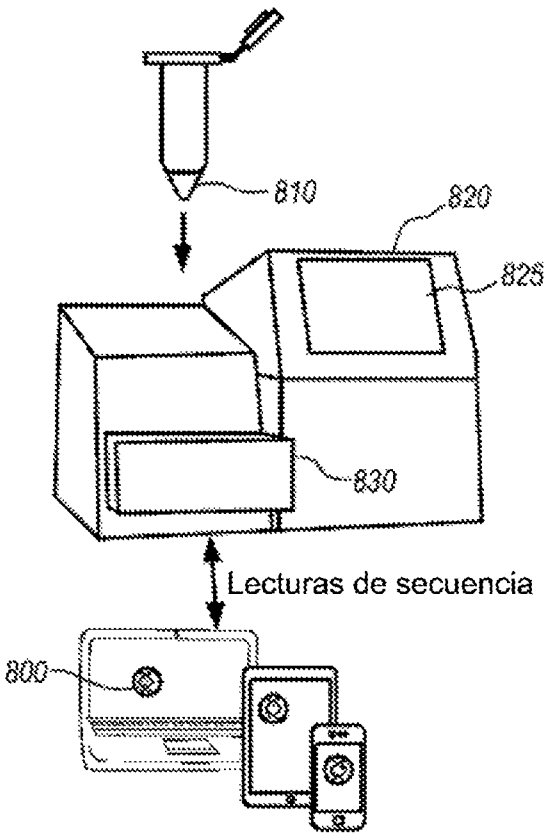

Fig. 12A

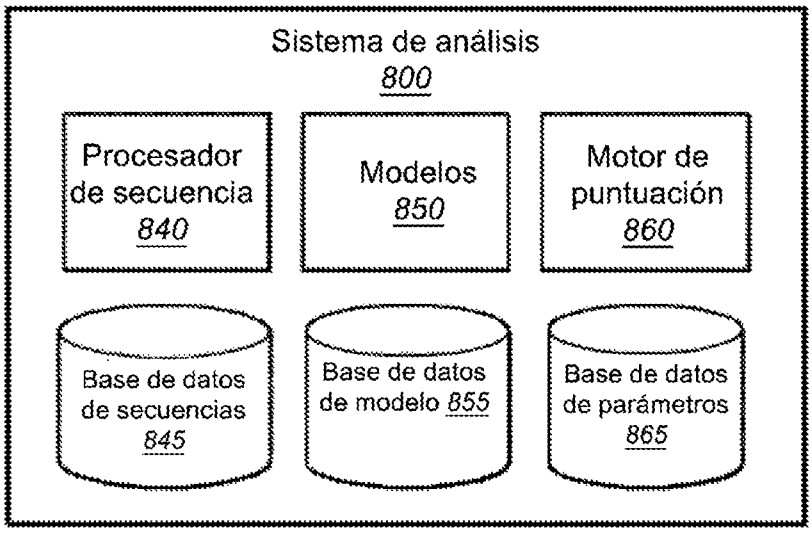

Fig. 12B

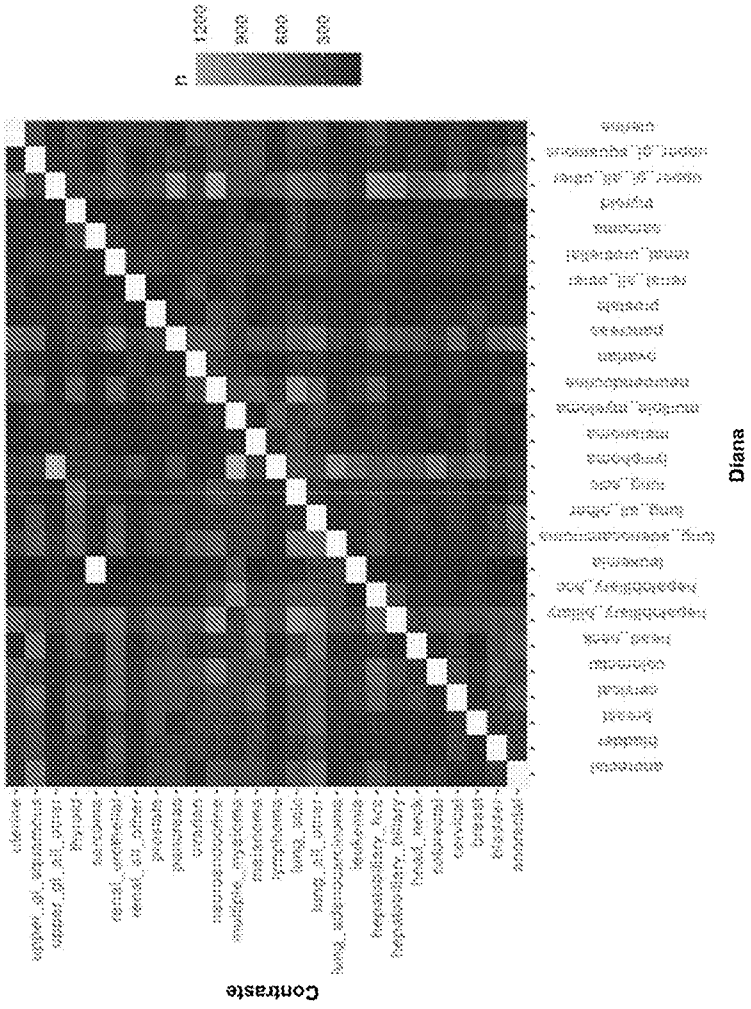

Fig. 13

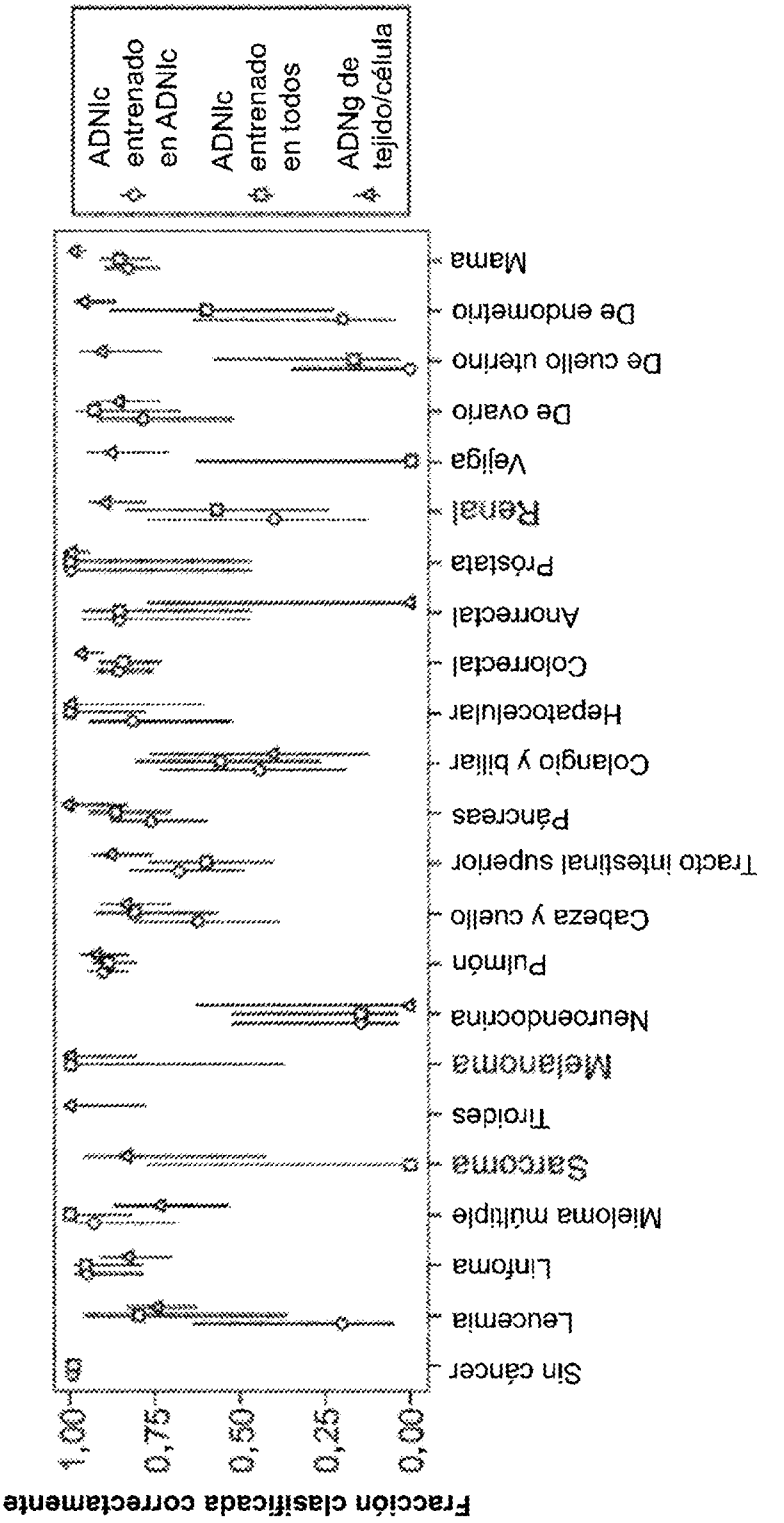
ADNlc entrenado en ADNlc
ADNlc entrenado en todos
ADNg de tejido/célula
Sin cáncer
Leucemia
Linfoma
Mieloma múltiple
Sarcoma
Tiroides
Melanoma
Neuroendocrina
Pulmón
Cabeza y cuello
Tracto intestinal superior
Páncreas
Colangio y biliar
Hepatocelular
Colorrectal
Anorrectal
Próstata
Renal
Vejiga
De ovario
De cuello uterino
De endometrio
Mama
1,00
0,75
0,50
0,25
0,00
Fracción clasificada correctamente

