(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 991 960**

(51) Int. Cl.:

G16B 20/20 (2009.01)
C12Q 1/6869 (2008.01)
C12Q 1/6886 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **16.05.2018 PCT/US2018/033038**

(87) Fecha y número de publicación internacional: **22.11.2018 WO18213498**

(96) Fecha de presentación y número de la solicitud europea: **16.05.2018 E 18802961 (5)**

(97) Fecha y número de publicación de la concesión europea: **25.09.2024 EP 3625341**

(54) Título: **Identificación del origen somático o germinal del ADN libre de células**

(30) Prioridad:

16.05.2017 US 201762507127 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.12.2024

(73) Titular/es:

GUARDANT HEALTH, INC. (50.0%)
3100 Hanover Street
Palo Alto, CA 94304, US y
DANA-FARBER CANCER INSTITUTE, INC. (50.0%)

(72) Inventor/es:

LANMAN, RICHARD B. y
OXNARD, GEOFFREY R.

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Identificación del origen somático o germinal del ADN libre de células

FONDO

La comparación del genoma de un sujeto y un genoma de referencia (por ejemplo, GRCh38.p4), mostrará típicamente diferencias (variación genética) en torno al 0,01% de bases. Las variantes genéticas en la línea germinal pueden representar SNP transferidos a través de la herencia normal o a través de mutaciones germinales. Las variaciones pueden existir en forma homocigótica o heterocigótica.

Ciertos estados patológicos, como el cáncer, se caracterizan por variaciones genéticas en los genomas de las células patológicas en comparación con el genoma de la línea germinal. Estas variaciones son el resultado de mutaciones en células somáticas, y se denominan mutaciones somáticas.

Los polinucleótidos que albergan mutaciones somáticas pueden detectarse en el ADN libre de células (ADNcf), donde se mezclan con el ADN de células que tienen el genoma de la línea germinal. Cuando en el ADNcf hay un gran fondo (línea germinal), ningún proceso implementado por ordenador puede diferenciar automáticamente las variantes de la línea germinal de las mutaciones somáticas. En cambio, los sistemas convencionales se basan en la pericia de un experto humano individual o de un consorcio de expertos (en ambos casos denominado Junta de Tumores) para distinguir las mutaciones somáticas de las de la línea germinal.

Si no hubiera ruido ni sesgos, las variantes de la línea germinal serían aquellas con una fracción alélica del 50% (en el caso de loci heterocigotos) o del 100% (en el caso de loci homocigotos). Sin embargo, en la práctica, la existencia de ruido y sesgos en el sistema hace que estos números nítidos sean difusos. En otras palabras, los loci het u homo no se detectan exactamente al 50% o al 100%, sino que se sitúan entre los límites de confianza inferior y superior para cada una de las categorías het y homo. Por ejemplo, un locus het podría estar entre el 40% y el 60%, mientras que un locus homo podría estar entre el 98% y el 100%.

Lanman et al. (2015), "Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA", PLOS ONE, vol. 10, n.º 10, página e0140712, divulga la evaluación cuantitativa del ADN tumoral circulante libre de células mediante secuenciación digital, en la que las fracciones de alelos mutantes se notifican cuantitativamente para las variantes de nucleótido único somáticas de importancia clínica y se distinguen de las variantes de nucleótido único de línea germinal. Sin embargo, no se divulga una clasificación basada en un umbral de frecuencia alélica y un umbral de variabilidad.

El documento US 2017/073774 Al enseña la detección precisa de mutaciones somáticas en el plasma u otras muestras que contienen ADN libre de células de pacientes con cáncer, incluida la aplicación de varios filtros, pero no habla de las variantes de línea germinal ni de la necesidad de diferenciar entre variantes de línea germinal y somáticas en casos ambiguos.

RESUMEN

La invención se define en las reivindicaciones adjuntas. Otros aspectos y ventajas de la presente divulgación resultarán evidentes para los expertos en la materia a partir de la siguiente descripción detallada, en la que sólo se muestran y describen realizaciones ilustrativas de la presente divulgación. Como se dará cuenta, la presente divulgación es capaz de otras y diferentes realizaciones, y sus diversos detalles son capaces de modificaciones en varios aspectos obvios, todo ello sin apartarse de la divulgación. Por consiguiente, los dibujos y la descripción deben considerarse de carácter ilustrativo y no restrictivo.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, dicha información de secuenciación comprende lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar medidas cuantitativas de fracción alélica (FA) para cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf; determinar una desviación estándar (STDEV) para cada una de dichas medidas de FA; proporcionar un umbral STDEV y un umbral AF; determinar si cada una de dichas medidas FA tiene una STDEV por encima o por debajo de dicho umbral STDEV; determinar si cada una de dichas medidas FA está por encima o por debajo de dicho umbral FA; y clasificar cada locus con una STDEV por debajo de dicho umbral STDEV y una medida FA por debajo de dicho umbral FA como de origen somático.

En un aspecto, la presente divulgación proporciona un método para identificar el origen de línea germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, dicha información de secuenciación comprende lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar medidas cuantitativas de fracción alélica (FA) para cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf;

determinar una desviación estándar (STDEV) para cada una de dichas medidas de FA; proporcionar un umbral STDEV y un umbral FA; determinar si cada una de dichas medidas FA tiene una STDEV por encima o por debajo de dicho umbral STDEV; determinar si cada una de dichas medidas FA está por encima o por debajo de dicho umbral FA; y clasificar cada locus con una STDEV por debajo de dicho umbral STDEV y una medida FA por encima de dicho umbral AF como de origen de línea germinal.

En algunos aspectos, una medida de FA para un locus genómico por debajo de dicho umbral STDEV indica una baja variación del número de copias (CNV) para dicho locus genómico.

En algunos aspectos, una medida de FA para un locus genómico por encima de dicho umbral STDEV indica una alta variación del número de copias (CNV) para el locus genómico asociado.

En algunos aspectos, el umbral de FA se determina empíricamente.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto con cáncer, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto en un primer punto temporal antes del tratamiento con un terapéutico contra el cáncer, comprendiendo dicha información de secuenciación un primer conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; recibir información de secuenciación de dicho ADNcf de dicho sujeto en un segundo punto temporal después del tratamiento con un terapéutico contra el cáncer, comprendiendo dicha información de secuenciación un segundo conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar medidas de fracción alélica cuantitativa (FA) para cada uno de dicha pluralidad de loci genómicos basándose en dichas lecturas de secuenciación de ADNcf en dicho primer punto temporal y basándose en dichas lecturas de secuenciación de ADNcf en dicho segundo punto temporal; comparar dichas medidas de FA de dicho primer punto temporal y de dicho segundo punto temporal; en el que dicho cáncer responde a dicho terapéutico contra el cáncer; e identificar un locus genómico como de origen somático si una medida de FA de dicho locus genómico disminuye entre dicho primer punto temporal y dicho segundo punto temporal.

En un aspecto, la presente divulgación proporciona un método para identificar el origen de línea germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto con cáncer, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto en un primer punto temporal antes del tratamiento con un terapéutico contra el cáncer, dicha información de secuenciación comprende un primer conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; recibir secuenciación de información de dicho ADNcf de dicho sujeto en un segundo punto temporal después del tratamiento con un terapéutico contra el cáncer, dicha información de secuenciación que comprende un segundo conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar medidas de fracción alélica cuantitativa (FA) para cada uno de dicha pluralidad de loci genómicos basándose en dichas lecturas de secuenciación de ADNcf en dicho primer punto temporal y basándose en dichas lecturas de secuenciación de ADNcf en dicho segundo punto temporal; comparar dichas medidas de FA de dicho primer punto temporal y de dicho segundo punto temporal; en el que dicho cáncer responde a dicho tratamiento del cáncer; e identificar un locus genómico como de origen germinal si una medida de FA de dicho locus genómico no disminuye entre dicho primer punto temporal y dicho segundo punto temporal.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático o de línea germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto recogida en un primer punto temporal, dicha información de secuenciación comprende primeras lecturas de secuenciación de ADNcf; proporcionar información de secuencia de dicha pluralidad de loci genómicos; realizar un binning de cada uno de dicha pluralidad de loci genómicos, en donde dicho binning comprende asignar una clasificación inicial para cada locus genómico en dicha pluralidad de loci genómicos, dicha clasificación inicial seleccionada del grupo que consiste en: a) presunto origen somático; b) presunto origen de línea germinal; o c) origen indeterminado; generando así una primera bandeja que comprende los loci genómicos de presunto origen somático, una segunda bandeja que comprende los loci genómicos de presunto origen de línea germinal, y una tercera bandeja que comprende los loci genómicos de origen indeterminado; determinar, para cada una de dichas regiones genómicas en dicho primer cajón, dicho segundo cajón y dicho tercer cajón, una medida de fracción alélica cuantitativa (FA) basada en dichas primeras lecturas de secuenciación de ADNcf para generar un primer conjunto de FA, un segundo conjunto de FA y un tercer conjunto de FA, respectivamente; generar una primera distribución de frecuencias basada en dicho primer conjunto de FA y una segunda distribución de frecuencias basada en dicho segundo conjunto de FA, en la que no exista solapamiento entre dicha primera distribución de frecuencias y dicha segunda distribución de frecuencias; identificar un valor umbral de FA basado en dichas primera y segunda distribuciones de frecuencia, cuyo valor umbral de FA es (i) no menor que la mayor medida cuantitativa de FA entre dicho primer conjunto de FA y (ii) no mayor que la menor medida cuantitativa de FA entre dicho segundo conjunto de FA; y asignar una clasificación final para cada uno de dicho tercer conjunto de loci genómicos, cuya clasificación final es (A) presunto origen somático si dicho locus genómico tiene una medida de FA cuantitativa no superior a dicho valor umbral de FA, o (B) presunto origen de línea germinal si dicha región genómica tiene una medida de FA cuantitativa no inferior a dicho valor umbral de FA.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático o de línea

germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto recogida en un primer punto temporal, dicha información de secuenciación comprende primeras lecturas de secuenciación de ADNcf; proporcionar información de secuencia de dicha pluralidad de loci genómicos; realizar un binning de cada uno de dicha pluralidad de loci genómicos en dicha pluralidad de loci genómicos, en donde dicho binning comprende asignar una clasificación inicial para cada locus genómico en dicha pluralidad de loci genómicos, dicha clasificación inicial seleccionada del grupo que consiste en: a) presunto origen somático; b) presunto origen de línea germinal; o c) origen indeterminado; generando así una primera bandeja que comprende loci genómicos de presunto origen somático, una segunda bandeja que comprende loci genómicos de presunto origen de línea germinal, y una tercera bandeja que comprende loci genómicos de origen indeterminado; determinar, para cada una de dichas regiones genómicas en dicho primer cajón, dicho segundo cajón y dicho tercer cajón, una medida de fracción alélica cuantitativa (FA) basada en dichas primeras lecturas de secuenciación de ADNcf para generar un primer conjunto de FA, un segundo conjunto de FA y un tercer conjunto de FA, respectivamente; generar una primera distribución de frecuencias basada en dicho primer conjunto de FA y una segunda distribución de frecuencias basada en dicho segundo conjunto de FA, en la que existe un solapamiento entre dicha primera distribución de frecuencias y dicha segunda distribución de frecuencias; identificar un primer valor umbral de FA basado en dichas primera y segunda distribuciones de frecuencia, cuyo primer valor umbral de FA es la mayor medida cuantitativa de FA entre dicho primer conjunto de FA; identificar un segundo valor umbral de FA basado en dichas primera y segunda distribuciones de frecuencia, cuyo segundo valor umbral de FA es la menor medida cuantitativa de FA entre dicho segundo conjunto de FA; y asignar una clasificación final para cada uno de dicha pluralidad de loci genómicos, donde dicha clasificación final es (A) presunto origen somático si dicha región genómica tiene una medida cuantitativa de FA no mayor que dicho primer valor umbral de FA, (B) presunto origen de línea germinal si dicha región genómica tiene una medida cuantitativa de FA no menor que dicho segundo valor umbral de FA, o (C) ambigua si dicha región genómica tiene una medida cuantitativa de FA mayor que dicho primer valor umbral de FA y menor que dicho segundo valor umbral de FA.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, comprendiendo dicha información de secuenciación un conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar un primer conjunto de medidas de fracción alélica cuantitativa (FA), comprendiendo dicho primer conjunto de medidas de FA medidas de FA para cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf; proporcionar un segundo conjunto de medidas de FA, dicho segundo conjunto de medidas de FA comprende medidas de FA para cada una de una o más variantes somáticas conocidas; comparar una medida de FA de un locus genómico de dicho primer conjunto de medidas de FA con una medida de FA de dicho segundo conjunto de medidas de FA; identificar un locus genómico como de origen somático si hay una diferencia del 10% o menos entre dicha medida de FA de dicho primer conjunto de medidas de FA para un locus genómico y dicha medida de FA de dicho segundo conjunto de medidas de FA.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una segunda pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una pluralidad de loci genómicos de ADNcf de una pluralidad de sujetos de control.

En un aspecto, la presente divulgación proporciona un método para identificar el origen de línea germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, dicha información de secuenciación comprende un conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar un primer conjunto de medidas de fracción alélica cuantitativa (FA), dicho primer conjunto de medidas de FA comprende medidas de FA para cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf; proporcionar un segundo conjunto de medidas de FA, dicho segundo conjunto de medidas de FA comprende medidas de FA para cada una de una o más variantes somáticas conocidas; comparar una medida de FA de un locus genómico de dicho primer conjunto de medidas de FA con una medida de FA de dicho segundo conjunto de medidas de FA; identificar un locus genómico como de origen de línea germinal si hay una diferencia superior al 10% entre dicha medida de FA de dicho primer conjunto de medidas de FA para un locus genómico y dicha medida de FA de dicho segundo conjunto de medidas de FA.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una segunda pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una pluralidad de loci genómicos de ADNcf de una pluralidad de sujetos de control.

En un aspecto, la presente divulgación proporciona un método para identificar el origen de línea germinal de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, dicha información de secuenciación comprende un conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar un primer conjunto de medidas de fracción alélica cuantitativa (FA), dicho primer conjunto de medidas de FA comprende medidas de FA para

cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf; proporcionar un segundo conjunto de medidas de FA, dicho segundo conjunto de medidas de FA comprende medidas de FA para cada una de una o más variantes de línea germinal conocidas; comparar una medida de FA de un locus genómico de dicho primer conjunto de medidas de FA con una medida de FA de dicho segundo conjunto de medidas de FA; identificar un locus genómico como de origen de línea germinal si hay una diferencia del 10% o menos entre dicha medida de FA de dicho primer conjunto de medidas de FA para un locus genómico y dicha medida de FA de dicho segundo conjunto de medidas de FA.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una segunda pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una pluralidad de loci genómicos de ADNcf de una pluralidad de sujetos de control.

En un aspecto, la presente divulgación proporciona un método para identificar el origen somático de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto, dicho método comprende: recibir información de secuenciación de dicho ADNcf de dicho sujeto, comprendiendo dicha información de secuenciación un conjunto de lecturas de secuenciación de ADNcf de dicha pluralidad de loci genómicos; determinar un primer conjunto de medidas de fracción alélica cuantitativa (FA), comprendiendo dicho primer conjunto de medidas de FA medidas de FA para cada uno de dicha pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf; proporcionar un segundo conjunto de medidas de FA, dicho segundo conjunto de medidas de FA comprende medidas de FA para cada una de una o más variantes de línea germinal conocidas; comparar una medida de FA de un locus genómico de dicho primer conjunto de medidas de FA con una medida de FA de dicho segundo conjunto de medidas de FA; identificar un locus genómico como de origen somático si hay una diferencia superior al 10% entre dicha medida de FA de dicho primer conjunto de medidas de FA para un locus genómico y dicha medida de FA de dicho segundo conjunto de medidas de FA.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una segunda pluralidad de loci genómicos basadas en dichas lecturas de secuenciación de ADNcf.

En algunos aspectos, dicho segundo conjunto de medidas de FA comprende medidas de FA de una pluralidad de loci genómicos de ADNcf de una pluralidad de sujetos de control.

En algunos aspectos, uno o más de dichos loci genómicos es un locus de un gen BRCA.

La invención se refiere a un método implementado por ordenador que comprende: a) proporcionar un conjunto de lecturas de secuencias de moléculas de ADNcf, en el que las lecturas de secuencias corresponden a una región genómica seleccionada de un genoma de referencia (p. ej., un gen, un exón, un intrón, una porción de un gen (p. ej., al menos 100 nucleótidos, al menos 500 nucleótidos, o al menos 1000 nucleótidos)); b) determinar la frecuencia alélica de un conjunto que comprende una pluralidad de variantes genéticas (por ejemplo, nucleótidos que difieren de la secuencia de referencia) dentro de la región genómica, donde el conjunto incluye una variante de interés; c) determinar una medida de variabilidad (por ejemplo, desviación estándar o varianza) de la frecuencia alélica de las variantes genéticas del conjunto; d) proporcionar un umbral de medida de variabilidad y un umbral de frecuencia alélica; e) determinar si la medida de variabilidad está por debajo del umbral de variabilidad; y f) si la medida de variabilidad está por debajo del umbral de variabilidad: (i) calificar la variante de interés como de origen germinal si la frecuencia alélica de la variante de interés está por encima del umbral de frecuencia alélica, y (ii) calificar la variante de interés como de origen somático si la frecuencia alélica de la variante de interés está por debajo del umbral de frecuencia alélica.

La invención también se refiere al uso de dicho método implementado por ordenador:

(i) para diagnosticar una enfermedad o afección como el cáncer o una afección inflamatoria;
(ii) en el pronóstico de una enfermedad o afección como el cáncer o una afección inflamatoria;
(iii) para evaluar la eficacia del tratamiento de una enfermedad o afección como el cáncer o una afección inflamatoria; y/o
(iv) para controlar la progresión o regresión de una enfermedad o afección como el cáncer o una afección inflamatoria.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Las características novedosas de la divulgación se exponen con particularidad en las reivindicaciones adjuntas. Una mejor comprensión de las características y ventajas de la presente divulgación se obtendrá por referencia a la siguiente descripción detallada que expone realizaciones ilustrativas, en las que se utilizan los principios de la divulgación, y los dibujos adjuntos de los cuales:

FIG. 1 muestra un sistema de control por ordenador que está programado o configurado de otro modo para implementar los métodos aquí proporcionados.
FIG. 2A muestra que las mutaciones T790M de la línea germinal (201 - puntos negros) están presentes en una

concentración similar a las mutaciones T790M somáticas (202 - puntos grises), pero en una fracción alélica (FA) más alta.

FIG. 2B muestra que las concentraciones de mutaciones somáticas del EGFR en cuatro pacientes en tratamiento disminuyen, mientras que la concentración de EGFR T790M de línea germinal permanece constante (203 representa la mutación conductora del EGFR y 204 representa EGFR T790M).

FIG. 2C muestra que la distribución de FA para EGFR T790M (gráfico inferior) a través de los resultados de NGS en plasma para 950 casos incluye un pico somático que también se observa con las mutaciones conductoras de EGFR (gráfico superior), así como un pico heterocigoto (flecha) que también se observa más claramente con un SNP común (EGFR Q787Q, gráfico central).

FIG. 3A muestra la NGS plasmática para muestras de plasma pretratamiento y en tratamiento de tres casos iniciales, todos positivos para EGFR T790M de línea germinal. Entre todas las variantes codificantes y no codificantes detectadas se observaron tres grupos de variantes, correspondientes al FA esperado de variantes homocigóticas, heterocigóticas, y derivadas de tumores. Las variantes del grupo derivado del tumor respondieron al tratamiento, mientras que las variantes de los grupos homocigoto y heterocigoto se mantuvieron en una FA relativamente constante.

FIG. 3B muestra los resultados de NGS en plasma de 102 casos adicionales, para un total de 105 casos. Se observó una distribución trimodal con picos cercanos al 0% (probablemente derivado del tumor), 49% (probablemente heterocigoto), y 100% (probablemente homocigoto) para todas las variantes codificantes y no codificantes detectadas en los 105 casos.

FIG. 3C muestra que, para las variantes missense y nonsense, hubo un enriquecimiento en FA bajas (flechas), donde se esperaría encontrar variantes derivadas de tumores. En cambio, las variantes sinónimas, probablemente reflejo de polimorfismos benignos de la línea germinal, se enriquecieron en torno al 50% y el 100% de FA.

FIG. 4A muestra el FA de todas las variantes encontradas en plasma NGS de 105 casos positivos para mutaciones EGFR, en orden creciente de FA de mutación conductora EGFR (401 - puntos negros), con un SNP EGFR común mostrado (402 - puntos grises más grandes).

FIG. 4B muestra que, para las variantes de FA entre el 25% y el 75%, la desviación estándar y la diferencia absoluta entre la media de los casos y la de la población aumentan con el incremento de la FA impulsora del EGFR.

FIG. 5 muestra la distinción entre variantes codificantes heterocigotas y derivadas del tumor en casos con baja variación del número de copias. FIG. 5, también muestra que, cuando hay una menor variación en el número de copias, es posible distinguir visualmente qué casos de EGFR T790M germinal (501) son probablemente germinales.

FIG. 6A muestra que los resultados de la NGS revelan que 48 (0,15%) pacientes con cáncer de una base de datos de 31.414 pacientes únicos con cáncer eran portadores de una mutación germinal EGFR T790M, siendo el cáncer de pulmón no escamoso de células no pequeñas (NSCLC) el diagnóstico dominante en estos pacientes.

FIG. 6B muestra que, en comparación con la prevalencia poblacional de la línea germinal del EGFR T790M en una cohorte de referencia (0,008%), existe una mayor prevalencia en sujetos con NSCLC no escamoso (0,34%) pero no en sujetos con otros cánceres (0,03%, p = 0,06), lo que sugiere que la línea germinal del EGFR T790M es una variante de riesgo para el cáncer de pulmón.

FIG. 7 muestra un gráfico de FA de variantes codificantes y no codificantes detectadas en tres momentos (FA de la mutación TP53 en el momento 2 imputada al 0%). La mutación EGFR T790M se observa dentro de una banda de variantes que incluye el SNP común (EGFR Q787Q), sospechoso de un EGFR T790M de línea germinal detectado incidentalmente.

FIG. 8A muestra que se puede ajustar una curva para la desviación estándar, incluso con la presencia de valores atípicos.

FIG. 8B muestra que se puede ajustar una curva para la media, incluso con la presencia de valores atípicos.

FIG. 9 muestra que, entre 11 casos de NGS en plasma (Cohorte A) designados con baja variación del número de copias y alta FA de EGFR T790M (izquierda), se confirmó que los 11 eran de línea germinal (valor predictivo positivo del 100%). Entre 10 casos (Cohorte B) con alta variación del número de copias y alta FA de EGFR T790M (derecha), uno fue positivo para una mutación T790M de línea germinal.

DESCRIPCIÓN DETALLADA

Los títulos de sección utilizados en el presente documento tienen únicamente fines organizativos y no deben interpretarse como una limitación de la materia descrita.

En esta descripción detallada de las diversas realizaciones, a efectos de explicación, se exponen numerosos detalles específicos para proporcionar una comprensión completa de las realizaciones divulgadas. Un experto en la materia apreciará, sin embargo, que estas diversas realizaciones pueden practicarse con o sin estos detalles específicos. En otros casos, las estructuras y los dispositivos se muestran en forma de diagrama de bloques. A menos que se describa de otro modo, todos los términos técnicos y científicos utilizados en el presente documento tienen el significado que comúnmente entiende una persona con conocimientos ordinarios de la técnica a la que pertenecen las diversas realizaciones descritas en el presente documento.

Se apreciará que hay un "aproximadamente" implícito antes de las temperaturas, concentraciones, tiempos, número de bases, cobertura, etc. discutidos en las presentes enseñanzas, de tal manera que las equivalencias leves e

insustanciales están dentro del alcance de las presentes enseñanzas. En esta solicitud, el uso del singular incluye el plural a menos que se indique específicamente lo contrario. Asimismo, el uso de "comprender", "comprende", "que comprende", "contener", "contiene", "que contiene", "incluir", "incluye", e "incluyendo" no pretende ser limitativo. Debe entenderse que tanto la descripción general precedente como la descripción detallada siguiente son meramente ejemplificativas y explicativas y no restrictivas de las presentes enseñanzas.

Tal y como se utiliza aquí, "un" o "una" también puede referirse a "al menos uno" o a "uno o más". Además, el uso de "o" es inclusivo, de modo que la frase "A o B" es verdadera cuando "A" es verdadera, "B" es verdadera, o ambas "A" y "B" son verdaderas.

Además, salvo que el contexto exija lo contrario, los términos en singular incluirán los plurales y los términos en plural incluirán los singulares. En general, las nomenclaturas utilizadas en relación con el cultivo de células y tejidos, la biología molecular y la química e hibridación de proteínas y oligos o polinucleótidos descritos en el presente documento son las conocidas y comúnmente utilizadas en la técnica. Se utilizan técnicas estándar, por ejemplo, para la purificación y preparación de ácidos nucleicos, el análisis químico, el ácido nucleico recombinante, y la síntesis de oligonucleótidos. Las reacciones enzimáticas y las técnicas de purificación se llevan a cabo de acuerdo con las especificaciones del fabricante o como se hace habitualmente en la técnica o como se describe en el presente documento. Las técnicas y procedimientos descritos en el presente documento se llevan a cabo generalmente de acuerdo con métodos convencionales bien conocidos en la técnica y como se describe en diversas referencias generales y más específicas que se citan y discuten a lo largo de la presente especificación. Véase, por ej., Sambrook et al., Molecular Cloning: A Laboratory Manual (Tercera ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. 2000). Las nomenclaturas utilizadas en relación con los procedimientos y técnicas de laboratorio aquí descritos son las conocidas y comúnmente utilizadas en la técnica.

Un "sistema" establece un conjunto de componentes, reales o abstractos, que comprenden un todo en el que cada componente interactúa o está relacionado con al menos otro componente dentro del todo.

Una "biomolécula" puede referirse a cualquier molécula producida por un organismo biológico, incluyendo grandes moléculas poliméricas como proteínas, polisacáridos, lípidos, y ácidos nucleicos (ADN y ARN), así como pequeñas moléculas como metabolitos primarios, metabolitos secundarios, y otros productos naturales.

Tal como se utiliza en el presente documento, el término "secuenciación" se refiere a cualquiera de las diversas tecnologías utilizadas para determinar la secuencia de una biomolécula, por ejemplo, un ácido nucleico como el ADN o el ARN. Entre los métodos de secuenciación ejemplares se incluyen, entre otros, la secuenciación dirigida, la secuenciación en tiempo real de molécula única, la secuenciación de exones, la secuenciación basada en microscopía electrónica, la secuenciación de panel, la secuenciación mediada por transistor, la secuenciación directa, la secuenciación shotgun aleatoria, secuenciación por terminación dideoxídica Sanger, secuenciación del genoma completo, secuenciación por hibridación, pirosecuenciación, electroforesis capilar, electroforesis en gel, secuenciación dúplex, secuenciación por ciclos, secuenciación por extensión monobase, secuenciación en fase sólida, secuenciación de alto rendimiento, secuenciación de firmas masivamente paralela, PCR en emulsión, co-amplificación a baja temperatura de desnaturalización-PCR (COLD-PCR), PCR multiplex, secuenciación por terminador de colorante reversible, secuenciación de extremo pareado, secuenciación de corto plazo, secuenciación por exonucleasa, secuenciación por ligación, secuenciación de lectura corta, secuenciación de molécula única, secuenciación por síntesis, secuenciación en tiempo real, secuenciación por terminador inverso, secuenciación nanopore, secuenciación 454, secuenciación Solexa Genome Analyzer, secuenciación SOLiD™, secuenciación MS-PET, y una combinación de las mismas. En algunas realizaciones, la secuenciación puede llevarse a cabo mediante un analizador de genes como, por ejemplo, los analizadores de genes disponibles comercialmente en Illumina o Applied Biosystems.

La expresión "secuenciación de nueva generación" o NGS hace referencia a las tecnologías de secuenciación con un mayor rendimiento en comparación con los enfoques tradicionales basados en Sanger y la electroforesis capilar, por ejemplo, con la capacidad de generar cientos de miles de lecturas de secuencias relativamente pequeñas a la vez. Algunos ejemplos de técnicas de secuenciación de nueva generación incluyen, entre otras, la secuenciación por síntesis, la secuenciación por ligación, y la secuenciación por hibridación.

La frase "ejecución de secuenciación" se refiere a cualquier paso o porción de un experimento de secuenciación realizado para determinar alguna información relacionada con al menos una biomolécula (por ejemplo, una molécula de ácido nucleico como ADN o ARN).

El ADN (ácido desoxirribonucleico) es una cadena de nucleótidos compuesta por cuatro tipos de nucleótidos: adenina (A), timina (T), citosina (C), y guanina (G). El ARN (ácido ribonucleico) es una cadena de nucleótidos compuesta por cuatro tipos de nucleótidos: A, uracilo (U), G, y C. Determinados pares de nucleótidos se unen específicamente entre sí de forma complementaria (lo que se denomina emparejamiento de bases complementarias). En el ADN, la adenina (A) se empareja con la timina (T) y la citosina (C) con la guanina (G). En el ARN, la adenina (A) se empareja con el uracilo (U) y la citosina (C) con la guanina (G). Cuando una primera cadena de ácido nucleico se une a una segunda cadena de ácido nucleico formada por nucleótidos complementarios a los de la primera cadena, las dos cadenas se unen para formar una doble cadena. Tal y como se utiliza en el presente documento, "datos de secuenciación de ácidos nucleicos",

"información de secuenciación de ácidos nucleicos", "secuencia de ácidos nucleicos", "secuencia de nucleótidos", "secuencia genómica" secuencia genómica", "secuencia genética", "secuencia de fragmentos", o "lectura de secuenciación de ácidos nucleicos" denota cualquier información o dato indicativo del orden de las bases nucleotídicas (p. ej., adenina, guanina, citosina, y timina o uracilo) en una molécula (por ejemplo, un genoma completo, un transcriptoma completo, un exoma, un oligonucleótido, un polinucleótido, o un fragmento) de un ácido nucleico como el ADN o el ARN. Debe entenderse que las presentes enseñanzas contemplan la información de secuencias obtenida utilizando todas las variedades disponibles de técnicas, plataformas o tecnologías, incluyendo, pero sin limitarse a: electroforesis capilar, microarrays, sistemas basados en ligación, sistemas basados en polimerasa, sistemas basados en hibridación, sistemas de identificación directa o indirecta de nucleótidos, pirosecuenciación, sistemas de detección basados en iones o pH, y sistemas basados en firma electrónica.

Un "polinucleótido", "ácido nucleico", u "oligonucleótido" se refiere a un polímero lineal de nucleósidos (incluyendo desoxirribonucleósidos, ribonucleósidos, o análogos de los mismos) unidos por enlaces internucleosídicos. Típicamente, un polinucleótido comprende al menos tres nucleósidos. Los oligonucleótidos a menudo varían en tamaño desde unas pocas unidades monoméricas, por ejemplo 3-4, hasta cientos de unidades monoméricas. Siempre que un polinucleótido se represente mediante una secuencia de letras, como "ATGCCTG", se entenderá que los nucleótidos están en orden 5' -> 3' de izquierda a derecha y que "A" significa desoxiadenosina, "C" significa desoxicitidina, "G" significa desoxiguanosina, y "T" significa timidina, a menos que se indique lo contrario. Las letras A, C, G, y T pueden utilizarse para referirse a las propias bases, a nucleósidos o a nucleótidos que comprenden las bases, como es habitual en la técnica.

Los términos "adaptador(es)", y "marcación(es)" se utilizan como sinónimos a lo largo de esta especificación. Un adaptador o marcación puede acoplarse a una secuencia polinucleotídica para ser "marcada" por cualquier método, incluyendo ligación, hibridación, u otros métodos.

Tal y como se utiliza aquí, una variante común tiene al menos un 5% de GMAF (frecuencia alélica menor global), mientras que una variante de baja frecuencia tiene alrededor de un 0,1-5% de GMAF, y una variante rara tiene un 0,5% o menos de GMAF, donde GMAF es una frecuencia en la que se produce el alelo menos común en una población determinada.

Tal como se utiliza aquí, "genotipo" se refiere a la identidad alélica en un locus genético en uno o más cromosomas de la línea germinal. Esto incluye el genotipo completo (identidad alélica en todos los cromosomas), el genotipo parcial (identidad alélica en al menos un cromosoma), y el genotipo nulo (alelo(s) inexistente(s) en uno o más cromosomas o en todos ellos), incluida la determinación de la homocigosidad o heterocigosidad en un locus ("designación alélica").

Tal y como se utiliza aquí, una variante somática indica que la fuente es un tejido canceroso. Tal y como se utiliza aquí, el origen somático de una variante genética se refiere a una variante genética que se produce por primera vez en una célula somática y no en la línea germinal. Esto contrasta con las variantes de línea germinal, que tienen como fuente células normales. La variación puede transmitirse a las células hijas a través de la división mitótica. Esto puede dar lugar a que un grupo de células tenga una diferencia genética con el resto de las células de un organismo. Además, como la variación no se produce en una célula de la línea germinal, la mutación no será heredada por los organismos de la progenie.

SNP puede referirse a polimorfismo o variación de un solo nucleótido en la población, normalmente en el contexto de variantes de la línea germinal, mientras que SNV puede referirse a variante de un solo nucleótido y SSNV puede referirse a variante somática de un solo nucleótido (normalmente utilizada en el contexto de variantes asociadas al cáncer). Para un individuo, el término SNV se utiliza para las variaciones detectadas tanto en el ADNcf somático (canceroso) como en la línea germinal (normal).

CNV puede referirse a una variante en el número de copias (mutación en el número de copias a nivel genético, normalmente resultado de un evento de duplicación).

El ADN libre de células (ADNcf) de un sujeto con cáncer comprende ADN de células con el genoma de la línea germinal (por ejemplo, de "células sanas") ("ADN de la línea germinal") y ADN de células cancerosas, normalmente con mutaciones somáticas ("ADN canceroso"). Las cantidades relativas de ADN de la línea germinal y de ADN canceroso en una muestra de ADN libre de células depende de lo avanzado que esté el cáncer. En las fases iniciales, sólo una pequeña cantidad del ADN es ADN canceroso. Esto podría ser, por ejemplo, alrededor del l%-5% del ADN total. Por lo tanto, la detección de una pequeña cantidad (por ejemplo, alrededor del l%-5% del ADNcf de una muestra) de una variante genética, como un alelo menor, puede ser indicativa de una mutación somática y, por lo tanto, de la presencia de ADN canceroso. Sin embargo, a medida que la enfermedad progresa y el tumor se expande, la cantidad de ADN canceroso en una muestra de ADN libre de células puede aumentar significativamente, por ejemplo hasta más del 25% del total de ADN libre de células. Cuando el porcentaje de moléculas de ADN portadoras de una variante genética alcanza niveles elevados, puede resultar ambiguo si la variante representa una mutación somática procedente de células cancerosas o representa heterocigosidad en el ADN de la línea germinal.

El análisis genómico del ADNcf plasmático puede ser una herramienta para el descubrimiento genómico y para

ayudar a la administración de la medicina oncológica de precisión, pero la liberación de ADN derivado del cáncer en el plasma puede ser muy variable y depende del estadio del cáncer, el grado de diseminación metastásica, y si el cáncer está respondiendo o progresando. Además, los niveles plasmáticos de alteraciones genómicas somáticas pueden ser muy dinámicos en respuesta a la terapia, llegando en ocasiones a ser indetectables en dos semanas. En consecuencia, en muchos pacientes la mayor parte del ADNcf plasmático es ADN de la línea germinal, en gran parte procedente de células hematopoyéticas o endoteliales benignas. La presente divulgación proporciona un enfoque que puede distinguir las variantes de la línea germinal de las variantes somáticas derivadas del cáncer dentro de los perfiles de secuenciación de próxima generación (NGS) de ADNcf, proporcionando así tanto el genotipado tumoral para la selección de la terapia como la caracterización de la línea germinal para la evaluación del riesgo hereditario con un único ensayo.

El enfoque proporcionado por la presente divulgación tiene varias aplicaciones. La NGS en plasma identificará, en ocasiones, mutaciones incidentales de la línea germinal en pacientes con cáncer que tienen posibles implicaciones clínicas para los pacientes y sus familias. Ciertas mutaciones de la línea germinal del EGFR descritas en el presente documento son alelos de riesgo poco frecuentes que se cree que están asociados con el riesgo de cáncer hereditario, y los enfoques descritos en el presente documento pueden aplicarse a otras mutaciones de la línea germinal. Otros genes relacionados con el cáncer (por ejemplo, TP53 o BRCAI/2 y genes de reparación de emparejamientos erróneos) también pueden secuenciarse con NGS en plasma, y las mutaciones de línea germinal en estos genes pueden tener profundas implicaciones clínicas. La presente divulgación describe la presencia de una presunta mutación de la línea germinal mediante un algoritmo bioinformático que predice una mutación de la línea germinal confirmada con un alto valor predictivo positivo, una importante característica de rendimiento diagnóstico para una prueba de inclusión.

El discernimiento de las variantes somáticas y de línea germinal en el ADNcf plasmático también puede influir en la comprensión de la biología del cáncer. Con la NGS tumoral, puede ser difícil determinar si una variante de significado desconocido en un oncogén representa una posible mutación impulsora o un polimorfismo de línea germinal. En los casos de NGS en plasma sin alta variación en el número de copias, la presente divulgación permite diferenciar entre estos dos tipos de alteraciones genómicas con una sola muestra (de sangre), reduciendo así el riesgo de que un polimorfismo de la línea germinal se dirija terapéuticamente por error. Además, en los casos en los que se utiliza el genotipado plasmático en serie a lo largo del tiempo para controlar la respuesta y la resistencia a la terapia, la capacidad de distinguir las variantes de la línea germinal y las somáticas en el ADNcf plasmático puede facilitar el seguimiento preciso de los niveles de ADN tumoral.

La carga de mutación tumoral (TMB) es un biomarcador emergente para comprender la sensibilidad y la resistencia a los inhibidores de puntos de control inmunitarios. Un mayor número de mutaciones dentro de un cáncer puede dar lugar a más neoantígenos de superficie celular para la estimulación inmunitaria. Sin embargo, puede ser difícil calcular la carga mutacional utilizando la NGS tumoral porque los polimorfismos de la línea germinal pueden confundirse con mutaciones somáticas potencialmente antigénicas. La presente divulgación ofrece la posibilidad de superar este reto, permitiendo distinguir bioinformáticamente las variantes de línea germinal de las somáticas e identificar más claramente las variantes somáticas antigénicas, reduciendo así la posibilidad de que un polimorfismo de línea germinal se confunda con una mutación somática potencialmente antigénica.

El ADN de la línea germinal de un sujeto puede ser homocigótico o heterocigótico en cualquier locus genético. Las mediciones en un locus pueden adoptar la forma de fracciones alélicas (FA), que miden la frecuencia con la que se observa un alelo en una muestra. Por diversos motivos (incluidos, por ejemplo, errores en la secuenciación del ADN), en un conjunto de lecturas de secuencias generadas a partir de ADNcf de un sujeto no oncológico, los recuentos de lecturas para una forma alélica (por ejemplo, una SNV) que se asigna a un locus genético para el que el sujeto es homocigoto pueden no ser exactamente del 100%. Del mismo modo, el recuento de lecturas para una forma alélica que se corresponde con un locus genético para el que el sujeto es heterocigoto puede no ser exactamente del 50%. Si un individuo es homocigótico en la línea germinal para una variante genética (que no coincide con el alelo en un genoma de referencia, el porcentaje de llamadas de bases que llevan la variante genética será normalmente cercano, pero no siempre idéntico, al 100% de las llamadas. Del mismo modo, si un individuo es heterocigoto en la línea germinal para una variante genética, el porcentaje de llamadas de bases que portan la variante genética será normalmente cercano al 50%, pero podría oscilar, por ejemplo, entre el 30% y el 70%. Las medidas en este rango son consistentes con la heterocigosidad en el locus. Sin embargo, la medición puede hacer que la determinación sea ambigua. En este caso, se puede calificar un genotipo en un locus de heterocigoto u homogencigoto con un cierto nivel de confianza o probabilidad.

En consecuencia, si un sujeto tiene cáncer, y una variante genética de un locus se mide en un rango consistente con la heterocigosidad, la confianza de que la variante es el resultado de una mutación somática puede disminuir en comparación con una medición en un rango entre la homocigosidad y la heterocigosidad. Por ejemplo, una medición en el rango del 5% al 20% puede indicar que un locus contiene una variante genética en una cantidad demasiado alta para ser contabilizada por homocigosis y demasiado baja para ser contabilizada por heterocigosis. Por lo tanto, es probable que la medida sea el resultado de una mutación somática. Por el contrario, una medición en torno al 40% puede indicar heterocigosidad o puede indicar una abundancia de ADN que contiene una mutación somática (por ejemplo, si esa mutación somática ha causado un tumor que ha aportado una gran cantidad relativa de ADN a la muestra).

Esta divulgación proporciona, entre otras cosas, métodos para determinar si el origen de una variante genética detectada en una muestra de ADN libre de células es más probablemente de línea germinal (por ejemplo, representando

heterocigosidad en la línea germinal), o somática (por ejemplo, de cáncer). En particular, la presente divulgación proporciona métodos que utilizan FAs para hacer esta determinación.