Fig. 14

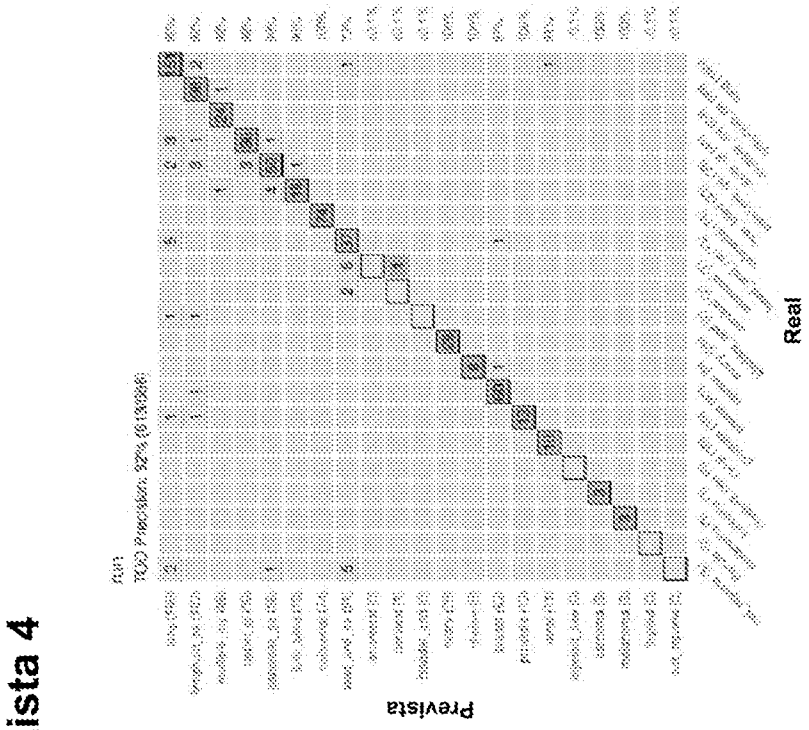

Fig. 15B

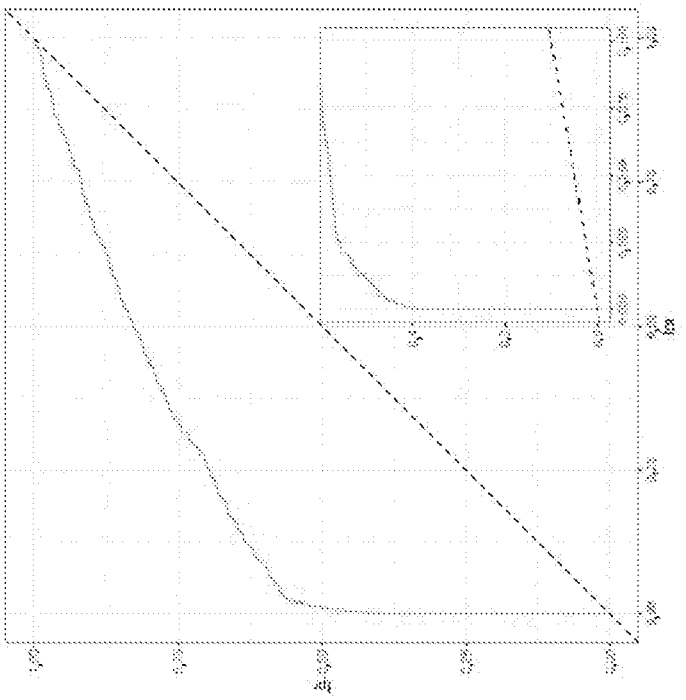

Fig. 15A

Lista 5

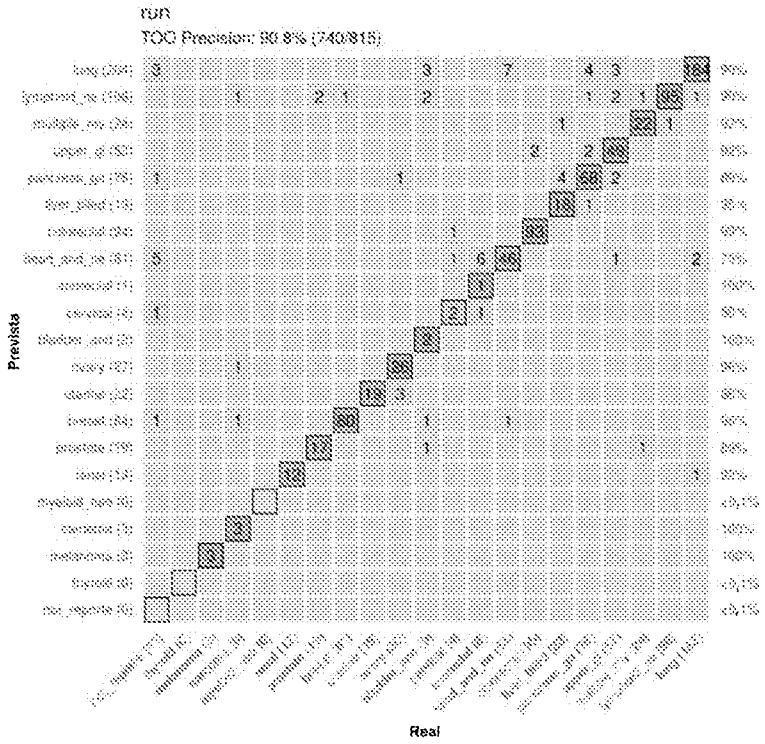

Fig. 16B

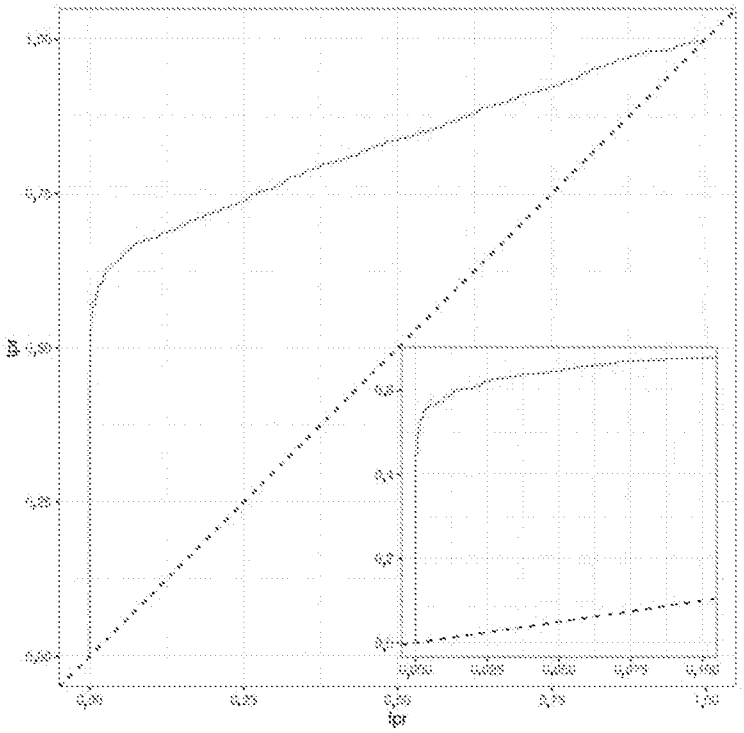

Fig. 16A

Lista 6

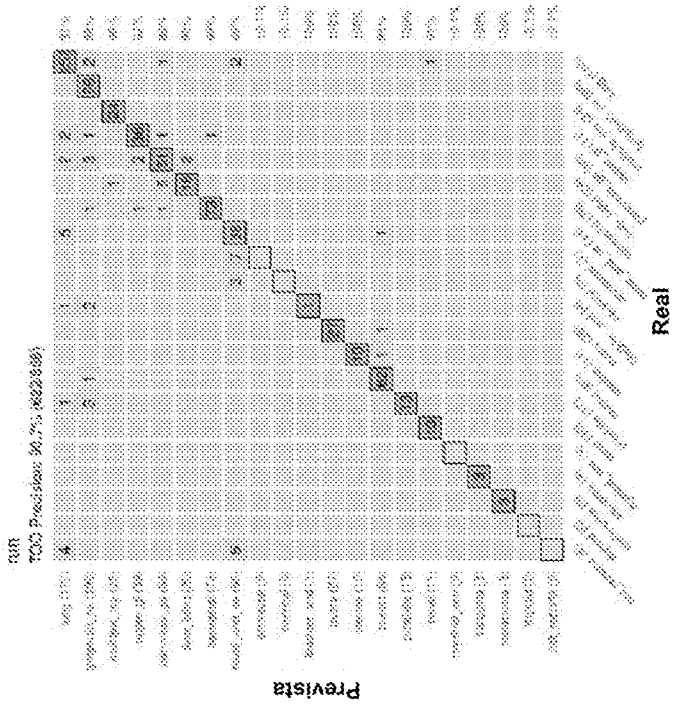