En algunas realizaciones, la presente divulgación proporciona un método para identificar el origen germinal o somático de cada uno de una pluralidad de loci genómicos en ADN libre de células (ADNcf) de un sujeto con uno o más umbrales que pueden utilizarse para determinar si una variante en un locus es de origen germinal o somático. Un umbral ejemplar que puede utilizarse es un umbral de desviación estándar (STDEV). Por ejemplo, el trabajador cualificado, tras determinar una fracción alélica cuantitativa (FA) para un locus genómico, puede determinar un STDEV para la medida de FA. A medida que aumenta la variación del número de copias (CNV), es de esperar que también aumenten los STDEV. Por lo tanto, se puede suponer que las STDEV bajas tienen una CNV baja, lo que facilita el procesamiento de esos datos. Se puede utilizar un umbral STDEV para separar la CNV alta de la baja, lo que aumenta el poder predictivo del método. Se puede utilizar un segundo umbral de FA en combinación con el umbral de CNV o en sustitución de éste. Dado que se espera que las medidas de FA sean mayores en las variantes derivadas de la línea germinal que en las variantes somáticas, las medidas de FA por encima de un umbral de FA pueden clasificarse como derivadas de la línea germinal, mientras que las medidas de FA por debajo de un umbral de FA pueden clasificarse como derivadas de la línea somática. Los umbrales de FA ejemplares incluyen, pero no se limitan a, alrededor del 10%, alrededor del 11%, alrededor del 12%, alrededor del 13%, alrededor del 14%, alrededor del 15%, alrededor del 16%, alrededor del 17%, alrededor del 18%, alrededor del 19%, alrededor del 20%, alrededor del 21%, alrededor del 22%, alrededor del 23%, alrededor del 24%, alrededor del 25%, alrededor del 26%, alrededor del 27%, alrededor del 28%, alrededor del 29%, alrededor del 30%, alrededor del 31%, alrededor del 32%, alrededor del 33%, alrededor del 34% y alrededor del 35%. En algunas realizaciones, el umbral de FA se determina empíricamente.

Los métodos aquí descritos también pueden utilizarse para determinar si un locus del ADNcf es de origen germinal o somático en función de la respuesta al tratamiento. Por ejemplo, se puede obtener información de la secuencia de un sujeto con cáncer antes y después del tratamiento con una terapia contra el cáncer. Si el cáncer responde a la terapia oncológica, y una variante asociada al cáncer en un locus es de origen somático, su FA debería disminuir. Así, se puede medir la FA antes y después del tratamiento, y comparar esos valores, para determinar el origen somático o germinal. Si la medida de FA disminuye, la variante puede identificarse como de origen somático. Si la medida de FA no disminuye (es decir, permanece igual o aumenta), la variante puede identificarse como de origen germinal.

En algunas realizaciones, los métodos aquí descritos pueden usarse para determinar si los loci del ADNcf son de origen somático o de línea germinal clasificando los loci según una clasificación inicial de presunto origen somático, presunto origen de línea germinal, u origen indeterminado. A continuación, pueden determinarse medidas cuantitativas de FA para los loci de cada bin con el fin de generar conjuntos de FA, que posteriormente se utilizan para generar distribuciones de frecuencias para los loci de presunto origen somático o germinal. Las distribuciones pueden utilizarse para establecer un valor umbral de FA, por ejemplo, un valor umbral no inferior a la mayor medida cuantitativa de FA entre el conjunto de FA "presuntamente somáticas" y no superior a la menor medida cuantitativa de FA entre el conjunto de FA "presuntamente de línea germinal". Así, los loci de "origen indeterminado" pueden clasificarse como de línea germinal o somáticos en función de si el FA de un locus está por encima del valor umbral de FA (y es, por tanto, de línea germinal) o por debajo del valor umbral de FA (y es, por tanto, somático). Alternativamente, cuando existe un solapamiento entre una distribución de frecuencias para medidas de FA "presuntamente somáticas" y una distribución de frecuencias para medidas de FA "presuntamente de línea germinal", pueden determinarse dos valores umbral de tal manera que un primer valor umbral de FA sea la medida cuantitativa de FA más grande entre el conjunto de FA "presuntamente somáticas" y un segundo valor umbral de FA sea la medida cuantitativa de FA más pequeña entre el conjunto de FA "presuntamente de línea germinal". En tales realizaciones, los loci con medidas cuantitativas de FA por debajo del umbral "presuntamente somático" se clasifican como somáticos, los loci con medidas cuantitativas de FA por encima del umbral "presuntamente de línea germinal" se clasifican como de línea germinal, y los loci con medidas cuantitativas de FA por debajo del entre los dos umbrales se clasifican como ambiguos. A estos loci ambiguos se les puede entonces, por ejemplo, asignar una probabilidad para saber si son de origen germinal o somático basándose en la posición de su medida de FA en la distribución de frecuencias.

En algunas realizaciones, la presente divulgación proporciona un método para identificar el origen somático o de línea germinal de un locus genómico en ADNcf comparando una medida de FA de un locus genómico en una muestra con una o más medidas de FA de variantes somáticas o de línea germinal conocidas. Si, por ejemplo, se utilizan medidas de FA de variantes somáticas conocidas, los loci genómicos con medidas de FA que son similares (por ejemplo, dentro del 30%, dentro del 25%, dentro del 20%, dentro del 15%, dentro del 10%, dentro del 9%, dentro del 8%, dentro del 7%, dentro del 6%, dentro del 5%, dentro del 4%, dentro del 3%, dentro del 2%, dentro del 1%, o dentro del 0,1%) pueden clasificarse como de origen somático, mientras que los loci genómicos con medidas de FA que no son similares (por ejemplo, no dentro del 30%, no dentro del 25%, no dentro del 20%, no dentro del 15%, no dentro del 10%, no dentro del 9%, no dentro del 8%, no dentro del 7%, no dentro del 6%, no dentro del 5%, no dentro del 4%, no dentro del 3%, no dentro del 2%, no dentro del 1%, o no dentro del 0,1%), pueden clasificarse como de origen germinal. Del mismo modo, si se utilizan medidas de variantes conocidas de la línea germinal, los loci genómicos con medidas de FA que son similares (por ejemplo, dentro del 30%, dentro del 25%, dentro del 20%, dentro del 15%, dentro del 10%, dentro del 9%, dentro del 8%, dentro del 7%, dentro del 6%, dentro del 5%, dentro del 4%, dentro del 3%, dentro del 2%, dentro del 1%, o dentro del 0,1%) pueden clasificarse como de origen de la línea germinal, mientras que los loci genómicos con medidas de FA que no son similares (por ejemplo, no dentro del 30%, no dentro del 25%, no dentro del 20%, no dentro del 15%, no dentro del 10%, no dentro del 9%, no dentro del 8%, no dentro del 7%, no dentro del 6%, no dentro del 5%, no dentro del 4%, no

dentro del 3%, no dentro del 2%, no dentro del 1%, o no dentro del 0,1%), pueden clasificarse como de origen somático. Las medidas de FA de variantes somáticas o de línea germinal conocidas pueden proceder de lecturas de secuenciación de ADNcf del sujeto sometido a prueba o de una pluralidad de sujetos de control.

En algunas realizaciones, el ADN libre de células de un sujeto es secuenciado y una o más variantes genéticas son detectadas y cuantificadas. Por ejemplo, se determina la cantidad relativa de lecturas totales (número de recuentos de lecturas) que corresponden a un locus que contiene la variante. Si la cantidad relativa es coherente con la homocigosis, se puede tener una alta confianza en que la variante está presente en la línea germinal. Dicha cantidad podría ser, por ejemplo, superior al 95%, superior al 96%, superior al 97%, superior al 98%, superior al 99%, o al 100%. Esta llamada puede compararse con un genotipo determinado para su confirmación.

Si la cantidad relativa es inconsistente con un genotipo homocigoto o heterocigoto en el locus, se puede tener una alta confianza en que la variante es el resultado de una mutación somática y no está presente en la línea germinal. Dicha cantidad podría ser, por ejemplo, inferior al 30%, inferior al 25%, inferior al 20%, inferior al 15%, inferior al 10%, inferior al 9%, inferior al 8%, inferior al 7%, inferior al 6%, inferior al 5%, inferior al 4%, inferior al 3%, inferior al 2%, o inferior al 1%. Una vez más, esta llamada puede compararse con un genotipo determinado para su confirmación.

Alternativamente, la cantidad relativa puede ser consistente con la heterocigosidad en el locus. Dicha cantidad podría estar, por ejemplo, entre el 30% y el 70%, por ejemplo, entre el 40% y el 60%, entre el 45% y el 55%, entre el 46% y el 54%, entre el 47% y el 53%, entre el 48% y el 52%, o entre el 49% y el 51%. En algunas realizaciones, se determina el genotipo probable de la línea germinal (por ejemplo, obtenido a partir del ADNg) del sujeto en el locus. En algunas realizaciones, el genotipo se compara con la identidad de la variante encontrada en el ADN libre de células. En ciertas realizaciones, si el genotipo es homocigoto, entonces se puede concluir con alta confianza que la variante representa una mutación somática, y muy probablemente en cantidades elevadas. Si se determina que el genotipo es heterocigoto, y la variante coincide con uno de los alelos heterocigotos, entonces se puede concluir que la variante no es una mutación somática, sino que representa la heterocigosidad en el genotipo de la línea germinal.

En algunas realizaciones, un genotipo homocigoto puede ser descartado con alta confianza pero un genotipo heterocigoto no puede ser determinado con alta confianza, resultando en un genotipo potencialmente ambiguo. Por ejemplo, la variante puede medirse en el ADN genómico en un extremo del intervalo, por ejemplo, en el 30%. En tal caso, es posible que no se pueda determinar, con un alto grado de confianza, si la cantidad de variante detectada en el ADNcf es o no más probable que represente una mutación somática o una heterocigosidad de la línea germinal. Dicha medida puede surgir cuando en una muestra hay una abundancia de ADN que contiene la mutación somática debido, por ejemplo, al rápido crecimiento de las células tumorales. Cabe señalar que, en cualquier nivel de medición, existe cierta probabilidad de que una variante detectada en el ADN genómico no represente heterocigosidad. Sin embargo, lo más probable es que entre el 30% y el 70% de la detección de una variante en la línea germinal represente heterocigosidad, y la variante detectada en el ADNcf puede medirse en función de ésta.

En tales casos, se puede utilizar otra información de forma bayesiana para aumentar o disminuir la probabilidad de que una variante en el ADNcf represente una mutación somática o una heterocigosidad en la línea germinal. Por ejemplo, los estudios de población pueden indicar la prevalencia de una variante en la línea germinal de diversos grupos, por ejemplo, basándose en la ascendencia genética. Así, por ejemplo, si se tiene poca confianza en la determinación de un genotipo heterocigoto en un individuo, y la variante se encuentra con alta incidencia en las personas que comparten la ascendencia genética del sujeto, entonces se puede determinar con mayor confianza que la persona es, de hecho, heterocigota y que la variante en el ADNcf no representa una mutación somática. Por el contrario, si la variante sólo se encuentra con una incidencia muy baja en las personas que comparten la ascendencia genética del sujeto, entonces se puede determinar con mayor confianza que la persona no es heterocigota y que la variante en el ADNcf representa una mutación somática.

Esta divulgación contempla varias formas de determinar si una cantidad (por ejemplo, de recuentos de lecturas) es consistente o inconsistente con un genotipo heterocigoto. En algunas realizaciones, se utilizan valores de corte. Por ejemplo, se puede establecer un límite del 30% del recuento total de lecturas para una variante genética concreta en un locus. En algunas realizaciones, se presume que los valores por debajo de la cantidad de corte representan mutaciones somáticas. En algunas realizaciones, los valores por encima de la cantidad de corte, y, típicamente por debajo de un corte para la homocigosidad, pueden presumirse consistentes con la heterocigosidad y, por lo tanto, requieren un análisis adicional antes de calificar la variante como una mutación somática.

En algunas realizaciones, se utiliza una función probabilística (por ejemplo, una función bayesiana) para calcular una probabilidad de que una cantidad represente heterocigosidad. Las probabilidades por encima de ciertos niveles pueden activar el genotipo de comparación.

En algunas realizaciones, la determinación de un genotipo se realiza como parte rutinaria del análisis. En algunas realizaciones, la determinación de un genotipo se determina sólo si una abundancia de una variante es consistente con una interpretación de heterocigosidad.

En algunas realizaciones, los métodos de la presente divulgación reducen las tasas de error y el sesgo que

pueden ser órdenes de magnitud superiores a lo que se requiere para detectar de forma fiable alteraciones genómicas de novo asociadas con el cáncer. En algunas realizaciones, los métodos capturan primero la información genética recogiendo muestras de fluidos corporales como fuentes de material genético (sangre, saliva, sudor, entre otros), seguido de la secuenciación de los materiales. Por ejemplo, los polinucleótidos de una muestra pueden secuenciarse, produciendo una pluralidad de lecturas de secuencia. La carga tumoral en una muestra que comprende polinucleótidos puede estimarse como una relación como el número relativo de lecturas de secuencias que presentan una variante y el número total de lecturas de secuencias generadas a partir de la muestra. Además, en el caso de las variantes del número de copias, la carga tumoral puede estimarse como el exceso relativo (en el caso de la duplicación de genes) o el déficit relativo (en el caso de la eliminación de genes) del número total de lecturas de secuencias en los loci de prueba y de control. Así, por ejemplo, una ejecución puede producir 1000 lecturas asignadas a un locus oncogénico, de las cuales 900 corresponden al tipo salvaje y 100 a un mutante cancerígeno, lo que indica una variante del número de copias en este gen. A continuación, se procesa la información genética y se identifican las variantes genéticas. Las variantes genéticas incluyen variantes de secuencia, variantes de número de copias y variantes de modificación de nucleótidos. Una variante de secuencia es una variación en una secuencia genética de nucleótidos. Una variante del número de copias es una desviación del tipo salvaje en el número de copias de una porción de un genoma. Las variantes genéticas incluyen, por ejemplo, variaciones de nucleótido único (SNP), inserciones, deleciones, inversiones, transversiones, translocaciones, fusiones génicas, fusiones cromosómicas, truncamientos génicos, variaciones en el número de copias (por ejemplo, aneuploidía, aneuploidía parcial, poliploidía, amplificación génica), cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos y cambios anormales en la metilación de los ácidos nucleicos. A continuación, el proceso determina la frecuencia de las variantes genéticas en la muestra que contiene el material genético. Como este proceso es ruidoso, el proceso separa la información del ruido.

Los métodos de secuenciación tienen tasas de error. Por ejemplo, el sistema mySeq de Illumina puede producir porcentajes de error de un solo dígito. Así, para 1000 lecturas de secuencias asignadas a un locus, cabe esperar que unas 50 lecturas (alrededor del 5%) incluyan errores. Ciertas metodologías, como las descritas en el documento WO 2014/149134 (Talasaz y Eltoukhy) pueden reducir significativamente la tasa de error. Los errores crean ruido que puede oscurecer las señales de cáncer presentes a bajos niveles en una muestra. Así, si una muestra tiene una carga tumoral a un nivel en torno a la tasa de error del sistema de secuenciación, por ejemplo, en torno al 0,1%-5%, puede ser difícil distinguir una señal correspondiente a una variante genética debida al cáncer de otra debida al ruido.

El diagnóstico del cáncer puede realizarse analizando las variantes genéticas, incluso en presencia de ruido. El análisis puede basarse en la frecuencia de las variantes de secuencia o en el nivel de CNV y puede establecerse una indicación o un nivel de confianza en el diagnóstico para detectar variantes genéticas en el intervalo de ruido. A continuación, el proceso aumenta la confianza del diagnóstico. Esto puede hacerse utilizando una pluralidad de mediciones para aumentar la confianza del diagnóstico, o alternativamente utilizando mediciones en una pluralidad de puntos temporales para determinar si el cáncer está avanzando, en remisión o estabilizado. La confianza diagnóstica puede utilizarse para identificar estados de enfermedad. Por ejemplo, los polinucleótidos libres de células tomados de un sujeto pueden incluir polinucleótidos derivados de células normales, así como polinucleótidos derivados de células enfermas, como células cancerosas. Los polinucleótidos de las células cancerosas pueden presentar variantes genéticas, como mutaciones somáticas celulares y variantes en el número de copias. Cuando se secuencian polinucleótidos libres de células de una muestra de un sujeto, estos polinucleótidos cancerígenos se detectan como variantes de secuencia o como variantes del número de copias. La cantidad relativa de polinucleótidos tumorales en una muestra de polinucleótidos libres de células se denomina "carga tumoral".

Las mediciones de un parámetro, estén o no en el intervalo de ruido, pueden proporcionarse con un intervalo de confianza. Si se comprueba a lo largo del tiempo, se puede determinar si un cáncer está avanzando, estabilizado o en remisión comparando los intervalos de confianza a lo largo del tiempo. Cuando los intervalos de confianza no se solapan, esto indica la dirección de la enfermedad.

A continuación, el proceso genera un Informe/Diagnóstico genético. El proceso recibe SNP de línea germinal y mutaciones somáticas del cáncer y marca las mutaciones somáticas del cáncer y genera un informe para anotar las mutaciones somáticas similar al análisis de la junta de tumores humanos y proporcionar opciones de tratamiento para ser revisadas y aprobadas por el director del laboratorio.

Volviendo ahora al proceso para generar recomendaciones de la junta tumoral, en algunas realizaciones, el sistema utiliza datos de SNV del portal cBio para los 68 genes en GH2.7, donde GH2.7 es el panel de Guardant Health y los procesos de prueba relacionados publicados en febrero de 2015 (prueba Guardane60). El cBioPortal for Cancer Genomics (http://cbioportal.org) ofrece un recurso web para explorar, visualizar, y analizar datos multidimensionales de genómica del cáncer. El portal reduce los datos de perfiles moleculares de tejidos y líneas celulares cancerosos a eventos genéticos, epigenéticos, de expresión génica, y proteómicos fácilmente comprensibles. La interfaz de consulta combinada con el almacenamiento de datos personalizado permite a los investigadores explorar de forma interactiva las alteraciones genéticas en muestras, genes, y vías y, cuando están disponibles en los datos subyacentes, vincularlas a los resultados clínicos. El portal ofrece resúmenes gráficos de datos a nivel de genes procedentes de múltiples plataformas, visualización y análisis de redes, análisis de supervivencia, consultas centradas en el paciente, y acceso a programas informáticos. El sistema proporciona llamadas a nivel de variante así como llamadas a nivel de muestra para determinar si el director 3 debe revisar la prueba en profundidad.

Los métodos y sistemas aquí descritos permiten detectar numerosos tipos de cáncer. Las células cancerosas, como la mayoría de las células, pueden caracterizarse por una tasa de recambio, en la que las células viejas mueren y son sustituidas por células nuevas. Generalmente, las células muertas, en contacto con la vasculatura de un sujeto determinado, pueden liberar ADN o fragmentos de ADN en el torrente sanguíneo. Lo mismo ocurre con las células cancerosas en las distintas fases de la enfermedad. Las células cancerosas también pueden caracterizarse, en función del estadio de la enfermedad, por diversas aberraciones genéticas, como la variación del número de copias y las mutaciones. Este fenómeno puede utilizarse para detectar la presencia o ausencia de individuos con cáncer mediante los métodos y sistemas aquí descritos.

En algunas realizaciones, los métodos de la presente divulgación pueden utilizarse para diagnosticar una enfermedad o afección como el cáncer o una afección inflamatoria. El término "diagnóstico", tal como se utiliza en el presente documento, se refiere a los métodos mediante los cuales el trabajador especializado puede estimar y/o determinar si un paciente padece o no una enfermedad o afección determinada. En algunas realizaciones, los métodos de la presente divulgación se pueden utilizar en el pronóstico si una enfermedad de una enfermedad o condición tal como cáncer o una condición inflamatoria. El término "pronóstico", tal como se utiliza aquí, se refiere a la probabilidad de progresión de una enfermedad o afección, incluida la recurrencia de una enfermedad o afección. En algunas realizaciones, los métodos de la presente divulgación pueden utilizarse para evaluar el riesgo de desarrollar una enfermedad o afección como el cáncer o una afección inflamatoria. En algunas realizaciones, los métodos de la presente divulgación pueden utilizarse para evaluar la eficacia del tratamiento de una enfermedad o afección como el cáncer o una afección inflamatoria. Por ejemplo, los métodos de la presente divulgación pueden utilizarse antes y después de tratar a un paciente con la enfermedad o afección (por ejemplo, antes y después de administrar un fármaco como un agente quimioterapéutico). En algunas realizaciones, los métodos de la presente divulgación pueden utilizarse para monitorizar la progresión o regresión de una enfermedad o afección como el cáncer o una afección inflamatoria. Por ejemplo, los métodos de la presente divulgación pueden realizarse en diferentes puntos temporales para monitorizar la progresión o regresión. En algunas realizaciones, los métodos de la presente divulgación pueden utilizarse para identificar un compuesto para mejorar o tratar una enfermedad o afección como el cáncer o una afección inflamatoria. Por ejemplo, los métodos de la presente divulgación pueden utilizarse antes y después de administrar el compuesto para determinar si el compuesto mejora o trata la enfermedad.

Tal como se utiliza aquí, "tratar" una enfermedad o afección se refiere a tomar medidas para obtener resultados beneficiosos o deseados, incluidos los resultados clínicos. Los resultados clínicos beneficiosos o deseados incluyen, entre otros, el alivio o la mejora de uno o más síntomas asociados a enfermedades o afecciones. Tal como se utiliza en el presente documento, la "administración" o "administración de" un compuesto o un agente a un sujeto puede llevarse a cabo utilizando uno de los diversos métodos conocidos por los expertos en la técnica. Por ejemplo, un compuesto o un agente puede administrarse por vía intravenosa, arterial, intradérmica, intramuscular, intraperitoneal, intravenosa, subcutánea, ocular, sublingual, oral (por ingestión), intranasal (por inhalación), intraespinal, intracerebral, y transdérmica (por absorción, por ejemplo, a través de un conducto cutáneo). Un compuesto o agente también puede introducirse adecuadamente mediante dispositivos poliméricos recargables o biodegradables u otros dispositivos, por ejemplo, parches y bombas, o formulaciones, que permiten la liberación prolongada, lenta, o controlada del compuesto o agente. La administración también puede realizarse, por ejemplo, una vez, varias veces y/o durante uno o varios periodos prolongados. En algunos aspectos, la administración incluye tanto la administración directa, incluida la autoadministración, como la administración indirecta, incluido el acto de prescribir un fármaco. Por ejemplo, tal y como se utiliza en el presente documento, un médico que indica a un paciente que se autoadministre un fármaco o que se lo administre otro y/o que proporciona a un paciente una receta para un fármaco está administrando el fármaco al paciente. En algunas realizaciones, un compuesto o un agente se administra por vía oral, por ejemplo, a un sujeto por ingestión, o por vía intravenosa, por ejemplo, a un sujeto por inyección. En algunas realizaciones, el compuesto o agente administrado por vía oral se encuentra en una formulación de liberación prolongada o lenta, o se administra utilizando un dispositivo para dicha liberación lenta o prolongada.

En algunas realizaciones, la sangre de sujetos con riesgo de cáncer puede extraerse y prepararse como se describe en el presente documento para generar una población de polinucleótidos libres de células. En un ejemplo, podría tratarse de ADN libre de células. Los sistemas y métodos de la divulgación pueden emplearse para detectar mutaciones o variaciones en el número de copias que puedan existir en determinados cánceres presentes. El método puede ayudar a detectar la presencia de células cancerosas en el organismo, a pesar de la ausencia de síntomas u otros signos distintivos de la enfermedad.

Tal como se utiliza aquí, el término "cáncer" incluye, pero no se limita a, varios tipos de neoplasias malignas, la mayoría de las cuales pueden invadir los tejidos circundantes, y pueden hacer metástasis en diferentes lugares (véase, por ejemplo, el Diccionario Médico PDR, 1.ª edición (1995). Los términos "neoplasia" y "tumor" se refieren a un tejido anormal que crece por proliferación celular más rápidamente de lo normal y que continúa creciendo después de eliminar el estímulo que inició la proliferación. Este tejido anormal muestra una falta parcial o total de organización estructural y de coordinación funcional con el tejido normal, que puede ser benigno (como un tumor benigno) o maligno (como un tumor maligno). Algunos ejemplos de categorías generales de cáncer son, entre otros, los carcinomas (tumores malignos derivados de células epiteliales como, por ejemplo, las formas comunes de cáncer de mama, próstata, pulmón y colon), los sarcomas (tumores malignos derivados del tejido conjuntivo o células mesenquimales), linfomas (tumores malignos

derivados de células hematopoyéticas), leucemias (tumores malignos derivados de células hematopoyéticas), y tumores de células germinales (tumores derivados de células totipotentes, que en los adultos se encuentran con mayor frecuencia en el testículo o el ovario; en fetos, bebés y niños pequeños, se encuentran con mayor frecuencia en la línea media del cuerpo, sobre todo en la punta de la rabadilla), tumores blásticos (tumor típicamente maligno que se asemeja a un tejido inmaduro o embrionario) y similares. Ejemplos de los tipos de neoplasias que pretende abarcar la presente divulgación incluyen, pero no se limitan a, aquellas neoplasias asociadas con cánceres de tejido neural, tejido hematopoyético, mama, piel, hueso, próstata, ovarios, útero, cuello uterino, hígado, pulmón, cerebro, laringe, vesícula biliar, páncreas, recto, paratiroides, tiroides, glándula suprarrenal, sistema inmunitario, cabeza y cuello, colon, estómago, bronquios, y/o riñones. En realizaciones particulares, los tipos y el número de cánceres que pueden detectarse incluyen, pero no se limitan a, cánceres de sangre, cánceres de cerebro, cánceres de pulmón, cánceres de piel, cánceres de nariz, cánceres de garganta, cánceres de hígado, cánceres de hueso, linfomas, cánceres de páncreas, cánceres de piel, cánceres de intestino, cánceres de recto, cánceres de tiroides, cánceres de vejiga, cánceres de riñón, cánceres de boca, cánceres de estómago, tumores en estado sólido, tumores heterogéneos, tumores homogéneos y similares.

En algunas realizaciones, el sistema y los métodos pueden usarse para detectar cualquier número de aberraciones genéticas que puedan causar o resultar de cánceres. Pueden incluir, entre otros, mutaciones, indels, variaciones del número de copias, transversiones, translocaciones, inversiones, deleciones, aneuploidías, aneuploidías parciales, poliploidías, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones génicas fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de los ácidos nucleicos infección y cáncer.

Además, los sistemas y métodos aquí descritos también pueden utilizarse para ayudar a caracterizar ciertos cánceres. Los datos genéticos producidos a partir del sistema y los métodos de esta divulgación pueden permitir a los profesionales ayudar a caracterizar mejor una forma específica de cáncer. A menudo, los cánceres son heterogéneos tanto en su composición como en su estadificación. Los datos del perfil genético pueden permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes para el diagnóstico o el tratamiento de ese subtipo específico. Esta información también puede proporcionar al sujeto o al médico pistas sobre el pronóstico de un tipo específico de cáncer.

En algunas realizaciones, los sistemas y métodos aquí proporcionados se utilizan para monitorizar cánceres ya conocidos, u otras enfermedades en un sujeto en particular. Esto puede permitir a un sujeto o a un profesional adaptar las opciones de tratamiento en función de la evolución de la enfermedad. En este ejemplo, los sistemas y métodos aquí descritos pueden utilizarse para construir perfiles genéticos de un sujeto particular del curso de la enfermedad. En algunos casos, los cánceres pueden progresar, volviéndose más agresivos y genéticamente inestables. En otros ejemplos, los cánceres pueden permanecer benignos, inactivos o latentes. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad.

Además, los sistemas y métodos aquí descritos pueden ser útiles para determinar la eficacia de una opción de tratamiento particular. En algunas realizaciones, las opciones de tratamiento exitosas pueden en realidad aumentar la cantidad de variación del número de copias o mutaciones detectadas en la sangre del sujeto si el tratamiento tiene éxito, ya que más cánceres pueden morir y desprender ADN. En otras realizaciones, esto puede no ocurrir. En algunas realizaciones, ciertas opciones de tratamiento se correlacionan con los perfiles genéticos de los cánceres a lo largo del tiempo. Esta correlación puede ser útil a la hora de seleccionar una terapia. Además, si se observa que un cáncer está en remisión después del tratamiento, los sistemas y métodos aquí descritos pueden ser útiles para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

Los métodos y sistemas aquí descritos no se limitan a la detección de mutaciones y variaciones del número de copias asociadas únicamente con cánceres. Otras enfermedades e infecciones pueden dar lugar a otros tipos de afecciones que pueden ser adecuadas para la detección precoz y el seguimiento. Por ejemplo, en determinados casos, los trastornos genéticos o las enfermedades infecciosas pueden provocar un determinado mosaicismo genético en un sujeto. Este mosaicismo genético puede causar variaciones en el número de copias y mutaciones que podrían observarse. En algunas realizaciones, el sistema y los métodos de la divulgación también pueden utilizarse para monitorizar los genomas de las células inmunitarias del organismo. Las células inmunitarias, como los linfocitos B, pueden experimentar una rápida expansión clonal ante la presencia de determinadas enfermedades. Las expansiones clonales pueden controlarse mediante la detección de variaciones en el número de copias y pueden controlarse determinados estados inmunitarios. En este ejemplo, el análisis de la variación del número de copias puede realizarse a lo largo del tiempo para producir un perfil de cómo puede estar progresando una enfermedad concreta.

En algunas realizaciones, los métodos de la presente divulgación son aplicables a enfermedades o afecciones autoinmunes o relacionadas con la inmunidad. Tal como se utiliza en el presente documento, "enfermedad o afección autoinmune o relacionada con el sistema inmunitario" puede referirse a cualquier enfermedad, trastorno, o afección que afecte al sistema inmunitario o esté asociada a él. Ejemplos de enfermedades o afecciones autoinmunes o relacionadas con la inmunidad incluyen, entre otras, la inflamación, el síndrome antifosfolípido, el lupus eritematoso sistémico, la artritis reumatoide, la vasculitis autoinmune, la enfermedad celíaca, la tiroiditis autoinmune, la inmunización postransfusional, la incompatibilidad materno-fetal, las reacciones a transfusiones, la deficiencia inmunológica como la deficiencia de IgA, la

inmunodeficiencia común variable, el lupus inducido por fármacos, diabetes mellitus, diabetes de tipo I, diabetes de tipo II, diabetes juvenil, artritis reumatoide juvenil, artritis psoriásica, esclerosis múltiple, inmunodeficiencia, alergias, asma, psoriasis, dermatitis atópica, dermatitis alérgica de contacto, enfermedades crónicas de la piel, esclerosis lateral amiotrófica, lesiones inducidas por quimioterapia, enfermedades de injerto contra huésped, rechazo de trasplante de médula ósea, espondilitis anquilosante, eczema atópico, pénfigo, enfermedad de Behcet's, síndrome de fatiga crónica, fibromialgia, lesiones inducidas por la quimioterapia, miastenia gravis, glomerulonefritis, retinitis alérgica, esclerosis sistémica, lupus eritematoso cutáneo subagudo, lupus eritematoso cutáneo, incluido el lupus eritematoso chilblain, síndrome de Sjogren, nefropatía autoinmunitaria, vasculitis autoinmune, hepatitis autoinmune, carditis autoinmune, encefalitis autoinmune, enfermedades hematológicas mediadas por autoinmunidad, lc-SSc (forma cutánea limitada de la esclerodermia), dc-SSc (forma cutánea difusa de la esclerodermia), tiroiditis autoinmune (AT), enfermedad de Grave' s (GD), miastenia grave, esclerosis múltiple (MS), espondilitis anquilosante, rechazo de trasplantes, envejecimiento inmunitario, enfermedades reumáticas/autoinmunitarias, enfermedad mixta del tejido conjuntivo, espondiloartropatía, psoriasis, artritis psoriásica, miositis, esclerodermia, dermatomiositis, vasculitis autoinmunitaria, enfermedad mixta del tejido conjuntivo, púrpura trombocitopénica idiopática, enfermedad de Crohn, enfermedad adyuvante humana, osteoartritis, artritis crónica juvenil, una espondiloartropatía, una miopatía inflamatoria idiopática, vasculitis sistémica, sarcoidosis, anemia hemolítica autoinmune, trombocitopenia autoinmune, tiroiditis, enfermedad renal inmunomediada, una enfermedad desmielinizante del sistema nervioso central o periférico, polineuropatía desmielinizante idiopática, síndrome de Guillain-Barré, polineuropatía desmielinizante inflamatoria crónica, enfermedad hepatobiliar, hepatitis infecciosa o autoinmune crónica activa, cirrosis biliar primaria, hepatitis granulomatosa, colangitis esclerosante, enfermedad inflamatoria intestinal, enteropatía sensible al gluten, enfermedad de Whipple, una enfermedad autoinmune o inmunomediada de la piel, una enfermedad bullosa de la piel, eritema multiforme, rinitis alérgica, dermatitis atópica, hipersensibilidad alimentaria, urticaria, una enfermedad inmunológica del pulmón, neumonías eosinofílicas, fibrosis pulmonar idiopática, neumonitis por hipersensibilidad, una enfermedad asociada a un trasplante, rechazo de injerto o enfermedad de injerto contra huésped, artritis psoriásica, psoriasis, dermatitis, polimiositis/dermatomiositis, necrólisis epidérmica tóxica, esclerodermia y esclerosis sistémicas, respuestas asociadas a la enfermedad inflamatoria intestinal, enfermedad de Crohn, colitis ulcerosa, síndrome de dificultad respiratoria, síndrome de dificultad respiratoria del adulto (ARDS), meningitis, encefalitis, uveítis, colitis, glomerulonefritis, afecciones alérgicas, eczema, asma, afecciones que implican infiltración de células T y respuestas inflamatorias crónicas, aterosclerosis, miocarditis autoinmune, deficiencia de adhesión de leucocitos, encefalomielitis alérgica, respuestas inmunitarias asociadas a hipersensibilidad aguda y retardada mediada por citocinas y linfocitos T, tuberculosis, sarcoidosis, granulomatosis, incluida la granulomatosis de Wegener, agranulocitosis, vasculitis (incluyendo ANCA), anemia aplásica, anemia de Diamond Blackfan, anemia hemolítica inmune incluyendo anemia hemolítica autoinmune (AIHA), anemia perniciosa, aplasia pura de células rojas (PRCA), deficiencia de Factor VIII, hemofilia A, neutropenia autoinmune, pancitopenia, leucopenia, enfermedades que implican diapedesis de leucocitos, trastornos inflamatorios del sistema nervioso central (CNS), síndrome de lesión multiorgánica, misatenia gravis, enfermedades mediadas por complejos antígeno-anticuerpo, enfermedad de la membrana basal antiglomerular, síndrome de anticuerpos antifosfolípidos, neuritis alérgica, enfermedad de Bechet, síndrome de Castleman, síndrome de Goodpasture, síndrome miasténico de Lambert-Eaton, síndrome de Reynaud, síndrome de Sjorgen, síndrome de Stevens-Johnson, penfigoide bulloso, pénfigo, polendocrinopatías autoinmunes, enfermedad de Reiter, síndrome del hombre rígido, arteritis de células gigantes, nefritis por complejos inmunes, nefropatía IgA, polineuropatías IgM o neuropatía mediada por IgM, púrpura trombocitopénica idiopática (ITP), púrpura trombótica pulsátil (TTP), trombocitopenia autoinmune, enfermedades autoinmunes de testículos y ovarios, como orquitis y ooforitis autoinmunes, hipotiroidismo primario, enfermedades endocrinas autoinmunes, como tiroiditis autoinmune, tiroiditis crónica (tiroiditis de Hashimoto), tiroiditis subcutánea, hipotiroidismo idiopático, enfermedad de Addison, enfermedad de Grave's, síndromes poliglandulares autoinmunes (o síndromes de endocrinopatía poliglandular), síndrome de Sheehan, hepatitis autoinmune, neumonitis intersticial linfoide (HIV), bronquiolitis obliterante (no trasplante) frente a NSIP, síndrome de Guillain-Barré, vasculitis de grandes vasos (incluida la polimialgia reumática y la arteritis de células gigantes (Takayasu' s), vasculitis de vasos sanguíneos medianos (incluida la enfermedad de Kawasaki y la poliarteritis nodosa), espondilitis anquilosante, enfermedad de Berger (nefropatía IgA), glomerulonefritis rápidamente progresiva, cirrosis biliar primaria, esprúe celíaco (enteropatía por gluten), crioglobulinemia, y esclerosis lateral amiotrófica (ALS). En ciertas realizaciones, los métodos de la presente divulgación son aplicables a afecciones inflamatorias que incluyen, entre otras, asma, esclerosis múltiple (por ejemplo, esclerosis múltiple remitente recidivante y esclerosis múltiple secundaria progresiva), artritis (por ejemplo, artritis reumatoide, osteoartritis y artritis psoriásica), lupus eritematoso, y psoriasis.

En algunas realizaciones, los sistemas y métodos de la presente divulgación pueden utilizarse para monitorizar infecciones sistémicas propiamente dichas, como las que pueden ser causadas por un patógeno como una bacteria o un virus. La variación del número de copias o incluso la detección de mutaciones pueden utilizarse para determinar cómo cambia una población de patógenos durante el curso de la infección. Esto puede ser especialmente importante durante las infecciones crónicas, como las infecciones por VIH/sida o hepatitis, en las que los virus pueden cambiar de estado de ciclo vital y/o mutar a formas más virulentas durante el curso de la infección.

En algunas realizaciones, el sistema y los métodos de esta divulgación pueden utilizarse para monitorizar sujetos trasplantados. Por lo general, los tejidos trasplantados sufren cierto grado de rechazo por parte del organismo en el momento del trasplante. Los métodos de la presente divulgación pueden utilizarse para determinar o perfilar las actividades de rechazo del organismo huésped, ya que las células inmunitarias intentan destruir el tejido trasplantado. Esto puede ser útil para controlar el estado del tejido trasplantado, así como para modificar el curso del tratamiento o prevenir el rechazo.

Además, en algunas realizaciones, los métodos de la divulgación pueden usarse para caracterizar la heterogeneidad de una condición anormal en un sujeto, comprendiendo el método generar un perfil genético de polinucleótidos extracelulares en el sujeto, en el que el perfil genético comprende una pluralidad de datos resultantes de análisis de variación de número de copias y de mutación. En algunos casos, incluido el cáncer pero sin limitarse a él, una enfermedad puede ser heterogénea. Las células enfermas pueden no ser idénticas. En el ejemplo del cáncer, se sabe que algunos tumores comprenden diferentes tipos de células tumorales, algunas células en diferentes etapas del cáncer. En algunas realizaciones, la heterogeneidad comprende múltiples focos de enfermedad. De nuevo, en el ejemplo del cáncer, puede haber múltiples focos tumorales, tal vez cuando uno o más focos son el resultado de metástasis que se han extendido desde un sitio primario.

Los métodos de la presente divulgación pueden utilizarse para generar un perfil, huella dactilar o conjunto de datos que sea una suma de información genética derivada de diferentes células en una enfermedad heterogénea. Este conjunto de datos puede incluir análisis de variación del número de copias y de mutaciones, solos o combinados.

Además, los sistemas y métodos de la divulgación pueden utilizarse para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades de origen fetal. Es decir, estas metodologías pueden emplearse en un sujeto embarazado para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades en un sujeto nonato cuyo ADN y otros polinucleótidos pueden co-circular con moléculas maternas. En algunas realizaciones, los sistemas y métodos son útiles para diagnosticar, pronosticar, monitorizar u observar una enfermedad o condición prenatal o relacionada con el embarazo. Tal como se utiliza aquí, el término "enfermedad o afección prenatal o relacionada con el embarazo" se refiere a cualquier enfermedad, trastorno, o afección que afecte a una mujer embarazada, un embrión, o un feto. Las afecciones prenatales o relacionadas con el embarazo también pueden referirse a cualquier enfermedad, trastorno, o afección que se asocie o surja, directa o indirectamente, como consecuencia del embarazo. Estas enfermedades o afecciones pueden incluir todos y cada uno de los defectos de nacimiento, afecciones congénitas, o enfermedades o afecciones hereditarias. Algunos ejemplos de enfermedades prenatales o relacionadas con el embarazo son, entre otros, la enfermedad de Rhesus, la enfermedad hemolítica del recién nacido, la beta-talasemia, la determinación del sexo, la determinación del embarazo, un trastorno genético mendeliano hereditario, aberraciones cromosómicas, una aneuploidía cromosómica fetal, trisomía cromosómica fetal, monosomía cromosómica fetal, trisomía 8, trisomía 13 (Síndrome de Patau), trisomía 16, trisomía 18 (Síndrome de Edwards), trisomía 21 (síndrome de Down), trastornos ligados al cromosoma X, trisomía X (síndrome XXX), monosomía X (síndrome de Turner), síndrome XXY, síndrome XYY, síndrome XYY, síndrome XXXY, síndrome XXYY, síndrome XYY, síndrome XXXXX, síndrome XXXXY, síndrome XXXYY, síndrome XXYYY, síndrome X frágil, retraso del crecimiento fetal, fibrosis quística, una hemoglobinopatía, muerte fetal, síndrome alcohólico fetal, anemia falciforme, hemofilia, síndrome de Klinefelter, dup(17)(pl l.2pl.2), endometriosis, enfermedad de Pelizaeus-Merzbacher, síndrome dup(22)(ql l.2ql l.2), síndrome del ojo de gato, síndrome cri-du-chat, síndrome de Wolf-Hirschhorn, síndrome de Williams-Beuren, enfermedad de Charcot-Marie-Tooth, neuropatía con tendencia a las parálisis por presión, síndrome de Smith-Magenis, neurofibromatosis, síndrome de Alagille, síndrome velocardiofacial, síndrome de DiGeorge, deficiencia de esteroides sulfatasa, síndrome de Prader-Willi, síndrome de Kallmann, microftalmia con defectos cutáneos lineales, hipoplasia suprarrenal, deficiencia de glicerol quinasa, enfermedad de Pelizaeus-Merzbacher, factor determinante de los testículos en Y, azospermia (factor a), azospermia (factor b), azospermia (factor c), deleción lp36, fenilcetonuria, enfermedad de Tay-Sachs, hiperplasia suprarrenal, anemia de Fanconi, atrofia muscular espinal, distrofia muscular de Duchenne, distrofia muscular de Huntington, distrofia miotónica, translocación robertsoniana, síndrome de Angelman, esclerosis tuberosa, ataxia telangieltasia, espina bífida abierta, defectos del tubo neural, defectos de la pared ventral, pequeño para la edad gestacional, citomegalovirus congénito, acondroplasia, síndrome de MarfanSíndrome de Marfan, hipotiroidismo congénito, toxoplasmosis congénita, deficiencia de biotinidasa, galactosemia, enfermedad de la orina con olor a jarabe de arce, homocistinuria, deficiencia de acil Co-A deshidrogenasa de cadena media, defectos congénitos estructurales, defectos cardíacos, extremidades anormales, pie zambo, anencefalia, rinencefalia/holoprosencefalia, hidrocefalia, anoftalmos/microftalmos, anotia/microtia, transposición de grandes vasos, tetralogía de Fallot, síndrome del corazón izquierdo hipoplásico, coartación de aorta, paladar hendido sin labio leporino, labio leporino con o sin paladar hendido, atresia/estenosis esofágica con o sin fístula, atresia/estenosis del intestino delgado, atresia/estenosis anorrectal, hipospadias, sexo indeterminado, agenesia renal, riñón quístico, polidactilia preaxial, defectos de reducción de las extremidades, hernia diafragmática, ceguera, cataratas, problemas visuales, hipoacusia, sordera, adrenoleucodistrofia ligada al cromosoma X, síndrome de Rett, trastornos lisosomales, parálisis cerebral, autismo, aglosia, albinismo, albinismo ocular, albinismo oculocutáneo, diabetes gestacional, malformación de Arnold-Chiari, síndrome de CHARGE, hernia diafragmática congénita, braquidactilia, aniridia, pie y mano hendidos, heterocromía, oreja de Dwarn, síndrome de Ehlers Danlos, epidermólisis bullosa, enfermedad de Gorham, enfermedad de Hashimoto, hidropesía fetal, hipotonía, síndrome de Klippel-Feil, distrofia muscular, osteogénesis imperfecta, progeria, síndrome de Smith Lemli Opitz, cromatelopsia, enfermedad linfoproliferativa ligada al cromosoma X, onfalocele, gastrosquisis, preeclampsia, eclampsia, parto prematuro, nacimiento prematuro, aborto espontáneo, retraso del crecimiento intrauterino, embarazo ectópico, hiperémesis gravídica, náuseas matutinas, o probabilidad de éxito de la inducción del parto.