Fig. 17B

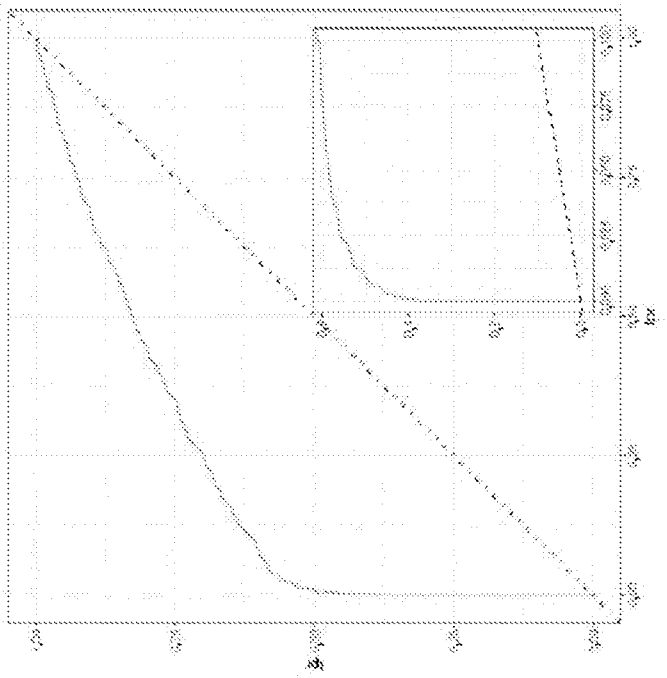

Fig. 17A

Lista 7

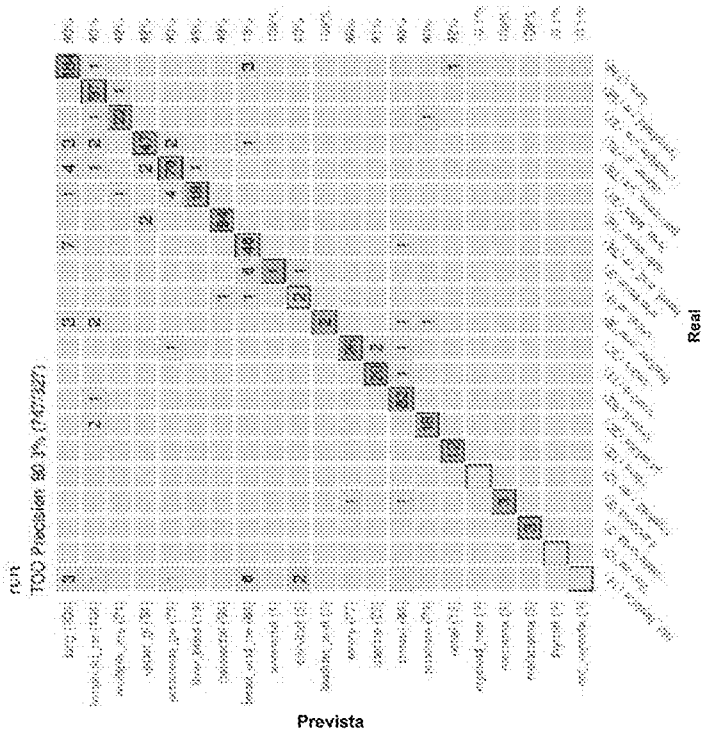

Fig. 18B

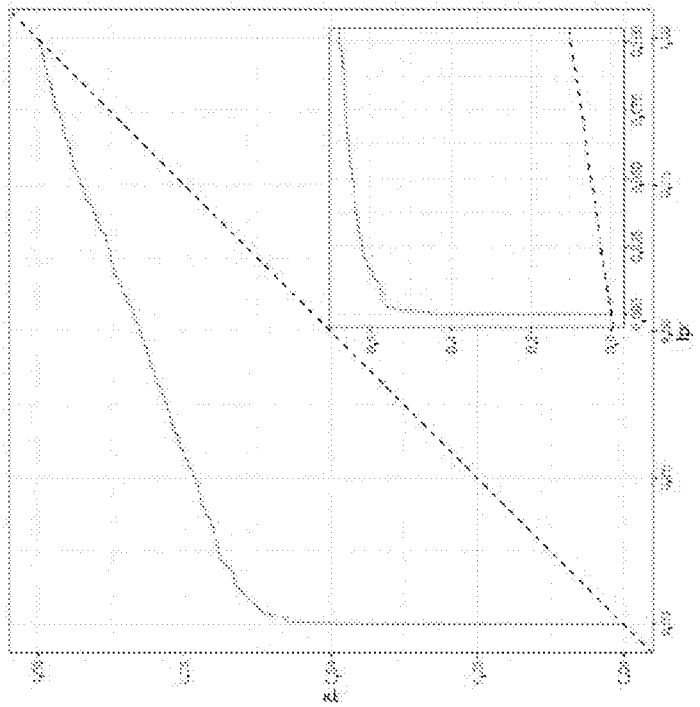

Fig. 18A

Lista 8

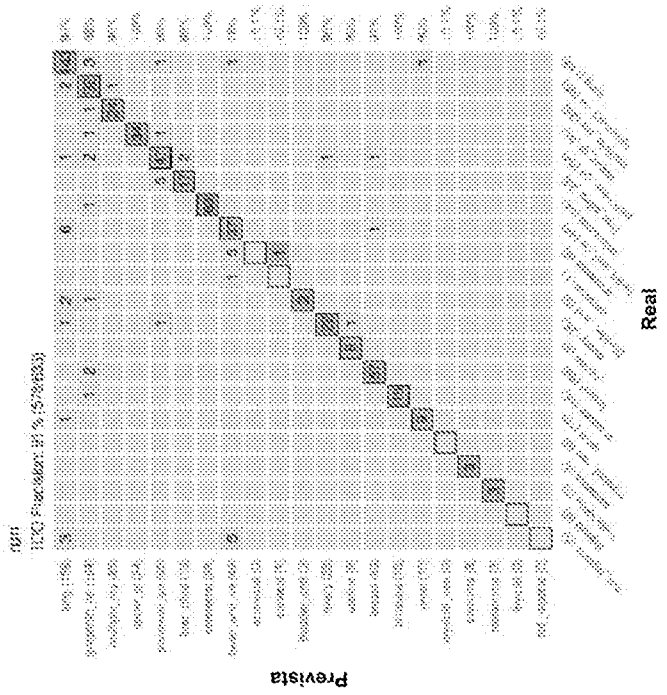

Fig. 19B