Además, en algunas realizaciones, los informes se envían y se accede a ellos electrónicamente a través de Internet. En ciertas realizaciones, el análisis de los datos de la secuencia se realiza en un lugar distinto al del sujeto. El informe se genera y se transmite al lugar donde se encuentra el sujeto. A través de un ordenador con conexión a Internet, el sujeto accede a los informes que reflejan su carga tumoral.

La información anotada puede ser utilizada por un proveedor de asistencia sanitaria para seleccionar otras opciones de tratamiento farmacológico y/o proporcionar información sobre opciones de tratamiento farmacológico a una compañía de seguros. El método puede incluir la anotación de las opciones de tratamiento farmacológico para una afección en, por ejemplo, las Guías de Práctica Clínica en Oncología de la NCCN o las guías de práctica clínica de la Sociedad Americana de Oncología Clínica (ASCO).

Las opciones de tratamiento farmacológico estratificadas en un informe pueden anotarse en el informe enumerando opciones adicionales de tratamiento farmacológico. Un tratamiento farmacológico adicional puede ser un medicamento aprobado por la FDA para un uso en una indicación no autorizada. Una disposición de la Ley Ómnibus de Reconciliación Presupuestaria (OBRA) de 1993 obliga a Medicare a cubrir los usos en una indicación no autorizada de los fármacos contra el cáncer incluidos en los compendios médicos estándar. Los fármacos utilizados para anotar las listas pueden encontrarse en compendios aprobados por los CMS, incluidos el National Comprehensive Cancer Network (NCCN) Drugs and Biologies Compendium", Thomson Micromedex DrugDex®, el compendio de farmacología clínica de Elsevier Gold Standard y el American Hospital Formulary Service-Drug Information Compendium®.

Las opciones de tratamiento farmacológico pueden anotarse enumerando un fármaco experimental que puede ser útil en el tratamiento de un cáncer con uno o más marcadores moleculares de un estado particular. El fármaco experimental puede ser un fármaco para el que se disponga de datos in vitro, datos in vivo, datos de modelos animales, datos de ensayos preclínicos o datos de ensayos clínicos. Los datos pueden publicarse en la literatura médica revisada por pares que se encuentra en las revistas enumeradas en el Manual de políticas de prestaciones de Medicare de los CMS, incluidas, por ejemplo, American Journal of Medicine, Annals of Internal Medicine, Annals of Oncology, Annals of Surgical Oncology, Biology of Blood and Marrow Transplantation, Blood, Bone Marrow Transplantation, British Journal of Cancer, British Journal of Hematology, British Medical Journal, Cancer, Clinical Cancer Research, Drugs, European Journal of Cancer (antes European Journal of Cancer and Clinical Oncology), Gynecologic Oncology, International Journal of Radiation, Oncology, Biology, and Physics, The Journal of the American Medical Association, Journal of Clinical Oncology, Journal of the National Cancer Institute, Journal of the National Comprehensive Cancer Network (NCCN), Journal of Urology, Lancet, Lancet Oncology, Leukemia, The New England Journal of Medicine y Radiation Oncology.

Las opciones de tratamiento farmacológico pueden anotarse proporcionando un enlace en un informe electrónico que conecte un fármaco de la lista con información científica relativa al fármaco. Por ejemplo, se puede proporcionar un enlace a información sobre un ensayo clínico de un medicamento (clinicaltrials.gov). Si el informe se facilita a través de un sitio web de ordenador o ordenador, el enlace puede ser una nota a pie de página, un hipervínculo a un sitio web, un cuadro emergente, o un cuadro volante con información, etc. El informe y la información anotada pueden facilitarse en un formulario impreso, y las anotaciones pueden ser, por ejemplo, una nota a pie de página de una referencia.

La información para anotar una o más opciones de tratamiento farmacológico en un informe puede ser proporcionada por una entidad comercial que almacene información científica. Un profesional sanitario puede tratar a un sujeto, como un paciente con cáncer, con un fármaco experimental incluido en la información anotada, y el profesional sanitario puede acceder a la opción de tratamiento farmacológico anotada, recuperar la información científica (por ejemplo, imprimir un artículo de una revista médica) y enviarla (por ejemplo, un artículo impreso de una revista) a una compañía de seguros junto con una solicitud de reembolso por proporcionar el tratamiento farmacológico. Los médicos pueden utilizar cualquiera de los diversos códigos de grupos relacionados por el diagnóstico (DRG) para permitir el reembolso.

Una opción de tratamiento farmacológico en un informe también se puede anotar con información relativa a otros componentes moleculares en una vía a la que afecta un fármaco (por ejemplo, información sobre un fármaco que se dirige a una quinasa corriente abajo de un receptor de superficie celular que es una diana farmacológica). La opción de tratamiento farmacológico puede anotarse con información sobre fármacos dirigidos a uno o más componentes de otras vías moleculares. La identificación y/o anotación de la información relacionada con las vías puede externalizarse o subcontratarse a otra empresa.

La información anotada puede ser, por ejemplo, el nombre de un fármaco (por ejemplo, un fármaco aprobado por la FDA para uso en una indicación no autorizada; un fármaco que se encuentra en un compendio aprobado por la CMS, y/o un fármaco descrito en un artículo de una revista científica (médica)), información científica relativa a una o más opciones de tratamiento farmacológico, uno o más enlaces a información científica relativa a uno o más fármacos, información de ensayos clínicos relativos a uno o más fármacos (por ejemplo, información de clinicaltrials.gov/), uno o más enlaces a citas de información científica relativa a fármacos, etc.

La información anotada puede insertarse en cualquier lugar de un informe. La información anotada puede insertarse en varios lugares de un informe. La información anotada puede insertarse en un informe cerca de una sección sobre opciones de tratamiento farmacológico estratificado. La información anotada puede insertarse en un informe en una página separada de las opciones de tratamiento farmacológico estratificadas. Un informe que no contenga opciones estratificadas de tratamiento farmacológico puede anotarse con información.

El sistema también puede incluir informes sobre los efectos de los fármacos en la muestra (por ejemplo, células tumorales) aislada de un sujeto (por ejemplo, paciente con cáncer). Puede establecerse un cultivo in vitro a partir de un

tumor de un paciente con cáncer utilizando diversas técnicas. El sistema también puede incluir el cribado de alto rendimiento de fármacos no aprobados por la FDA o fármacos experimentales utilizando dicho cultivo in vitro y/o modelo de xenoinjerto. El sistema también puede incluir la monitorización del antígeno tumoral para la detección de recurrencia.

El sistema puede proporcionar acceso por Internet a los informes de un sujeto con cáncer. El sistema puede utilizar un secuenciador de ADN portátil o un secuenciador de ADN de sobremesa. El secuenciador de ADN es un instrumento científico utilizado para automatizar el proceso de secuenciación del ADN. Dada una muestra de ADN, se utiliza un secuenciador de ADN para determinar el orden de las cuatro bases: adenina, guanina, citosina, y timina. El orden de las bases de ADN se presenta como una cadena de texto, denominada lectura. Algunos secuenciadores de ADN también pueden considerarse instrumentos ópticos, ya que analizan señales luminosas procedentes de fluorocromos unidos a nucleótidos.

El secuenciador de ADN puede aplicar el método de secuenciación de Gilbert basado en la modificación química del ADN seguida de la escisión en bases específicas, o puede aplicar la técnica de Sanger que se basa en la terminación de la cadena de dideoxinucleótidos. El método Sanger se popularizó debido a su mayor eficacia y baja radiactividad. El secuenciador de ADN puede utilizar técnicas que no requieren la amplificación del ADN (reacción en cadena de la polimerasa - PCR), lo que acelera la preparación de la muestra antes de la secuenciación y reduce los errores. Además, se recogen datos de secuenciación de las reacciones provocadas por la adición de nucleótidos en la cadena complementaria en tiempo real. Por ejemplo, los secuenciadores de ADN pueden utilizar un método denominado molécula única en tiempo real (SMRT), en el que los datos de secuenciación se producen mediante la luz (captada por una cámara) emitida cuando se añade un nucleótido a la cadena complementaria mediante enzimas que contienen colorantes fluorescentes. Como alternativa, los secuenciadores de ADN pueden utilizar sistemas electrónicos basados en tecnologías de detección de nanoporos.

Los datos son enviados por los secuenciadores de ADN a través de una conexión directa o por internet a un ordenador para su procesamiento. Los aspectos de procesamiento de datos del sistema pueden implementarse en circuitos electrónicos digitales, o en hardware informático, firmware, software, o en combinaciones de ellos. El aparato de procesamiento de datos de la presente divulgación puede implementarse en un producto de programa de ordenador tangiblemente incorporado en un dispositivo de almacenamiento legible por máquina para su ejecución por un procesador programable; y los pasos del método de procesamiento de datos de la presente divulgación pueden ser realizados por un procesador programable que ejecuta un programa de instrucciones para realizar funciones de la presente divulgación operando sobre datos de entrada y generando salida. Los aspectos de procesamiento de datos de la presente divulgación pueden implementarse ventajosamente en uno o más programas de ordenador que son ejecutables en un sistema programable que incluye al menos un procesador programable acoplado para recibir datos e instrucciones desde y para transmitir datos e instrucciones a un sistema de almacenamiento de datos, al menos un dispositivo de entrada y al menos un dispositivo de salida. Cada programa informático puede implementarse en un lenguaje de programación de alto nivel procedimental u orientado a objetos, o en lenguaje ensamblador o de máquina, si se desea; y, en cualquier caso, el lenguaje puede ser un lenguaje compilado o interpretado. Los procesadores adecuados incluyen, a modo de ejemplo, microprocesadores de propósito general y especial. Generalmente, un procesador recibirá instrucciones y datos de una memoria de sólo lectura y/o de una memoria de acceso aleatorio. Los dispositivos de almacenamiento adecuados para incorporar de forma tangible instrucciones y datos de programas informáticos incluyen todas las formas de memoria no volátil, incluyendo, a modo de ejemplo, dispositivos de memoria semiconductores, como EPROM, EEPROM y dispositivos de memoria flash; discos magnéticos, como discos duros internos y discos extraíbles; discos magneto-ópticos; y discos CD-ROM. Todo lo anterior puede complementarse con ASIC (circuitos integrados de aplicación específica) o incorporarse a ellos.

Para permitir la interacción con un usuario, la presente divulgación puede implementarse utilizando un sistema informático que tenga un dispositivo de visualización, como un monitor o una pantalla LCD (pantalla de cristal líquido) para mostrar información al usuario, y dispositivos de entrada mediante los cuales el usuario pueda proporcionar información al sistema informático, como un teclado, un dispositivo señalador bidimensional, como un ratón o una bola rastreadora, o un dispositivo señalador tridimensional, como un guante de datos o un ratón giroscópico. El sistema informático puede programarse para proporcionar una interfaz gráfica de usuario a través de la cual los programas informáticos interactúan con los usuarios. El sistema informático puede programarse para proporcionar una interfaz de visualización tridimensional de realidad virtual.

Muestras de Ensayo

Los métodos aquí divulgados pueden comprender el aislamiento de uno o más polinucleótidos.

Un polinucleótido puede comprender cualquier tipo de ácido nucleico, como ADN y/o ARN. Por ejemplo, si un polinucleótido es ADN, puede ser ADN genómico, ADN complementario (ADNc), o cualquier otro ácido desoxirribonucleico. Un polinucleótido también puede ser un ácido nucleico libre de células, como el ADN libre de células (ADNcf). Por ejemplo, el polinucleótido puede ser ADNcf circulante. El ADNcf circulante puede comprender ADN desprendido de células corporales por apoptosis o necrosis. El ADNcf desprendido por apoptosis o necrosis puede proceder de células corporales normales. Cuando se produce un crecimiento anormal del tejido, como en el caso del cáncer, puede desprenderse ADN tumoral. El ADNcf circulante puede comprender ADN tumoral circulante (ADNct). Como

se describe en el presente documento, los métodos de la presente divulgación permiten al trabajador cualificado determinar si el origen de un locus genético (por ejemplo, una variante en un locus genético) es germinal o somático a partir del ADNcf, sin necesidad de información de secuencia separada del ADN genómico.

Un polinucleótido puede ser bicatenario o monocatenario. Alternativamente, un polinucleótido puede comprender una combinación de una porción bicatenaria y una porción monocatenaria.

Una muestra puede ser cualquier muestra biológica aislada de un sujeto. Por ejemplo, una muestra puede comprender, sin limitación, fluido corporal, sangre entera, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejido, fluido sinovial, fluido linfático, fluido de ascitis, líquido intersticial o extracelular, el líquido que se encuentra en los espacios entre las células, incluido el líquido crevicular gingival, médula ósea, líquido cefalorraquídeo, saliva, mucosa, esputo, semen, sudor, orina, líquido del cepillado nasal, líquido de una citología vaginal, o cualquier otro fluido corporal. Un fluido corporal puede incluir saliva, sangre, o suero. Por ejemplo, un polinucleótido puede ser ADN libre de células aislado de un fluido corporal, por ejemplo, sangre o suero. Una muestra también puede ser una muestra tumoral, que puede obtenerse de un sujeto por diversos enfoques, incluyendo, entre otros, venopunción, excreción, eyaculación, masaje, biopsia, aspiración con aguja, lavado, raspado, incisión quirúrgica, o intervención u otros enfoques. Una muestra puede ser una muestra libre de células (por ejemplo, que no contenga células).

Una muestra puede comprender un volumen de plasma que contenga moléculas de ADN libres de células. Una muestra puede comprender un volumen de plasma suficiente para alcanzar una profundidad de lectura determinada. Un volumen de plasma muestreado puede ser de al menos 0,5 mililitros (mL), 1 mL, 5 mL 10 mL, 20 mL, 30 mL, o 40 mL. Un volumen de plasma muestreado como máximo de 0,5 mL, 1 mL, 5 mL 10 mL, 20 mL, 30 mL, o 40 mL. El volumen de plasma muestreado puede ser de 5 a 20 mL. El volumen de plasma muestreado puede ser de 10 ml a 20 mL.

Una muestra puede comprender diversas cantidades de ácido nucleico que contengan equivalentes genómicos. Por ejemplo, una muestra de unos 30 ng de ADN puede contener unos 10.000 (10*) equivalentes de genoma humano haploide y, en el caso del ADNcf, unos 200.000 millones (2x10) de moléculas de polinucleótidos individuales. Del mismo modo, una muestra de unos 100 ng de ADN puede contener unos 30.000 equivalentes de genoma humano haploide y, en el caso del ADNcf, unos 600.000 millones de moléculas individuales.

Una muestra puede comprender ácidos nucleicos de diferentes fuentes. Por ejemplo, una muestra puede comprender ADN de la línea germinal o ADN somático. Una muestra puede comprender ácidos nucleicos portadores de mutaciones. Por ejemplo, una muestra puede comprender ADN portador de mutaciones germinales y/o mutaciones somáticas. Una muestra también puede comprender ADN portador de mutaciones asociadas al cáncer (por ejemplo, mutaciones somáticas asociadas al cáncer). En algunas realizaciones, una muestra comprende uno o más de: una sustitución de una sola base, una variación del número de copias, un indel, una fusión génica, una transversión, una translocación, una inversión, una deleción, una aneuploidía, una aneuploidía parcial, una poliploidía, una inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones cromosómicas, un truncamiento génico, una amplificación génica, una duplicación génica, una lesión cromosómica, una lesión de ADN, cambios anormales en las modificaciones químicas del ácido nucleico, cambios anormales en los patrones epigenéticos, cambios anormales en las distribuciones de fragmentos de ácido nucleico (por ejemplo, cfDNA) a través de regiones genómicas, cambios anormales en las distribuciones de longitudes de fragmentos de ácido nucleico (por ejemplo, cfDNA), y cambios anormales en la metilación del ácido nucleico.

Los métodos aquí descritos pueden comprender la obtención de cierta cantidad de moléculas de ácido nucleico, por ejemplo, moléculas de ácido nucleico libres de células a partir de una muestra. Por ejemplo, el método puede comprender la obtención de hasta aproximadamente 600 ng, hasta aproximadamente 500 ng, hasta aproximadamente 400 ng, hasta aproximadamente 300 ng, hasta aproximadamente 200 ng, hasta aproximadamente 100 ng, hasta aproximadamente 50 ng, o hasta aproximadamente 20 ng de moléculas de ácido nucleico libres de células a partir de una muestra. El método puede comprender la obtención de al menos 1 femtogramo (fg), al menos 10 fg, al menos 100 fg, al menos 1 picogramo (pg), al menos 10 pg, al menos 100 pg, al menos 1 ng, al menos 10 ng, al menos 100 ng, al menos 150 ng, o al menos 200 ng de moléculas de ácido nucleico libres de células. El método puede comprender la obtención de como máximo 1 femtogramo (fg), como máximo 10 fg, como máximo 100 fg, como máximo 1 picogramo (pg), como máximo 10 pg, como máximo 100 pg, como máximo 1 ng, como máximo 10 ng, como máximo 100 ng, como máximo 150 ng, o como máximo 200 ng de moléculas de ácido nucleico libres de células. El método puede comprender la obtención de 1 femtogramo (fg) a 200 ng, 1 picogramo (pg) a 200 ng, 1 ng a 100 ng, 10 ng a 150 ng, 10 ng a 200 ng, 10 ng a 300 ng, 10 ng a 400 ng, 10 ng a 500 ng, 10 ng a 600 ng, 10 ng a 700 ng, 10 ng a 800 ng, 10 ng a 900 ng, o 10 ng a 1000 ng de moléculas de ácido nucleico libres de células. Una cantidad de moléculas de ácido nucleico libres de células puede ser equivalente a un número de copias del genoma haploide. Dado que una copia del genoma haploide tiene una masa de unos 3,3 picogramos (pg), cada nanogramo (ng) de moléculas nucleicas libres de células puede equivaler a unas 300 copias del genoma haploide. Por ejemplo, 5 ng de moléculas de ácido nucleico libres de células pueden equivaler a 1500 copias del genoma.

Un ácido nucleico libre de células puede ser cualquier ácido nucleico extracelular que no esté unido a una célula. Un ácido nucleico libre de células puede ser un ácido nucleico circulante en la sangre. Alternativamente, un ácido nucleico libre de células puede ser un ácido nucleico en otro fluido corporal divulgado en el presente documento, por ejemplo, la

orina. Un ácido nucleico libre de células puede ser un ácido desoxirribonucleico ("ADN"), por ejemplo, ADN genómico, ADN mitocondrial, o un fragmento del mismo. Un ácido nucleico libre de células puede ser un ácido ribonucleico ("ARN"), por ejemplo ARNm, ARN de interferencia corta (siARN), microARN (miARN), ARN circulante (ARNc), ARN de transferencia (ARNt), ARN ribosómico (ARNr), ARN nucleolar pequeño (snoARN), ARN que interactúa con Piwi (piARN), ARN no codificante largo (ARNnc largo), o un fragmento del mismo. En algunos casos, un ácido nucleico libre de células es un híbrido ADN/ARN. Un ácido nucleico libre de células puede ser bicatenario, monocatenario, o un híbrido de los mismos. Un ácido nucleico libre de células puede liberarse al fluido corporal a través de la secreción o de procesos de muerte celular, por ejemplo, necrosis celular y apoptosis.

Un ácido nucleico libre de células puede comprender una o más modificaciones epigenéticas. Por ejemplo, un ácido nucleico libre de células puede estar acetilado, metilado, ubiquitilado, fosforilado, sumoilado, ribosilado, y/o citrulinado. Por ejemplo, un ácido nucleico libre de células puede ser ADN libre de células metilado.

El ADN libre de células suele tener una distribución de tamaños de entre 110 y 230 nucleótidos, con una moda de 168 nucleótidos. Un segundo pico menor detectado en los ensayos que cuantifican la longitud de la molécula de ácido nucleico libre de células tiene un rango entre 240 y 440 nucleótidos. También hay picos adicionales de nucleótidos de orden superior en longitudes mayores.

En algunas realizaciones de la presente divulgación, los ácidos nucleicos libres de células pueden tener como máximo 1000 nucleótidos (nt) de longitud, como máximo 500 nucleótidos de longitud, como máximo 400 nucleótidos de longitud, como máximo 300 nucleótidos de longitud, como máximo 250 nucleótidos de longitud, como máximo 225 nucleótidos de longitud, como máximo 200 nucleótidos de longitud, como máximo 190 nucleótidos de longitud, como máximo 180 nucleótidos de longitud, como máximo 170 nucleótidos de longitud, como máximo 160 nucleótidos de longitud, como máximo 150 nucleótidos de longitud, como máximo 140 nucleótidos de longitud, como máximo 130 nucleótidos de longitud, como máximo 120 nucleótidos de longitud, como máximo 110 nucleótidos de longitud, o como máximo 100 nucleótidos de longitud.

En algunas realizaciones de la presente divulgación, los ácidos nucleicos libres de células pueden tener al menos 1000 nucleótidos de longitud, al menos 500 nucleótidos de longitud, al menos 400 nucleótidos de longitud, al menos 300 nucleótidos de longitud, al menos 250 nucleótidos de longitud, al menos 225 nucleótidos de longitud, al menos 200 nucleótidos de longitud, al menos 190 nucleótidos de longitud, al menos 180 nucleótidos de longitud, al menos 170 nucleótidos de longitud, al menos 160 nucleótidos de longitud, al menos 150 nucleótidos de longitud, al menos 140 nucleótidos de longitud, al menos 130 nucleótidos de longitud, al menos 120 nucleótidos de longitud, al menos 110 nucleótidos de longitud, o al menos 100 nucleótidos de longitud. Los ácidos nucleicos libres de células pueden tener entre 140 y 180 nucleótidos de longitud.

En algunas realizaciones de la presente divulgación, los ácidos nucleicos libres de células en un sujeto pueden derivar de un tumor. Por ejemplo, el ADN libre de células aislado de un sujeto puede incluir ADN tumoral circulante (ADNct). La secuenciación de nueva generación permite detectar y medir mutaciones raras. La detección de mutaciones relativas a la secuencia de la línea germinal en una fracción de ADN libre de células puede indicar la presencia de ADNct, indicando así la presencia de un tumor. La secuenciación del ADN libre de células puede permitir la detección de una variante genética que se sabe que indica la presencia de cáncer. Por ejemplo, la secuenciación del ADN libre de células puede permitir la detección de mutaciones en genes relacionados con el cáncer.

Aislamiento y Extracción

Los polinucleótidos libres de células pueden ser de origen fetal (a través de fluido tomado de un sujeto embarazado), o pueden derivarse de tejido del propio sujeto. Los polinucleótidos libres de células pueden proceder de tejido sano, de tejido enfermo, como tejido tumoral, o de un órgano trasplantado.

En algunas realizaciones, los polinucleótidos libres de células se derivan de una muestra de sangre o de una fracción de la misma. Por ejemplo, se puede tomar una muestra de sangre (por ejemplo, de unos 10 a unos 30 mis) de un sujeto, centrifugarla para eliminar las células, y utilizar el plasma resultante para la extracción de ADNcf.

El aislamiento y la extracción de polinucleótidos pueden realizarse mediante la recogida de fluidos corporales utilizando diversas técnicas. En algunos casos, la recogida puede comprender la aspiración de un fluido corporal de un sujeto utilizando una jeringa. En otros casos, la recogida puede realizarse mediante pipeteo o recogida directa del fluido en un recipiente colector.

Después de la recolección del fluido corporal, los polinucleótidos pueden aislarse y extraerse utilizando una variedad de técnicas utilizadas en el arte. En algunos casos, el ADN libre de células puede aislarse, extraerse y prepararse utilizando kits disponibles en el mercado, como el protocolo Qiamp® Circulating Nucleic Acid Kit de Qiagen. En otros ejemplos, puede utilizarse el protocolo del kit Qiagen Qubit™ dsDNA HS Assay, el kit Agilent™ DNA 1000 o el protocolo TruSeq™ Sequencing Library Preparation; Low- Throughput (LT).

Generalmente, los polinucleótidos libres de células pueden ser extraídos y aíslan de fluidos corporales a través

de una etapa de partición en el que el ADN libre de células, como se encuentra en solución, se separa de las células y otros componentes no solubles del fluido corporal. La separación puede incluir, entre otras, técnicas como la centrifugación o la filtración. En otros casos, es posible que las células no se separen primero del ADN libre de células, sino que se lisen. Por ejemplo, el ADN genómico de células intactas puede dividirse mediante precipitación selectiva. La partición de muestras puede combinarse con el marcado de ácidos nucleicos con identificadores (como identificadores que comprenden códigos de barras), o puede realizarse en un método sin el uso de un identificador. Una muestra puede dividirse en particiones de forma que cada una de ellas pueda llevar un código de barras independiente (por ejemplo, con un código de barras único por partición), y los datos de secuenciación de las particiones puedan recombinarse posteriormente. Una muestra puede dividirse en particiones, y las moléculas de ácido nucleico marcarse de forma no única entre sí dentro de una partición, o entre particiones. En algunas realizaciones, una muestra puede dividirse en particiones sin utilizar identificadores. En un ejemplo, una muestra de ADNcf se divide en 4 o más particiones, en las que cada partición es una ubicación espacialmente direccionable. La preparación de la muestra y la secuenciación se realizan en cada partición espacialmente direccionable, y la bioinformática puede utilizar la ubicación direccionable para seguir identificando una molécula única. En un ejemplo, las moléculas de ácido nucleico pueden dividirse en particiones, por ejemplo, que contengan diferentes tipos de moléculas de ácido nucleico (por ejemplo, ácidos nucleicos bicatenarios como el ADN y/o ácidos nucleicos monocatenarios como el ARN y/o el ADN monocatenario). Los polinucleótidos libres de células, incluido el ADN, pueden permanecer solubles y separarse del ADN genómico insoluble y extraerse. Generalmente, tras la adición de tampones y otras etapas de lavado específicas de los distintos kits, el ADN puede precipitarse mediante precipitación con isopropanol. Pueden utilizarse otros pasos de limpieza, como columnas de sílice o perlas (como perlas magnéticas) para eliminar contaminantes o sales. Las etapas generales pueden optimizarse para aplicaciones específicas. Los polinucleótidos portadores a granel no específicos, por ejemplo, pueden añadirse a lo largo de la reacción para optimizar ciertos aspectos del procedimiento, como el rendimiento.

En algunas realizaciones, una muestra de plasma es tratada para degradar la proteinasa K y el ADN es precipitado con isopropanol y posteriormente capturado en una columna Qiagen. A continuación, el ADN puede eluirse (por ejemplo, utilizando 100 microlitros (µl) de eluyente, como agua o tampón de elución Tris-EDTA (TE)). En algunas realizaciones, una parte del ADN puede seleccionarse en función del tamaño (por ejemplo, ADN de 500 nucleótidos o menos de longitud), por ejemplo, utilizando perlas de inmovilización reversible en fase sólida (SPRI), como las perlas AgenCourt®AMPure®. En algunas realizaciones, el ADN puede resuspenderse en un volumen más pequeño, como 30 µl de agua, y comprobar la distribución de tamaños del ADN (por ejemplo, para comprobar si hay un pico mayor en 166 nucleótidos y un pico menor en 330 nucleótidos). Aproximadamente 5 ng de ADN pueden equivaler a unos 1500 equivalentes de genoma haploide ("HGE").

Tras la extracción, las muestras pueden producir hasta 1 microgramo (µg) de ADN, hasta 800 ng de ADN, hasta 500 ng de ADN, hasta 300 ng de ADN, hasta 250 ng de ADN, hasta 200 ng de ADN, hasta 180 ng de ADN, hasta 160 ng de ADN, hasta 140 ng de ADN, hasta 120 ng de ADN, hasta 100 ng de ADN, hasta 90 ng de ADN, hasta 80 ng de ADN, hasta 70 ng de ADN, hasta 60 ng de ADN, hasta 50 ng de ADN, hasta 40 ng de ADN, hasta 30 ng de ADN, hasta 20 ng de ADN, hasta 10 ng de ADN, hasta 9 ng de ADN, hasta 8 ng de ADN, hasta 7 ng de ADN, hasta 6 ng de ADN, hasta 5 ng de ADN, hasta 4 ng de ADN, hasta 3 ng de ADN, hasta 2 ng de ADN, o hasta 1 ng de ADN.

Después de la extracción, las muestras pueden producir al menos 1 ng de ADN, al menos 3 ng de ADN, al menos 5 ng de ADN, al menos 7 ng de ADN, al menos 10 ng de ADN, al menos 20 ng de ADN, al menos 30 ng de ADN, al menos 40 ng de ADN, al menos 50 ng de ADN, al menos 70 ng de ADN, al menos 100 ng de ADN, al menos 150 ng de ADN, al menos 200 ng de ADN, al menos 250 ng de ADN, al menos 300 ng de ADN, al menos 400 ng de ADN, al menos 500 ng de ADN, o al menos 700 ng de ADN.

Uno o más de los ácidos nucleicos libres de células pueden aislarse de un fragmento celular de una muestra. En algunos casos, uno o más de los ácidos nucleicos libres de células se aíslan de membrana, orgánulos celulares, nucleosomas, exosomas, o núcleo, mitocondrias, retículo endoplásmico rugoso, ribosomas, retículo endoplásmico liso, cloroplastos, aparato de Golgi, cuerpos de Golgi, glicoproteínas, glicolípidos, cisternas, liposomas, peroxisomas, glioxisomas, centríolo, citoesqueleto, lisosomas, cilios, flagelo, vacuola contráctil, vesículas, envolturas nucleares, vacuolas, microtúbulos, nucléolos, membrana plasmática, endosomas, cromatinas, o una combinación de los mismos. Uno o más ácidos nucleicos libres de células pueden aislarse de uno o más exosomas. En algunos casos, uno o más de los ácidos nucleicos libres de células se aíslan de uno o más ácidos nucleicos unidos a la superficie celular.

Purificación del ADN libre de células puede realizarse utilizando cualquier metodología, incluyendo, pero sin limitarse a, el uso de kits comerciales y protocolos proporcionados por empresas como Sigma Aldrich, Life Technologies, Promega, Affymetrix, IBI o similares. También puede haber kits y protocolos no comercializados.

Tras el aislamiento, en algunos casos, los polinucleótidos libres de células pueden premezclarse con uno o más materiales adicionales, como uno o más reactivos (por ejemplo, ligasa, proteasa, polimerasa) antes de la secuenciación.

El ADN libre de células puede secuenciarse con una profundidad de lectura suficiente para detectar una variante genética con una frecuencia en una muestra tan baja como el 0,0005%. El ADN libre de células puede secuenciarse con una profundidad de lectura suficiente para detectar una variante genética con una frecuencia en una muestra tan baja como el 0,001%. El ADN libre de células puede secuenciarse con una profundidad de lectura suficiente para detectar una

variante genética con una frecuencia en una muestra tan baja como 1,0%, 0,75%, 0,5%, 0,25%, 0,1%, 0,075%, 0,05%, 0,025%, 0,01%, o 0,005%. Así, la secuenciación del ADN libre de células permite una detección muy sensible del cáncer en un sujeto.

Los métodos aquí descritos pueden utilizarse para detectar cáncer en un sujeto. El ADN libre de células puede secuenciarse en sujetos de los que no se sabe si tienen cáncer, o de los que se sospecha que lo tienen, para diagnosticar la presencia o ausencia de un cáncer. La secuenciación del ADN libre de células proporciona un método no invasivo para la detección precoz del cáncer o para la "biopsia" de un cáncer conocido. El ADN libre de células puede secuenciarse en sujetos diagnosticados de cáncer para obtener información sobre éste. El ADN libre de células puede secuenciarse en sujetos antes y después del tratamiento contra el cáncer para determinar la eficacia del tratamiento.

Un sujeto puede ser sospechoso de tener cáncer o puede no serlo. Un sujeto puede haber experimentado síntomas compatibles con un diagnóstico de cáncer. Un sujeto puede no haber experimentado ningún síntoma, o puede haber mostrado síntomas no compatibles con el cáncer. Un sujeto puede haber sido diagnosticado con un cáncer basado en métodos de imagen biológica. Un sujeto puede no tener un cáncer detectable por métodos de imagen. Los métodos de obtención de imágenes pueden ser tomografía por emisión de positrones, resonancia magnética, rayos X, tomografía axial computerizada, ultrasonidos, o una combinación de los mismos.

Un sujeto puede presentar un cáncer. Alternativamente, un sujeto puede no presentar un cáncer detectable. En algunos casos, un sujeto que no presenta un cáncer detectable puede tener un cáncer, pero no tener síntomas detectables. Los sujetos de los que no se sabe si tienen cáncer, o de los que se sospecha que lo tienen, pueden tener un cáncer no detectable mediante diversos métodos de cribado del cáncer. La ausencia de cáncer puede detectarse mediante diversos métodos de diagnóstico por imagen. Los métodos de obtención de imágenes pueden incluir, por ejemplo, tomografía por emisión de positrones, resonancia magnética, rayos X, tomografía axial computerizada, endoscopia, ultrasonidos, o una combinación de los mismos. En el caso de un sujeto que no se sabe que tiene cáncer o del que se sospecha que lo tiene, pruebas como la biopsia de tejido, la aspiración de médula ósea, las pruebas de Papanicolaou, las pruebas de sangre oculta en heces, la detección de biomarcadores proteicos, por ejemplo, la prueba del antígeno prostático específico, la prueba de sangre de alfa-fetoproteína, o la prueba CA-125, o una combinación de las mismas, pueden indicar que un sujeto no tiene cáncer, por ejemplo, no detectar cáncer en el sujeto. En otros casos, un sujeto que no presenta un cáncer detectable puede no tener ningún cáncer.

El sujeto puede tener mayor riesgo de padecer cáncer que la población general. El sujeto puede tener antecedentes familiares de cáncer. El sujeto puede tener fuentes genéticas conocidas de riesgo de cáncer. El sujeto puede haber estado expuesto a condiciones ambientales que se sabe que aumentan o causan el riesgo de cáncer. Los sujetos pueden ser pacientes cuyos únicos factores de riesgo de cáncer son la edad y/o el sexo. El sujeto puede no tener factores de riesgo de cáncer conocidos.

El sujeto puede haber sido diagnosticado de un cáncer. El cáncer puede estar en fase inicial o avanzada. El cáncer puede ser metastásico o no metastásico. Los tipos de cáncer que se le pueden haber diagnosticado a un sujeto incluyen, entre otros: carcinomas, sarcomas, linfomas, leucemias, tumores de células germinales y blastomas. Los tipos de cáncer que se le pueden haber diagnosticado a un sujeto incluyen, entre otros: Leucemia linfoblástica aguda (ALL), Leucemia mieloide aguda, Carcinoma corticosuprarrenal, Leucemia mieloide aguda del adulto, Carcinoma de sitio primario desconocido del adulto, Mesotelioma maligno del adulto, Cánceres relacionados con el SIDA, Linfoma relacionado con el SIDA, Cáncer anal, Cáncer de apéndice, Astrocitoma cerebeloso o cerebral infantil, Carcinoma basocelular, Cáncer de vías biliares, Cáncer de vejiga, Tumor óseo, osteosarcoma/histiocitoma fibroso maligno, Cáncer cerebral, Glioma de tronco cerebral, Cáncer de mama, Adenomas/carcinoides bronquiales, Linfoma de Burkitt, Tumor carcinoide, Carcinoma de sitio primario desconocido, Linfoma del sistema nervioso central, Astrocitoma cerebeloso, Astrocitoma cerebral/Glioma maligno, Cáncer de cuello uterino, Leucemia mieloide aguda infantil, Cáncer infantil de sitio primario desconocido, Cánceres infantiles, Astrocitoma cerebral infantil, Mesotelioma infantil, Condrosarcoma, Leucemia linfocítica crónica, Leucemia mielógena crónica, Trastornos mieloproliferativos crónicos, Cáncer de colon, Linfoma cutáneo de células T, Tumor desmoplásico de células redondas pequeñas, Cáncer de endometrio, Cáncer de útero endometrial, Ependimoma, Hemangioendotelioma epiteliode (EHE), Cáncer de esófago, Sarcoma de la familia de tumores de Ewing, Sarcoma de Ewing 's sarcoma de la familia de tumores de Ewing, Tumor extracraneal de células germinales, Tumor extragonadal de células germinales, Cáncer de vías biliares extrahepáticas, Cáncer de ojo, Melanoma intraocular, Cáncer de vesícula biliar, Cáncer gástrico (de estómago), Carcinoide gástrico, Tumor carcinoide gastrointestinal, Tumor del estroma gastrointestinal (GIST), Tumor trofoblástico gestacional, Glioma del tronco encefálico, Glioma, Leucemia de células pilosas, Cáncer de cabeza y cuello, Cáncer de corazón, Cáncer hepatocelular (hígado), Linfoma de Hodgkin, Cáncer hipofaríngeo, Glioma hipotalámico y de la vía visual, Carcinoma de células de los islotes (páncreas endocrino), Sarcoma de Kaposi, Cáncer de riñón (cáncer de células renales), Cáncer de laringe, Leucemia linfoblástica aguda (también denominada leucemia linfocítica aguda), Leucemia mieloide aguda (también denominada leucemia mielógena aguda), Leucemia linfocítica crónica (también llamada leucemia linfocítica crónica), Leucemias, Leucemia mielógena crónica (también llamada leucemia mieloide crónica), Leucemia de células pilosas, Cáncer de labio y cavidad oral, Liposarcoma, Cáncer de hígado (primario), Cáncer de pulmón, de células no pequeñas, Cáncer de pulmón, de células pequeñas, Linfoma (relacionado con el SIDA), Linfomas, Macroglobulinemia, Waldenstrom, Cáncer de mama masculino, Histiocitoma fibroso maligno óseo/osteosarcoma, meduloblastoma, Melanoma, Cáncer de células de Merkel, Cáncer escamoso metastásico de cuello con tumor primario oculto, Cáncer de boca, Síndrome de neoplasia endocrina múltiple infantil,

Mieloma múltiple (cáncer de la médula ósea), Mieloma múltiple/neoplasia de células plasmáticas, Micosis fungoide, Síndromes mielodisplásicos, Enfermedades mielodisplásicas/mieloproliferativas, Leucemia mielógena crónica, Mixoma, Cáncer de cavidad nasal y senos paranasales, Carcinoma nasofaríngeo, Neuroblastoma, Linfomas no Hodgkin, Cáncer de pulmón no microcítico, Oligodendroglioma, Cáncer oral, Cáncer orofaríngeo, Osteosarcoma/histiocitoma fibroso maligno óseo, Cáncer de ovario, Cáncer epitelial de ovario (tumor epitelial-estromal de superficie), Tumor de células germinales de ovario, Tumor de ovario de bajo potencial maligno, Cáncer de páncreas, Cáncer de páncreas de células de los islotes, Cáncer de seno paranasal y cavidad nasal, Cáncer de paratiroides, Cáncer de pene, Cáncer de faringe, Feocromocitoma, Astrocitoma pineal, Germinoma pineal, Pineoblastoma y tumores neuroectodérmicos primitivos supratentoriales, Adenoma hipofisario, Neoplasia de células plasmáticas/Mieloma múltiple, Blastoma pleuropulmonar, Linfoma primario del sistema nervioso central, Cáncer de próstata, Cáncer de recto, Carcinoma de células renales (cáncer de riñón), Cáncer de células transicionales de pelvis renal y uréter, Retinoblastoma, Rabdomiosarcoma, Cáncer de glándulas salivales, Síndrome de Sezary, Cáncer de piel (melanoma), Cáncer de piel (no melanoma), Carcinoma de piel, Células de Merkel, Cáncer de pulmón microcítico, Cáncer de intestino delgado, tejidos blandos Sarcoma, Carcinoma de células escamosas, Cáncer escamoso de cuello con primario oculto, metastásico, Cáncer de estómago, Tumor neuroectodérmico primitivo supratentorial, Linfoma cutáneo de células T, Cáncer de testículo, Cáncer de garganta, Timoma y carcinoma tímico, Timoma, Cáncer de tiroides, Cáncer de células de transición de pelvis renal y uréter, Uréter y pelvis renal, cáncer de células de transición, Cáncer de uretra, Sarcoma uterino, Cáncer de vagina, Glioma de la vía visual y del hipotálamo, Glioma de la vía visual y del hipotálamo, infantil, Cáncer de vulva, Macroglobulinemia de Waldenstrom, y Tumor de Wilms (cáncer de riñón).

El sujeto puede haber recibido previamente tratamiento para un cáncer. El sujeto puede haber recibido tratamiento quirúrgico, radioterapia, quimioterapia, terapia dirigida contra el cáncer o inmunoterapia contra el cáncer. El sujeto puede haber sido tratado con una vacuna contra el cáncer. El sujeto puede haber sido tratado con un tratamiento experimental contra el cáncer. El sujeto puede no haber recibido un tratamiento contra el cáncer. El sujeto puede estar en remisión del cáncer. El sujeto puede haber recibido previamente un tratamiento contra el cáncer y no presentar ningún síntoma detectable.

Análisis genético

Algunos métodos de secuenciación del ADN utilizan la captura de secuencias para enriquecer las secuencias de interés. La captura de secuencias suele implicar el uso de sondas de oligonucleótidos que se hibridan con la secuencia de interés. Una estrategia de conjunto de sondas puede consistir en colocar las sondas en mosaico a lo largo de una región de interés. Dichas sondas pueden tener, por ejemplo, entre 60 y 120 bases de longitud. El conjunto puede tener una profundidad de aproximadamente 2x, 3x, 4x, 5x, 6x, 8x, 9x, lOx, 15x, 20x, 50x o más. La eficacia de la captura de secuencias depende, en parte, de la longitud de la secuencia en la molécula diana que es complementaria (o casi complementaria) a la secuencia de la sonda. Las moléculas de ácido nucleico enriquecidas pueden ser representativas de más de 5.000 bases del genoma humano, más de 10.000 bases del genoma humano, más de 15.000 bases del genoma humano, más de 20.000 bases del genoma humano, más de 25.000 bases del genoma humano, más de 30.000 bases del genoma humano, más de 35.000 bases del genoma humano, más de 40.000 bases del genoma humano, más de 45.000 bases del genoma humano, más de 50.000 bases del genoma humano, más de 60.000 bases del genoma humano, más de 65.000 bases del genoma humano, más de 70.000 bases del genoma humano,000 bases del genoma humano, más de 55.000 bases del genoma humano, más de 60.000 bases del genoma humano, más de 65.000 bases del genoma humano, más de 70.000 bases del genoma humano, más de 75.000 bases del genoma humano, más de 80.000 bases del genoma humano, más de 85.000 bases del genoma humano, más de 90.000 bases del genoma humano, más de 95.000 bases del genoma humano, o más de 100.000 bases del genoma humano. Las moléculas de ácido nucleico enriquecidas pueden ser representativas de no más de 5.000 bases del genoma humano, no más de 10.000 bases del genoma humano, no más de 15.000 bases del genoma humano, no más de 20.000 bases del genoma humano, no más de 25.000 bases del genoma humano, no más de 30.000 bases del genoma humano, no más de 35.000 bases del genoma humano, no más de 40.000 bases del genoma humano, no más de 45.000 bases del genoma humano, no más de 50,000 bases del genoma humano, no más de 55.000 bases del genoma humano, no más de 60.000 bases del genoma humano, no más de 65.000 bases del genoma humano, no más de 70.000 bases del genoma humano, no más de 75.000 bases del genoma humano, no más de 80.000 bases del genoma humano, no más de 85.000 bases del genoma humano, no más de 90.000 bases del genoma humano, no más de 95.000 bases del genoma humano, o no más de 100.000 bases del genoma humano. Las moléculas de ácido nucleico enriquecidas pueden ser representativas de 5.000- 100.000 bases del genoma humano, 5.000-50.000 bases del genoma humano, 5.000-30.000 bases del genoma humano, 10.000-100.000 bases del genoma humano, 10.000-50.000 bases del genoma humano, o 10.000-30.000 bases del genoma humano. Las moléculas de ácido nucleico enriquecidas pueden ser representativas de diversas características del ácido nucleico, incluidas variantes genéticas como variantes de nucleótidos (SNV), variantes del número de copias (CNV), inserciones o deleciones (por ejemplo, indels), regiones nucleosómicas asociadas al cáncer, fusiones de genes, e inversiones.

Generalmente, los métodos y sistemas aquí proporcionados son útiles para la preparación de secuencias de polinucleótidos libres de células para una reacción de secuenciación de aplicación down-stream. El método de secuenciación puede ser una secuenciación paralela masiva, es decir, secuenciar simultáneamente (o en rápida sucesión) cualquiera de al menos 100, 1000, 10.000, 100.000, 1 millón, 10 millones, 100 millones, 1.000 millones, o 10.000 millones de moléculas de polinucleótidos. Los métodos de secuenciación pueden incluir, entre otros: secuenciación de alto

rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de molécula única, secuenciación por nanoporos, secuenciación por semiconductores, secuenciación por ligación, secuenciación por hibridación, RNA- Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, secuenciación de molécula única por síntesis (SMSS) (Helicos), secuenciación masiva en paralelo, Clonal Single Molecule Array (Solexa), secuenciación shotgun, secuenciación Maxam-Gilbert o Sanger, primer walking, secuenciación mediante plataformas PacBio, SOLiD, Ion Torrent o Nanopore y cualquier otro método de secuenciación conocido en la técnica.

Los fragmentos polinucleotídicos individuales de una muestra de ácido nucleico genómico (por ejemplo, una muestra de ADN genómico) pueden identificarse de forma única marcándolos con identificadores no únicos, por ejemplo, marcando de forma no única los fragmentos polinucleotídicos individuales.

Panel de secuenciación

Para mejorar la probabilidad de detectar mutaciones indicadoras de tumores, la región de ADN secuenciada puede comprender un panel de genes o regiones genómicas. La selección de una región limitada para la secuenciación (por ejemplo, un panel limitado) puede reducir la secuenciación total necesaria (por ejemplo, una cantidad total de nucleótidos secuenciados. Un panel de secuenciación puede dirigirse a una pluralidad de genes o regiones diferentes para detectar un único cáncer, un conjunto de cánceres, o todos los cánceres.

En algunos aspectos, se selecciona un panel dirigido a una pluralidad de genes o regiones genómicas diferentes de manera que una proporción determinada de sujetos que tienen un cáncer presenten una variante genética o marcador tumoral en uno o más genes o regiones genómicas diferentes del panel. El panel puede seleccionarse para limitar una región de secuenciación a un número fijo de pares de bases. El panel puede seleccionarse para secuenciar una cantidad deseada de ADN. El panel puede seleccionarse para conseguir la profundidad de lectura deseada. El panel puede seleccionarse para lograr una profundidad de lectura de secuencia o una cobertura de lectura de secuencia deseada para una cantidad de pares de bases secuenciados. El panel puede seleccionarse para alcanzar una sensibilidad teórica, una especificidad teórica y/o una precisión teórica para detectar una o más variantes genéticas en una muestra.

Las sondas para la detección del panel de regiones pueden incluir aquellas para la detección de regiones hotspots, así como sondas conscientes del nucleosoma (por ejemplo, los codones 12 y 13 de KRAS) y pueden diseñarse para optimizar la captura basándose en el análisis de la cobertura del ADNcf y la variación del tamaño del fragmento afectada por los patrones de unión del nucleosoma y la composición de la secuencia GC. Las regiones aquí utilizadas también pueden incluir regiones no hotspot optimizadas en función de las posiciones de los nucleosomas y los modelos de GC. El panel puede comprender una pluralidad de subpaneles, incluidos los subpaneles para identificar el tejido de origen (por ejemplo, el uso de la literatura publicada para definir 50-100 cebos que representan genes con el perfil de transcripción más diverso en los tejidos (no necesariamente promotores)), el andamiaje del genoma completo (por ejemplo, para identificar el contenido genómico ultraconservador y formar mosaicos dispersos en los cromosomas con un puñado de sondas con el fin de alinear las bases del número de copias), sitios de inicio de la transcripción (TSS)/islas CpG (por ejemplo, para capturar regiones metiladas diferenciales (por ejemplo, regiones metiladas diferencialmente (DMR)) en, por ejemplo, los promotores de genes supresores de tumores (por ejemplo, SEPT9/VIM en el cáncer colorrectal)). En algunas realizaciones, los marcadores de un tejido de origen son marcadores epigenéticos específicos de un tejido.

En la Tabla 1 y la Tabla 2 se pueden encontrar listados ejemplares de localizaciones genómicas de interés. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos una porción de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, o 97 de los genes de la Tabla 1. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, o 70 de los SNV de la Tabla 1. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 11, al menos 12, al menos 13, al menos 14, al menos 15, al menos 16, al menos 17, o 18 de las CNV de la Tabla 1. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, o 6 de las fusiones de la Tabla 1. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos una porción de al menos 1, al menos 2, o 3 de los indels de la Tabla 1. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos una porción de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 105, al menos 110, o 115 de los genes de la Tabla 2. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, o 73 de los SNV de la Tabla 2. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 11, al menos 12, al menos 13, al menos 14, al menos 15, al menos 16, al menos

17, o 18 de las CNVs de la Tabla 2. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, o 6 de las fusiones de la Tabla 2. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos una porción de al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 11, al menos 12, al menos 13, al menos 14, al menos 15, al menos 16, al menos 17, o 18 de las indels de la Tabla 2. Cada una de estas localizaciones genómicas de interés puede identificarse como una región troncal o una región caliente para un determinado panel de cebos. En la Tabla 3 figura una lista de ejemplos de puntos calientes genómicos de interés. En algunas realizaciones, las regiones genómicas utilizadas en los métodos de la presente divulgación comprenden al menos una porción de al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 11, al menos 12, al menos 13, al menos 14, al menos 15, al menos 16, al menos 17, al menos 18, al menos 19, o al menos 20 de los genes de la Tabla 3. Cada región genómica de interés se enumera con varias características, como el gen asociado, el cromosoma en el que reside, la posición de inicio y fin del genoma que representa el locus del gen, la longitud del locus del gen en pares de bases, los exones cubiertos por el gen y la característica crítica (por ejemplo, el tipo de mutación) que una determinada región genómica de interés puede tratar de captar.

Tabla 1

Mutaciones Puntuales (SNVs)							Amplificaciones (CNVs)		Fusiones	Indels
AKT1	ALK	APC	AR	ARAF	ARID1A	AR	BRAF	ALK	EGFR (exones 19 & 20)
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	CCND1	CCND2	FGFR2	
CCNE1	CDH1	CDK4	CDK6	CDKN2A	CDKN2B	CCNE1	CDK4	FGFR3	
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	CDK6	EGFR	NTRK1	ERBB2 (exones 19 & 20)
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	ERBB2	FGFR1	RET	
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	FGFR2	KIT	ROS1	
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	KRAS	MET	MET (salto de exón 14)	
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1	MYC	PDGFRA		
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	PIK3CA	RAF1		
PTPN11	RAF1	RB1	RET	RHEB	RHOA				
RIT1	ROS1	SMAD4	SMO	SRC	STK11				
TERT	TP53	TSC1	VHL						

Tabla 2

Mutaciones Puntuales (SNVs)							Amplificaciones (CNVs)		Fusiones	Indels
AKT1	ALK	APC	AR	ARAF	ARID1A		AR	BRAF	ALK	EGFR (exones 19 & 20)
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2		CCND1	CCND2	FGFR2	
CCNE1	CDH1	CDK4	CDK6	CDKN2A	DDR2		CCNE1	CDK4	FGFR3	
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7		CDK6	EGFR	NTRK1	ERBB2 (exones 19 & 20)
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ		ERBB2	FGFR1	RET	
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2		FGFR2	KIT	ROS1	
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET		KRAS	MET		MET (salto de exón 14)
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1		MYC	PDGFRA		
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN		PIK3CA	RAF1		
PTPN11	RAF1	RB1	RET	RHEB	RHOA					ATM
RIT1	ROS1	SMAD4	SMO	MAPK1	STK11					
TERT	TP53	TSC1	VHL	MAPK3	MTOR					APC
NTRK3										ARID1A
										BRCA1
										BRCA2

(continuación)

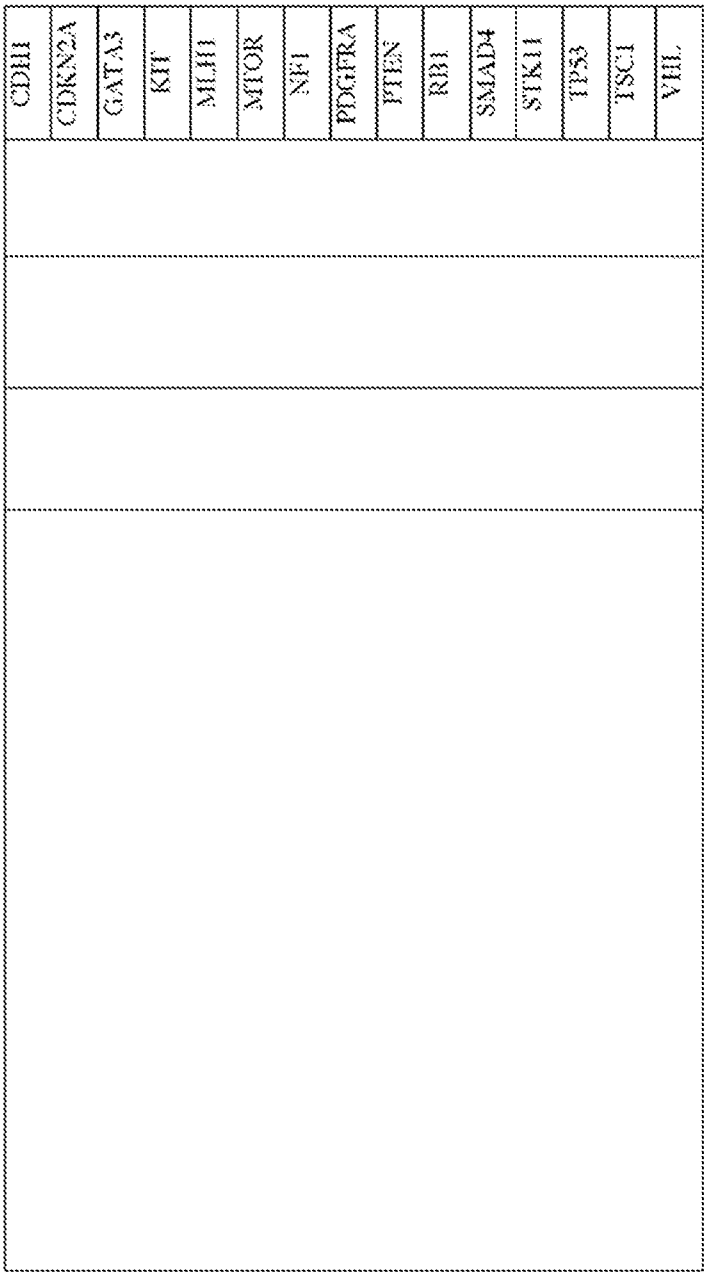

Tabla 3

Gen	Cromosoma	Posición Inicial	Posición Parada	Longitud (bp)	Exones Cubiertos	Rasgo Crítico
ALK	chr2	29446405	29446655	250	intrón 19	Fusión
ALK	chr2	29446062	29446197	135	intrón 20	Fusión
ALK	chr2	29446198	29446404	206	20	Fusión
ALK	chr2	29447353	29447473	120	intrón 19	Fusión
ALK	chr2	29447614	29448316	702	intrón 19	Fusión
ALK	chr2	29448317	29448441	124	19	Fusión
ALK	chr2	29449366	29449777	411	intrón 18	Fusión
ALK	chr2	29449778	29449950	172	18	Fusión
BRAF	chr7	140453064	140453203	139	15	BRAF V600
CTNNB1	chr3	41266007	41266254	247	3	S37
EGFR	chr7	55240528	55240827	299	18 y 19	G719 y deleciones
EGFR	chr7	55241603	55241746	143	20	Inserciones/T790M
EGFR	chr7	55242404	55242523	119	21	L858R
ERBB2	chr17	37880952	37881174	222	20	Inserciones
ESR1	chr6	152419857	152420111	254	10	V534, P535, L536, Y537, D538
FGFR2	chr10	123279482	123279693	211	6	S252
GATA3	chr10	8111426	8111571	145	5	SS / Indels
GATA3	chr10	8115692	8116002	310	6	SS / Indels
GNAS	chr20	57484395	57484488	93	8	R844
IDH1	chr2	209113083	209113394	311	4	R132
IDH2	chr15	90631809	90631989	180	4	R140, R172
KIT	chr4	55524171	55524258	87	1	
KIT	chr4	55561667	55561957	290	2	
KIT	chr4	55564439	55564741	302	3	

(continuación)

KIT	chr4	55565785	55565942	157	4	
KIT	chr4	55569879	55570068	189	5	
KIT	chr4	55573253	55573463	210	6	
KIT	chr4	55575579	55575719	140	7	
KIT	chr4	55589739	55589874	135	8	
KIT	chr4	55592012	55592226	214	9	
KIT	chr4	55593373	55593718	345	10 y 11	557, 559, 560, 576
KIT	chr4	55593978	55594297	319	12 y 13	V654
KIT	chr4	55595490	55595661	171	14	T670, S709
KIT	chr4	55597483	55597595	112	15	D716
KIT	chr4	55598026	55598174	148	16	L783
KIT	chr4	55599225	55599368	143	17	C809, R815, D816, L818, D820, S821F, N822, Y823
KIT	chr4	55602653	55602785	132	18	A829P
KIT	chr4	55602876	55602996	120	19	
KIT	chr4	55603330	55603456	126	20	
KIT	chr4	55604584	55604733	149	21	
KRAS	chr12	25378537	25378717	180	4	A146
KRAS	chr12	25380157	25380356	199	3	Q61
KRAS	chr12	25398197	25398328	131	2	G12/G13
MET	chr7	116411535	116412255	720	13, 14, intrón 13, intrón 14	MET exón 14 SS
NRAS	chr1	115256410	115256609	199	3	Q61
NRAS	chr1	115258660	115258791	131	2	G12/G13
PIK3CA	chr3	178935987	178936132	145	10	E545K
PIK3CA	chr3	178951871	178952162	291	21	H1047R

(continuación)

PTEN	chr10	89692759	89693018	259	5	R130
SMAD4	chr18	48604616	48604849	233	12	D537
TERT	chr5	1294841	1295512	671	promotor	chr5:1295228
TP53	chr17	7573916	7574043	127	11	Q331, R337, R342
TP53	chr17	7577008	7577165	157	8	R273
TP53	chr17	7577488	7577618	130	7	R248
TP53	chr17	7578127	7578299	172	6	R213/Y220
TP53	chr17	7578360	7578564	204	5	R175 /Deleciones
TP53	chr17	7579301	7579600	299	4	
				12574 (región diana total)		
				16330 (cob. total de sonda)		

En algunas realizaciones, la una o más regiones en el panel comprenden uno o más loci de uno o una pluralidad de genes para detectar cáncer residual después de la cirugía. Esta detección puede ser más temprana de lo que permiten los métodos actuales de detección del cáncer. En algunas realizaciones, la una o más regiones del panel comprenden uno o más loci de uno o una pluralidad de genes para detectar cáncer en una población de pacientes de alto riesgo. Por ejemplo, los fumadores tienen tasas de cáncer de pulmón mucho más elevadas que la población general. Además, los fumadores pueden desarrollar otras afecciones pulmonares que dificultan la detección del cáncer, como la aparición de nódulos irregulares en los pulmones. En algunas realizaciones, los métodos aquí descritos detectan el cáncer en pacientes de alto riesgo antes de lo que permiten los métodos existentes de detección del cáncer.

Una región puede seleccionarse para su inclusión en un panel de secuenciación basándose en un número de sujetos con un cáncer que tienen un marcador tumoral en ese gen o región. Una región puede seleccionarse para su inclusión en un panel de secuenciación basándose en la prevalencia de sujetos con un cáncer y un marcador tumoral presente en ese gen. La presencia de un marcador tumoral en una región puede ser indicativa de que un sujeto padece cáncer.

En algunos casos, el panel puede seleccionarse utilizando información de una o más bases de datos. La información relativa a un cáncer puede proceder de biopsias de tumores cancerosos o de ensayos de ADNcf. Una base de datos puede comprender información que describa una población de muestras tumorales secuenciadas. Una base de datos puede incluir información sobre la expresión de ARNm en muestras tumorales. Una base de datos puede incluir información sobre elementos reguladores en muestras tumorales. La información relativa a las muestras tumorales secuenciadas puede incluir la frecuencia de diversas variantes genéticas y describir los genes o regiones en los que se producen las variantes genéticas. Las variantes genéticas pueden ser marcadores tumorales. Un ejemplo no limitativo de este tipo de base de datos es COSMIC. COSMIC es un catálogo de mutaciones somáticas encontradas en diversos tipos de cáncer. Para un cáncer concreto, COSMIC clasifica los genes en función de la frecuencia de mutación. Un gen puede ser seleccionado para su inclusión en un panel por tener una alta frecuencia de mutación dentro de un gen determinado. Por ejemplo, COSMIC indica que el 33% de una población de muestras secuenciadas de cáncer de mama tiene una mutación en TP53 y el 22% de una población de muestras de cáncer de mama tiene una mutación en KRAS. Otros genes clasificados, incluido el APC, presentan mutaciones que sólo se han encontrado en aproximadamente el 4% de una población de muestras de cáncer de mama secuenciadas. TP53 y KRAS pueden incluirse en un panel de secuenciación

basándose en que tienen una frecuencia relativamente alta entre los cánceres de mama muestreados (en comparación con APC, por ejemplo, que ocurre con una frecuencia de alrededor del 4%). COSMIC se proporciona como ejemplo no limitativo, sin embargo, puede utilizarse cualquier base de datos o conjunto de información que asocie un cáncer con un marcador tumoral localizado en un gen o región genética. En otro ejemplo, proporcionado por COSMIC, de 1156 muestras de cáncer del tracto biliar, 380 muestras (33%) presentaban mutaciones en TP53. Otros genes, como el APC, presentan mutaciones en el 4-8% de todas las muestras. Así, TP53 puede seleccionarse para su inclusión en el panel basándose en una frecuencia relativamente alta en una población de muestras de cáncer del tracto biliar.

Se puede seleccionar un gen o región para un panel cuando la frecuencia de un marcador tumoral es significativamente mayor en el tejido tumoral muestreado o en el ADN tumoral circulante que la encontrada en una población de fondo determinada. Se puede seleccionar una combinación de regiones para su inclusión en un panel de forma que al menos una mayoría de los sujetos que padecen un cáncer tengan un marcador tumoral presente en al menos una de las regiones o genes del panel. La combinación de regiones puede seleccionarse basándose en datos que indiquen que, para un determinado cáncer o conjunto de cánceres, la mayoría de los sujetos tienen uno o más marcadores tumorales en una o más de las regiones seleccionadas. Por ejemplo, para detectar el cáncer 1, se puede seleccionar un panel que comprenda las regiones A, B, C, y/o D basándose en datos que indiquen que el 90% de los sujetos con cáncer 1 tienen un marcador tumoral en las regiones A, B, C, y/o D del panel. Alternativamente, se puede demostrar que los marcadores tumorales aparecen de forma independiente en dos o más regiones en sujetos que padecen un cáncer de tal forma que, combinados, un marcador tumoral en las dos o más regiones está presente en la mayoría de una población de sujetos que padecen un cáncer. Por ejemplo, para detectar el cáncer 2, se puede seleccionar un panel que comprenda las regiones X, Y, y Z basándose en datos que indiquen que el 90% de los sujetos tienen un marcador tumoral en una o más regiones, y que en el 30% de dichos sujetos se detecta un marcador tumoral sólo en la región X, mientras que en el resto de los sujetos en los que se detectó un marcador tumoral sólo se detectan marcadores tumorales en las regiones Y y/o Z. Los marcadores tumorales presentes en una o más regiones que previamente han demostrado estar asociadas con uno o más cánceres pueden ser indicativos o predictivos de que un sujeto tiene cáncer si se detecta un marcador tumoral en una o más de esas regiones el 50% o más de las veces. Los enfoques computacionales, como los modelos que emplean probabilidades condicionales de detectar cáncer dada una frecuencia de cáncer conocida para un conjunto de marcadores tumorales dentro de una o más regiones, pueden utilizarse para predecir qué regiones, solas o en combinación, pueden ser predictoras de cáncer. Otros enfoques para la selección de paneles implican el uso de bases de datos que describen información de estudios que emplean perfiles genómicos completos de tumores con grandes paneles y/o secuenciación del genoma completo (WGS, RNA-seq, Chip-seq, secuenciación de bisulfato, ATAC-seq, y otros). La información obtenida de la literatura también puede describir vías comúnmente afectadas y mutadas en ciertos tipos de cáncer. La selección de paneles puede basarse en el uso de ontologías que describan la información genética.

Los genes incluidos en el panel para secuenciación pueden incluir la región completamente transcrita, la región promotora, las regiones potenciadoras, los elementos reguladores, y/o la secuencia corriente abajo. Para aumentar aún más la probabilidad de detectar mutaciones indicadoras de tumores, sólo pueden incluirse en el panel los exones. El panel puede comprender todos los exones de un gen seleccionado, o sólo uno o más de los exones de un gen seleccionado. El panel puede comprender exones de cada uno de una pluralidad de genes diferentes. El panel puede comprender al menos un exón de cada uno de la pluralidad de genes diferentes.

En algunos aspectos, un panel de exones de cada uno de una pluralidad de genes diferentes se selecciona de tal manera que una proporción determinada de sujetos que padecen un cáncer presentan una variante genética en al menos un exón del panel de exones.

Se puede secuenciar al menos un exón completo de cada gen diferente de un panel de genes. El panel secuenciado puede comprender exones de una pluralidad de genes. El panel puede comprender exones de 2 a 100 genes diferentes, de 2 a 70 genes, de 2 a 50 genes, de 2 a 30 genes, de 2 a 15 genes, o de 2 a 10 genes.

Un panel seleccionado puede comprender un número variable de exones. El panel puede comprender de 2 a 3000 exones. El panel puede comprender de 2 a 1000 exones. El panel puede comprender de 2 a 500 exones. El panel puede comprender de 2 a 100 exones. El panel puede comprender de 2 a 50 exones. El panel no puede tener más de 300 exones. El panel no puede tener más de 200 exones. El panel puede comprender no más de 100 exones. El panel no puede tener más de 50 exones. El panel no puede tener más de 40 exones. El panel no puede tener más de 30 exones. El panel no puede comprender más de 25 exones. El panel no puede tener más de 20 exones. El panel no puede tener más de 15 exones. El panel no puede comprender más de 10 exones. El panel no puede tener más de 9 exones. El panel no puede tener más de 8 exones. El panel no puede tener más de 7 exones.

El panel puede comprender uno o más exones de una pluralidad de genes diferentes. El panel puede comprender uno o más exones de cada uno de una proporción de la pluralidad de genes diferentes. El panel puede comprender al menos dos exones de cada uno de al menos el 25%, 50%, 75%, o 90% de los diferentes genes. El panel puede comprender al menos tres exones de cada uno de al menos el 25%, 50%, 75% o 90% de los diferentes genes. El panel puede comprender al menos cuatro exones de cada uno de al menos el 25%, 50%, 75% o 90% de los diferentes genes.

Los tamaños del panel de secuenciación pueden variar. Un panel de secuenciación puede hacerse más grande o más pequeño (en términos de tamaño de nucleótidos) dependiendo de varios factores incluyendo, por ejemplo, la

cantidad total de nucleótidos secuenciados o un número de moléculas únicas secuenciadas para una región particular en el panel. El panel de secuenciación puede tener un tamaño de 5 kb a 50 kb. El panel de secuenciación puede tener un tamaño de 10 kb a 30 kb. El panel de secuenciación puede tener un tamaño de 12 kb a 20 kb. El panel de secuenciación puede tener un tamaño de 12 kb a 60 kb. El panel de secuenciación puede tener un tamaño de al menos lOkb, 12 kb, 15 kb, 20 kb, 25 kb, 30 kb, 35 kb, 40 kb, 45 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 110 kb, 120 kb, 130 kb, 140 kb, o 150 kb. El panel de secuenciación puede tener un tamaño inferior a 100 kb, 90 kb, 80 kb, 70 kb, 60 kb, o 50 kb.

El panel seleccionado para la secuenciación puede comprender al menos 1, 5, 10, 15, 20, 25, 30, 40, 50, 60, 80, o 100 regiones. En algunos casos, las regiones del panel se seleccionan de forma que el tamaño de las regiones sea relativamente pequeño. En algunos casos, las regiones del panel tienen un tamaño de unas 10 kb o menos, unas 8 kb o menos, unas 6 kb o menos, unas 5 kb o menos, unas 4 kb o menos, unas 3 kb o menos, unas 2,5 kb o menos, unas 2 kb o menos, unas 1,5 kb o menos, o unas 1 kb o menos. En algunos casos, las regiones del panel tienen un tamaño de aproximadamente 0,5 kb a aproximadamente 10 kb, de aproximadamente 0,5 kb a aproximadamente 6 kb, de aproximadamente 1 kb a aproximadamente 11 kb, de aproximadamente 1 kb a aproximadamente 15 kb, de aproximadamente 1 kb a aproximadamente 20 kb, de aproximadamente 0,1 kb a aproximadamente 10 kb, o de aproximadamente 0,2 kb a aproximadamente 1 kb. Por ejemplo, las regiones del panel pueden tener un tamaño de aproximadamente 0,1 kb a aproximadamente 5 kb.

El panel aquí seleccionado puede permitir una secuenciación profunda que sea suficiente para detectar variantes genéticas de baja frecuencia (por ejemplo, en moléculas de ácido nucleico libres de células obtenidas de una muestra). La cantidad de variantes genéticas en una muestra puede referirse a la frecuencia alélica menor de una variante genética determinada. La frecuencia de alelos menores puede referirse a la frecuencia con la que se producen alelos menores (por ejemplo, no el alelo más común) en una población dada de ácidos nucleicos, como una muestra. Las variantes genéticas con una baja frecuencia alélica menor pueden tener una frecuencia de presencia relativamente baja en una muestra. En algunos casos, el panel permite la detección de variantes genéticas con una frecuencia alélica menor de al menos 0,0001%, 0,001%, 0,005%, 0,01%, 0,05%, 0,1%, o 0,5%. El panel puede permitir la detección de variantes genéticas con una frecuencia alélica menor del 0,001% o superior. El panel puede permitir la detección de variantes genéticas con una frecuencia alélica menor del 0,01% o superior. El panel puede permitir la detección de una variante genética presente en una muestra con una frecuencia tan baja como 0,0001%, 0,001%, 0,005%, 0,01%, 0,025%, 0,05%, 0,075%, 0,1%, 0,25%, 0,5%, 0,75%, o 1,0%. El panel puede permitir la detección de marcadores tumorales presentes en una muestra con una frecuencia de al menos 0,0001%, 0,001%, 0,005%, 0,01%, 0,025%, 0,05%, 0,075%, 0,1%, 0,25%, 0,5%, 0,75%, o 1,0%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 1,0%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,75%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,5%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,25%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,1%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,075%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,05%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,025%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,01%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,005%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,001%. El panel puede permitir la detección de marcadores tumorales con una frecuencia en una muestra tan baja como el 0,0001%. El panel puede permitir la detección de marcadores tumorales en ADNcf secuenciado a una frecuencia en una muestra tan baja como 1,0% a 0,0001%. El panel puede permitir la detección de marcadores tumorales en ADNcf secuenciado a una frecuencia en una muestra tan baja como 0,01% a 0,0001%.

Una variante genética puede presentarse en un porcentaje de una población de sujetos que padecen una enfermedad (por ejemplo, cáncer). En algunos casos, al menos el 1%, 2%, 3%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, o 99% de una población con cáncer presenta una o más variantes genéticas en al menos una de las regiones del panel. Por ejemplo, al menos el 80% de una población con cáncer puede presentar una o más variantes genéticas en al menos una de las regiones del panel.

El panel puede comprender una o más regiones de cada uno de uno o más genes. En algunos casos, el panel puede comprender una o más regiones de cada uno de al menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, u 80 genes. En algunos casos, el panel puede comprender una o más regiones de cada uno de, como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, u 80 genes. En algunos casos, el panel puede comprender una o más regiones de cada uno de aproximadamente 1 a aproximadamente 80, de 1 a aproximadamente 50, de aproximadamente 3 a aproximadamente 40, de 5 a aproximadamente 30, de 10 a aproximadamente 20 genes diferentes.

Las regiones del panel pueden seleccionarse de forma que se detecten una o más regiones modificadas epigenéticamente. La una o más regiones modificadas epigenéticamente pueden estar acetiladas, metiladas, ubiquitiladas, fosforiladas, sumoiladas, ribosiladas, y/o citrulinadas. Por ejemplo, las regiones del panel pueden seleccionarse de forma que se detecten una o más regiones metiladas.

Las regiones del panel pueden seleccionarse de modo que comprendan secuencias transcritas diferencialmente en uno

o más tejidos. En algunos casos, las regiones pueden comprender secuencias transcritas en determinados tejidos a un nivel superior en comparación con otros tejidos. Por ejemplo, las regiones pueden comprender secuencias transcritas en determinados tejidos pero no en otros.

Las regiones del panel pueden comprender secuencias codificantes y/o no codificantes. Por ejemplo, las regiones del panel pueden comprender una o más secuencias en exones, intrones, promotores, regiones no traducidas 3', regiones no traducidas 5', elementos reguladores, sitios de inicio de transcripción, y/o sitios de empalme. En algunos casos, las regiones del panel pueden comprender otras secuencias no codificantes, incluidos pseudogenes, secuencias repetidas, transposones, elementos virales, y telómeros. En algunos casos, las regiones del panel pueden comprender secuencias de ARN no codificante, por ejemplo, ARN ribosómico, ARN de transferencia, ARN que interactúa con Piwi, y microARN.

Las regiones del panel pueden seleccionarse para detectar (diagnosticar) un cáncer con un nivel deseado de sensibilidad (por ejemplo, mediante la detección de una o más variantes genéticas). Por ejemplo, las regiones del panel pueden seleccionarse para detectar el cáncer (por ejemplo, mediante la detección de una o más variantes genéticas) con una sensibilidad de al menos el 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones del panel pueden seleccionarse para detectar el cáncer con una sensibilidad del 100%.

Las regiones del panel pueden seleccionarse para detectar (diagnosticar) un cáncer con un nivel deseado de especificidad (por ejemplo, mediante la detección de una o más variantes genéticas). Por ejemplo, las regiones del panel pueden seleccionarse para detectar cáncer (por ejemplo, mediante la detección de una o más variantes genéticas) con una especificidad de al menos el 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones del panel pueden seleccionarse para detectar una o más variantes genéticas con una especificidad del 100%.

Las regiones del panel pueden seleccionarse para detectar (diagnosticar) un cáncer con un valor predictivo positivo deseado. El valor predictivo positivo puede aumentarse incrementando la sensibilidad (por ejemplo, probabilidad de detectar un positivo real) y/o la especificidad (por ejemplo, probabilidad de no confundir un negativo real con un positivo). Como ejemplo no limitativo, las regiones del panel pueden seleccionarse para detectar una o más variantes genéticas con un valor predictivo positivo de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones del panel pueden seleccionarse para detectar una o más variantes genéticas con un valor predictivo positivo del 100%.

Las regiones del panel pueden seleccionarse para detectar (diagnosticar) un cáncer con la precisión deseada. En el presente documento, el término "precisión" puede referirse a la capacidad de una prueba para discriminar entre una enfermedad (por ejemplo, cáncer) y la salud. La precisión puede cuantificarse utilizando medidas como la sensibilidad y la especificidad, los valores predictivos, los cocientes de probabilidad, el área bajo la curva ROC, el índice de Youden y/o la odds ratio diagnóstica.

La precisión puede presentarse en forma de porcentaje, que se refiere a una relación entre el número de pruebas que dan un resultado correcto y el número total de pruebas realizadas. Las regiones del panel pueden seleccionarse para detectar el cáncer con una precisión de al menos el 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones del panel pueden seleccionarse para detectar el cáncer con una precisión del 100%.

Un panel puede seleccionarse para ser altamente sensible y detectar variantes genéticas de baja frecuencia. Por ejemplo, un panel puede seleccionarse de forma que una variante genética o marcador tumoral presente en una muestra con una frecuencia tan baja como 0,01%, 0,05%, o 0,001% pueda detectarse con una sensibilidad de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones de un panel pueden seleccionarse para detectar un marcador tumoral presente en una frecuencia del 1% o inferior en una muestra con una sensibilidad del 70% o superior. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,1% con una sensibilidad de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,01% con una sensibilidad de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,001% con una sensibilidad de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%.

Un panel puede seleccionarse para ser altamente específico y detectar variantes genéticas de baja frecuencia. Por ejemplo, un panel puede seleccionarse de forma que una variante genética o marcador tumoral presente en una muestra con una frecuencia tan baja como 0,01%, 0,05%, o 0,001% pueda detectarse con una especificidad de al menos 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones de un panel pueden seleccionarse para detectar un marcador tumoral presente en una frecuencia del 1% o inferior en una muestra con una especificidad del 70% o superior. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,1% con una especificidad de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,01% con una especificidad de al menos 70%, 75%, 80%, 85%, 90%, 95%,

96%, 97%, 98%, 99%, 99,5%, o 99,9%. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,001% con una especificidad de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%.

Un panel puede seleccionarse para ser muy preciso y detectar variantes genéticas de baja frecuencia. Un panel puede seleccionarse de forma que una variante genética o marcador tumoral presente en una muestra con una frecuencia tan baja como 0,01%, 0,05%, o 0,001% pueda detectarse con una precisión de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Las regiones de un panel pueden seleccionarse para detectar un marcador tumoral presente en una frecuencia del 1% o inferior en una muestra con una precisión del 70% o superior. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,1% con una precisión de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Un panel puede seleccionarse para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,01% con una precisión de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%. Se puede seleccionar un panel para detectar un marcador tumoral en una frecuencia en una muestra tan baja como 0,001% con una precisión de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%.

Un panel puede seleccionarse para ser altamente predictivo y detectar variantes genéticas de baja frecuencia. Un panel puede seleccionarse de forma que una variante genética o marcador tumoral presente en una muestra con una frecuencia tan baja como 0,01%, 0,05%, o 0,001% pueda tener un valor predictivo positivo de al menos 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99,5%, o 99,9%.

La concentración de sondas o cebos utilizados en el panel puede aumentarse (2 a 6 ng/ ^L) para capturar más molécula de ácido nucleico dentro de una muestra. La concentración de las sondas o cebos utilizados en el panel puede ser de al menos 2 ng/ ^L, 3 ng/ µï^, 4 ng/ µï^, 5 ng/ ^L, 6 ng/ ^L, o mayor. La concentración de sondas puede ser de unos 2 g/µL- a unos 3 ng/ ^L, de unos 2 g/µL- a unos 4 ng/ ^L, de unos 2 g/µL- a unos 5 ng/ ^L, de unos 2 g/µL- a unos 6 ng/ ^L. La concentración de las sondas o cebos utilizados en el panel puede ser de 2 g/µL- o más a 6 g/µL- o menos. En algunos casos, esto puede permitir que se analicen más moléculas dentro de un biológico, lo que permite detectar alelos de menor frecuencia.

Profundidad de secuenciación

El ADN enriquecido a partir de una muestra de moléculas de ADNcf puede secuenciarse con distintas profundidades de lectura para detectar variantes genéticas de baja frecuencia en una muestra. Para una posición determinada, la profundidad de lectura puede referirse a un número de todas las lecturas de todas las moléculas de una muestra que corresponden a una posición, incluidas las moléculas originales y las moléculas generadas mediante la amplificación de moléculas originales. Así, por ejemplo, una profundidad de lectura de 50.000 lecturas puede referirse al número de lecturas de 5.000 moléculas, con 10 lecturas por molécula. Las moléculas originales que corresponden a una posición pueden ser únicas y no redundantes (por ejemplo, no amplificadas, ADNcf de muestra).

Para evaluar la profundidad de lectura de las moléculas de la muestra en una posición determinada, se pueden rastrear las moléculas de la muestra. Las técnicas de rastreo molecular pueden comprender diversas técnicas de marcación de moléculas de ADN, como el marcado con códigos de barras, para identificar de forma única moléculas de ADN en una muestra. Por ejemplo, una o más secuencias únicas de código de barras pueden unirse a uno o más extremos de una molécula de ADNcf de muestra. Al determinar la profundidad de lectura en una posición dada, el número de moléculas distintas de ADNcf marcadas con el código de barras que corresponden a esa posición puede ser indicativo de la profundidad de lectura para esa posición. En otro ejemplo, ambos extremos de las moléculas de ADNcf de la muestra pueden marcarse con una de las ocho secuencias de código de barras. La profundidad de lectura en una posición dada puede determinarse cuantificando el número de moléculas de ADNcf originales en una posición dada, por ejemplo, mediante el colapso de las lecturas redundantes de la amplificación y la identificación de moléculas únicas basadas en las marcaciones del código de barras y la información de la secuencia endógena.