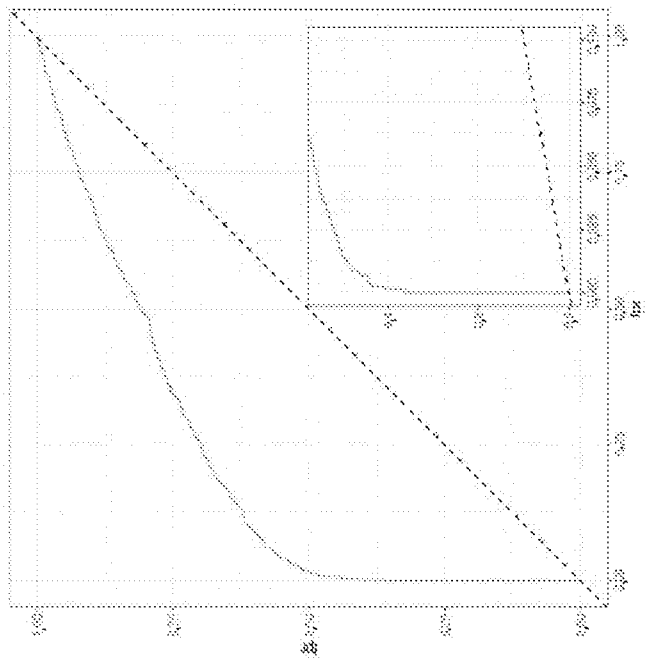

Fig. 19A

Lista 9

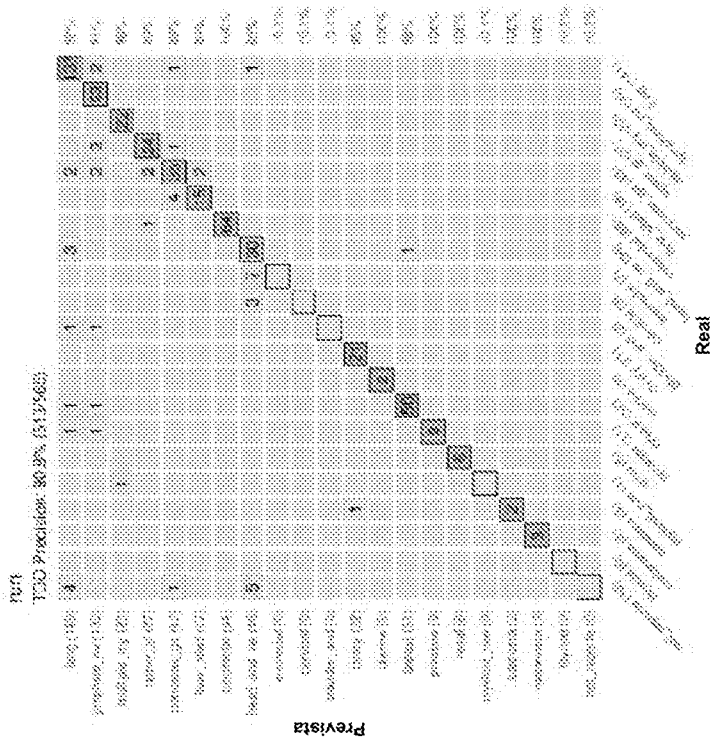

Fig. 20B

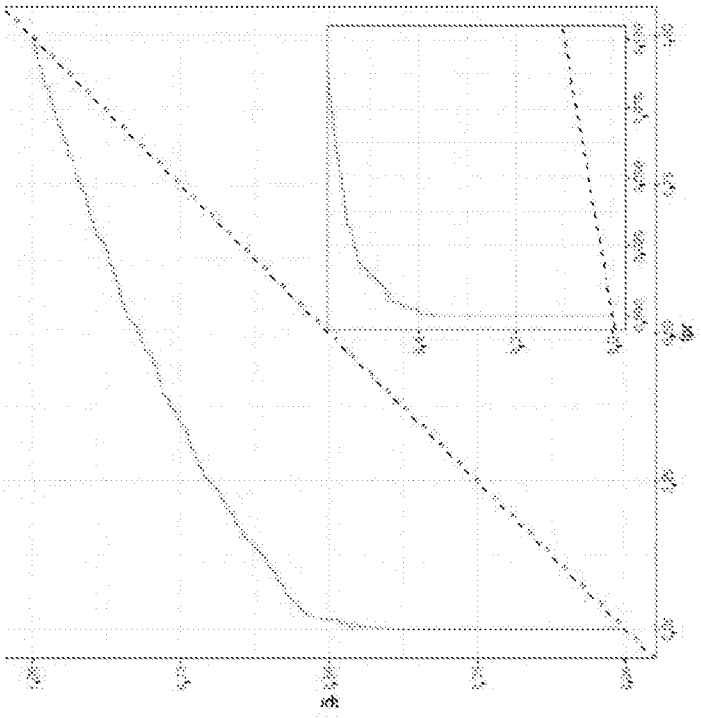

Fig. 20A

Lista 10

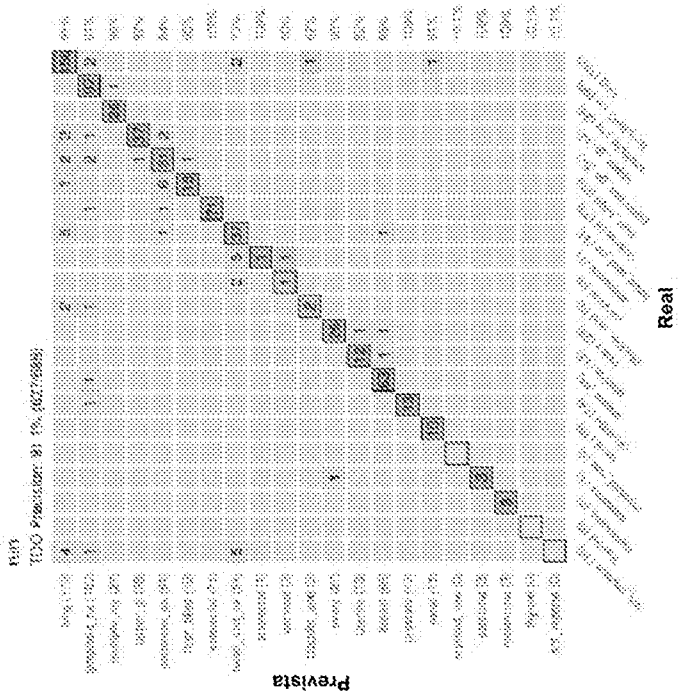

Fig. 21B

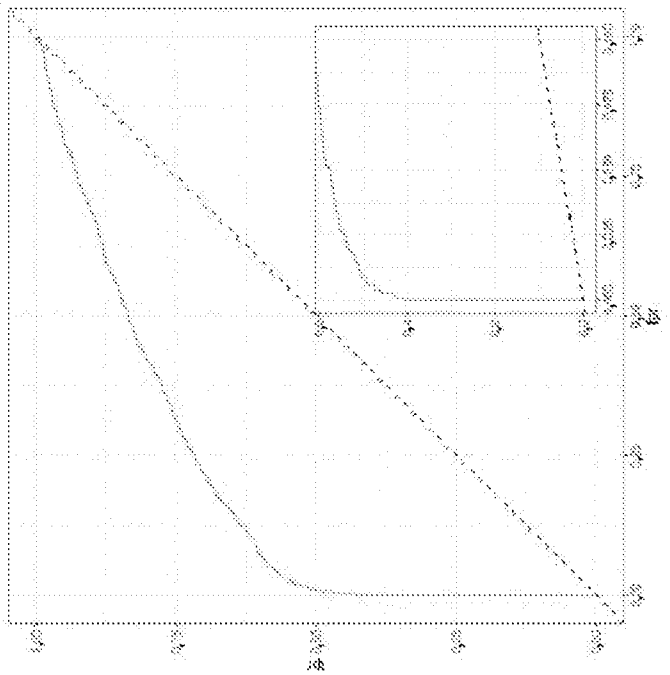

Fig. 21A

Lista 11

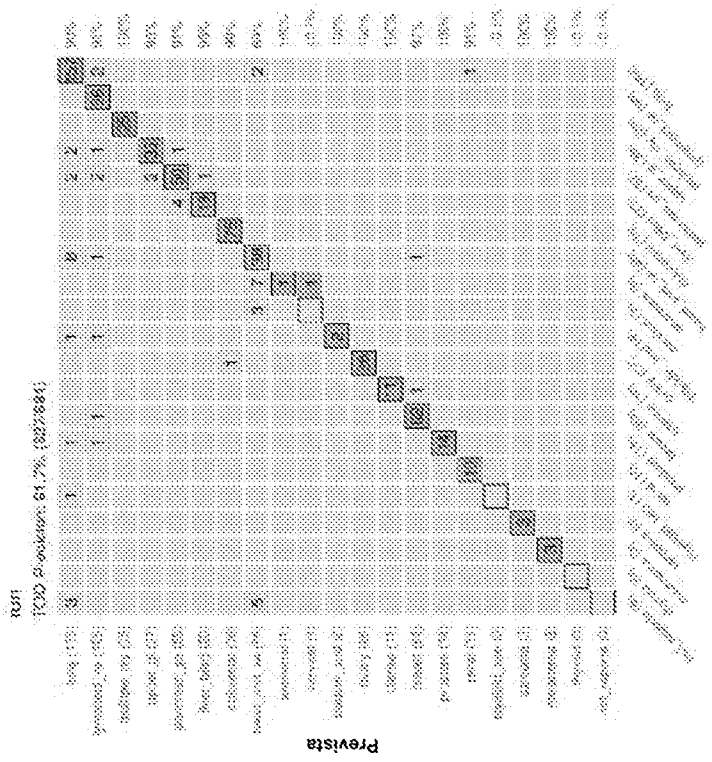

Fig. 22B

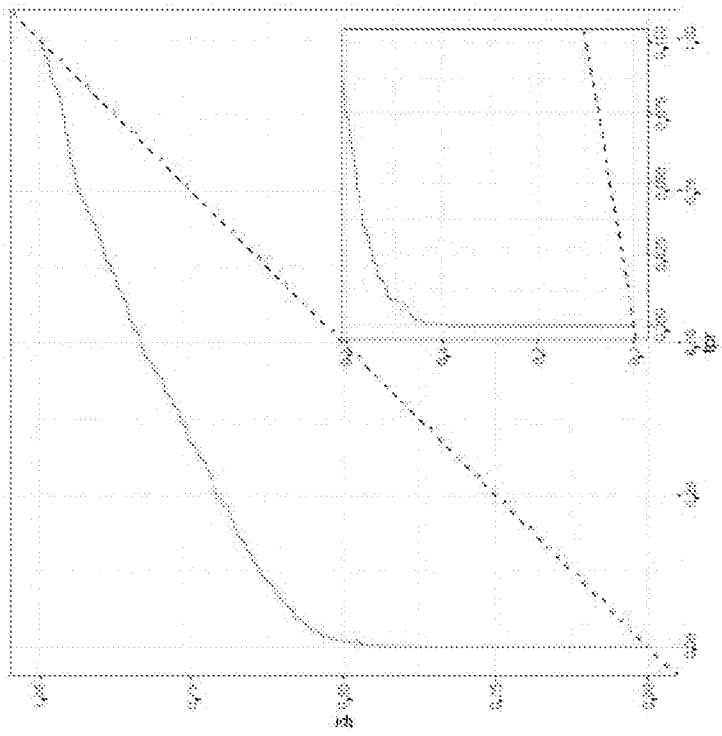

Fig. 22A

Lista 12

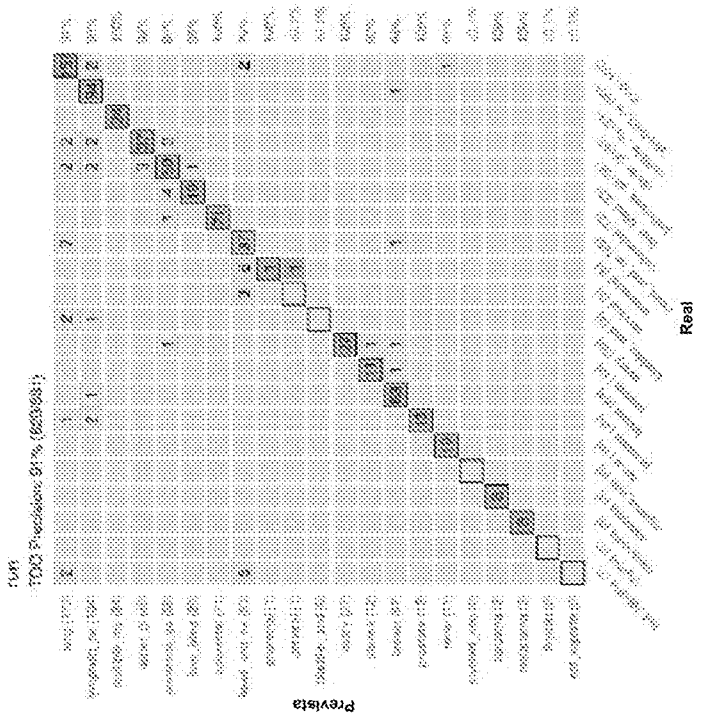

Fig. 23B

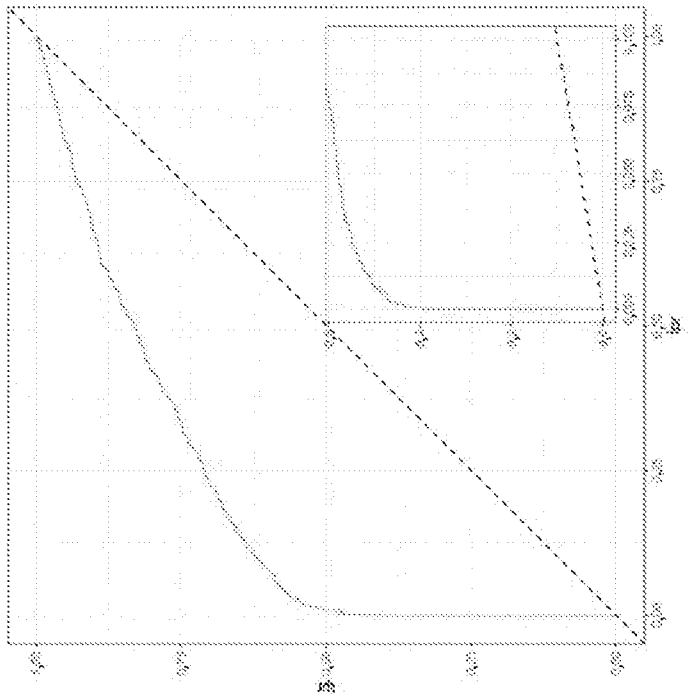

Fig. 23A

Lista 13