El ADN puede secuenciarse con una profundidad de lectura de al menos 3000 lecturas por base, al menos 4000 lecturas por base, al menos 5000 lecturas por base, al menos 6000 lecturas por base, al menos 7000 lecturas por base, al menos 8000 lecturas por base, al menos 9000 lecturas por base, en al menos 10 000 lecturas por base, al menos 15 000 lecturas por base, al menos 20 000 lecturas por base, al menos 25 000 lecturas por base, al menos 30 000 lecturas por base, al menos 40 000 lecturas por base, al menos 50 000 lecturas por base, al menos 60 000 lecturas por base, al menos 70 000 lecturas por base, al menos 80 000 lecturas por base, al menos 90 000 lecturas por base, al menos 100 000 lecturas por base, al menos 110 000 lecturas por base, al menos 120 000 lecturas por base, al menos 130 000 lecturas por base, al menos 140 000 lecturas por base, al menos 150 000 lecturas por base, al menos 160 000 lecturas por base, al menos 170 000 lecturas por base, al menos 180 000 lecturas por base, al menos 190 000 lecturas por base, al menos 200 000 lecturas por base, al menos 250.000 lecturas por base, al menos 500.000 lecturas por base, al menos 1.000.000 lecturas por base, o al menos 2.000.000 lecturas por base. El ADN puede secuenciarse con una profundidad de lectura de aproximadamente 3000 lecturas por base, aproximadamente 4000 lecturas por base, aproximadamente 5000 lecturas por base, aproximadamente 6000 lecturas por base, aproximadamente 7000 lecturas por base, aproximadamente 8000 lecturas por base, aproximadamente 9000 lecturas por base, aproximadamente 10 000 lecturas por base, aproximadamente 15 000 lecturas por base, aproximadamente 20 000 lecturas por base, aproximadamente 25 000

lecturas por base, aproximadamente 30 000 lecturas por base, aproximadamente 40 000 lecturas por base, aproximadamente 50 000 lecturas por base, aproximadamente 60 000 lecturas por base, aproximadamente 70 000 lecturas por base, aproximadamente 80 000 lecturas por base, aproximadamente 90.000 lecturas por base, aproximadamente 100.000 lecturas por base, aproximadamente 110.000 lecturas por base, aproximadamente 120.000 lecturas por base, aproximadamente 130.000 lecturas por base, aproximadamente 140.000 lecturas por base, aproximadamente 150.000 lecturas por base, aproximadamente 160.000 lecturas por base, aproximadamente 170.000 lecturas por base, aproximadamente 180.000 lecturas por base, aproximadamente 190.000 lecturas por base, aproximadamente 200.000 lecturas por base, aproximadamente 250.000 lecturas por base, aproximadamente 500.000 lecturas por base, aproximadamente 1.000.000 lecturas por base o aproximadamente 2.000.000 lecturas por base. El ADN puede secuenciarse con una profundidad de lectura de entre 10.000 y 30.000 lecturas por base, de entre 10.000 y 50.000 lecturas por base, de entre 10.000 y 5.000.000 lecturas por base, de entre 50.000 y 3.000.000 lecturas por base, de entre 100.000 y 2.000.000 lecturas por base, o de entre 500.000 y 1.000.000 lecturas por base. En algunas realizaciones, el ADN puede secuenciarse a cualquiera de las profundidades de lectura anteriores en un tamaño de panel seleccionado entre: menos de 70.000 bases, menos de 65.000 bases, menos de 60.000 bases, menos de 55.000 bases, menos de 50.000 bases, menos de 45.000 bases, menos de 40.000 bases, menos de 35.000 bases, menos de 30.000 bases, menos de 25.000 bases, menos de 20.000 bases, menos de 15.000 bases, menos de 10.000 bases, menos de 5.000 bases, y menos de 1.000 bases. Por ejemplo, el número total de lecturas para un panel puede ser tan bajo como 600.000 (3.000 lecturas por base para 1.000 bases) y tan alto como 1,4 x 10" (2.000.000 de lecturas por base para 70.000 bases). En algunas realizaciones, el ADN puede secuenciarse a cualquiera de las profundidades de lectura anteriores en un tamaño de panel seleccionado entre: 5.000 bases a 70.000 bases, 5.000 bases a 60.000 bases, 10.000 bases a 70.000 bases, o 10.000 bases a 70.000 bases.

La cobertura de lectura puede incluir lecturas de una o ambas cadenas de una molécula de ácido nucleico. Por ejemplo, la cobertura de lectura puede incluir lecturas de ambas cadenas de al menos 5.000, al menos 10.000, al menos 15.000, al menos 20.000, al menos 25.000, al menos 30.000, al menos 35.000, al menos 40.000, al menos 45.000, o al menos 50.000 moléculas de ADN de la muestra que se corresponden con cada nucleótido del panel.

Se puede seleccionar un panel para optimizar una profundidad de lectura deseada dada una cantidad fija de lecturas de base.

Marcación

En algunas realizaciones de la presente divulgación, se prepara una biblioteca de ácidos nucleicos antes de la secuenciación. Por ejemplo, fragmentos individuales de polinucleótidos en una muestra de ácido nucleico genómico (por ejemplo, una muestra de ADN genómico) pueden identificarse de forma única marcándolos con identificadores no únicos, por ejemplo, marcando de forma no única los fragmentos polinucleotídicos individuales. En algunas realizaciones, las moléculas de ácido nucleico están marcadas de forma no única entre sí.

Los polinucleótidos aquí divulgados pueden ser marcados. Por ejemplo, la doble cadena de polinucleótidos puede marcarse con marcaciones dúplex, es decir, marcaciones que marcan de forma diferente las cadenas complementarias. (es decir, las hebras "Watson" y "Crick") de una molécula de doble cadena. En algunos casos, las marcaciones dúplex son polinucleótidos con porciones complementarias y no complementarias.

Las marcaciones pueden ser cualquier tipo de moléculas unidas a un polinucleótido, incluyendo, pero sin limitarse a, ácidos nucleicos, compuestos químicos, sondas fluorescentes, o sondas radiactivas. Las marcaciones también pueden ser oligonucleótidos (por ejemplo, ADN o ARN). Las marcaciones pueden incluir secuencias conocidas, secuencias desconocidas, o ambas. Una marcación puede incluir secuencias aleatorias, secuencias predeterminadas, o ambas. Una marcación puede ser bicatenaria o monocatenaria. Una marcación de doble cadena puede ser una marcación dúplex. Una marcación de doble cadena puede incluir dos cadenas complementarias. Alternativamente, una marcación de doble cadena puede comprender una porción hibridada y una porción no hibridada. La marcación de doble cadena puede tener forma de Y, por ejemplo, la porción hibridada está en un extremo de la marcación y la porción no hibridada está en el extremo opuesto de la marcación. Un ejemplo son los "adaptadores Y" utilizados en la secuenciación Illumina. Otros ejemplos son los adaptadores en forma de horquilla o los adaptadores en forma de burbuja. Los adaptadores en forma de burbuja tienen secuencias no complementarias flanqueadas a ambos lados por secuencias complementarias. En algunas realizaciones, un adaptador en forma de Y comprende un código de barras de 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, o 32 nucleótidos de longitud. En algunas combinaciones. Esto puede combinarse con la reparación de extremos romos y la ligación.

El número de marcaciones diferentes puede ser mayor que un número estimado o predeterminado de moléculas en la muestra. Por ejemplo, para el marcado único, se pueden utilizar al menos dos veces más marcaciones diferentes que el número estimado o predeterminado de moléculas de la muestra.

El número de marcaciones de identificación diferentes utilizadas para marcar moléculas en una colección puede oscilar, por ejemplo, entre cualquiera de 2, 3, 4, 5, 6, 7, 8" 9, 10, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, o 49 en el extremo inferior del intervalo, y cualquiera de 50, 100, 500, 1000, 5000 y 10.000 en el extremo superior del intervalo. El número de marcaciones de identificación

utilizadas para marcar moléculas en una colección puede ser de al menos 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 o más. Así, por ejemplo, una colección de entre 100.000 millones y 1 billón de moléculas puede marcarse con entre 4 y 100 marcaciones de identificación diferentes. Una colección de entre 100.000 y 1.000 billones de moléculas puede marcarse con entre 8.000 y 10.000 marcaciones de identificación diferentes. Una colección de entre 100.000 y 1.000 billones de moléculas puede marcarse con entre 16 y 10.000 marcaciones de identificación diferentes. Una colección de entre 100.000 y 1.000 billones de moléculas puede marcarse con entre 16 y 5.000 marcaciones de identificación diferentes. Una colección de entre 100.000 y 1.000 billones de moléculas puede marcarse con entre 16 y 1.000 marcaciones de identificación diferentes.

Se puede considerar que una colección de moléculas está "no unívocamente marcada" si hay más moléculas en la colección que marcaciones. Se puede considerar que una colección de moléculas está "no marcada de forma única" si cada una de al menos el 1%, al menos el 5%, al menos el 10%, al menos el 15%, al menos el 20%, al menos el 25%, al menos el 30%, al menos el 35%, al menos el 40%, al menos el 45%, o al menos o aproximadamente el 50% de las moléculas de la colección lleva una marcación de identificación que es compartida por al menos otra molécula de la colección ("marcación no única" o "identificador no único"). Un identificador puede comprender un único código de barras o dos códigos de barras. Una población de moléculas de ácido nucleico puede marcarse de forma no única marcando las moléculas de ácido nucleico con menos marcaciones que el número total de moléculas de ácido nucleico de la población. Para una población no marcada de forma única, no más del 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, o 50% de las moléculas pueden estar marcadas de forma única. En algunas realizaciones, las moléculas de ácido nucleico se identifican mediante una combinación de marcaciones no únicas y las posiciones o secuencias de inicio y parada de las lecturas de secuencias. En algunas realizaciones, el número de moléculas de ácido nucleico que se secuencian es menor o igual que el número de combinaciones de identificadores y posiciones o secuencias de inicio y parada.

En algunos casos, las marcaciones del presente documento comprenden códigos de barras moleculares. Estos códigos de barras moleculares pueden utilizarse para diferenciar polinucleótidos en una muestra. Los códigos de barras moleculares pueden ser diferentes entre sí. Por ejemplo, los códigos de barras moleculares pueden tener una diferencia entre ellos que puede caracterizarse por una distancia de edición predeterminada o una distancia de Hamming. En algunos casos, los códigos de barras moleculares tienen una distancia de edición mínima de 1, 2, 3, 4, 5, 6, 7, 8, 9, o 10. Para mejorar aún más la eficacia de la conversión (por ejemplo, el marcado) de moléculas no marcadas en moléculas marcadas, se utilizan marcas cortas. Por ejemplo, una marcación adaptadora de biblioteca puede tener una longitud de hasta 65, 60, 55, 50, 45, 40, o 35 bases nucleotídicas. Una colección de tales códigos de barras cortos de biblioteca puede incluir un número de códigos de barras moleculares diferentes, por ejemplo, al menos 2, 4, 6, 8, 10, 12, 14, 16, 18 o 20 códigos de barras diferentes con una distancia de edición mínima de 1, 2, 3, o más.

Así, una colección de moléculas puede incluir una o más marcaciones. En algunos casos, algunas moléculas de una colección pueden incluir una marcación identificativa ("identificador") como un código de barras molecular que no comparte ninguna otra molécula de la colección. Por ejemplo, en algunos casos de una colección de moléculas, el 100% o al menos el 50%, 60%, 70%, 80%, 90%, 95%, 97%, 98%, o 99% de las moléculas de la colección pueden incluir un identificador o código de barras molecular que no es compartido por ninguna otra molécula de la colección. Tal y como se utiliza en el presente documento, se considera que una colección de moléculas está "marcada de forma única" si cada una de al menos el 95% de las moléculas de la colección lleva un identificador que no comparte ninguna otra molécula de la colección ("marcación única" o "identificador único"). En algunas realizaciones, las moléculas de ácido nucleico están marcadas de forma única entre sí. Se considera que una colección de moléculas está "no unívocamente marcada" si cada una de al menos el 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, o 50% de las moléculas de la colección lleva una marcación identificadora o un código de barras molecular que es compartido por al menos otra molécula de la colección ("marcación no única" o "identificador no único"). En algunas realizaciones, las moléculas de ácido nucleico están marcadas de forma no única entre sí. Por consiguiente, en una población no marcada de forma única, no más del 1% de las moléculas están marcadas de forma única. Por ejemplo, en una población no marcada de forma única, no más del 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, o 50% de las moléculas pueden estar marcadas de forma única.

En función del número estimado de moléculas de una muestra, pueden utilizarse diferentes marcaciones. En algunos métodos de marcación, el número de marcaciones diferentes puede ser al menos el mismo que el número estimado de moléculas en la muestra. En otros métodos de marcación, el número de marcaciones diferentes puede ser al menos dos, tres, cuatro, cinco, seis, siete, ocho, nueve, diez, cien o mil veces mayor que el número estimado de moléculas en la muestra. En la marcación único, se pueden utilizar al menos dos veces (o más) marcaciones diferentes que el número estimado de moléculas de la muestra.

Los fragmentos de polinucleótidos (antes del marcado) pueden comprender secuencias de cualquier longitud. Por ejemplo, los fragmentos polinucleotídicos (antes del marcado) pueden comprender al menos 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 o más nucleótidos de longitud. El fragmento polinucleotídico puede tener aproximadamente la longitud media del ADN libre de células. Por ejemplo, los fragmentos polinucleotídicos pueden comprender unas 160 bases de longitud. El fragmento polinucleotídico también puede fragmentarse a partir de un fragmento mayor en fragmentos más pequeños de unas 160 bases de longitud.

Se pueden conseguir mejoras en la secuenciación siempre que al menos algunos de los polinucleótidos duplicados o afines lleven identificadores únicos entre sí, es decir, lleven marcaciones diferentes. Sin embargo, en ciertas realizaciones, el número de marcaciones utilizadas se selecciona de forma que haya al menos un 95% de posibilidades de que todas las moléculas duplicadas que comiencen en cualquier posición lleven identificadores únicos. Por ejemplo, en una muestra que comprenda unos 10.000 equivalentes de genoma humano haploide de ADN genómico fragmentado, por ejemplo, ADNcf, se espera que z esté entre 2 y 8. Dicha población puede ser marcada con entre unos 10 y 100 identificadores diferentes, por ejemplo, unos 2 identificadores, unos 4 identificadores, unos 9 identificadores, unos 16 identificadores, unos 25 identificadores, unos 36 identificadores diferentes, unos 49 identificadores diferentes, unos 64 identificadores diferentes, unos 81 identificadores diferentes, o unos 100 identificadores diferentes.

Los códigos de barras de ácidos nucleicos con secuencias identificables, incluidos los códigos de barras moleculares, pueden utilizarse para el marcado. Por ejemplo, una pluralidad de códigos de barras de ADN puede comprender varios números de secuencias de nucleótidos. Puede utilizarse una pluralidad de códigos de barras de ADN que tengan 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 o más secuencias identificables de nucleótidos. Cuando se une a un solo extremo de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 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 o más identificadores diferentes. Alternativamente, cuando se une a ambos extremos de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 o más identificadores diferentes (que es el 2 de cuando el código de barras de ADN se une a sólo 1 extremo de un polinucleótido). En un ejemplo, puede utilizarse una pluralidad de códigos de barras de ADN que tengan 6, 7, 8, 9 o 10 secuencias identificables de nucleótidos. Cuando se unen a ambos extremos de un polinucleótido, producen 36, 49, 64, 81 o 100 posibles identificadores diferentes, respectivamente. En un ejemplo particular, la pluralidad de códigos de barras de ADN puede comprender 8 secuencias identificables de nucleótidos. Cuando se unen a un solo extremo de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 8 identificadores diferentes. Alternativamente, cuando se unen a ambos extremos de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 64 identificadores diferentes. Las muestras marcadas de esta forma pueden ser aquellas con un intervalo de aproximadamente 10 ng a cualquiera de aproximadamente 200 ng, aproximadamente 1 µ, aproximadamente 10 µ de polinucleótidos fragmentados, por ejemplo, ADN genómico, por ejemplo, ADNcf.

Un polinucleótido puede identificarse de forma única de varias maneras. Un polinucleótido puede identificarse unívocamente mediante un código de barras único. Por ejemplo, a dos polinucleótidos cualesquiera de una muestra se les adjuntan dos códigos de barras diferentes. Un código de barras puede ser de ADN o de ARN. Por ejemplo, un código de barras puede ser un código de barras de ADN.

Alternativamente, un polinucleótido puede ser identificado unívocamente por la combinación de un código de barras y una o más secuencias endógenas del polinucleótido. El código de barras puede ser una marcación no única o una marcación única. En algunos casos, el código de barras es una marcación no única. Por ejemplo, dos polinucleótidos de una muestra pueden unirse a códigos de barras que contengan el mismo código de barras, pero los dos polinucleótidos pueden seguir identificándose por secuencias endógenas diferentes. Los dos polinucleótidos pueden identificarse por la información de las diferentes secuencias endógenas. Dicha información incluye la secuencia de las secuencias endógenas o una porción de las mismas, la longitud de las secuencias endógenas, la ubicación de las secuencias endógenas, una o más modificaciones epigenéticas de las secuencias endógenas, o cualquier otra característica de las secuencias endógenas. En algunas realizaciones, los polinucleótidos pueden identificarse mediante un identificador (que comprende un código de barras o que comprende dos códigos de barras) en combinación con las secuencias de inicio y parada de la secuencia leída.

Se puede utilizar una combinación de marcaciones no únicas e información de secuencia endógena para detectar inequívocamente moléculas de ácido nucleico. Por ejemplo, las moléculas de ácido nucleico no marcadas de forma única de una muestra ("polinucleótidos parentales") pueden amplificarse para generar polinucleótidos progenie. A continuación, los polinucleótidos padre y progenie pueden secuenciarse para producir lecturas de secuencias. Para reducir el error, las lecturas de secuencias pueden agruparse para generar un conjunto de secuencias de consenso. Para generar secuencias de consenso, las lecturas de secuencias pueden colapsarse basándose en la información de secuencia de la marcación no única y la información de secuencia endógena, incluida la información de secuencia en una región inicial de una lectura de secuencia, la información de secuencia en una región final de una lectura de secuencia, y la longitud de una lectura de secuencia. En algunas realizaciones, la secuencia consenso se genera mediante secuenciación circular, en la que la misma cadena de ácido nucleico se secuencia múltiples veces en un círculo rodante para obtener la secuencia consenso. Una secuencia de consenso puede determinarse molécula a molécula (en la que la secuencia de consenso se determina sobre un tramo de bases) o base a base (en la que se determina un nucleótido de consenso para una base en una posición determinada). En algunas realizaciones, se construye un modelo probabilístico para modelar los perfiles de error de amplificación y secuenciación y se utiliza para estimar las probabilidades de nucleótido verdadero en cada posición de la molécula. En algunas realizaciones, las estimaciones de los parámetros del modelo probabilístico se actualizan basándose en los perfiles de error observados en la muestra individual o en el lote de muestras que se procesan conjuntamente o en un conjunto de muestras de referencia. En algunas realizaciones, se determina una secuencia consenso utilizando códigos de barras que marcan moléculas individuales de cfNA (por ejemplo, cfDNA) de un sujeto.

Una secuencia endógena puede estar en un extremo de un polinucleótido. Por ejemplo, la secuencia endógena

puede ser adyacente (por ejemplo, base intermedia) al código de barras adjunto. En algunos casos, la secuencia endógena puede tener al menos 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, o 100 bases de longitud. La secuencia endógena puede ser una secuencia terminal del fragmento/polinucleótidos a analizar. La secuencia endógena puede ser la longitud de la secuencia. Por ejemplo, una pluralidad de códigos de barras que comprenda 8 códigos de barras diferentes puede adherirse a ambos extremos de cada polinucleótido de una muestra. Cada polinucleótido de la muestra puede identificarse mediante la combinación de los códigos de barras y una secuencia endógena de aproximadamente 10 pares de bases en un extremo del polinucleótido. Sin estar limitado por la teoría, la secuencia endógena de un polinucleótido también puede ser la secuencia completa del polinucleótido.

También se divulgan en el presente documento composiciones de polinucleótidos marcados. El polinucleótido marcado puede ser monocatenario. Alternativamente, el polinucleótido marcado puede ser bicatenario (por ejemplo, polinucleótidos marcados dúplex). En consecuencia, esta divulgación también proporciona composiciones de polinucleótidos marcados dúplex. Los polinucleótidos pueden comprender cualquier tipo de ácido nucleico (ADN y/o ARN). Los polinucleótidos comprenden cualquier tipo de ADN descrito en el presente documento. Por ejemplo, los polinucleótidos pueden comprender ADN, por ejemplo, ADN fragmentado o ADNcf. Un conjunto de polinucleótidos en la composición que mapean a una posición de base mapeable en un genoma puede ser marcado de forma no única, es decir, el número de identificadores diferentes puede ser al menos 2 y menor que el número de polinucleótidos que mapean a la posición de base mapeable. El número de identificadores diferentes también puede ser al menos 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 y menor que el número de polinucleótidos que mapean a la posición de base mapeable.

En algunos casos, a medida que una composición pasa de aproximadamente 1 ng a aproximadamente 10 µ o más, puede utilizarse un conjunto mayor de códigos de barras moleculares diferentes. Por ejemplo, pueden utilizarse entre 5 y 100 adaptadores de biblioteca diferentes para marcar polinucleótidos en una muestra de ADNcf.

Los códigos de barras moleculares pueden asignarse a cualquier tipo de polinucleótidos divulgados en la presente divulgación. Por ejemplo, los códigos de barras moleculares pueden asignarse a polinucleótidos libres de células (por ejemplo, ADNcf). A menudo, un identificador divulgado en el presente documento puede ser un oligonucleótido de código de barras que se utiliza para marcar el polinucleótido. El identificador del código de barras puede ser un oligonucleótido de ácido nucleico (por ejemplo, un oligonucleótido de ADN). El identificador del código de barras puede ser monocatenario. Alternativamente, el identificador del código de barras puede ser de doble cadena. El identificador de código de barras puede unirse a polinucleótidos mediante cualquier método descrito en el presente documento. Por ejemplo, el identificador del código de barras se puede unir al polinucleótido mediante ligación utilizando una enzima. El identificador del código de barras también puede incorporarse al polinucleótido mediante PCR. En otros casos, la reacción puede comprender la adición de un isótopo metálico, ya sea directamente al analito o mediante una sonda marcada con el isótopo. En general, la asignación de identificadores únicos o no únicos o códigos de barras moleculares en las reacciones de la presente divulgación puede seguir los métodos y sistemas descritos, por ejemplo, en las solicitudes de patente de EE.UU. 2001/0053519, 2003/0152490, 2011/0160078 y Pat. de EE. UU. N.º 6,582,908.

Los identificadores o códigos de barras moleculares aquí utilizados pueden ser completamente endógenos, para lo cual puede realizarse la ligación circular de fragmentos individuales seguida de cizallamiento aleatorio o amplificación dirigida. En este caso, la combinación de un nuevo punto de inicio y fin de la molécula y el punto de ligación intramolecular original puede formar un identificador específico.

Los identificadores o códigos de barras moleculares utilizados en el presente documento pueden comprender cualquier tipo de oligonucleótidos. En algunos casos, los identificadores pueden ser oligonucleótidos de secuencia predeterminada, aleatoria, o semialeatoria. Los identificadores pueden ser códigos de barras. Por ejemplo, se puede utilizar una pluralidad de códigos de barras de tal forma que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. Alternativamente, se puede utilizar una pluralidad de códigos de barras de forma que cada código de barras sea único con respecto a cualquier otro código de barras de la pluralidad. Los códigos de barras pueden comprender secuencias específicas (por ejemplo, secuencias predeterminadas) que pueden rastrearse individualmente. Además, los códigos de barras pueden fijarse (por ejemplo, mediante ligación) a moléculas individuales de forma que la combinación del código de barras y la secuencia a la que puede ligarse cree una secuencia específica que pueda rastrearse individualmente. Como se describe en el presente documento, la detección de códigos de barras en combinación con los datos de secuencia de las partes inicial (inicio) y/o final (parada) de las lecturas de secuencia puede permitir la asignación de una identidad única a una molécula concreta. La longitud, o número de pares de bases, de una lectura de secuencia individual también puede utilizarse para asignar una identidad única a dicha molécula. Como se describe en el presente documento, los fragmentos de una única cadena de ácido nucleico a los que se ha asignado una identidad única pueden permitir la identificación posterior de fragmentos de la cadena original. De este modo, los polinucleótidos de la muestra pueden marcarse de forma única o sustancialmente única. Una marcación dúplex puede incluir una secuencia de nucleótidos degenerada o semidegenerada, por ejemplo, una secuencia degenerada aleatoria. La secuencia de nucleótidos puede incluir cualquier número de nucleótidos. Por ejemplo, la secuencia nucleotídica puede comprender 1 (si se utiliza un nucleótido no natural), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 o más nucleótidos. En un ejemplo particular, la secuencia puede comprender 7 nucleótidos. En otro ejemplo, la secuencia puede comprender 8 nucleótidos. La secuencia también puede comprender 9 nucleótidos. La secuencia puede comprender 10 nucleótidos.

Un código de barras puede comprender secuencias contiguas o no contiguas. Un código de barras que comprende al menos 1, 2, 3, 4, 5 o más nucleótidos es una secuencia contigua o no contigua, si los 4 nucleótidos no están interrumpidos por ningún otro nucleótido. Por ejemplo, si un código de barras comprende la secuencia TTGC, un código de barras es contiguo si el código de barras es TTGC. Por otra parte, un código de barras es no contiguo si el código de barras es TTXGC, donde X es una base de ácido nucleico.

Un identificador o código de barras molecular puede tener una secuencia n-mer que puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 o más nucleótidos de longitud. Una marcación puede tener cualquier longitud de nucleótidos. Por ejemplo, la secuencia puede tener una longitud de entre 2 y 100, 10 y 90, 20 y 80, 30 y 70, 40 y 60, o unos 50 nucleótidos. Una población de códigos de barras puede comprender códigos de barras de la misma longitud o de longitudes diferentes.

La marcación puede comprender una secuencia de referencia fija de doble cadena corriente abajo del identificador o código de barras molecular. Alternativamente, la marcación puede comprender una secuencia de referencia fija de doble cadena aguas arriba o aguas abajo del identificador o código de barras molecular. Cada hebra de una secuencia de referencia fija bicatenaria puede tener, por ejemplo, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleótidos de longitud.

El marcado aquí descrito puede realizarse mediante cualquier método. Un polinucleótido puede marcarse con un adaptador mediante hibridación. Por ejemplo, el adaptador puede tener una secuencia de nucleótidos complementaria a al menos una parte de una secuencia del polinucleótido. Como alternativa, un polinucleótido puede marcarse con un adaptador mediante ligación.

Los códigos de barras o marcaciones pueden fijarse mediante diversas técnicas. La fijación puede realizarse por métodos que incluyen, por ejemplo, la ligación (blunt-end o sticky-end) o el recocido optimizado de sondas de inversión molecular. Por ejemplo, el marcado puede comprender el uso de una o más enzimas. La enzima puede ser una ligasa. La ligasa puede ser una ADN ligasa. Por ejemplo, la ADN ligasa puede ser una ADN ligasa T4, una ADN ligasa de E. coli y/o una ligasa de mamífero. La ligasa de mamífero puede ser ADN ligasa I, ADN ligasa III, o ADN ligasa IV. La ligasa también puede ser una ligasa termoestable. Las marcaciones pueden ligarse al extremo romo de un polinucleótido (ligación de extremo romo). Alternativamente, las marcaciones pueden ligarse a un extremo pegajoso de un polinucleótido (ligación de extremo pegajoso). La eficacia de la ligación puede aumentarse optimizando diversas condiciones. La eficacia de la ligación puede aumentarse optimizando el tiempo de reacción de la ligación. Por ejemplo, el tiempo de reacción de ligación puede ser inferior a 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, o 20 horas. En un ejemplo particular, el tiempo de reacción de la ligación es inferior a 20 horas. La eficacia de la ligación puede aumentarse optimizando la concentración de ligasa en la reacción. Por ejemplo, la concentración de ligasa puede ser de al menos 10, 50, 100, 150, 200, 250, 300, 400, 500, o 600 unidades/microlitro. La eficacia también puede optimizarse añadiendo o variando la concentración de una enzima adecuada para la ligación, cofactores enzimáticos u otros aditivos, y/o optimizando la temperatura de una solución que contenga la enzima. La eficacia también puede optimizarse variando el orden de adición de los distintos componentes de la reacción. El extremo de la secuencia de la marcación puede incluir un dinucleótido para aumentar la eficacia de la ligación. Cuando la marcación comprende una porción no complementaria (por ejemplo, un adaptador en forma de Y), la secuencia en la porción complementaria del adaptador de la marcación puede comprender una o más secuencias seleccionadas que promuevan la eficiencia de ligación. Estas secuencias se sitúan en el extremo terminal de la marcación. Dichas secuencias pueden comprender 1, 2, 3, 4, 5, o 6 bases terminales. También puede utilizarse una solución de reacción con alta viscosidad (por ejemplo, un número de Reynolds bajo) para aumentar la eficacia de la ligación. Por ejemplo, la solución puede tener un número de Reynolds inferior a 3000, 2000, 1000, 900, 800, 700, 600, 500, 400, 300, 200, 100, 50, 25, o 10. También se contempla la posibilidad de utilizar una distribución más o menos unificada de los fragmentos (por ejemplo, una desviación estándar ajustada) para aumentar la eficacia de la ligación. Por ejemplo, la variación en el tamaño de los fragmentos puede ser inferior al 20%, 15%, 10%, 5%, o 1%. El marcado también puede comprender la extensión del cebador, por ejemplo, mediante la reacción en cadena de la polimerasa (PCR). El marcado también puede incluir cualquiera de las siguientes técnicas: PCR ligada, PCR multiplex, ligación de una sola cadena o circularización de una sola cadena. La eficiencia del marcado (por ejemplo, por ligación) puede aumentarse hasta una eficiencia de moléculas marcadas (eficiencia de conversión) de al menos el 20%, 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 al menos el 98%.

Se puede realizar una reacción de ligación en la que los polinucleótidos parentales de una muestra se mezclan con una mezcla de reacción que comprende y oligonucleótidos de código de barras diferentes, donde y = una raíz cuadrada de n. La ligación puede dar lugar a la unión aleatoria de oligonucleótidos de código de barras a polinucleótidos parentales de la muestra. A continuación, la mezcla de reacción puede incubarse en condiciones de ligación suficientes para efectuar la ligación de los oligonucleótidos del código de barras a los polinucleótidos parentales de la muestra. En algunas realizaciones, los códigos de barras aleatorios seleccionados entre los y oligonucleótidos de código de barras diferentes se ligan a ambos extremos de los polinucleótidos parentales. La ligación aleatoria de los códigos de barras y a uno o ambos extremos de los polinucleótidos parentales puede dar lugar a la producción de identificadores únicos y^2. Por ejemplo, una muestra que comprenda unos 10.000 equivalentes de genoma humano haploide de ADNcf puede marcarse

con unos 36 identificadores únicos. Los identificadores únicos pueden comprender seis códigos de barras de ADN únicos. La ligación de 6 códigos de barras únicos a ambos extremos de un polinucleótido puede dar lugar a que se produzcan 36 posibles identificadores únicos.

En algunas realizaciones, una muestra que comprende alrededor de 10.000 equivalentes de genoma humano haploide de ADN se marca con un número de identificadores únicos producidos por ligación de un conjunto de códigos de barras únicos a ambos extremos de los polinucleótidos parentales. Por ejemplo, se pueden producir 64 identificadores únicos ligando 8 códigos de barras únicos a ambos extremos de los polinucleótidos parentales. Asimismo, pueden producirse 100 identificadores únicos mediante ligación de 10 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 225 identificadores únicos mediante ligación de 15 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 400 identificadores únicos mediante ligación de 20 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 625 identificadores únicos mediante ligación de 25 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 900 identificadores únicos pueden producirse mediante ligación de 30 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 1225 identificadores únicos pueden producirse mediante ligación de 35 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 1600 identificadores únicos pueden producirse mediante ligación de 40 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, 2025 identificadores únicos pueden producirse mediante ligación de 45 códigos de barras únicos a ambos extremos de los polinucleótidos parentales, y 2500 identificadores únicos pueden producirse mediante ligación de 50 códigos de barras únicos a ambos extremos de los polinucleótidos parentales. La eficacia de ligación de la reacción puede ser superior al 10%, superior al 20%, superior al 30%, superior al 40%, superior al 50%, superior al 60%, superior al 70%, superior al 80%, o superior al 90%. Las condiciones de ligación pueden incluir el uso de adaptadores bidireccionales que puedan unirse a cualquiera de los extremos del fragmento y seguir siendo amplificables. Las condiciones de ligación pueden comprender adaptadores de ligación de extremo pegajoso, cada uno con un saliente de al menos una base nucleotídica. En algunos casos, las condiciones de ligación pueden comprender adaptadores con diferentes bases voladizas para aumentar la eficacia de la ligación. Como ejemplo no limitativo, las condiciones de ligación pueden comprender adaptadores con salientes de citosina (C) de base única (es decir, adaptadores de cola C), salientes de timina (T) de base única (adaptadores de cola T), salientes de adenina (A) de base única (adaptadores de cola A), y/o salientes de guanina (G) de base única (adaptadores de cola G). Las condiciones de ligación pueden incluir ligación de extremo romo, en contraposición a la ligación de cola. Las condiciones de ligación pueden comprender la titulación cuidadosa de una cantidad de oligonucleótidos adaptadores y/o de código de barras. Las condiciones de ligación pueden comprender el uso de un exceso molar de más de 2X, más de 5X, más de 10X, más de 20X, más de 40X, más de 60X, más de 80X, (por ejemplo, ~100X) de oligonucleótidos adaptadores y/o de código de barras en comparación con una cantidad de fragmentos de polinucleótidos parentales en la mezcla de reacción. Las condiciones de ligación pueden comprender el uso de una ADN ligasa T4 (por ejemplo, NEBNExt Ultra Ligation Module). En un ejemplo, se utilizan 18 microlitros de mezcla maestra de ligasa con 90 microlitros de ligación (18 partes de los 90) y potenciador de ligación. En consecuencia, la marcación de polinucleótidos parentales con n identificadores únicos puede comprender el uso de un número y de códigos de barras diferentes, donde y= una raíz cuadrada de n. Las muestras marcadas de este modo pueden ser aquellas con un rango de aproximadamente 10 ng a cualquiera de aproximadamente 100 ng, aproximadamente 200 ng, aproximadamente 300 ng, aproximadamente 400 ng, aproximadamente 500 ng, aproximadamente 1 µg, o aproximadamente 10 µg de polinucleótidos fragmentados, por ejemplo, ADN genómico, por ejemplo, ADNcf. El número y de códigos de barras utilizados para identificar polinucleótidos parentales en una muestra puede depender de la cantidad de ácido nucleico en la muestra.

Un método para aumentar la eficiencia de la conversión implica el uso de una ligasa diseñada para una reactividad óptima en ADN monocatenario, como un derivado de ligasa de ADN monocatenario (ssADN) de ThermoPhage. Dichas ligasas evitan las etapas tradicionales en la preparación de bibliotecas de reparación de extremos y de cola A, que pueden tener eficiencias pobres y/o pérdidas acumuladas debido a las etapas intermedias de limpieza, y permiten el doble de probabilidad de que el polinucleótido de partida sentido o antisentido se convierta en un polinucleótido marcado adecuadamente. También convierte polinucleótidos de doble cadena que pueden poseer salientes que no pueden ser lo suficientemente despuntados por la reacción típica de reparación de extremos. Las condiciones de reacción óptimas para esta reacción ssADN son: 1 x tampón de reacción (50 milimolar (mM) MOPS (pH 7,5), 1 mM DTT, 5 mM MgC12, 10 mM KC1). Con 50 mM de ATP, 25 mg/ml de BSA, 2,5 mM de MnC12, 200 pmol de 85 nt de ssADN oligómero y 5 U de ssADN ligasa incubados a 65°C durante 1 hora. La amplificación posterior mediante PCR puede convertir aún más la biblioteca monocatenaria marcada en una biblioteca bicatenaria y obtener una eficiencia de conversión global muy superior al 20%. Otros métodos para aumentar la tasa de conversión, por ejemplo, por encima del 10%, incluyen, por ejemplo, cualquiera de los siguientes, solos o en combinación: sondas de inversión molecular de recocido optimizado, ligación de extremo romo con un rango de tamaño de polinucleótido bien controlado, selección de una polimerasa de alta eficiencia, ligación de extremo pegajoso o un paso de amplificación múltiplex por adelantado con o sin el uso de cebadores de fusión, optimización de las bases finales en una secuencia diana, optimización de las condiciones de reacción (incluido el tiempo de reacción), y la introducción de uno o más pasos para limpiar una reacción (p. ej., de fragmentos de ácido nucleico no deseados) durante la ligación, y optimización de la temperatura de las condiciones tampón. La ligación de extremos pegajosos puede realizarse utilizando salientes de múltiples nucleótidos. La ligación de extremos pegajosos puede realizarse utilizando salientes de un solo nucleótido que comprendan una base A, T, C, o G.

La presente divulgación también proporciona composiciones de polinucleótidos marcados. Los polinucleótidos pueden comprender ADN fragmentado, por ejemplo, ADNcf. Un conjunto de polinucleótidos en la composición que

mapean a una posición de base mapeable en un genoma puede ser marcado de forma no única, es decir, el número de identificadores diferentes puede ser al menos 2 y menor que el número de polinucleótidos que mapean a la posición de base mapeable. Una composición de entre aproximadamente 10 ng a aproximadamente 10 µg (por ejemplo, cualquiera de entre aproximadamente 10 ng- 1 µg, aproximadamente 10 ng- 100 ng, aproximadamente 100 ng- 10 µg, aproximadamente 100 ng- 1 µg, aproximadamente 1 / g- 10 µ) puede llevar entre 2, 5, 10, 50 o 100 a cualquiera de 100, 1000, 10.000 o 100.000 identificadores diferentes. Por ejemplo, pueden utilizarse entre 5 y 100 identificadores diferentes para marcar los polinucleótidos de dicha composición.

Secuenciación

Los polinucleótidos marcados pueden secuenciarse para generar lecturas de secuencias. Por ejemplo, se puede secuenciar un polinucleótido dúplex marcado. Las lecturas de secuencias pueden generarse a partir de una sola hebra de un polinucleótido dúplex marcado. Alternativamente, ambas hebras de un polinucleótido dúplex marcado pueden generar lecturas de secuencias. Las dos cadenas del polinucleótido dúplex marcado pueden incluir las mismas marcaciones. Alternativamente, las dos hebras del polinucleótido dúplex marcado pueden incluir marcaciones diferentes. Cuando las dos hebras del polinucleótido dúplex marcado están marcadas de forma diferente, las lecturas de secuencias generadas a partir de una hebra (por ejemplo, una hebra Watson) pueden distinguirse de las lecturas de secuencias generadas a partir de las otras hebras (por ejemplo, una hebra Crick). La secuenciación puede implicar la generación de múltiples lecturas de secuencias para cada molécula. Esto ocurre, por ejemplo, como resultado de la amplificación de cadenas polinucleotídicas individuales durante el proceso de secuenciación, por ejemplo, mediante PCR.

Los métodos aquí divulgados pueden comprender la amplificación de polinucleótidos. La amplificación puede realizarse antes del marcado, después del marcado, o en ambos casos. La amplificación de polinucleótidos puede dar lugar a la incorporación de nucleótidos en una molécula de ácido nucleico o cebador, formando así una nueva molécula de ácido nucleico complementaria a un ácido nucleico molde. La molécula polinucleotídica recién formada y su molde pueden utilizarse como plantillas para sintetizar polinucleótidos adicionales. Los polinucleótidos que se amplifican pueden ser cualquier ácido nucleico, por ejemplo, ácidos desoxirribonucleicos, incluidos ADN genómicos, ADNc (ADN complementario), ADNcf, y ADN tumoral circulante (ADNct). Los polinucleótidos amplificados también pueden ser ARN. Tal y como se utiliza aquí, una reacción de amplificación puede comprender muchas rondas de replicación del ADN. Las reacciones de amplificación del ADN pueden incluir, por ejemplo, la reacción en cadena de la polimerasa (PCR). Una reacción de PCR puede comprender de 2 a 100 "ciclos" de desnaturalización, recocido, y síntesis de una molécula de ADN. Por ejemplo, pueden realizarse de 2 a 7, de 5 a 10, de 6 a 11, de 7 a 12, de 8 a 13, de 9 a 14, de 10 a 15, de 11 a 16, de 12 a 17, de 13 a 18, de 14 a 19, o de 15 a 20 ciclos durante la etapa de amplificación. La condición de la PCR puede optimizarse en función del contenido en GC de las secuencias, incluidos los cebadores. Los cebadores de amplificación pueden elegirse para seleccionar una secuencia diana de interés. Los cebadores pueden diseñarse para optimizar o maximizar la eficacia de la conversión. En algunas realizaciones, los cebadores contienen una secuencia corta entre los cebadores para extraer una pequeña región de interés. En algunas realizaciones, los cebadores se dirigen a regiones nucleosómicas de modo que los cebadores hibridan con áreas donde están presentes nucleosomas, en contraposición a áreas entre nucleosomas, porque las áreas inter-nucleosómicas están más altamente escindidas y por lo tanto es menos probable que estén presentes como dianas.