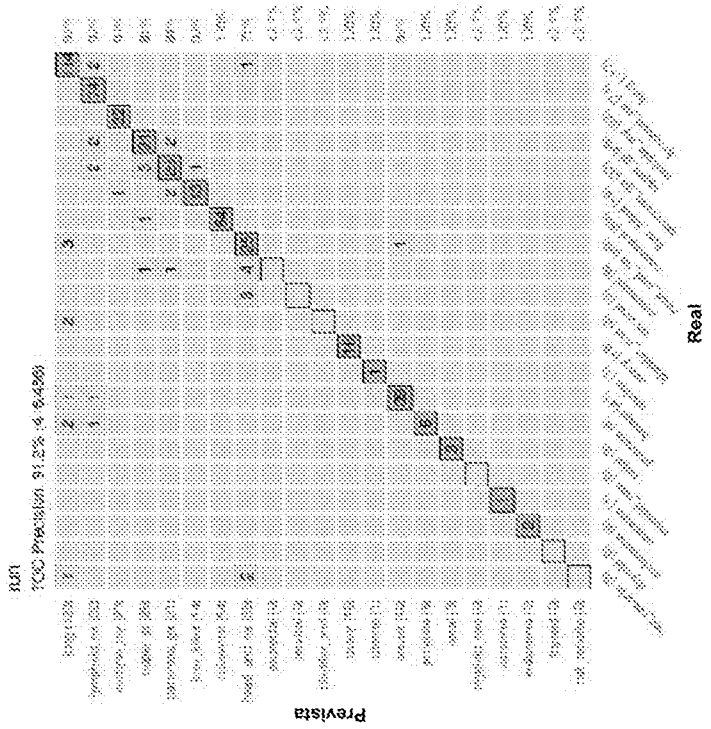

Fig. 24B

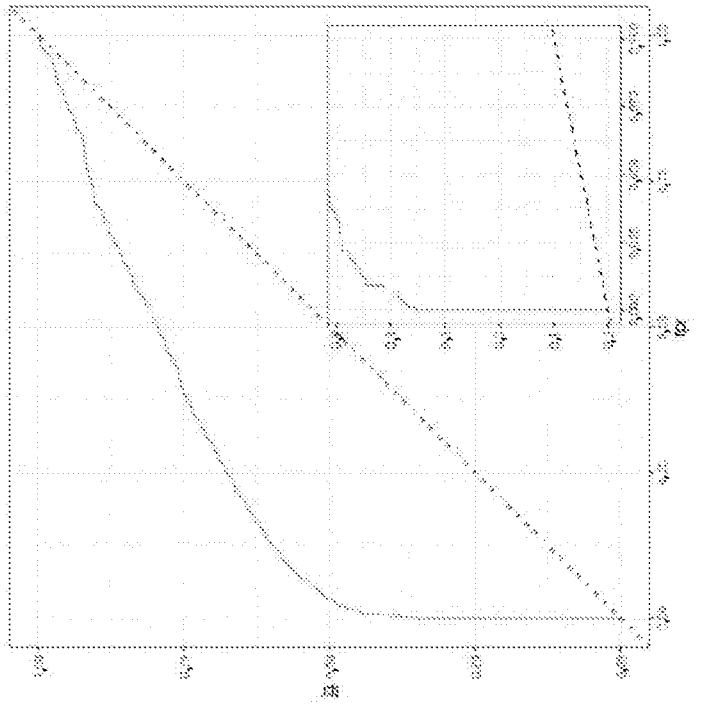

Fig. 24A

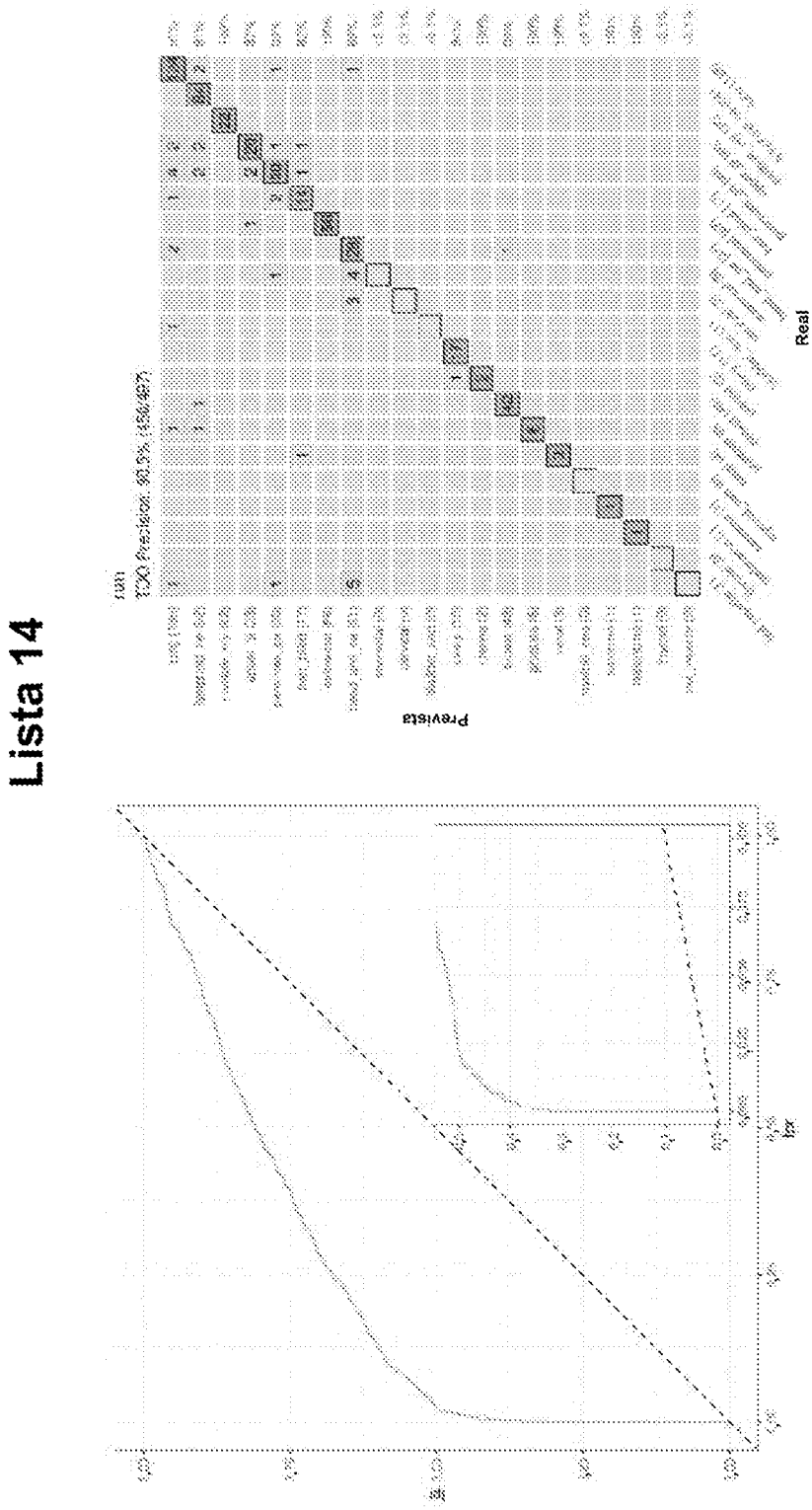

Fig. 25B

Fig. 25A

Lista 15

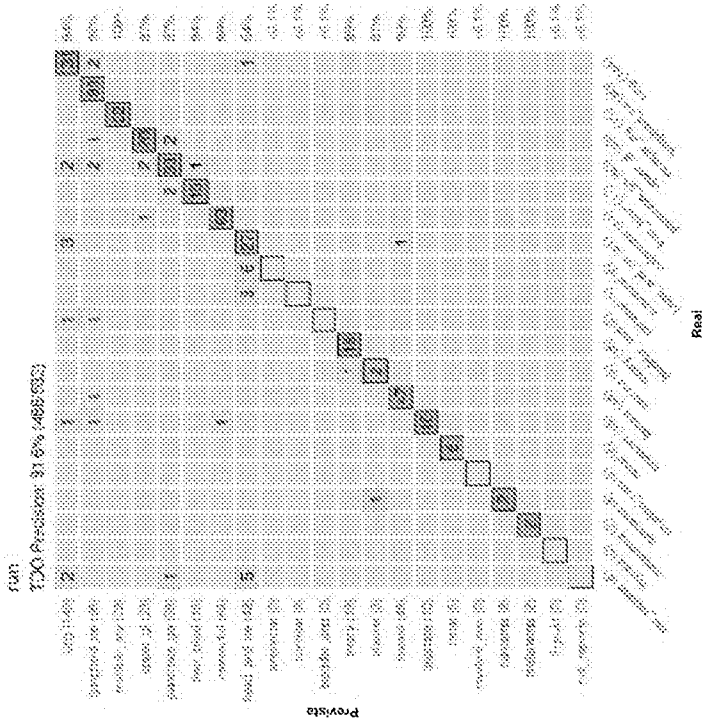

Fig. 26B

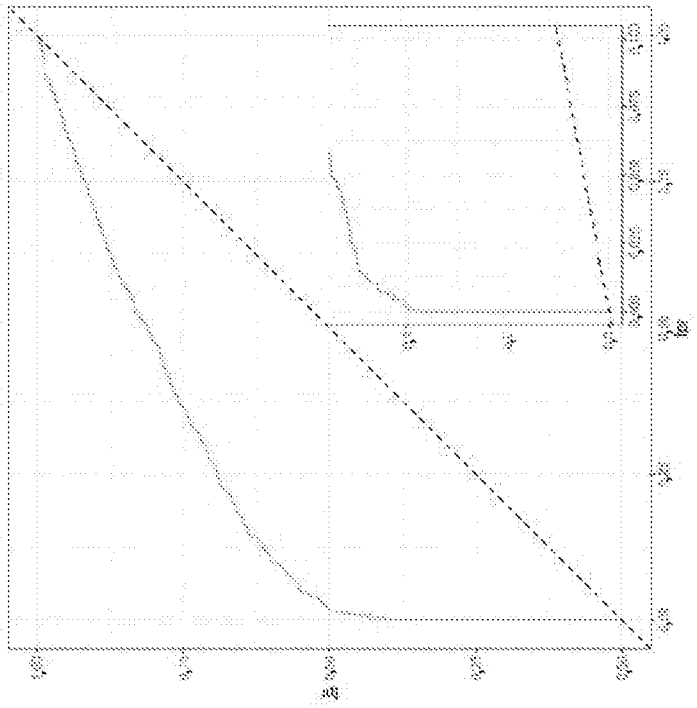

Fig. 26A

Lista 16

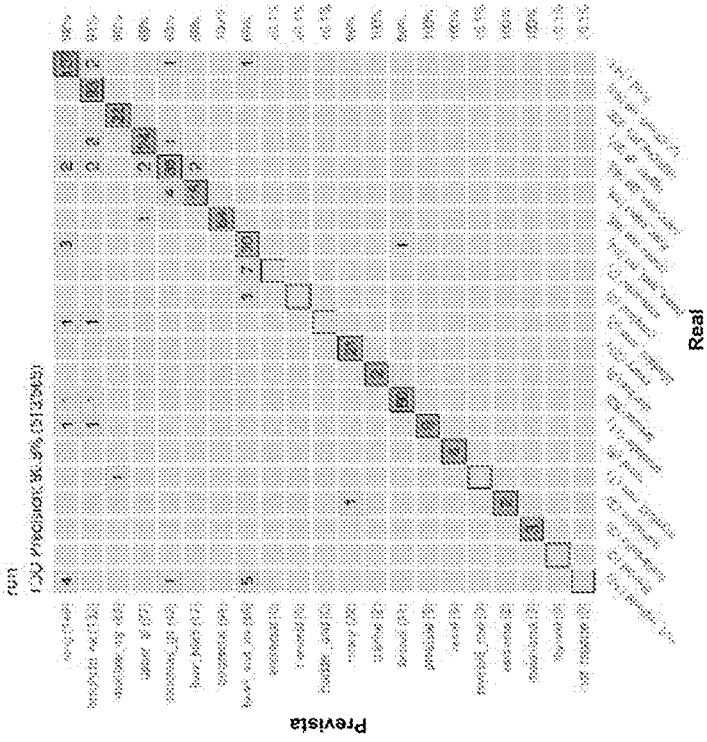

Fig. 27B

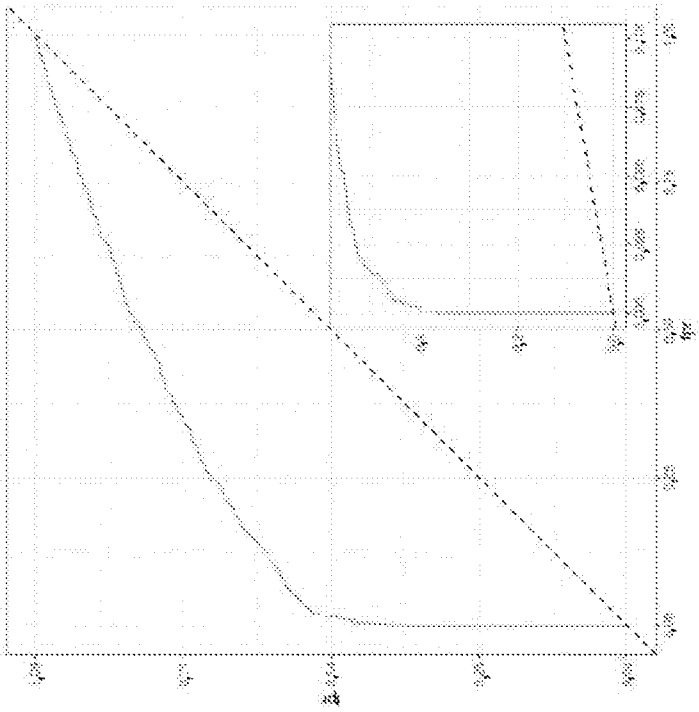

Fig. 27A

El 10 % aleatorio de la lista 12

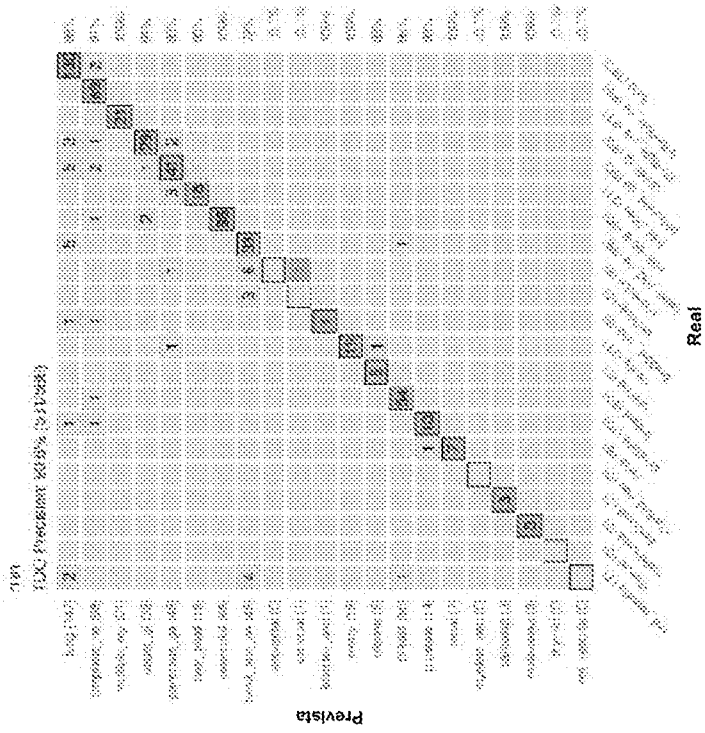

Fig. 28B

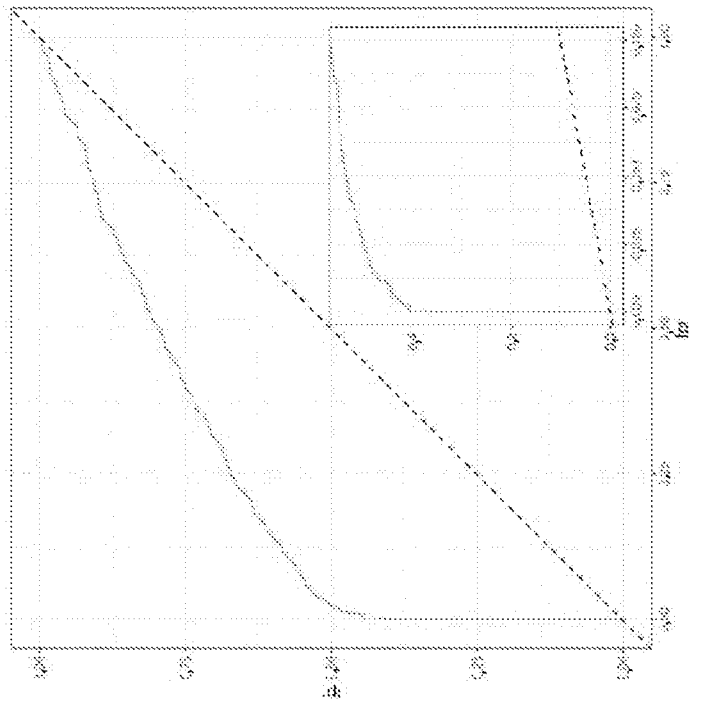

Fig. 28A