En algunas realizaciones, se atacan regiones del genoma que están protegidas diferencialmente por nucleosomas y otros mecanismos reguladores en células cancerosas, el microambiente tumoral, o componentes del sistema inmunitario (granulocitos, linfocitos infiltrantes de tumores, etc.). En algunas realizaciones, el objetivo son otras regiones que son estables y/o no están reguladas diferencialmente en las células tumorales. Dentro de estas regiones, las diferencias en la cobertura, los sitios de corte, la longitud del fragmento, el contenido de la secuencia, el contenido de la secuencia en los puntos finales del fragmento, o el contenido de la secuencia del contexto genómico cercano pueden utilizarse para inferir la presencia o ausencia de una determinada clasificación de células cancerosas (por ejemplo, cánceres mutantes EGFR, mutantes KRAS, amplificados ERBb2, o con expresión PD-1), o tipo de cáncer (por ejemplo, adenocarcinoma de pulmón, mama o colorrectal). Esta orientación también puede mejorar la sensibilidad y/o especificidad del ensayo al aumentar la cobertura en determinados sitios o la probabilidad de captura. Estos principios se aplican a los métodos de selección, incluidos, entre otros, la ligación más el enriquecimiento basado en la captura híbrida, el enriquecimiento basado en la amplificación, el enriquecimiento basado en el círculo rodante con cebadores de iniciación específicos de secuencia/localización genómica, y otros métodos. Las regiones que pueden ser objeto de tales métodos y análisis posteriores incluyen, entre otras, regiones intrónicas, regiones exónicas, regiones promotoras, regiones TSS, elementos reguladores distantes, regiones potenciadoras y regiones superpotenciadoras y/o uniones de las anteriores. Estos métodos también pueden utilizarse para inferir el tejido de origen del tumor y/o una medida de la carga tumoral en combinación con otras técnicas descritas en el presente documento para determinar variantes (por ejemplo, variantes de la línea germinal o somáticas) contenidas en la muestra. Por ejemplo, las variantes de la línea germinal pueden determinar la predisposición a ciertos tipos de cáncer, mientras que las variantes somáticas pueden correlacionarse con determinados tipos de cáncer basándose específicamente en los genes afectados, las vías y los porcentajes de las variantes. A continuación, esta información puede utilizarse en combinación con firmas epigenéticas relativas a mecanismos reguladores y/o modificaciones químicas como, por ejemplo, metilación, hidroximetilación, acetilación, y/o ARN. La biblioteca de ácidos nucleicos puede implicar el análisis combinado de ADN, modificaciones del ADN y ARN para aumentar la sensibilidad y especificidad en la detección del cáncer, el tipo de cáncer, las vías moleculares activadas en la enfermedad específica, el tejido de origen, así como una medida que corresponda a la carga tumoral. Los enfoques

para analizar cada uno de los anteriores se han descrito en otro lugar y pueden combinarse para el análisis de una o varias muestras del mismo paciente, pudiendo proceder la muestra de varios especímenes corporales.

Las técnicas de amplificación de ácidos nucleicos pueden utilizarse con los ensayos aquí descritos. Algunas técnicas de amplificación son las metodologías PCR que pueden incluir, entre otras, la PCR en solución y la PCR in situ. Por ejemplo, la amplificación puede comprender la amplificación basada en PCR. Alternativamente, la amplificación puede comprender amplificación no basada en PCR. La amplificación del ácido nucleico molde puede incluir el uso de una o más polimerasas. Por ejemplo, la polimerasa puede ser una ADN polimerasa o una ARN polimerasa. En algunos casos, se lleva a cabo una amplificación de alta fidelidad, como con el uso de polimerasas de alta fidelidad (por ejemplo, Phusion RTM High-Fidelity DNA Polymerase) o protocolos de PCR. En algunos casos, la polimerasa puede ser una polimerasa de alta fidelidad. Por ejemplo, la polimerasa puede ser KAPA HiFi ADN polimerasa. La polimerasa también puede ser la ADN polimerasa Phusion o una polimerasa Ultra II. La polimerasa puede utilizarse en condiciones de reacción que reduzcan o minimicen los sesgos de amplificación, por ejemplo, debidos a la longitud del fragmento y/o al contenido de GC.

La amplificación de una sola cadena de un polinucleótido por PCR generará copias tanto de esa cadena como de su complemento. Durante la secuenciación, tanto la cadena como su complemento generarán lecturas de secuencia. Sin embargo, las lecturas de secuencias generadas a partir del complemento de, por ejemplo, la cadena Watson, pueden identificarse como tales porque llevan el complemento de la parte de la marcación dúplex que marcó la cadena Watson original. Por el contrario, una lectura de secuencia generada a partir de una cadena Crick o su producto de amplificación llevará la parte de la marcación dúplex que marcó la cadena Crick original. De este modo, una lectura de secuencia generada a partir de un producto amplificado de un complemento de la cadena Watson puede distinguirse de una lectura de secuencia de complemento generada a partir de un producto de amplificación de la cadena Crick de la molécula original.

La amplificación, como la amplificación PCR, se realiza normalmente en rondas. Las rondas de amplificación ejemplares incluyen 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, o más rondas de amplificación. Las condiciones de amplificación pueden optimizarse, por ejemplo, para las condiciones del tampón y el tipo y las condiciones de la polimerasa. La amplificación también puede modificarse para reducir el sesgo en el procesamiento de la muestra, por ejemplo, reduciendo el sesgo de amplificación no específica, el sesgo de contenido de GC, y el sesgo de tamaño.

En algunas realizaciones, las secuencias pueden ser enriquecidas antes de la secuenciación. El enriquecimiento puede realizarse para regiones diana específicas o de forma inespecífica. En algunas realizaciones, las regiones genómicas de interés objetivo pueden enriquecerse con sondas de captura ("cebos") seleccionadas para uno o más paneles de conjuntos de cebos utilizando un esquema de mosaico y captura diferencial. Un esquema de mosaico y captura diferencial utiliza conjuntos de cebos de diferentes concentraciones relativas para mosaico diferencial (por ejemplo, a diferentes "resoluciones") a través de las regiones genómicas asociadas con los cebos, sujeto a un conjunto de restricciones (por ejemplo, restricciones del secuenciador como la carga de secuenciación, la utilidad de cada cebo, etc.), y capturarlos a un nivel deseado para la secuenciación posterior. Estas regiones genómicas de interés pueden incluir variantes de un solo nucleótido (SNV) e indels (es decir, inserciones o deleciones). Las regiones genómicas de interés objetivo pueden comprender regiones genómicas de interés troncales ("regiones troncales") o regiones genómicas de interés de puntos calientes ("regiones de puntos calientes" o "regiones de puntos calientes" o "puntos calientes" o "hotspots"). Mientras que los "puntos calientes" pueden referirse a loci concretos asociados a variantes de secuencia, las regiones "troncales" pueden referirse a regiones genómicas más amplias, cada una de las cuales puede tener una o más variantes de secuencia potenciales. Por ejemplo, una región troncal puede ser una región que contenga una o más mutaciones asociadas al cáncer, mientras que un punto caliente puede ser un locus con una mutación particular asociada al cáncer recurrente o un locus con una mutación recurrente particular asociada al cáncer. Las regiones genómicas de interés, tanto de la columna vertebral como de los puntos calientes, pueden incluir genes marcadores tumorales comúnmente incluidos en los ensayos de biopsia líquida (por ejemplo, BRAF, BRCA 1/2, EGFR, KRAS, PIK3CA, ROSI, TP53, y otros), para los que se puede esperar que se observen una o más variantes en sujetos con cáncer. En algunas realizaciones, se pueden utilizar microesferas marcadas con biotina con sondas para una o más regiones de interés para capturar secuencias diana, opcionalmente seguidas de amplificación de esas regiones, para enriquecer las regiones de interés.

La cantidad de datos de secuenciación que pueden obtenerse de una muestra es finita, y está limitada por factores como la calidad de las plantillas de ácido nucleico, el número de secuencias objetivo, la escasez de secuencias específicas, las limitaciones de las técnicas de secuenciación, y consideraciones prácticas como el tiempo y los gastos. Así pues, un "presupuesto de lectura" es una forma de conceptualizar la cantidad de información genética que puede extraerse de una muestra. Se puede seleccionar un presupuesto de lecturas por muestra que identifique el número total de lecturas de base que se asignarán a una muestra de prueba que comprende una cantidad predeterminada de ADN en un experimento de secuenciación. El presupuesto de lecturas puede basarse en el total de lecturas producidas, por ejemplo, incluyendo las lecturas redundantes producidas a través de la amplificación. Alternativamente, puede basarse en el número de moléculas únicas detectadas en la muestra. En ciertas realizaciones, el presupuesto de lectura puede reflejar la cantidad de soporte de doble cadena para una llamada en un locus. Es decir, el porcentaje de loci para los que se detectan lecturas de ambas cadenas de una molécula de ADN.

Los factores de un presupuesto de lectura incluyen la profundidad de lectura y la longitud del panel. Por ejemplo, un presupuesto de 3.000.000.000 de lecturas puede asignarse como 150.000 bases a una profundidad media de lectura de 20.000 lecturas/base. La profundidad de lectura puede referirse al número de moléculas que producen una lectura en un locus. En la presente divulgación, las lecturas en cada base pueden asignarse entre bases en la región de la columna vertebral del panel, a una primera profundidad media de lectura y bases en la región de puntos calientes del panel, a una profundidad de lectura mayor. En algunas realizaciones, una muestra se secuencia hasta una profundidad de lectura determinada por la cantidad de ácido nucleico presente en la muestra. En algunas realizaciones, una muestra se secuencia con una profundidad de lectura establecida, de forma que las muestras que comprenden diferentes cantidades de ácido nucleico se secuencian con la misma profundidad de lectura. Por ejemplo, una muestra de 300 ng de ácidos nucleicos puede secuenciarse con una profundidad de lectura de 1/10 de la de una muestra de 30 ng de ácidos nucleicos. En algunas realizaciones, los ácidos nucleicos de dos o más sujetos diferentes pueden añadirse juntos en una proporción basada en la cantidad de ácidos nucleicos obtenidos de cada uno de los sujetos.

A modo de ejemplo no limitativo, si un presupuesto de lecturas consiste en 100.000 recuentos de lecturas para una muestra dada, esos 100.000 recuentos de lecturas se dividirán entre lecturas de regiones troncales y lecturas de regiones hotspot. Si se asigna un gran número de esas lecturas (por ejemplo, 90.000 lecturas) a las regiones troncales, se asignará un pequeño número de lecturas (por ejemplo, las 10.000 lecturas restantes) a las regiones hotspot. A la inversa, si se asigna un gran número de lecturas (por ejemplo, 90.000 lecturas) a las regiones "hotspot", se asignará un pequeño número de lecturas (por ejemplo, las 10.000 lecturas restantes) a las regiones "backbone". Así, un trabajador cualificado puede asignar un presupuesto de lectura para proporcionar los niveles deseados de sensibilidad y especificidad. En ciertas realizaciones, el presupuesto de lecturas puede estar entre 100.000.000 de lecturas y 100.000.000.000 de lecturas, por ejemplo, entre 500.000.000 de lecturas y 50.000.000.000 de lecturas, o entre aproximadamente 1.000.000.000 de lecturas y 5.000.000.000 de lecturas a través de, por ejemplo, 20.000 bases a 100.000 bases.

Todos los polinucleótidos (por ejemplo, polinucleótidos amplificados) pueden ser sometidos a un dispositivo de secuenciación para ser secuenciados. Alternativamente, una muestra, o subconjunto, de todos los polinucleótidos amplificados se somete a un dispositivo de secuenciación para su secuenciación. Con respecto a cualquier polinucleótido original de doble cadena puede haber tres resultados con respecto a la secuenciación. En primer lugar, las lecturas de secuencias pueden generarse a partir de las dos cadenas complementarias de la molécula original (es decir, tanto de la cadena Watson como de la cadena Crick). En segundo lugar, las lecturas de secuencias sólo pueden generarse a partir de una de las dos cadenas complementarias (es decir, a partir de la cadena Watson o de la cadena Crick, pero no de ambas). En tercer lugar, no se puede generar ninguna lectura de secuencia a partir de ninguna de las dos cadenas complementarias. Por consiguiente, el recuento de las lecturas de secuencias únicas que corresponden a un locus genético subestimará el número de polinucleótidos de doble cadena de la muestra original que corresponden al locus. Aquí se describen métodos para estimar los polinucleótidos no vistos y no contados.

El método de secuenciación puede ser secuenciación paralela masiva, es decir, secuenciación simultánea (o en rápida sucesión) de al menos 100, 1000, 10.000, 100.000, 1 millón, 10 millones, 100 millones, o 1.000 millones de moléculas de polinucleótidos.

Los métodos de secuenciación pueden incluir, entre otros: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de molécula única, secuenciación por nanoporos, secuenciación por semiconductores, secuenciación por ligación, secuenciación por hibridación, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), secuenciación masiva en paralelo, Clonal Single Molecule Array (Solexa), secuenciación shotgun, secuenciación Maxam-Gilbert o Sanger, primer walking, secuenciación mediante plataformas PacBio, SOLiD, Ion Torrent, o Nanopore y cualquier otro método de secuenciación conocido en la técnica.

El método puede comprender la secuenciación de al menos 1 millón, 10 millones, 100 millones, 500 millones, 1.000 millones, 1.100 millones, 1.200 millones, 1.500 millones, 2.000 millones, 2.500 millones, 3.000 millones, 3.500 millones, 4.000 millones, 4.500 millones, 5.000 millones, 5.500 millones, 6.000 millones, 6.500 millones, 7.000 millones, 8.000 millones, 9.000 millones o 10.000 millones de pares de bases. En algunos casos, los métodos pueden comprender la secuenciación de aproximadamente 1.000 millones a aproximadamente 7.000 millones, de aproximadamente 1.100 millones a aproximadamente 6.800 millones, de aproximadamente 1.200 millones a aproximadamente 6.500 millones, de aproximadamente 1.100 millones a aproximadamente 6.400 millones, de aproximadamente 1.500 millones a aproximadamente 7.000 millones, de aproximadamente 2.000 millones a aproximadamente 6.000 millones, de aproximadamente 2.500 millones a aproximadamente 5.500 millones, de aproximadamente 3.000 millones a aproximadamente 5.000 millones de pares de bases. Por ejemplo, los métodos pueden comprender la secuenciación de aproximadamente 1.200 millones a aproximadamente 6.500 millones de pares de bases.

Marcadores Tumorales

Un marcador tumoral es una variante genética asociada a uno o más cánceres. Los marcadores tumorales pueden determinarse utilizando cualquiera de varios recursos o métodos. Un marcador tumoral puede haber sido descubierto previamente o puede ser descubierto de novo mediante técnicas experimentales o epidemiológicas. La

detección de un marcador tumoral puede ser indicativa de cáncer cuando el marcador tumoral está altamente correlacionado con un cáncer. La detección de un marcador tumoral puede ser indicativa de cáncer cuando un marcador tumoral en una región o gen se produce con una frecuencia que es mayor que una frecuencia para una determinada población de fondo o conjunto de datos.

Los recursos disponibles públicamente, como la literatura científica y las bases de datos, pueden describir en detalle las variantes genéticas que se han encontrado asociadas al cáncer. La literatura científica puede describir experimentos o estudios de asociación del genoma completo (GWAS) que asocian una o más variantes genéticas con el cáncer. Las bases de datos pueden agregar información obtenida de fuentes como la literatura científica para proporcionar un recurso más completo para determinar uno o más marcadores tumorales. Ejemplos no limitativos de bases de datos son FANTOM, GTex, GEO, Body Atlas, INSiGHT, OMIM (Online Mendelian Inheritance in Man, omim.org), cBioPortal (cbioportal.org), CIViC (Clinical Interpretations of Variants in Cancer, civic.genome.wustl.edu), DOCM (Database of Curated Mutations, docm.genome.wustl.edu), y ICGC Data Portal (dcc.icgc.org). Otro ejemplo es la base de datos COSMIC (Catalogue of Somatic Mutations in Cancer), que permite buscar marcadores tumorales por cáncer, gen, o tipo de mutación. Los marcadores tumorales también pueden determinarse de novo mediante la realización de experimentos como estudios de casos y controles o de asociación (por ejemplo, estudios de asociación de genoma completo).

Pueden detectarse uno o más marcadores tumorales en el panel de secuenciación. Un marcador tumoral puede ser una o más variantes genéticas asociadas al cáncer. Los marcadores tumorales pueden seleccionarse entre variantes de nucleótido único (SNV), variantes del número de copias (CNV), inserciones o deleciones (por ejemplo, indels), fusiones génicas e inversiones. Los marcadores tumorales pueden afectar al nivel de una proteína. Los marcadores tumorales pueden estar en un promotor o potenciador, y pueden alterar la transcripción de un gen. Los marcadores tumorales pueden afectar a la transcripción y/o a la eficacia traductora de un gen. Los marcadores tumorales pueden afectar a la estabilidad de un ARNm transcrito. El marcador tumoral puede dar lugar a un cambio en la secuencia de aminoácidos de una proteína traducida. El marcador tumoral puede afectar al splicing, puede cambiar el aminoácido codificado por un codón concreto, puede dar lugar a un desplazamiento de marco, o a un codón de parada prematuro. El marcador tumoral puede dar lugar a una sustitución conservadora de un aminoácido. Uno o más marcadores tumorales pueden dar lugar a una sustitución conservadora de un aminoácido. Uno o más marcadores tumorales pueden dar lugar a una sustitución no conservativa de un aminoácido.

Uno o más de los marcadores tumorales puede ser una mutación impulsora. Una mutación impulsora es una mutación que confiere una ventaja selectiva a una célula tumoral en su microentorno, ya sea aumentando su supervivencia o su reproducción. Ninguno de los marcadores tumorales puede ser una mutación impulsora. Uno o más de los marcadores tumorales puede ser una mutación pasajera. Una mutación pasajera es una mutación que no tiene ningún efecto sobre la aptitud de una célula tumoral, pero que puede estar asociada a una expansión clonal porque se produce en el mismo genoma que una mutación impulsora.

La frecuencia de un marcador tumoral puede ser tan baja como el 0,001%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,005%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,01%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,02%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,03%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,05%. La frecuencia de un marcador tumoral puede ser tan baja como el 0,1%. La frecuencia de un marcador tumoral puede ser tan baja como el 1%.

Ningún marcador tumoral puede estar presente en más del 50% de las personas con cáncer. Ningún marcador tumoral puede estar presente en más del 40% de los sujetos con cáncer. Ningún marcador tumoral puede estar presente en más del 30% de los sujetos con cáncer. Ningún marcador tumoral puede estar presente en más del 20% de los sujetos con cáncer. Ningún marcador tumoral puede estar presente en más del 10% de los sujetos con cáncer. Ningún marcador tumoral puede estar presente en más del 5% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,001% y el 50% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,01% y el 50% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,01% y el 30% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,01% y el 20% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,01% y el 10% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,1% y el 10% de los sujetos con cáncer. Un único marcador tumoral puede estar presente entre el 0,1% y el 5% de los sujetos con cáncer.

La detección de un marcador tumoral puede indicar la presencia de uno o más cánceres. La detección puede indicar la presencia de un cáncer seleccionado del grupo que comprende el cáncer de ovario, el cáncer de páncreas, el cáncer de mama, el cáncer colorrectal, el carcinoma pulmonar de células no pequeñas (por ejemplo, carcinoma de células escamosas o adenocarcinoma) o cualquier otro cáncer. La detección puede indicar la presencia de cualquier cáncer seleccionado del grupo que comprende el cáncer de ovario, el cáncer de páncreas, el cáncer de mama, el cáncer colorrectal, el carcinoma pulmonar de células no pequeñas (células escamosas o adenocarcinoma) o cualquier otro cáncer. La detección puede indicar la presencia de cualquiera de una pluralidad de cánceres seleccionados del grupo que comprende el cáncer de ovario, el cáncer de páncreas, el cáncer de mama, el cáncer colorrectal y el carcinoma pulmonar de células no pequeñas (células escamosas o adenocarcinoma), o cualquier otro cáncer. La detección puede indicar la presencia de uno o varios de los cánceres mencionados en esta solicitud.

Uno o más cánceres pueden presentar un marcador tumoral en al menos un exón del panel. Uno o más cánceres seleccionados del grupo que comprende el cáncer de ovario, el cáncer de páncreas, el cáncer de mama, el cáncer colorrectal, el carcinoma de pulmón de células no pequeñas (células escamosas o adenocarcinoma), o cualquier otro cáncer, presentan cada uno un marcador tumoral en al menos un exón del panel. Cada uno de al menos 3 de los cánceres puede presentar un marcador tumoral en al menos un exón del panel. Cada uno de al menos 4 de los cánceres puede presentar un marcador tumoral en al menos un exón del panel. Cada uno de al menos 5 de los cánceres puede presentar un marcador tumoral en al menos un exón del panel. Cada uno de al menos 8 de los cánceres puede presentar un marcador tumoral en al menos un exón del panel. Cada uno de al menos 10 de los cánceres puede presentar un marcador tumoral en al menos un exón del panel. Todos los cánceres pueden presentar un marcador tumoral en al menos un exón del panel.

Si un sujeto tiene un cáncer, el sujeto puede presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 85% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 90% de los sujetos con cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 92% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 95% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 96% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 97% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 98% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 99% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel. Al menos el 99,5% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos un exón o gen del panel.

Si un sujeto tiene un cáncer, el sujeto puede mostrar un marcador tumoral en al menos una región del panel. Al menos el 85% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 90% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 92% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 95% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 96% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 97% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 98% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 99% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel. Al menos el 99,5% de los sujetos que padecen un cáncer pueden presentar un marcador tumoral en al menos una región del panel.

La detección puede realizarse con una alta sensibilidad y/o una alta especificidad. La sensibilidad puede referirse a una medida de la proporción de positivos que se identifican correctamente como tales. En algunos casos, la sensibilidad se refiere al porcentaje de todos los marcadores tumorales existentes que se detectan. En algunos casos, la sensibilidad se refiere al porcentaje de personas enfermas a las que se identifica correctamente como portadoras de determinada enfermedad. La especificidad puede referirse a una medida de la proporción de negativos que se identifican correctamente como tales. En algunos casos, la especificidad se refiere a la proporción de bases inalteradas que se identifican correctamente. En algunos casos, la especificidad se refiere al porcentaje de personas sanas a las que se identifica correctamente como no portadoras de determinada enfermedad. El método de marcado no único descrito anteriormente aumenta significativamente la especificidad de la detección al reducir el ruido generado por los errores de amplificación y secuenciación, lo que reduce la frecuencia de falsos positivos. La detección puede realizarse con una sensibilidad de al menos 95%, 97%, 98%, 99%, 99,5%, o 99,9% y/o una especificidad de al menos 80%, 90%, 95%, 97%, 98% o 99%. La detección puede realizarse con una sensibilidad de al menos 90%, 95%, 97%, 98%, 99%, 99,5%, 99,6%, 99,98%, 99,9% o 99,95%. La detección puede realizarse con una especificidad de al menos 90%, 95%, 97%, 98%, 99%, 99,5%, 99,6%, 99,98%, 99,9% o 99,95%. La detección puede realizarse con una especificidad de al menos el 70% y una sensibilidad de al menos el 70%, una especificidad de al menos el 75% y una sensibilidad de al menos el 75%, una especificidad de al menos el 80% y una sensibilidad de al menos el 80%, una especificidad de al menos el 85% y una sensibilidad de al menos el 85%, una especificidad de al menos el 90% y una sensibilidad de al menos el 90%, una especificidad de al menos el 95% y una sensibilidad de al menos el 95%, una especificidad de al menos el 96% y una sensibilidad de al menos el 96%, una especificidad de al menos el 97% y una sensibilidad de al menos el 97%, una especificidad de al menos el 98% y una sensibilidad de al menos el 98%, una especificidad de al menos el 99% y una sensibilidad de al menos el 99%, o una especificidad del 100% y una sensibilidad del 100%. En algunos casos, los métodos pueden detectar un marcador tumoral con una sensibilidad de aproximadamente el 80% o superior. En algunos casos, los métodos pueden detectar un marcador tumoral con una sensibilidad de aproximadamente el 95% o superior. En algunos casos, los métodos pueden detectar un marcador tumoral con una sensibilidad de sensibilidad de alrededor del 80% o superior, y una sensibilidad de sensibilidad de alrededor del 95% o superior.

La detección puede ser muy precisa. La precisión puede aplicarse a la identificación de marcadores tumorales en el ADN libre de células, y/o al diagnóstico del cáncer. Pueden utilizarse herramientas estadísticas, como el análisis covariante descrito anteriormente, para aumentar y/o medir la precisión. Los métodos pueden detectar un marcador tumoral con una precisión de al menos el 80%, 90%, 95%, 97%, 98% o 99%, 99,5%, 99,6%, 99,98%, 99,9%, o 99,95%. En algunos casos, los métodos pueden detectar un marcador tumoral con una precisión de al menos el 95% o superior.

Límite de detección/intervalo de ruido

El ruido puede introducirse a través de errores en la copia y/o lectura de un polinucleótido. Por ejemplo, en un proceso de secuenciación, un polinucleótido único puede someterse primero a amplificación. La amplificación puede introducir errores, de modo que un subconjunto de los polinucleótidos amplificados puede contener, en un locus particular, una base que no es la misma que la base original en ese locus. Además, en el proceso de lectura, una base de un locus concreto puede leerse incorrectamente. Como consecuencia, la colección de lecturas de secuencias puede incluir un cierto porcentaje de llamadas de bases en un locus que no coinciden con la base original. En las tecnologías típicas de secuenciación, esta tasa de error puede ser de un solo dígito, por ejemplo, entre el 2% y el 3%. En algunos casos, la tasa de error puede ser de hasta aproximadamente el 10%, hasta aproximadamente el 9%, hasta aproximadamente el 8%, hasta aproximadamente el 7%, hasta aproximadamente el 6%, hasta aproximadamente el 5%, hasta aproximadamente el 4%, hasta aproximadamente el 3%, hasta aproximadamente el 2%, o hasta aproximadamente el 1%. Cuando se secuencia una colección de moléculas que se supone que tienen todas la misma secuencia, este ruido puede ser lo suficientemente pequeño como para poder identificar la base original con gran fiabilidad.

Sin embargo, si una colección de polinucleótidos parentales incluye un subconjunto de polinucleótidos que varían en un locus particular, el ruido puede ser un problema significativo. Este puede ser el caso, por ejemplo, cuando el ADN libre de células incluye no sólo ADN de la línea germinal, sino ADN de otra fuente, como ADN fetal o ADN de una célula cancerosa. En este caso, si la frecuencia de moléculas con variantes de secuencia puede estar en el mismo intervalo que la frecuencia de errores introducidos por el proceso de secuenciación, entonces las verdaderas variantes de secuencia pueden no ser distinguibles del ruido. Esto puede interferir, por ejemplo, con la detección de variantes de secuencia en una muestra. Por ejemplo, las secuencias pueden tener una tasa de error por base del 0,5-1%. El sesgo de amplificación y los errores de secuenciación introducen ruido en el producto final de la secuenciación. Este ruido puede disminuir la sensibilidad de la detección. Como ejemplo no limitativo, las variantes de secuencia cuya frecuencia es inferior a la tasa de error de secuenciación pueden confundirse con ruido.

Un intervalo de ruido o límite de detección se refiere a los casos en los que la frecuencia de moléculas con variantes de secuencia está en el mismo intervalo que la frecuencia de errores introducidos por el proceso de secuenciación. Un "límite de detección" también puede referirse a casos en los que se secuencian muy pocas moléculas portadoras de variantes para que la variante pueda detectarse. La frecuencia de moléculas con variantes de secuencia puede estar en el mismo intervalo que la frecuencia de errores como resultado de una pequeña cantidad de moléculas de ácido nucleico. Como ejemplo no limitativo, una cantidad muestreada de ácidos nucleicos, por ejemplo 100 ng, puede contener un número relativamente pequeño de moléculas de ácidos nucleicos libres de células, por ejemplo moléculas de ADN tumoral circulante, de tal forma que la frecuencia de una variante de secuencia puede ser baja, aunque la variante pueda estar presente en una mayoría de moléculas de ADN tumoral circulante. Alternativamente, la variante de secuencia puede ser rara o producirse sólo en una cantidad muy pequeña de los ácidos nucleicos muestreados, de forma que una variante detectada no se distinga del ruido y/o del error de secuenciación. Como ejemplo no limitativo, en un locus particular, un marcador tumoral puede detectarse sólo en el 0,1% al 5% de todas las lecturas en ese locus.

La distorsión puede manifestarse en el proceso de secuenciación como una diferencia en la intensidad de la señal, por ejemplo, el número total de lecturas de secuencia, producida por moléculas en una población parental con la misma frecuencia. La distorsión puede introducirse, por ejemplo, a través de un sesgo de amplificación, un sesgo de GC o un sesgo de secuenciación. Esto puede interferir en la detección de la variación del número de copias en una muestra. El sesgo de GC se traduce en una representación desigual de las zonas ricas o pobres en contenido de GC en la lectura de la secuencia. Además, al proporcionar lecturas de secuencias en mayor o menor cantidad que su número real en una población, el sesgo de amplificación puede distorsionar las mediciones de la variación del número de copias.

Una forma de reducir el ruido y/o la distorsión de una única molécula individual o de un conjunto de moléculas es agrupar las lecturas de secuencia en familias derivadas de moléculas individuales originales para reducir el ruido y/o la distorsión de una única molécula individual o de un conjunto de moléculas. La conversión eficiente de polinucleótidos individuales de una muestra de material genético inicial en polinucleótidos parentales marcados listos para secuencia puede aumentar la probabilidad de que los polinucleótidos individuales de una muestra de material genético inicial estén representados en una muestra lista para secuencia. Esto puede producir información sobre la secuencia de más polinucleótidos en la muestra inicial. Además, la generación de secuencias consenso de alto rendimiento para polinucleótidos parentales marcados mediante el muestreo a alta velocidad de polinucleótidos progenie amplificados a partir de los polinucleótidos parentales marcados, y el colapso de las lecturas de secuencias generadas en secuencias consenso que representan secuencias de polinucleótidos parentales marcados pueden reducir el ruido introducido por el sesgo de amplificación y/o los errores de secuenciación, y pueden aumentar la sensibilidad de la detección. Una forma de reducir el ruido en el mensaje recibido de una molécula es colapsar las lecturas de secuencia en una secuencia de consenso. El uso de funciones probabilísticas que convierten las frecuencias recibidas en estimaciones de probabilidad o posteriores de cada uno de los posibles nucleótidos verdaderos utilizando estimaciones definidas de los perfiles de error de amplificación y secuenciación es otra forma de reducir el ruido y/o la distorsión. Con respecto a un conjunto de moléculas, agrupar las lecturas en familias y determinar una medida cuantitativa de las familias reduce la distorsión, por ejemplo, en la cantidad de moléculas en cada uno de una pluralidad de loci diferentes. Una vez más, el colapso de las lecturas de secuencias de diferentes familias en secuencias de consenso elimina los errores introducidos por la amplificación y/o el error de secuenciación. Además, la determinación de las frecuencias de las llamadas base a partir de

probabilidades derivadas de la información familiar también reduce el ruido en el mensaje recibido de un conjunto de moléculas. Los informes de frecuencia o las llamadas de marcadores tumorales también pueden realizarse utilizando una pluralidad de secuencias de referencia y observaciones de cobertura, a partir de las cuales se determinará una frecuencia de observación de un marcador tumoral en una posición. Las secuencias de referencia pueden comprender secuencias o perfiles de marcadores de individuos sanos o de individuos que padezcan una enfermedad o afección, como el cáncer. Se puede utilizar una frecuencia de muestras de referencia "conocidas" para establecer una frecuencia umbral para realizar una llamada de detección de marcadores. Por ejemplo, una frecuencia del 0,1% para un nucleótido que tenga una "A" en una determinada posición puede utilizarse como umbral para determinar si se denomina o no "A" a una base en esa posición en un sujeto de ensayo. Por ejemplo, al menos 20, al menos 50, al menos 100, al menos 500, al menos 1.000, al menos 2.000, al menos 3.000, al menos 4.000, al menos 5.000, al menos 6.000, al menos 7.000, al menos 8.000, al menos 9.000, al menos 10.000, al menos 11.000, al menos 12.000, al menos 13.000, al menos 14.000, al menos 15.000, al menos 16.000, al menos 17.000, al menos 18.000, al menos 19.000, al menos 20.000, al menos 30.000, al menos 40.000, al menos 50.000, al menos 60.000, al menos 70.000, al menos 80.000, al menos 90.000, o al menos 100.000 secuencias de referencia.

El ruido y/o la distorsión pueden reducirse aún más identificando moléculas contaminantes de otras muestras procesadas, comparando la información de marcación y localización de moléculas con una colección de moléculas observadas dentro de la muestra que se está procesando o a través de lotes de muestras. El ruido y/o la distorsión pueden reducirse aún más comparando las variaciones genéticas de una secuencia leída con las variaciones genéticas de otras secuencias leídas. Una variación genética observada en una lectura de secuencia y de nuevo en otras lecturas de secuencia aumenta la probabilidad de que una variante detectada sea de hecho un marcador tumoral y no un mero error de secuenciación o ruido. Como ejemplo no limitativo, si se observa una variación genética en una primera lectura de secuencia y también se observa en una segunda lectura de secuencia, se puede hacer una inferencia bayesiana respecto a si la variación es de hecho una variación genética y no un error de secuenciación.

La detección repetida de una variante puede aumentar la probabilidad y/o confianza de que una variante se detecte con precisión. Una variante puede detectarse repetidamente comparando dos o más conjuntos de datos genéticos o variaciones genéticas. Los dos o más conjuntos de variaciones genéticas pueden detectarse tanto en muestras en múltiples puntos temporales como en diferentes muestras en el mismo punto temporal (por ejemplo, una muestra de sangre reanalizada). Al detectar una variante en el intervalo de ruido o por debajo del umbral de ruido, el remuestreo o la detección repetida de una variante de baja frecuencia hace más probable que la variante sea de hecho una variante y no un error de secuenciación. El remuestreo puede realizarse a partir de la misma muestra, como en el caso de una muestra que se vuelve a analizar o ejecutar, o a partir de muestras en diferentes momentos.

La detección de covariantes puede aumentar la probabilidad y/o confianza de que una variante sea detectada con precisión. En el caso de los marcadores tumorales covariantes, la presencia de un marcador tumoral se asocia con la presencia de otro u otros marcadores tumorales. Basándose en la detección de una variación genética covariante, puede ser posible inferir la presencia de una variación genética covariante asociada, incluso cuando la variación genética asociada esté presente por debajo de un límite de detección. Alternativamente, basándose en la detección de una variación genética covariante, puede aumentarse la indicación de confianza diagnóstica para la variación genética asociada. Además, en algunos casos en los que se detecta una variante covariante, puede disminuirse un umbral de detección para una variante covariante detectada por debajo de un límite de detección. Ejemplos no limitantes de variaciones o genes covariantes incluyen: mutaciones conductoras y mutaciones de resistencia, mutaciones conductoras y mutaciones pasajeras. Un ejemplo específico de covariantes o genes es la mutación activadora EGFR L858R y la mutación de resistencia EGFR T790M, que se encuentran en los cánceres de pulmón. Numerosas otras variantes covariantes y genes están asociados con diversas mutaciones de resistencia y serán reconocidas por un experto en la materia.

En una realización, utilizando mediciones de una pluralidad de muestras recogidas sustancialmente a la vez o a lo largo de una pluralidad de puntos temporales, la indicación de confianza diagnóstica para cada variante puede ajustarse para indicar una confianza de predecir la observación de la variación del número de copias (CNV) o mutación o marcador tumoral. La confianza puede aumentarse utilizando mediciones en una pluralidad de puntos temporales para determinar si el cáncer está avanzando, en remisión o estabilizado. La indicación de confianza en el diagnóstico puede asignarse mediante cualquiera de los métodos estadísticos y puede basarse, al menos en parte, en la frecuencia con la que se observan las mediciones durante un periodo de tiempo. Por ejemplo, puede hacerse una correlación estadística de los resultados actuales y los anteriores. Alternativamente, para cada diagnóstico, puede construirse un modelo oculto de Markov, de forma que pueda tomarse una decisión de máxima verosimilitud o máxima a posteriori basada en la frecuencia de aparición de un evento de prueba concreto a partir de una pluralidad de mediciones o de puntos temporales. Como parte de este modelo, también se puede emitir la probabilidad de error y la indicación de confianza de diagnóstico resultante para una decisión concreta. De este modo, las mediciones de un parámetro, estén o no en el intervalo de ruido, pueden estar provistas de un intervalo de confianza. Si se comprueba a lo largo del tiempo, se puede aumentar la confianza predictiva de si un cáncer está avanzando, estabilizado o en remisión comparando los intervalos de confianza a lo largo del tiempo. Dos puntos de tiempo de muestreo pueden estar separados por al menos aproximadamente 1 microsegundo, 1 milisegundo, 1 segundo, 10 segundos, 30 segundos, 1 minuto, 10 minutos, 30 minutos, 1 hora, 12 horas, 1 día, 1 semana, 2 semanas, 3 semanas, un mes, o un año. Dos puntos temporales pueden estar separados por aproximadamente un mes a aproximadamente un año, aproximadamente un año a aproximadamente 5 años, o no más

de aproximadamente tres meses, dos meses, un mes, tres semanas, dos semanas, una semana, un día, o doce horas. En algunas realizaciones, dos puntos temporales pueden estar separados por un evento terapéutico como la administración de un tratamiento o la realización de un procedimiento quirúrgico. Cuando los dos puntos temporales están separados por el evento terapéutico, la CNV o las mutaciones detectadas pueden compararse antes y después del evento.

Una vez recogidos los datos de secuenciación de las secuencias de polinucleótidos libres de células, pueden aplicarse uno o más procesos bioinformáticos a los datos de secuencia para detectar características o variaciones genéticas tales como características del ADNcf en elementos reguladores, patrones de espaciado nucleosómico/unión a nucleosomas, modificaciones químicas de los ácidos nucleicos, variación del número de copias y mutaciones o cambios en marcadores epigenéticos, incluidos, entre otros, perfiles de metilación, y variantes genéticas tales como SNV, CNV, indels, y/o fusiones. En algunos casos, en los que se desea analizar la variación del número de copias, los datos de la secuencia pueden ser: 1) se alinean con un genoma de referencia y se asignan a moléculas individuales; 2) se filtran; 4) se dividen en ventanas o intervalos de una secuencia; 5) se cuentan las lecturas de cobertura y las moléculas de cada ventana; 6) las moléculas de cobertura se pueden normalizar mediante un algoritmo de modelado estadístico; y 7) se puede generar un archivo de salida que refleje los estados discretos del número de copias en varias posiciones del genoma. En algunos casos, se cuenta el número de lecturas de cobertura / moléculas o lecturas de cobertura normalizadas que se alinean con un locus concreto del genoma de referencia. En otros casos, en los que se desea realizar un análisis de mutaciones, los datos de la secuencia pueden 1) alinearse con un genoma de referencia y mapearse a moléculas individuales; 2) filtrarse; 4) calcularse la frecuencia de bases variantes basándose en lecturas de cobertura para esa base específica; 5) normalizarse la frecuencia de bases variantes utilizando un algoritmo de modelado estocástico, estadístico o probabilístico; y 6) puede generarse un archivo de salida que refleje los estados de mutación en varias posiciones del genoma. Un genoma de referencia para el mapeo puede incluir el genoma de cualquier especie de interés. Las secuencias del genoma humano útiles como referencia pueden incluir el ensamblaje hgl9, GRCh38.p4, o cualquier ensamblaje hg anterior o disponible. Estas secuencias pueden consultarse mediante el navegador del genoma disponible en genome.ucsc.edu/index.html. Los genomas de otras especies incluyen, por ejemplo, PanTro2 (chimpancé) y mm9 (ratón).

En algunos casos, los identificadores (como los que incluyen códigos de barras) pueden utilizarse para agrupar lecturas de secuencias durante el análisis de mutaciones. En algunos casos, las lecturas de secuencias se agrupan en familias, por ejemplo, utilizando identificadores o una combinación de identificadores y posiciones de inicio/parada o secuencias. En algunos casos, se puede realizar una llamada de base comparando nucleótidos en una o más familias con una secuencia de referencia y determinando la frecuencia de una base particular 1) dentro de cada familia, y 2) entre las familias y las secuencias de referencia. Se puede realizar una llamada de base nucleotídica basándose en criterios como el porcentaje de familias que tienen una base en una posición. En algunos casos, se informa de una llamada base si su frecuencia es mayor que un umbral de ruido determinado por la frecuencia en una pluralidad de secuencias de referencia (por ejemplo, secuencias de individuos sanos). La información temporal de los análisis actuales y anteriores del paciente o sujeto se utiliza para mejorar el análisis y la determinación. En algunas realizaciones, la información de la secuencia del paciente o sujeto se compara con la información de la secuencia obtenida de una cohorte de individuos sanos, una cohorte de pacientes con cáncer o ADN de línea germinal del paciente o sujeto. El ADN de la línea germinal puede obtenerse, sin limitación, de fluidos corporales, sangre entera, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejidos, líquido sinovial, líquido linfático, líquido de ascitis, líquido intersticial o extracelular, el líquido de los espacios entre células, incluido el líquido crevicular gingival, médula ósea, líquido cefalorraquídeo, saliva, mucosa, esputo, semen, sudor, orina, o cualquier otro fluido corporal. Una cohorte de pacientes con cáncer puede tener el mismo tipo de cáncer que el paciente o sujeto, el mismo estadio de cáncer que el paciente o sujeto, ambos, o ninguno. En algunas realizaciones, se utiliza una cohorte de pacientes con cáncer, una cohorte de individuos sanos o ADN de línea germinal del sujeto para proporcionar una frecuencia de referencia de una base en una posición, y la frecuencia de referencia se utiliza para realizar una llamada de base en el sujeto. Sin limitación, una frecuencia para una base en una posición en una cohorte de individuos sanos, o ADN de línea germinal del sujeto puede compararse con la frecuencia de una base detectada entre lecturas de secuencia del sujeto.