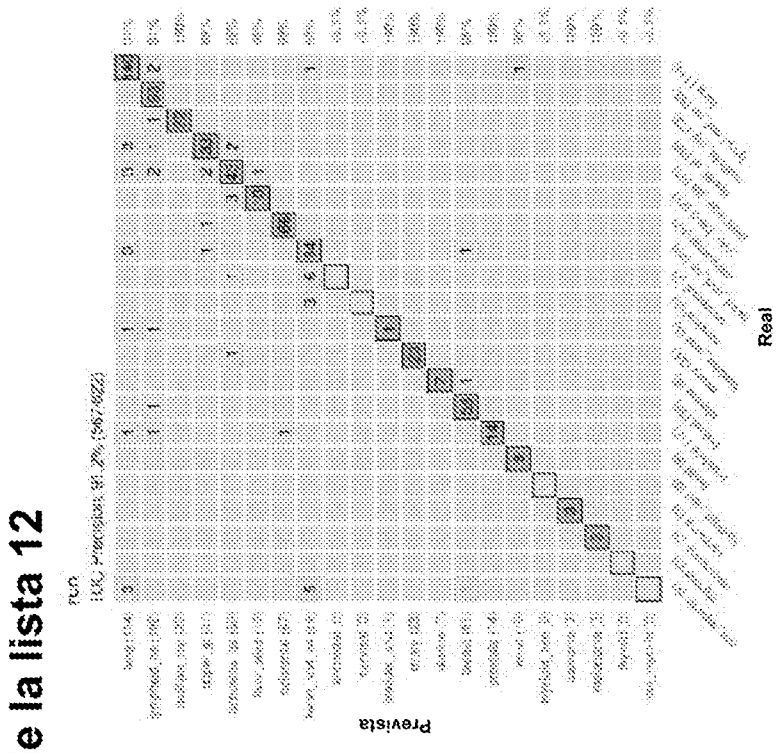

Fig. 29B

El 25 % aleatorio de la lista 12

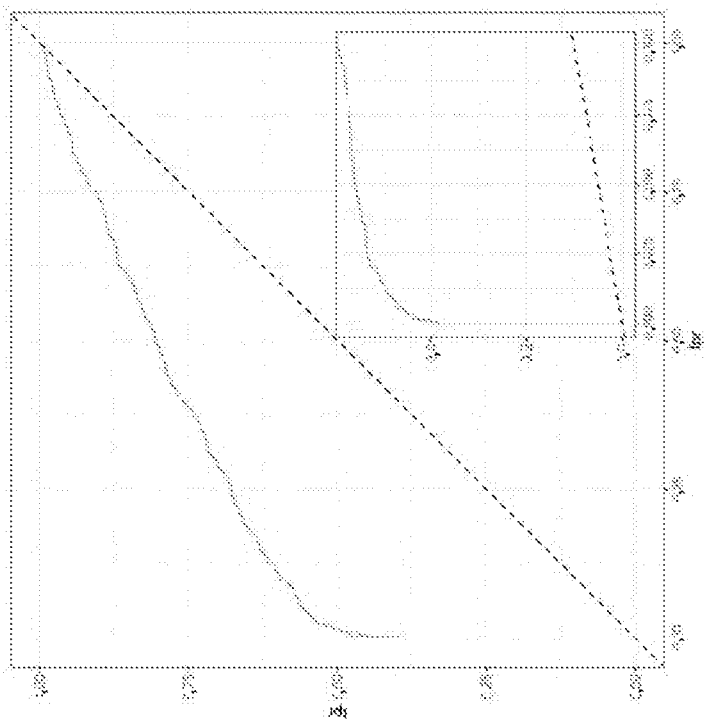

Fig. 29A

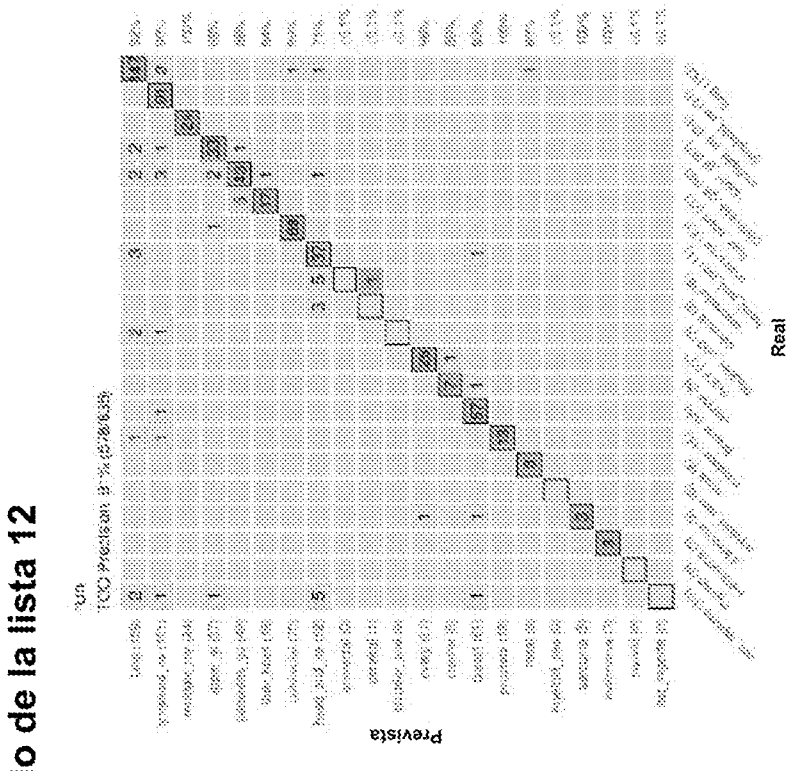

Fig. 30B

El 50 % aleatorio de la lista 12

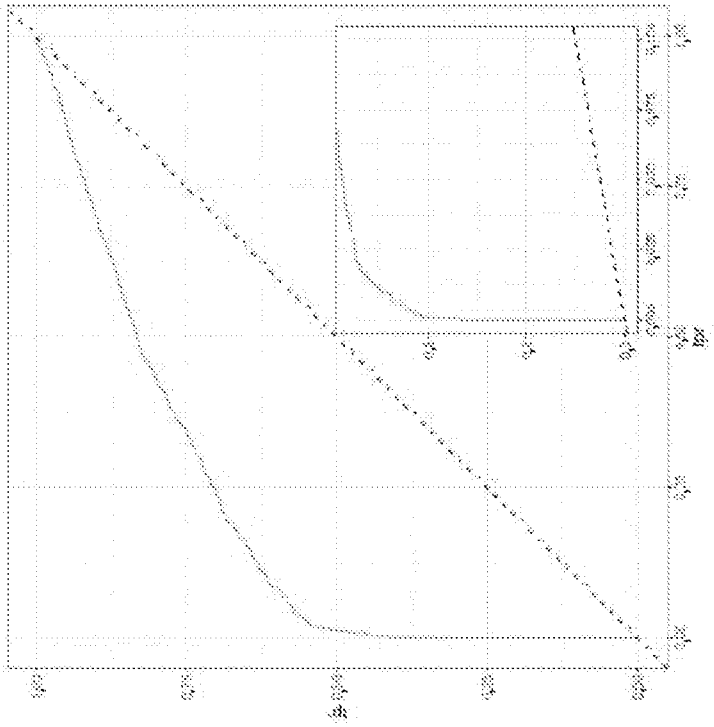

Fig. 30A