En algunas realizaciones, los métodos y sistemas de la presente divulgación pueden utilizarse para detectar una frecuencia alélica menor (MAF) de 0,025% o inferior, 0,05% o inferior, 0,075% o inferior, o 0,1% o inferior. La variación del número de copias puede medirse como una relación entre (1) recuentos de moléculas únicas (UMC) para un gen en una muestra de prueba y (2) UMC para ese gen en una muestra de referencia (por ejemplo, muestra de control). En algunas realizaciones, los métodos y sistemas de la presente divulgación pueden utilizarse para detectar una variación del número de copias que es una amplificación del número de copias (CNA). En algunas realizaciones, los métodos y sistemas de la presente divulgación pueden utilizarse para detectar un CNA de al menos 1,5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 55, 60, o más. En algunas realizaciones, los métodos y sistemas de la presente divulgación pueden utilizarse para detectar una variación del número de copias que es una pérdida del número de copias (CNL). En algunas realizaciones, los métodos y sistemas de la presente divulgación pueden utilizarse para detectar una CNL inferior a 0,9, 0,8, 0,7, 0,6, 0,5, 0,4, 0,3, 0,2, 0,1, o 0,05.

Una variedad de diferentes reacciones y/operaciones pueden ocurrir dentro de los sistemas y métodos aquí divulgados, incluyendo pero no limitado a: secuenciación de ácido nucleico, cuantificación de ácido nucleico, optimización de secuenciación, detección de expresión génica, cuantificación de expresión génica, perfil genómico, perfil de cáncer, o análisis de marcadores expresados. Además, los sistemas y métodos tienen numerosas aplicaciones médicas. Por

ejemplo, puede utilizarse para la identificación, detección, diagnóstico, tratamiento, seguimiento, estadificación o predicción del riesgo de diversas enfermedades y trastornos genéticos y no genéticos, incluido el cáncer. Puede utilizarse para evaluar la respuesta del sujeto a diferentes tratamientos de las enfermedades genéticas y no genéticas, o proporcionar información relativa a la progresión y el pronóstico de la enfermedad.

Sistemas informáticos de control

La presente divulgación proporciona sistemas de control por ordenador que están programados para implementar métodos de la divulgación. FIG. 1 muestra un sistema informático 101 que está programado o configurado de otro modo para analizar datos de secuenciación, detectar marcadores tumorales y determinar el estado del cáncer. El sistema informático 101 puede regular varios aspectos del análisis de secuencias de la presente divulgación, como, por ejemplo, cotejar datos con secuencias y variantes conocidas. El sistema informático 101 puede ser un dispositivo electrónico de un usuario o un sistema informático que se encuentra a distancia con respecto al dispositivo electrónico. El dispositivo electrónico puede ser un dispositivo electrónico móvil.

El sistema informático 101 incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador informático" en el presente documento) 105, que puede ser un procesador de núcleo único o multi-núcleo, o una pluralidad de procesadores para procesamiento paralelo. El Sistema Informático 101 también incluye memoria o ubicación de memoria 110 (por ejemplo, memoria de acceso aleatorio, memoria de sólo lectura, memoria flash), unidad de almacenamiento electrónico 115 (por ejemplo, disco duro), interfaz de comunicación 120 (por ejemplo, adaptador de red) para comunicarse con uno o más sistemas, y dispositivos periféricos 125, como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónica. La memoria 110, la unidad de almacenamiento 115, la interfaz 120 y los dispositivos periféricos 125 están en comunicación con la CPU 105 a través de un bus de comunicación (líneas continuas), como una placa base. La unidad de almacenamiento 115 puede ser una unidad de almacenamiento de datos (o repositorio de datos) para almacenar datos. El sistema informático 101 puede acoplarse operativamente a una red informática ("red") 130 con la ayuda de la interfaz de comunicación 120. La red 130 puede ser Internet, una Internet y/o extranet, o una intranet y/o extranet que esté en comunicación con Internet. En algunos casos, la red 130 es una red de telecomunicaciones y/o de datos. La red 130 puede incluir uno o más servidores informáticos, que pueden permitir la computación distribuida, como la computación en nube. La red 130, en algunos casos con la ayuda del sistema informático 101, puede implementar una red peer-to-peer, que puede permitir a los dispositivos acoplados al sistema informático 101 comportarse como un cliente o un servidor.

La CPU 105 puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden estar incorporadas en un programa o software. Las instrucciones pueden almacenarse en una ubicación de memoria, como la memoria 110. Las instrucciones pueden dirigirse a la CPU 105, que posteriormente puede programar o configurar de otro modo. La CPU 105 para implementar métodos de la presente divulgación. Ejemplos de operaciones realizadas por la CPU 105 pueden incluir búsqueda, decodificación, ejecución, y escritura.

La CPU 105 puede ser parte de un circuito, tal como un circuito integrado. Uno o más componentes del sistema 101 pueden incluirse en el circuito. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC).

La unidad de almacenamiento 115 puede almacenar archivos, como controladores, bibliotecas y programas guardados. La unidad de almacenamiento 115 puede almacenar datos de usuario, por ejemplo, preferencias de usuario y programas de usuario. En algunos casos, el sistema informático 101 puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático 101, como las ubicadas en un servidor remoto que está en comunicación con el sistema informático 101 a través de una intranet o Internet.

El Sistema Informático 101 puede comunicarse con uno o más Sistemas Informáticos remotos a través de la red 130. Por ejemplo, el sistema informático 101 puede comunicarse con un sistema informático remoto de un usuario (por ejemplo, un médico). Algunos ejemplos de sistemas informáticos remotos son los ordenadores personales (p. ej., PC portátiles), las tabletas (p. ej., Apple® iPad, Samsung® Galaxy Tab), los teléfonos, los teléfonos inteligentes (p. ej., Apple® iPhone, dispositivos con Android, Blackberry®), o los asistentes personales digitales. El usuario puede acceder al sistema informático 101 a través de la red 930.

Los métodos aquí descritos pueden implementarse mediante código ejecutable por máquina (por ejemplo, procesador de ordenador) almacenado en una ubicación de almacenamiento electrónico del sistema informático 101, como, por ejemplo, en la memoria 110 o en la unidad de almacenamiento electrónico 115. El código ejecutable o legible por máquina puede proporcionarse en forma de software. Durante su uso, el código puede ser ejecutado por el procesador 105. En algunos casos, el código puede recuperarse de la unidad de almacenamiento 115 y almacenarse en la memoria 110 para que el procesador 105 pueda acceder a él. En algunas situaciones, la unidad de almacenamiento electrónico 115 puede excluirse, y las instrucciones ejecutables por máquina se almacenan en la memoria 110.

El código puede ser pre-compilado y configurado para su uso con una máquina que tenga un procesador adaptado para ejecutar el código, o puede ser compilado durante el tiempo de ejecución. El código puede suministrarse en un lenguaje de programación que puede seleccionarse para permitir que el código se ejecute de forma precompilada o as-compilada.

Aspectos de los Sistemas y métodos aquí proporcionados, tales como el Sistema Informático 101, pueden ser incorporados en programación. Varios aspectos de la tecnología pueden considerarse "productos" o "artículos de fabricación", normalmente en forma de código ejecutable por máquina (o procesador) y/o datos asociados que se transportan o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina puede almacenarse en una unidad de almacenamiento electrónico, como una memoria (por ejemplo, memoria de sólo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir cualquiera o todas las memorias tangibles de los ordenadores, procesadores o similares, o módulos asociados a los mismos, como diversas memorias semiconductoras, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación del software. En ocasiones, la totalidad o parte del software puede comunicarse a través de Internet o de otras redes de telecomunicaciones. Dichas comunicaciones, por ejemplo, pueden permitir la carga del software desde un ordenador o procesador a otro, por ejemplo, desde un servidor de gestión u ordenador anfitrión a la plataforma informática de un servidor de aplicaciones. Así pues, otro tipo de medios que pueden portar los elementos de software incluyen las ondas ópticas, eléctricas y electromagnéticas, como las utilizadas a través de interfaces físicas entre dispositivos locales, a través de redes fijas cableadas y ópticas y a través de diversos enlaces aéreos. Los elementos físicos que transportan dichas ondas, como los enlaces por cable o inalámbricos, los enlaces ópticos o similares, también pueden considerarse medios portadores del software. Tal y como se utilizan en el presente documento, a menos que se restrinjan a medios de "almacenamiento" tangibles no transitorios, términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participe en el suministro de instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, como un código ejecutable por ordenador, puede adoptar muchas formas, incluyendo, pero sin limitarse a, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión físico. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, como cualquiera de los dispositivos de almacenamiento de cualquier ordenador(es) o similar(es), como los que pueden utilizarse para implementar las bases de datos, etc. que se muestran en los dibujos. Los medios de almacenamiento volátiles incluyen la memoria dinámica, como la memoria principal de dicha plataforma informática. Los medios de transmisión tangibles incluyen cables coaxiales, cables de cobre y fibra óptica, incluidos los cables que componen un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden adoptar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o luminosas como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarrojos (IR). Las formas comunes de medios legibles por ordenador incluyen por ejemplo: un disquete, un disco flexible, un disco duro, una cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, tarjetas perforadas, cinta de papel, cualquier otro medio de almacenamiento físico con patrones de agujeros, una RAM, una ROM, una PROM y EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporte datos o instrucciones, cables o enlaces que transporten dicha onda portadora, o cualquier otro medio a partir del cual un ordenador pueda leer código de programación y/o datos. Muchas de estas formas de medios legibles por ordenador pueden participar en el transporte de una o más secuencias de una o más instrucciones a un procesador para su ejecución.

El sistema informático 101 puede incluir o estar en comunicación con una pantalla electrónica 135 que comprende una interfaz de usuario (UI) 140 para proporcionar, por ejemplo, información sobre el diagnóstico del cáncer. Ejemplos de interfaces de usuario incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en web.

En un aspecto, se proporciona en el presente documento un sistema que comprende un ordenador con un procesador y una memoria de ordenador, en el que el ordenador está en comunicación con una red de comunicaciones, y en el que la memoria de ordenador comprende un código que, cuando es ejecutado por el procesador, (1) recibe datos de secuencia en la memoria de ordenador desde la red de comunicaciones; (2) determina si una variante genética en los datos de secuencia representa un mutante de línea germinal o un mutante de células somáticas, utilizando los métodos descritos en el presente documento; y (3) comunica, a través de la red de comunicaciones, la determinación.

La red de comunicaciones puede ser cualquier red disponible que se conecte a Internet. La red de comunicaciones puede utilizar, por ejemplo, una red de transmisión de alta velocidad que incluya, sin limitación, banda ancha sobre líneas eléctricas (BPL), módem por cable, línea de abonado digital (DSL), fibra, satélite, e inalámbrica.

En un aspecto, se proporciona en el presente documento un sistema que comprende: una red de área local; uno o más secuenciadores de ADN que comprenden una memoria de ordenador configurada para almacenar datos de secuencias de ADN que están conectados a la red de área local; un ordenador bioinformático que comprende una memoria de ordenador y un procesador, que está conectado a la red de área local; en el que el ordenador comprende además un código que, cuando se ejecuta, copia los datos de secuencias de ADN almacenados en un secuenciador de ADN, escribe los datos copiados en la memoria del ordenador bioinformático, y realiza los pasos descritos en el presente documento.

Los métodos y sistemas de la presente divulgación pueden implementarse mediante uno o más algoritmos. Un algoritmo puede implementarse mediante software al ser ejecutado por la unidad central de procesamiento 105. El algoritmo puede, por ejemplo, determinar si un cáncer está presente y/o progresando.

Mientras que las realizaciones preferidas de la divulgación presente se han demostrado y se han descrito adjunto, será obvio a ésos expertos en el arte que tales realizaciones están proporcionadas a modo de ejemplo solamente. No se pretende que la presente divulgación se vea limitada por los ejemplos específicos proporcionados en la especificación. Aunque la presente divulgación se ha descrito con referencia a la especificación antes mencionada, las descripciones e ilustraciones de las realizaciones no deben interpretarse en un sentido limitativo. Numerosas variaciones, cambios y sustituciones ocurrirán ahora a los expertos en la materia sin apartarse de la presente divulgación.

EJEMPLOS

Ejemplo 1: Ensayo de secuenciación de nueva generación para la detección de ADNct en pacientes con cáncer en fase inicial

En este estudio se incluyeron datos de secuenciación de ADNcf desidentificados de 10.288 pacientes con cáncer avanzado (pts) sometidos a pruebas clínicas de ADN tumoral circulante (73 genes de la Tabla 2). El ADNcf se extrajo del plasma y se cuantificó. Se preparó una biblioteca de ADN y se secuenció con una profundidad de lectura media de 15.000X. Mediante el análisis de variantes Ingenuity, se clasificaron las mutaciones puntuales y las pequeñas indels sospechosas de origen germinal (fracción alélica del 40-60%) siguiendo las directrices del American College of Medical Genetics and Genomics. Se estudiaron más de 50 tipos de cáncer, entre ellos los de pulmón (40%), mama (20%), colorrectal (8%), próstata (6%), y páncreas (3%). La edad media de los sujetos era de 63,6 años (intervalo: 18-95), y el 42% eran hombres. De las 34.873 variantes putativas de la línea germinal identificadas, 520 (1,5%) eran patogénicas o probablemente patogénicas (PV), 16.939 (49%) eran de significado incierto, y 17.414 (50%) eran benignas o probablemente benignas. De los 250 sujetos (2,4%) con PV del gen del síndrome del cáncer hereditario, 83 fueron excluidos debido al alto nivel de carga tumoral somática, lo que dejó 167 (1,6%) con PV de la línea germinal putativa; las tasas fueron mayores en pacientes menores de 50 años frente a los de 50 años o más en general (3,3% frente a 1,4%, p = 0,02) y en pacientes con cáncer de mama (4,3% frente a 1,5%, p = 0,03). Los resultados figuran en la Tabla 4.

Tabla 4: PV por tipo de cáncer (N)

Gen	TODOS	Ovarios	Páncreas	Próstata	Mama	Pulmón	CCR	Otros
BRCA2	**78**	3	7	18	26	17	1	6
BRCA1	**38**	11	1		9	12	1	4
TP53	**16**				4	8	2	2
CDKN2A	**10**		1		3	2		4
ATM	**5**		1		2	2		
KIT	**4**				1	2	1	
NF1	**4**			1		1		2
RET	**4**		1			2		1
APC	**4**						1	3
RB1	**2**							2
MLH1	**1**						1	
SMAD4	**1**						1	
TOTAL	**167**	**14**	**11**	**19**	**45**	**46**	**8**	**24**
# pacientes	10288	205	328	577	2047	4136	830	2165
% pacientes	**1.6%**	**6.8%**	**3.4%**	**3.3%**	**2.2%**	**1.1%**	**1.0%**	**1.1%**

[1]Genes Lynch no secuenciados excepto el exón 12 de MLH1

La frecuencia observada de PV de línea germinal putativa identificada incidentalmente fue inferior a la tasa de línea germinal verdadera, pero estos hallazgos ilustran que la detección a partir de ADNcf es clínicamente factible. Es importante destacar que los hallazgos incidentales en la línea germinal podrían influir en la planificación del tratamiento oncológico (por ejemplo, inhibidores de PARP para mutaciones BRCAI/2) y podrían beneficiar a las familias a través de una mayor vigilancia/prevención primaria.

Ejemplo 2: Discriminación de mutaciones germinales EGFR T790M en ADN libre de células

Para determinar si el análisis genómico del ADNcf plasmático podría permitir la determinación simultánea del genotipo tumoral y de la línea germinal, con una resolución precisa de las variantes derivadas del tumor y de las variantes de la línea germinal, se estudiaron las variantes somáticas y de la línea germinal en el gen EGFR, incluidas las variantes conocidas de la línea germinal y un grupo de mutaciones oncogénicas presentes en el 10-20% de los pacientes con NSCLC. Una mutación del EGFR, la T790M, puede detectarse raramente como variante de la línea germinal, donde su presencia se ha asociado con el cáncer de pulmón familiar. EGFR T790M es más comúnmente visto como una mutación somática adquirida después de que un paciente con NSCLC desarrolla resistencia a los inhibidores de la tirosina quinasa EGFR (TKIs). Los cánceres de pulmón que albergan resistencia mediada por T790M tras la terapia inicial muestran sensibilidad a un EGFR TKI de tercera generación, osimertinib.

Una mujer de 49 años, nunca fumadora y con antecedentes familiares de cáncer de pulmón, presentó un adenocarcinoma de pulmón metastásico con progresión primaria en tratamiento con afatinib, un inhibidor de la tirosina quinasa (ITC) EGFR de segunda generación. El genotipado tisular inicial había mostrado mutaciones EGFR L858R y T790M, así como otras alteraciones somáticas en CDKN2A, TP53, y CTNNB 1. Debido a la mutación L858R en EGFR, se inició afatinib de primera línea. Sin embargo, el paciente volvió con metástasis cerebrales progresivas tras sólo dos meses de tratamiento. En el momento de la remisión, la secuenciación plasmática de nueva generación (NGS) demostró las variantes EGFR L858R, TP53, y CTNNB 1 previamente observadas en una fracción alélica (FA) del 1,4-5,3%, mientras que el alelo EGFR T790M se detectó en una FA del 50,9%, como se observa en la Tabla 5.

Tabla 5

Alteración	Tiempo 1 (AF)	Tiempo 2 (AF)	Tiempo 3 (AF)
EGFR L858R	5.3%	0.6%	18.1%
EGFR T790M	50.9%	49.2%	54.4%
EGFR C797S	ND	ND	1.3%
EGFR Q787Q	51.5%	48.7%	54.8%
TP53 P278R	3.8%	ND	19.7%

Se inició el tratamiento con osimertinib, un inhibidor de la tirosina quinasa (ITC) del EGFR que es activo en el contexto de la resistencia mediada por EGFR T790M a los ITC iniciales del EGFR, y obtuvo un beneficio clínico durante nueve meses, momento en el que las exploraciones mostraron una progresión temprana en el pulmón. La repetición de la NGS en plasma mostró la variante EGFR L858R en un 0,6% de AF, pero T790M se mantuvo relativamente estable en un 49,2% de FA (FIG. 7, donde 701 es EGFR L858R; 702 es EGFR T790M; 703 es EGFR Q787Q y 704 es TP53 P278R). A continuación, el paciente recibió un tratamiento en investigación en un ensayo clínico y desarrolló una mayor progresión de la enfermedad. La repetición de la NGS en plasma en este momento demostró un aumento de los niveles de EGFR L858R en un 18% de FA, T790M en un 54% de AF, y una tercera mutación de EGFR que media la resistencia adquirida a osimertinib, C797S, en un 1,3% de FA. Esta mutación puede mediar la resistencia adquirida a osimertinib, y la presencia de una mutación EGFR T790M en el diagnóstico inicial del cáncer de pulmón, junto con su alta FA en el análisis de ADNcf, y los antecedentes familiares de cáncer de pulmón, hicieron sospechar que la mutación EGFR T790M podría haber representado un alelo de riesgo de línea germinal.

PCR Digital en Gotas

Se recogió sangre (6-10 ml) en tubos de vacío con tapón de lavanda y EDTA y se centrifugó durante 10 minutos a 1.200 g. El sobrenadante plasmático se aclaró mediante centrifugación durante 10 minutos a 3000 g. El segundo sobrenadante se almacenó en tubos criostáticos a -80 °C hasta su utilización. El ADN libre de células se aisló con el kit QIAmp Circulating Nucleic Acid Kit (Qiagen) y se realizó la PCR digital en gotitas (ddPCR). Brevemente, las mezclas de reacción TaqMan PCR se ensamblaron a partir de una 2x ddPCR Mastermix (Bio-Rad) y 40 sondas/primers TaqMan para cada ensayo. Las gotas se generaron utilizando un generador de gotas automatizado (Bio-RAD). La PCR se realizó hasta el punto final. Tras la PCR, las gotas se leyeron en un lector de gotas QX100 o QX200 (Bio-Rad). El análisis de los datos de la ddPCR se realizó con el software de análisis QuantaSoft (Bio-Rad). Todos los reactivos ddPCR se encargaron a Bio-Rad. Todos los cebadores y sondas se encargaron a Life Technologies. Los cebadores y las condiciones fueron los siguientes.

EGFR L858R cebador delantero, 5'- GC A GC AT GT C A A GAT C A C A GATT -3 ' (SEQ ID N.º 1); cebador inverso, 5'- CCTCCTTCTGCATGGTATTCTTTCT-3' (SEQ ID N.º 2); secuencias de sonda: 5'- VIC-AGTTTGGCCAGCCCAA-MGB-NFQ-3' (SEQ ID N.º 3), 5'-FAM-AGTTTGGCCCGCCCAA-MGB-NFQ-3' (SEQ ID N.º 4).

Condiciones ciclistas: 95 °C x 10 min (1 ciclo), 40 ciclos de 94 °C x 30 s y 58 °C x 1 min, y 10 °C de mantenimiento.

EGFR del 19 cebador directo, 5'-GTGAGAAAGTTAAAATTCCCGTC-3' (SEQ ID N.º 5); cebador inverso, 5'-CACACAGCAAAGCAGAAAC-3' (SEQ ID N.º 6); secuencias de sonda: 5'- VIC- ATCGAGGATTTCCTTGTTG-MGB-NFQ-3' (SEQ ID N.º 7), 5'-FAM-AGGAATTAAGAGAAGCAACATC-MGB-NFQ-3' (SEQ ID N.º 8). Condiciones ciclistas: 95 °C x 10 min (1 ciclo), 40 ciclos de 94 °C x 30 s y 55 °C x 1 min, seguidos de 10 °C de mantenimiento.

EGFR T790M, cebador delantero, 5'-GCCTGCTGGGCATCTG-3' (SEQ ID N.º 9), cebador inverso, 5'-TCTTTGTGTTCCCGGACATAGTC-3' (SEQ ID N.º 10); secuencias de sonda: 5'- VIC- ATGAGCTGCGTGATGAG-MGB-NFQ-3' (SEQ ID N.º 11), 5'-FAM-ATGAGCTGCATGATGAG-MGB-NFQ-3' (SEQ ID N.º 12). Condiciones ciclistas: 95 °C x 10 min (1 ciclo), 40 ciclos de 94 °C x 30 s y 58 °C x 1 min, seguidos de 10 °C de mantenimiento.

Secuenciación plasmática de nueva generación

El ADNcf se aisló a partir de 10 ml de sangre total extraída en tubos de ADN libre de células, se enriqueció mediante captura híbrida dirigida a los exones de 70 genes y a los intrones críticos de 6 genes, y se secuenció a una profundidad media de ~15.000X en un secuenciador Illumina NextSeq500.

Secuenciación de la línea germinal

Para los casos seleccionados, se proporcionaron muestras de buffy coat desidentificadas y se extrajo ADN genómico para la secuenciación Sanger del EGFR.

Análisis estadístico

La relación entre el FA de la mutación impulsora del EGFR y las medidas de variación del número de copias dentro del grupo heterocigoto de variantes se analizó mediante una regresión lineal. La función de densidad de probabilidad de la distribución de la desviación típica y la media de las FA de los casos individuales se estimó mediante una aproximación gaussiana, y los valores atípicos se identificaron mediante el método de Tukey. Se determinó un intervalo de confianza del 95% para la prevalencia de EGFR T790M en cada diagnóstico de interés. La prevalencia entre los distintos diagnósticos se comparó mediante una prueba exacta de Fisher de dos colas.

Resultados

De 85 pacientes con NSCLC avanzado que presentaban una mutación EGFR T790M, se sabía que tres eran portadores de una mutación EGFR T790M de línea germinal basada en una secuenciación de línea germinal previa, mientras que el resto había adquirido EGFR T790M tras el tratamiento con TKI. Al estudiar la concentración absoluta de alelos T790M en copias/mL de plasma, algunos casos con T790M somático presentaban una concentración aún mayor de alelos T790M mutantes en plasma que los tres casos con EGFR T790M de línea germinal **(FIG. 2A).** En cambio, con el AF de T790M, calculado como la proporción de copias de T790M mutante de todas las variantes mutantes o silvestres en ese locus, el FA de los tres casos de línea germinal rondó el 50%, superior al FA de los casos de T790M somática **(FIG. 2A).** A continuación, se estudiaron los cambios en los niveles de mutaciones somáticas frente a mutaciones germinales del EGFR en el ADNcf plasmático tras el tratamiento con TKIs del EGFR de tercera generación, como osimertinib. En los pacientes portadores de EGFR T790M adquirida tras una resistencia a los TKI de primera generación, la concentración tanto de las mutaciones EGFR T790M como de las mutaciones impulsoras (por ejemplo, L858R o deleción del exón 19) disminuyó drásticamente en respuesta al tratamiento **(FIG. 2B).** Por el contrario, en pacientes con mutaciones germinales EGFR T790M, la mutación conductora del EGFR respondió al tratamiento, mientras que los niveles de EGFR T790M se mantuvieron relativamente estables **(FIG. 2B).** Estos datos proporcionaron una prueba de concepto de que la cuantificación de los niveles de variantes en el ADNcf plasmático puede utilizarse para discriminar entre los orígenes somáticos y de línea germinal de las mutaciones asociadas a tumores.

La secuenciación de próxima generación (NGS) tiene el potencial de capturar una amplia gama de variantes en una serie de genes relacionados con el cáncer. Para investigar más a fondo el comportamiento de las mutaciones germinales y somáticas del EGFR en el ADNcf plasmático, se secuenciaron las regiones exónicas de 70 oncogenes y genes supresores de tumores, y las regiones intrónicas de 6 genes en los que se producen reordenamientos oncogénicos. Se consultó una base de datos de resultados de NGS de plasma clínico para estudiar la distribución de mutaciones somáticas y de línea germinal del EGFR, lo que dio lugar a la identificación de conjuntos de prueba de 950 muestras consecutivas de NSCLC para cada una de las siguientes: mutaciones somáticas conocidas (L858R y deleciones del exón 19), un polimorfismo de nucleótido único (SNP) de línea germinal común dentro del dominio tirosina quinasa del EGFR (Q787QX17), y T790M, y se trazó la distribución de FA de cada una **(FIG. 2C)**. La distribución de los SNP conocidos comprendía dos distribuciones de probabilidad discretas, normalmente distribuidas, centradas en FA del 50% y del 100%, compatibles con heterocigosidad y homocigosidad del alelo Q787Q. La distribución de las alteraciones somáticas conocidas, L858R y deleciones del exón 19, por el contrario, demostró una distribución de decaimiento exponencial que comenzaba en el límite de detección del ensayo, con la cola larga extendiéndose a FAs superiores al 90%, compatible con FAs somáticas que variaban sustancialmente pero que eran generalmente bajas (<5%). La distribución de T790M se ajustaba predominantemente a esta misma distribución somática. Sin embargo, existía una subpoblación menor pero

discreta, normalmente distribuida, centrada en el 50% de FA **(FIG. 2C).** Este patrón respalda el estudio de la variante FA en el ADNcf como método para categorizar variantes como EGFR T790M, que pueden ser de origen somático o de línea germinal.

La distribución del FA se estudió más a fondo realizando NGS plasmática en muestras de plasma antes del tratamiento y durante el tratamiento de tres casos con EGFR T790M de línea germinal que se sabía que albergaban mutaciones impulsoras del EGFR en su cáncer (dos con L858R, uno con L861Q). Al estudiar los FA de todas las variantes codificantes y no codificantes identificadas en la NGS plasmática, se visualizaron claramente tres grupos de variantes (FIG. 3A, donde 301 es EGFR T790M; 302 es EGFR driver mutation; 303 es TP53 mutation; 304 es other alterations; 305 es homozygous band; 306 es heterozygous band y 307 es tumor band). El grupo de variantes de menor FA incluye el driver EGFR y las mutaciones TP53, que representan variantes derivadas del cáncer. El grupo de variantes con mayor FA se centró en torno al 100% de AF, representando variantes homocigóticas de línea germinal. Por último, un grupo intermedio de variantes se centraba en torno al 50%, que incluía las mutaciones germinales conocidas del EGFR T790M y representaba variantes germinales heterocigóticas. En tratamiento con un TKI del EGFR de tercera generación (dos con osimertinib, uno con ASP7283), las variantes derivadas del cáncer con bajo AF disminuyeron o se volvieron indetectables (24% -> 0,2%, 3,7% - ND, 1,1% - ND), las variantes derivadas del cáncer con bajo FA disminuyeron o se volvieron indetectables. Por el contrario, el grupo intermedio de variantes heterocigotas de la línea germinal sólo cambió modestamente y permaneció centrado en torno al 50% de AF (56% - 49%, 52% - 49%, 49% - 50%). Curiosamente, algunas de estas variantes heterocigotas parecían presentar un aumento de la FA a medida que el cáncer respondía a la terapia, mientras que otras presentaban una disminución de la FA. Estos cambios en el grupo heterocigoto durante el tratamiento podrían representar una reducción en la variación del número de copias derivada del tumor, lo que llevaría a un cambio en la variante FA en el ADNcf.

A continuación se estudiaron todas las variantes codificantes y no codificantes de la NGS plasmática del caso inicial presentado anteriormente (Tabla 5). Esto reveló un patrón similar al de los casos de EGFR T790M de línea germinal estudiados, en los que la mutación EGFR T790M del paciente entraba dentro del grupo de variantes heterocigotas, y la FA cambiaba mínimamente con la terapia en comparación con la mutación EGFR L858R (FIG. 7).

Para estudiar más a fondo la relación entre el contenido tumoral en el ADNcf y la variación heterocigótica del número de copias, se consultó una base de datos para otros 63 casos de NGS en plasma que fueron positivos para EGFR T790M y 39 casos de NGS en plasma positivos para una mutación impulsora de EGFR sin T790M. En cada uno de estos 105 casos se detectó una mediana de 107 variantes codificantes y no codificantes. Observando la distribución de FA de las 10.702 variantes en total (FIG. 3B), se observa claramente una distribución trimodal, con tres picos de FA en ~0%, 49% y 100%. En comparación con las variantes exónicas e intrónicas no codificantes, las variantes codificantes con sentido erróneo y sin sentido se enriquecieron en el grupo de variantes de baja FA (FIG. 3C), lo que concuerda con la idea de que se trata de un grupo de variantes derivadas del cáncer.

Para estudiar la relación entre las posibles variantes de la línea germinal y las somáticas, cada caso de NGS en plasma se representó individualmente en orden de menor a mayor FA de las mutaciones impulsoras del EGFR (FIG. 4A, donde 401 (punto negro) es la mutación conductora del EGFR; 402 (punto gris más grande) es EGFR Q787Q (SNP conocido); 403 es la media de la banda heterocigota; 404 (punto gris de tamaño medio) son otras alteraciones codificantes y 405 (punto gris más pequeño) son alteraciones no codificantes). Aunque el FA de la mutación impulsora no es una medida perfecta del contenido tumoral en el ADNcf (debido a la presencia o ausencia de amplificación del gen EGFR en algunos casos), puede servir como estimación del contenido tumoral en el ADNcf en una cohorte. Al estudiar la distribución de los FA de las variantes en el grupo heterocigoto, aquí designado como todas las variantes con un FA entre el 25% y el 75%, la distribución cambiaba a medida que aumentaba el FA de la mutación impulsora del EGFR. Un aumento de la FA del driver EGFR se asoció con una mayor desviación estándar del grupo heterocigoto **(FIG. 4B)**, así como un aumento de la diferencia absoluta entre la media del caso y la de la población, lo que sugiere la presencia de una variación del número de copias derivada del cáncer. Al estudiar la desviación estándar de la FA de las variantes en el grupo heterocigoto, se observó que 94 casos se ajustaban a una distribución normal, mientras que 11 casos presentaban características atípicas **(FIG. 8A).** Del mismo modo, al estudiar la mediana de la FA para las variantes en el grupo heterocigoto, se ajustó una distribución normal a 94 casos, mientras que 11 casos presentaban características atípicas **(FIG. 8B).** Dado que estas poblaciones atípicas se solapaban, 16 casos en total presentaban una de estas dos características atípicas con evidencia de alta variación del número de copias en el ADNcf, lo que puede haberse debido a que los altos niveles de ADN tumoral causan variabilidad en el FA de las variantes de la línea germinal.

Dado que la elevada variación del número de copias puede dar lugar a una desviación sustancial del FA de las variantes de la línea germinal con respecto al 50% esperado, la discriminación germinal-somática podría verse afectada en estos casos atípicos. Así pues, estos 16 casos atípicos se segregaron de los 89 casos sin características atípicas **(FIG. 5).** Con la revisión visual de las variantes de codificación de los casos atípicos, puede resultar difícil distinguir una separación clara entre las variantes heterocigóticas de la línea germinal y las variantes somáticas derivadas del cáncer, pero, por el contrario, la revisión visual de las variantes de codificación de los casos sin estas características de alta variación del número de copias **(FIG. 5**, donde 501 (punto gris más grande) es EGFR T790M; 502 (punto negro) es EGFR driver mutation y 503 (punto gris más pequeño) son otras alteraciones de codificación), permite distinguir claramente un grupo de variantes heterocigotas con FAs en el rango del 35 al 60%, que no se solapan con un grupo de variantes derivadas del cáncer con FAs por debajo del 30%. Así, excluyendo los casos de NGS en plasma con alta variación en el

número de copias (y por tanto alto contenido tumoral), los resultados de NGS en plasma pueden diferenciarse con precisión en variantes somáticas dentro del grupo derivado del cáncer y alelos de riesgo de línea germinal identificados incidentalmente dentro del grupo heterocigoto.

Siguiendo la lógica de estos estudios de prueba de concepto, se desarrolló y evaluó un algoritmo bioinformático integrado para segregar alteraciones de la línea germinal y somáticas en los 70 genes analizados mediante NGS en plasma. Este algoritmo asignó primero a las variantes un presunto origen germinal o somático utilizando conocimientos a priori, incluidas bases de datos internas y externas de variantes germinales y somáticas conocidas (patogénicas y benignas). Por ejemplo, la alteración EGFR Q787Q es un polimorfismo benigno presente en ~52% de los exomas de línea germinal en la base de datos ExAC (http://exac.broadinstitute.org/), lo que permite designarlo como de presunto origen de línea germinal independientemente de la fracción alélica. Por el contrario, la alteración EGFR L858R es una mutación oncogénica relativamente frecuente en el NSCLC pero no aparece en las bases de datos de línea germinal, lo que permite designarla como de presunto origen somático. Este binning a priori suele dar como resultado una mediana de 78 variantes por caso asignadas como de línea germinal, lo que permite construir una distribución de probabilidad heterocigota por variante FA como se describe en los estudios anteriores. Si todas las presuntas mutaciones somáticas (generalmente menos numerosas) están presentes por debajo del límite inferior de esta distribución germinal heterocigótica, la discriminación germinal-somática de las variantes restantes no asignadas procede según su FA en relación con la distribución germinal descrita por la clasificación de variantes a priori. Sin embargo, si el FA de las presuntas variantes somáticas supera el FA del límite inferior de la distribución heterocigótica de la línea germinal, o si se detecta una inestabilidad cromosómica extrema (evaluada por la fracción diploide aparente del genoma), la discriminación línea germinal/somática se considera incierta para las variantes que permanecen dentro de esa región de solapamiento, y se presume que las variantes son de origen somático y se notifican como tales. Este enfoque permite la identificación de variantes sospechosas de línea germinal con un alto valor predictivo positivo, entendiendo que la sensibilidad para las variantes de origen germinal se reducirá en entornos de alto contenido de ADN tumoral.

A continuación, este algoritmo se aplicó a 21 muestras clínicas recogidas prospectivamente con mutaciones EGFR T790M de alta FA (30%-75%) detectadas en la NGS plasmática (FIG. 9, donde 901 (punto gris más grande) es EGFR T790M; 902 (punto negro) es mutación conductora de EGFR y 903 (punto gris más pequeño) son otras alteraciones de codificación). Los casos se segregaron en dos cohortes basándose en la segregación germinal-somática de EGFR T790M descrita anteriormente. La cohorte A incluyó 11 casos en los que la distribución de variantes somáticas frente a derivadas de la línea germinal llevó a predecir la presencia de una mutación T790M de la línea germinal. La cohorte B incluyó 10 casos en los que la determinación de línea germinal frente a somática fue complicada debido a la elevada variación del número de copias y a un amplio grupo heterocigoto. A continuación, las fracciones celulares que contenían ADN genómico de cada muestra se desidentificaron de forma irreversible y se enviaron a un laboratorio clínico con certificación CLIA para la secuenciación del EGFR de forma doble ciega, de modo que ningún resultado de la línea germinal fuera trazable a ningún paciente individual. Se confirmó que los 11 casos de la cohorte A albergaban un EGFR T790M de línea germinal (valor predictivo positivo del 100%, 11/11). De los 10 casos de la cohorte B, uno resultó ser de línea germinal, lo que dio lugar a una sensibilidad del 92% (11/12) y una precisión global del 95% (20/21). Se sospechó que la presencia de una muestra de línea germinal en la Cohorte B era un caso con un alto contenido tumoral, de modo que el FA de presuntas mutaciones somáticas se solapaba con la distribución de heterocigotos de línea germinal, lo que dificultaba discriminar con certeza las variantes de línea germinal.

Una vez validado un método para identificar casos de NGS en plasma portadores de EGFR T790M de línea germinal, se utilizaron los datos de NGS en plasma existentes para conocer la asociación de las variantes de línea germinal con tipos de cáncer específicos. Se consultó una base de datos de pruebas clínicas de 31.414 pacientes únicos consecutivos que representaban una amplia variedad de tipos de tumores sólidos en adultos para identificar 911 casos positivos para EGFR T790M, de los cuales 48 eran de origen germinal según la metodología anterior. Aunque el NSCLC no escamoso fue el diagnóstico de cáncer en una minoría de la cohorte total de pacientes (41%), éste fue el diagnóstico de cáncer en 43 de los 48 pacientes con EGFR T790M de línea germinal (90%, **FIG. 6A**). Además, de los 5 pacientes restantes con EGFR T790M de línea germinal, tres tenían un diagnóstico relacionado (NSCLC escamoso, cáncer de pulmón de células pequeñas, carcinoma de primario desconocido). La frecuencia poblacional de EGFR T790M de línea germinal en pacientes con NSCLC no escamoso (43/12.774, 0,34%) fue sustancialmente superior a la observada en pacientes con otro diagnóstico de cáncer (5/18.640, 0,03%, **FIG. 6B**), siendo esta última sólo moderadamente superior a la notificada por los esfuerzos de secuenciación de la población general (por ejemplo, la frecuencia alélica media de ExAC de 0,0082%). Estas observaciones son congruentes con el concepto de que los pacientes con T790M de línea germinal tienen un mayor riesgo específico de NSCLC, y sugieren que este alelo no confiere un riesgo sustancialmente mayor de otros cánceres aparte del de pulmón.

Los análisis anteriores demuestran el poder de la genómica del ADNcf como herramienta para la investigación de los alelos de riesgo de cáncer de la línea germinal. Utilizando datos y muestras existentes de investigaciones clínicas en curso, se desarrolló y validó un algoritmo bioinformático para distinguir las variantes de la línea germinal de las variantes somáticas derivadas del cáncer dentro de los perfiles de cfDNA NGS, proporcionando un único ensayo que puede ofrecer información sobre el genotipo del tumor para la selección de la terapia, así como la detección de alelos de riesgo hereditario. Se consultó una base de datos de pruebas clínicas para explorar el alelo de línea germinal poco frecuente, EGFR T790M, y se observó un enriquecimiento para esta mutación en pacientes con NSCLC no escamoso. Los datos anteriores ponen de relieve la capacidad del genotipado plasmático, utilizado actualmente para la atención clínica rutinaria,

para detectar variantes de la línea germinal y, en determinadas circunstancias, diferenciarlas de las variantes somáticas.

REIVINDICACIONES

1. Un método implementado por ordenador que comprende:

a) proporcionar un conjunto de lecturas de secuencias de moléculas de ADNcf, en el que las lecturas de secuencias corresponden a una región genómica seleccionada de un genoma de referencia;
b) determinar la frecuencia alélica de un conjunto que comprende una pluralidad de variantes genéticas dentro de la región genómica, donde el conjunto incluye una variante de interés;
c) determinar una medida de variabilidad de la frecuencia alélica de las variantes genéticas en el set;
d) proporcionar un umbral de medida de la variabilidad y un umbral de frecuencia alélica;
e) determinar si la medida de variabilidad está por debajo del umbral de variabilidad; y
f) si la medida de variabilidad está por debajo del umbral de variabilidad:

(i) calificar la variante de interés como de origen germinal si la frecuencia alélica de la variante de interés está por encima del umbral de frecuencia alélica, y
(ii) calificar la variante de interés como de origen somático si la frecuencia alélica de la variante de interés está por debajo del umbral de frecuencia alélica.

2. El método de la reivindicación 1, en el que la región genómica seleccionada es un gen, un exón, un intrón, una porción de un gen, opcionalmente de al menos 100 nucleótidos, al menos 500 nucleótidos, o al menos 1000 nucleótidos.

3. El método de la reivindicación 1 o de la reivindicación 2, en el que la medida de variabilidad es la desviación estándar o la varianza.

4. El método de cualquiera de las reivindicaciones 1 a 3, en el que el umbral de frecuencia alélica es de aproximadamente 10%, aproximadamente 11%, aproximadamente 12%, aproximadamente 13%, aproximadamente 14%, aproximadamente 15%, aproximadamente 16%, aproximadamente 17%, aproximadamente 18%, aproximadamente 19%, aproximadamente 20%, aproximadamente 21%, aproximadamente 22%, aproximadamente 23%, aproximadamente 24%, aproximadamente 25%, aproximadamente 26%, aproximadamente 27%, aproximadamente 28%, aproximadamente 29%, aproximadamente 30%, aproximadamente 31%, aproximadamente 32%, aproximadamente 33%, aproximadamente 34%, o aproximadamente 35%.

5. El método de cualquiera de las reivindicaciones 1 a 4, en el que el umbral de frecuencia alélica se determina empíricamente.

6. El método de cualquiera de las reivindicaciones 1 a 5, en el que la medida de variabilidad es la desviación estándar y el umbral de variabilidad es un umbral de desviación estándar (STDEV).

7. El método de la reivindicación 6, en el que una medida de FA para un locus genómico por debajo de dicho umbral STDEV indica baja variación del número de copias (CNV) para dicho locus genómico, mientras que una medida de FA para un locus genómico por encima de dicho umbral STDEV indica alta variación del número de copias (CNV) para el locus genómico asociado.

8. El método de cualquiera de las reivindicaciones precedentes, en el que la región genómica es una o más regiones genómicas y las regiones genómicas comprenden:

(i) al menos una parte de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95 o 97 de los genes de Tabla 1;
(ii) al menos una parte de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 105, al menos 110, o 115 de los genes de la Tabla 2; o
(iii) al menos una porción de al menos 1, al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 11, al menos 12, al menos 13, al menos 14, al menos 15, al menos 16, al menos 17, al menos 18, al menos 19, o al menos 20 de los genes de la Tabla 3.

9. El método de cualquiera de las reivindicaciones anteriores, en el que las moléculas de ADNcf se aíslan de un fluido corporal, como sangre o suero.

10. El método de cualquiera de las reivindicaciones anteriores, en el que las moléculas de ADNcf comprenden ADN tumoral circulante.

11. El método de cualquiera de las reivindicaciones anteriores, en el que el suministro de un conjunto de lecturas de

secuencias de moléculas de ADNcf comprende secuenciar ADNcf de un sujeto y detectar y cuantificar una o más variantes genéticas.

12. El método de la reivindicación 11, en el que se prepara una biblioteca de ácidos nucleicos antes de la secuenciación.

13. El método de la reivindicación 12, en el que una molécula de ADNcf se identifica unívocamente mediante la combinación de un código de barras y una o más secuencias endógenas del polinucleótido.

14. Uso de los métodos de cualquiera de las reivindicaciones anteriores:

(i) para diagnosticar una enfermedad o afección como el cáncer o una afección inflamatoria;
(ii) en el pronóstico de una enfermedad o afección como el cáncer o una afección inflamatoria;
(iii) para evaluar la eficacia del tratamiento de una enfermedad o afección como el cáncer o una afección inflamatoria; y/o
(iv) para controlar la progresión o regresión de una enfermedad o afección como el cáncer o una afección inflamatoria.

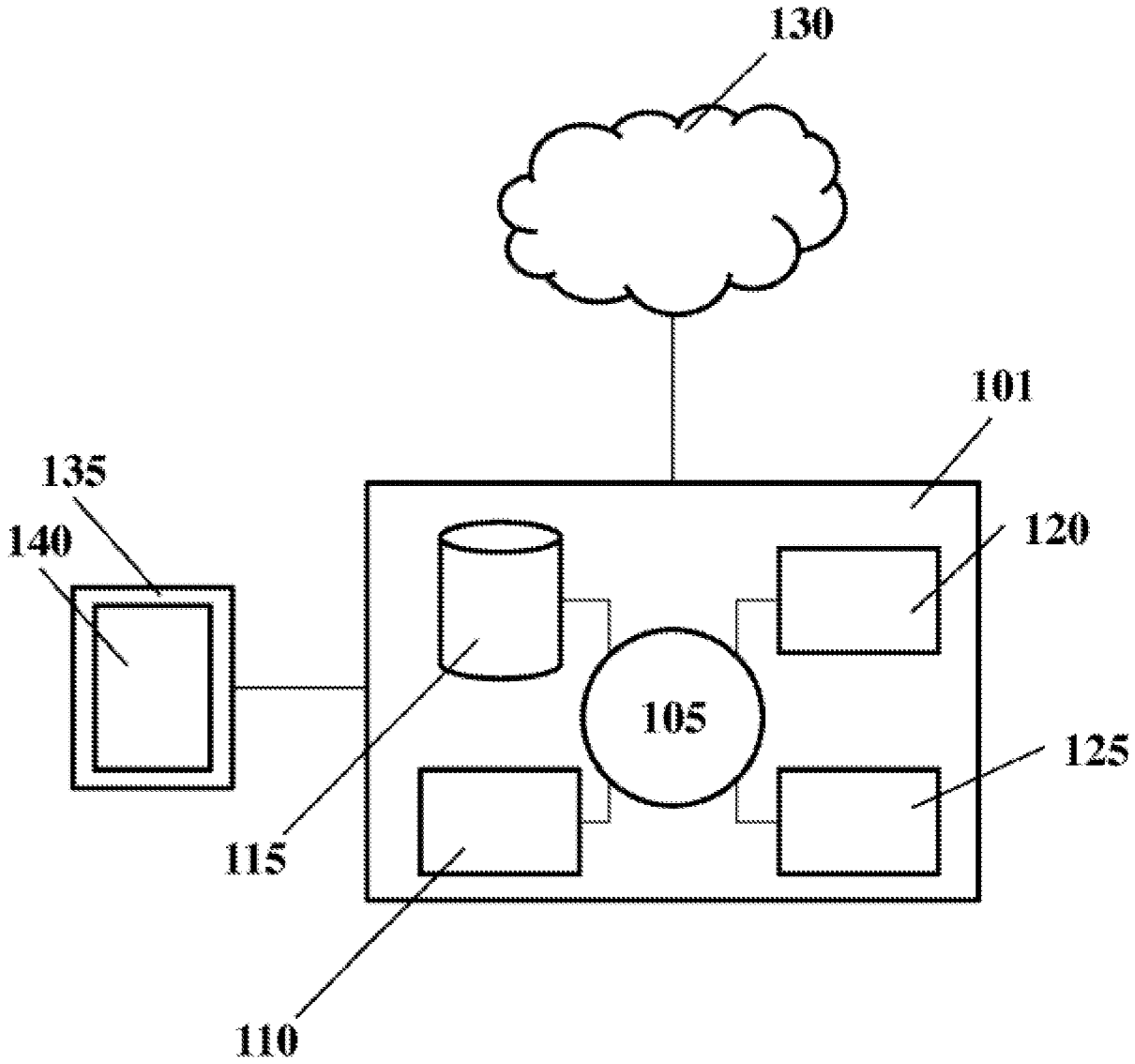

FIG. 1

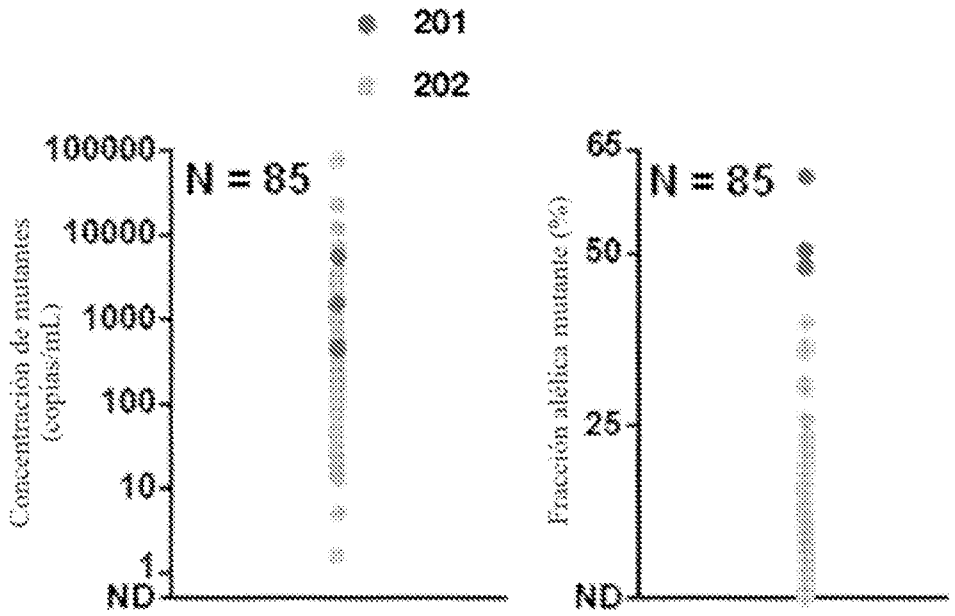
201
202
100000
10000
1000
100
10
1
ND
N = 85
Concentración de mutantes (copias/mL)
65
50
25
ND
N = 85
Fracción alélica mutante (%)

FIG. 2A

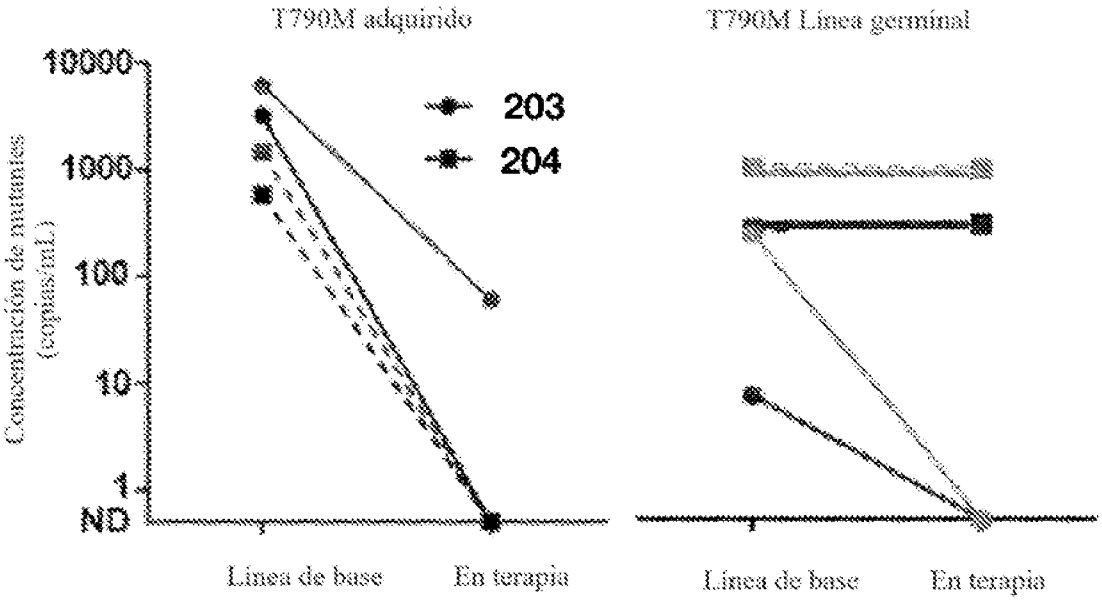
T790M adquirido
T790M Línea germinal
10000
1000
100
10
1
ND
Concentración de mutantes (copias/mL)
203
204
Línea de base
En terapia
Línea de base
En terapia

FIG. 2B

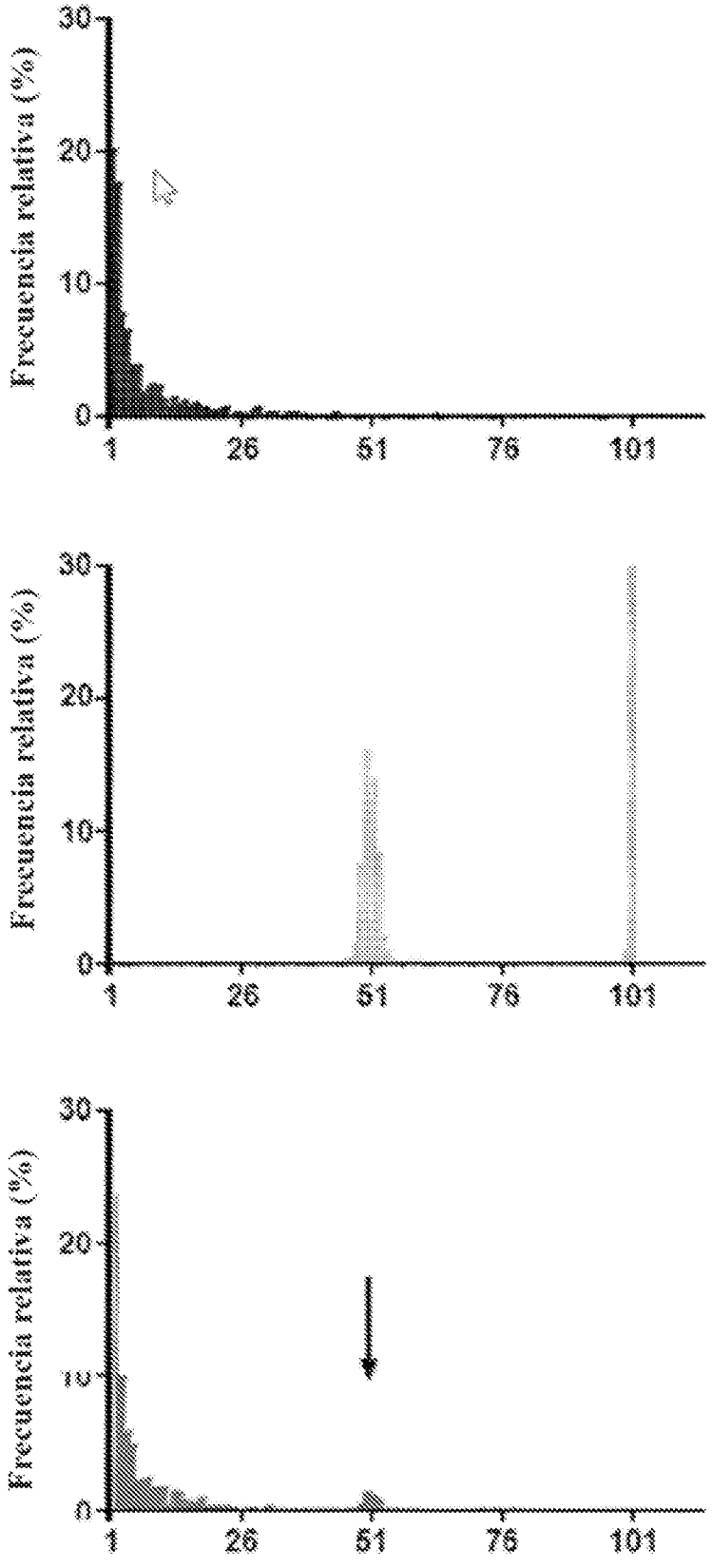
Frecuencia relativa (%)
30
20
10
0
1
26
51
76
101
Frecuencia relativa (%)
30
20
10
0
1
26
51
76
101
Frecuencia relativa (%)
30
20
10
0
1
26
51
76
101

FIG. 2C

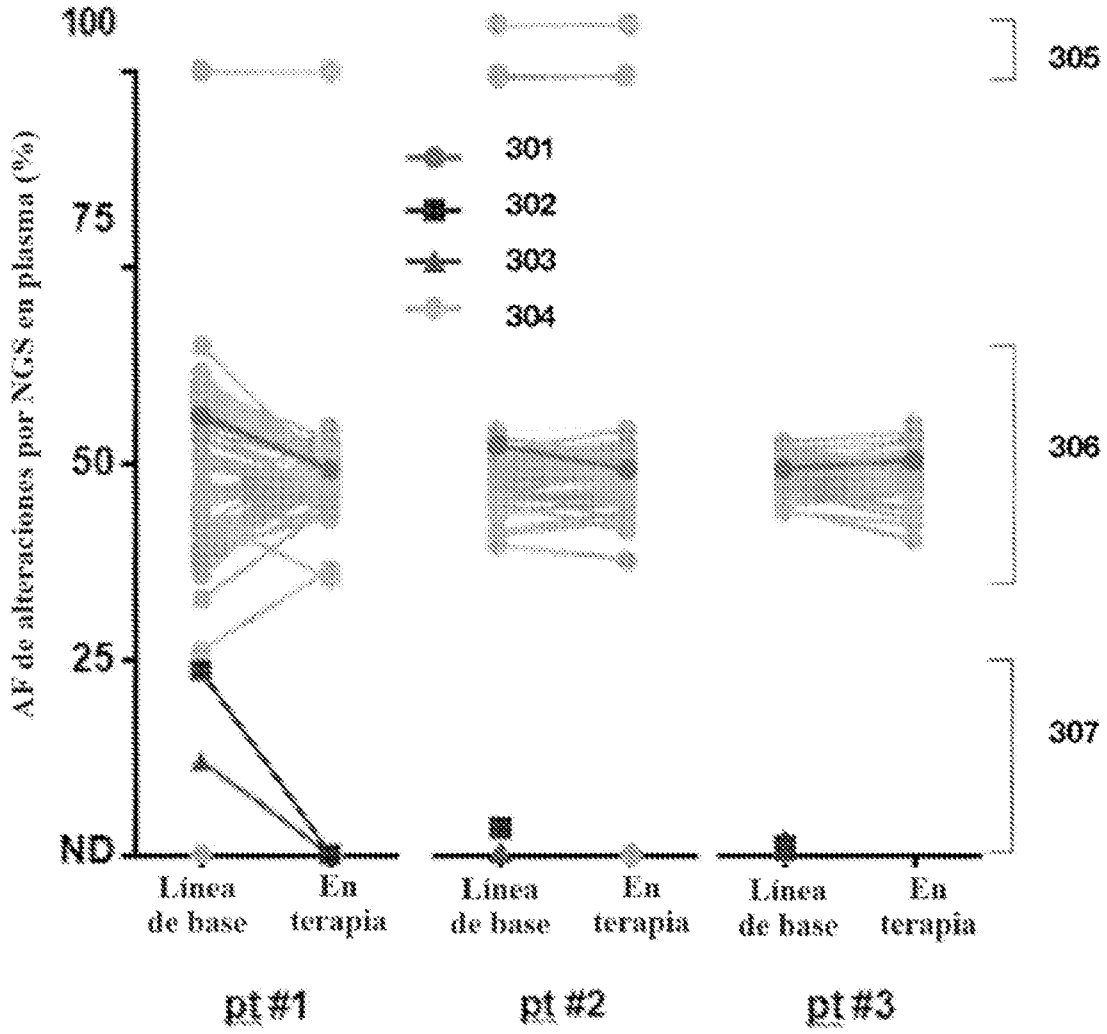
100
75
50
25
ND
AF de alteraciones por NGS en plasma (%)
301
302
303
304
305
306
307
Línea de base
En terapia
Línea de base
En terapia
Línea de base
En terapia
pt #1
pt #2
pt #3

FIG. 3A

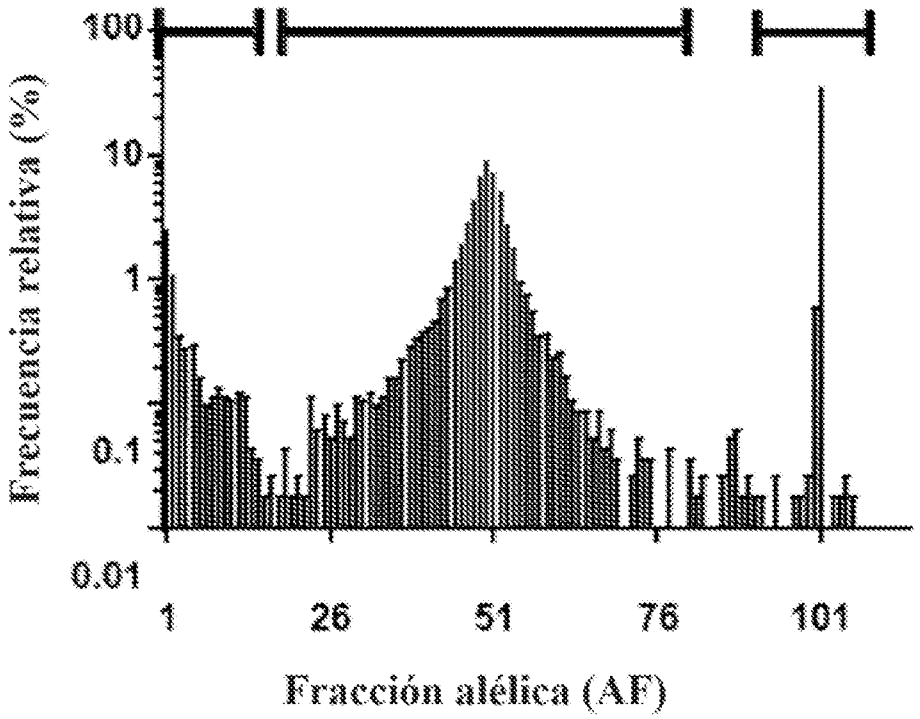
100
10
1
0.1
0.01
Frecuencia relativa (%)
1
26
51
76
101
Fracción alélica (AF)

FIG. 3B

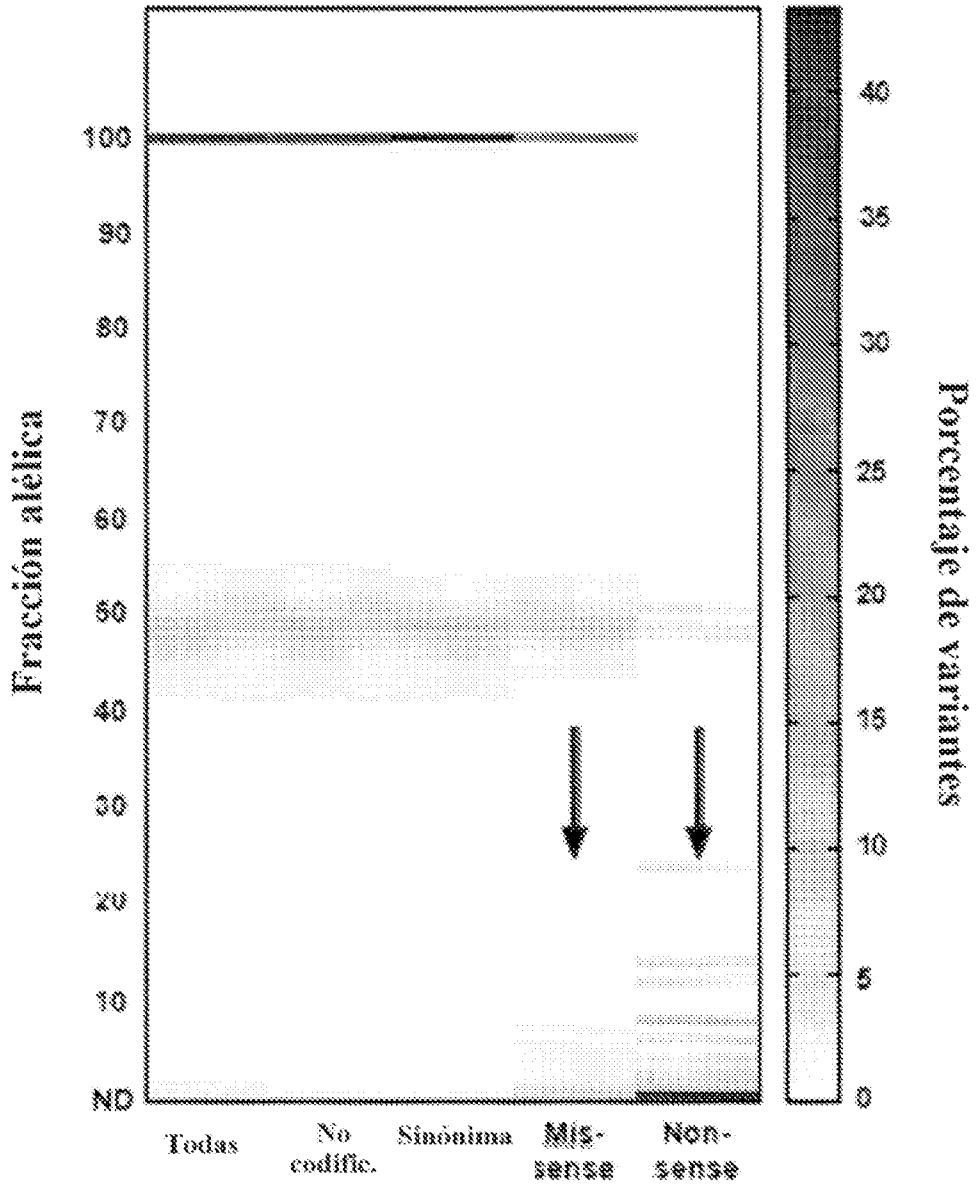
Fracción alélica
100
90
80
70
60
50
40
30
20
10
ND
Todas
No codific.
Sinónima
Mis-sense
Non-sense
Porcentaje de variantes
45
35
30
25
20
15
10
5
0

FIG. 3C

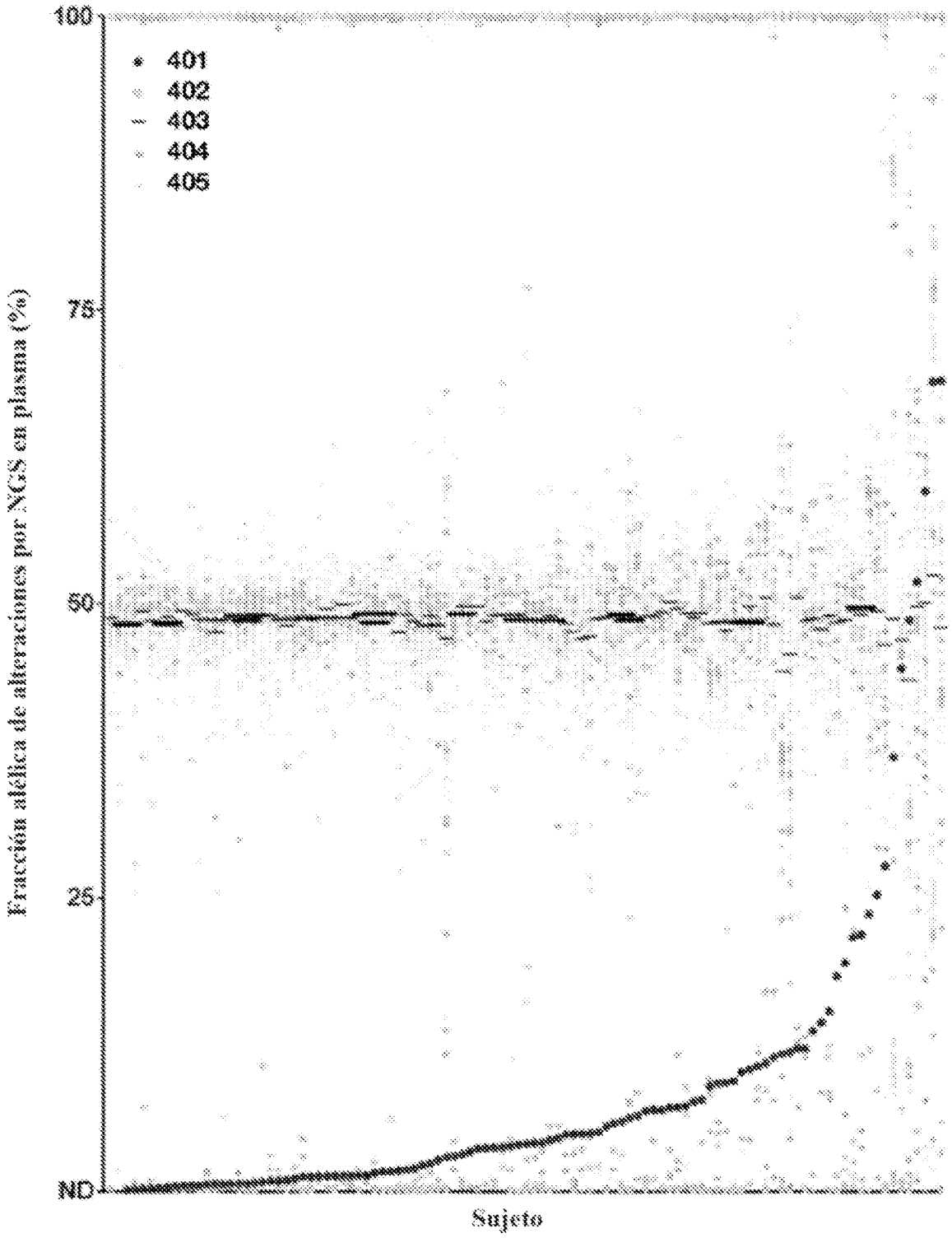
100
75
50
25
ND
401
402
403
404
405
Fracción alélica de alteraciones por NGS en plasma (%)
Sujeto

FIG. 4A

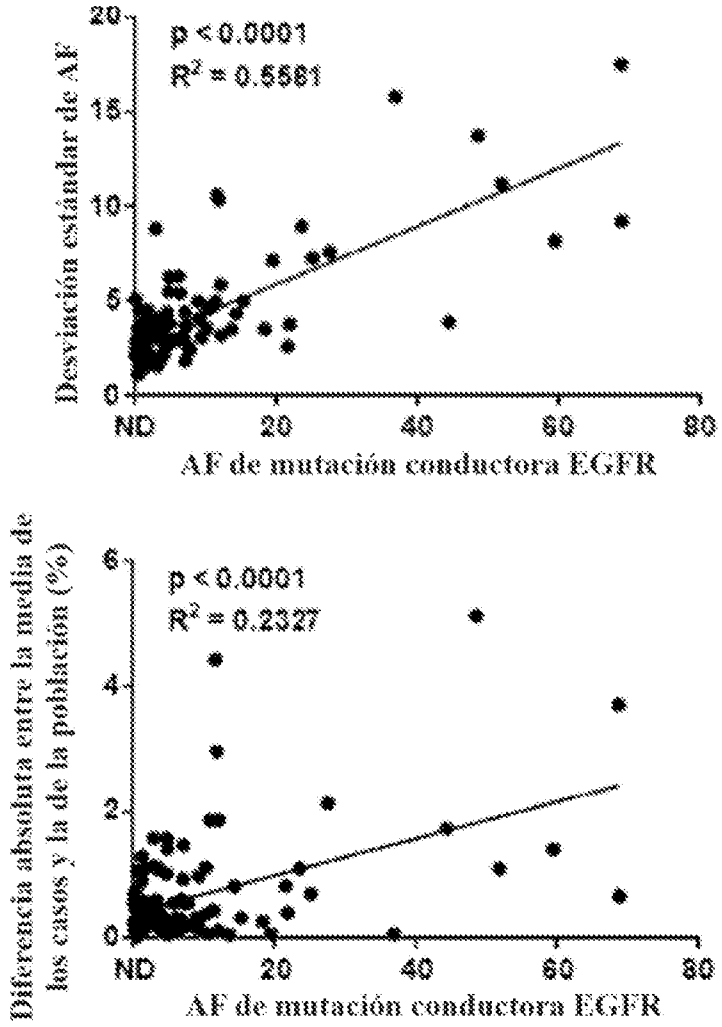
p < 0.0001
R^2 = 0.5561
Desviación estándar de AF
AF de mutación conductora EGFR
p < 0.0001
R^2 = 0.2327
Diferencia absoluta entre la media de los casos y la de la población (%)
AF de mutación conductora EGFR

FIG. 4B

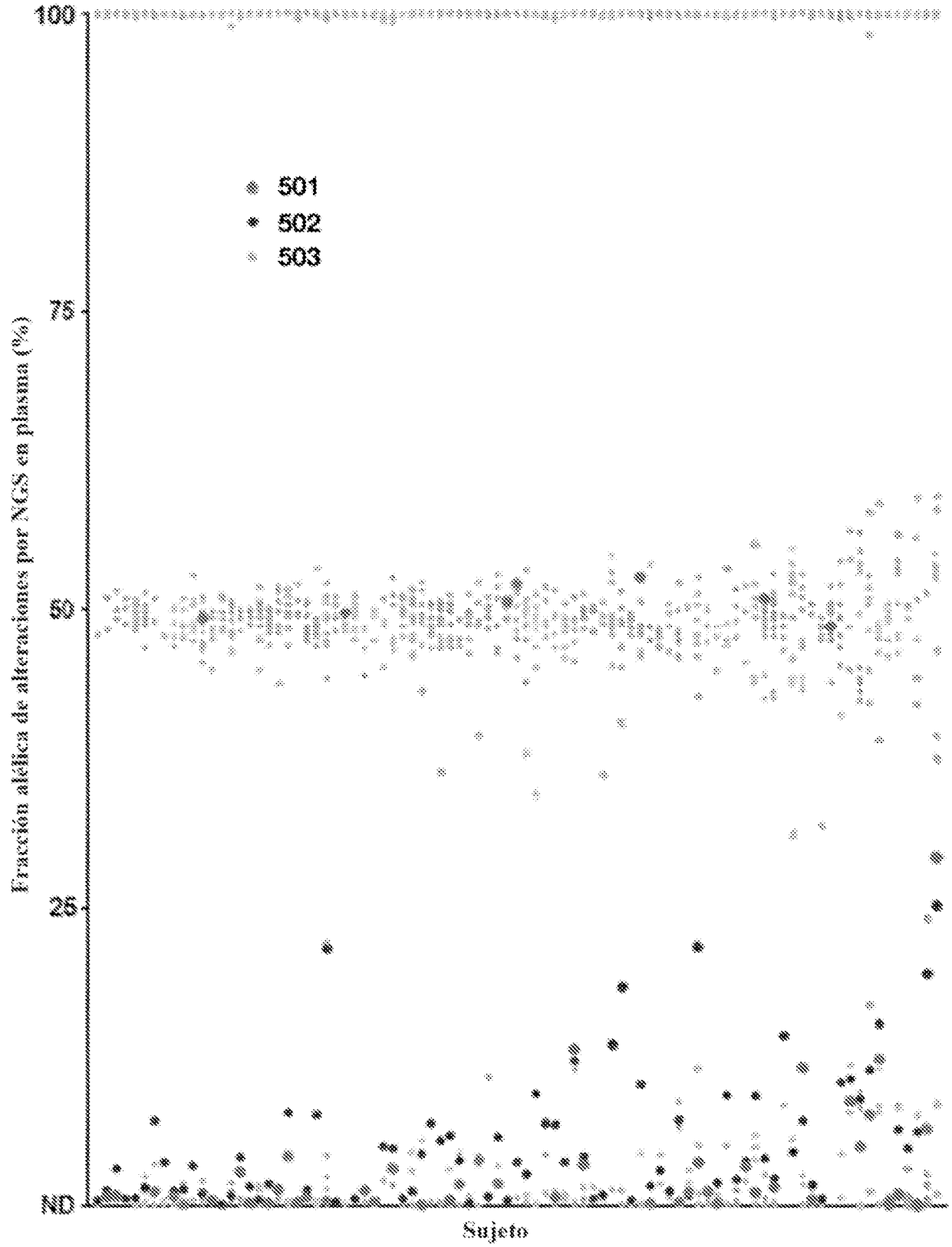
100
75
50
25
ND
501
502
503
Fracción alélica de alteraciones por NGS en plasma (%)
Sujeto

FIG. 5

	Línea germinal T790M detectada	Sin línea germinal T790M detectada
Total	48	31,366
NSCLC no escamoso	43 (90%)	12,731 (41%)
Otros cánceres	5 (10%)	18,635 (59%)

FIG. 6A

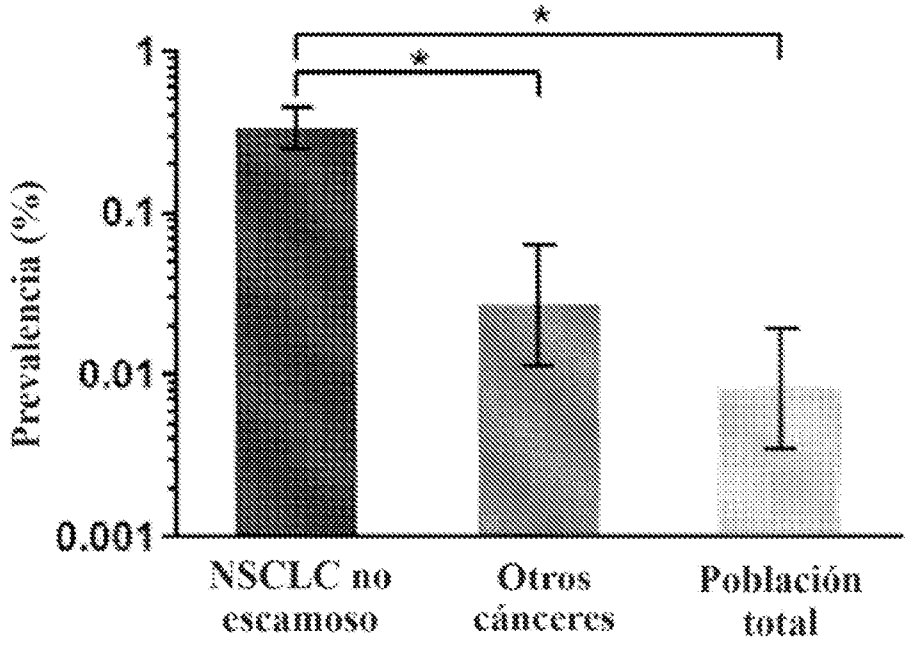
*
*
1
0.1
0.01
0.001
Prevalencia (%)
NSCLC no escamoso
Otros cánceres
Población total

FIG. 6B

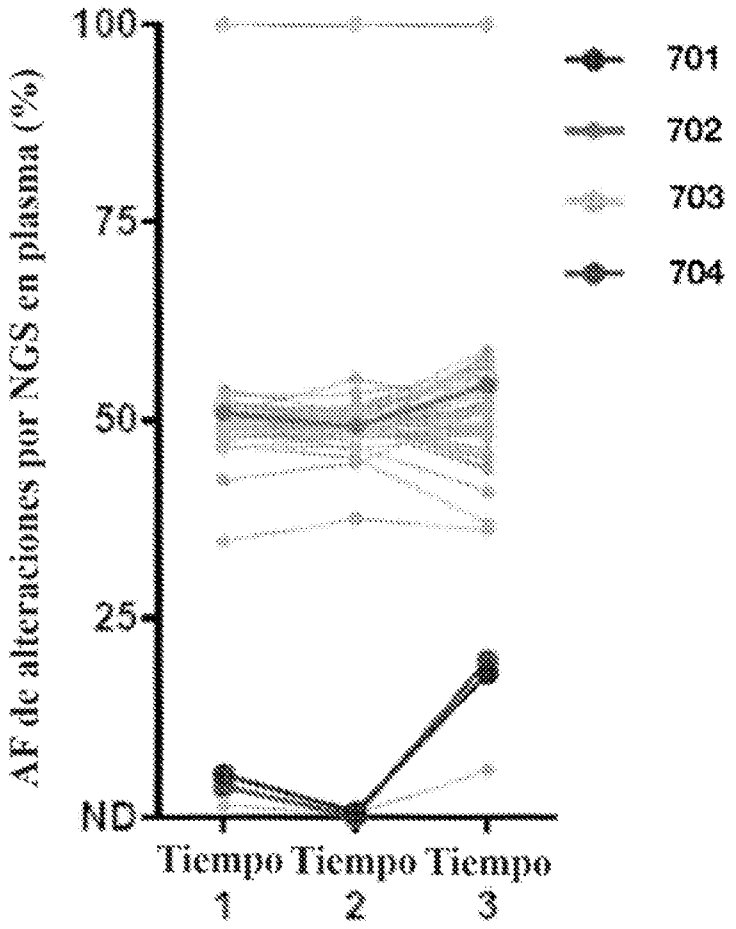
100
75
50
25
ND
AF de alteraciones por NGS en plasma (%)
Tiempo 1
Tiempo 2
Tiempo 3
701
702
703
704

FIG. 7

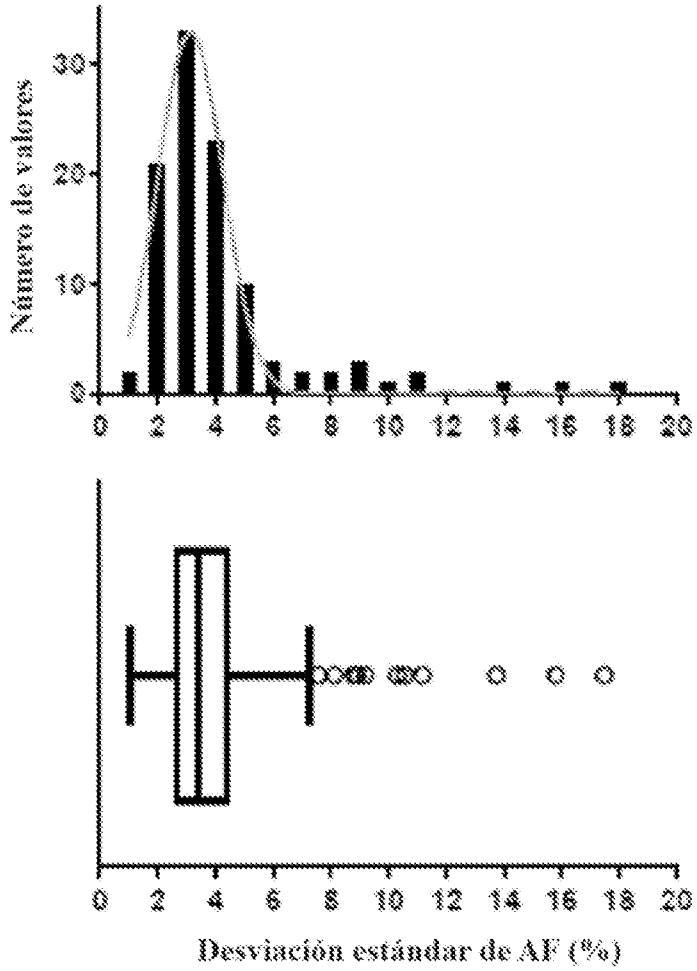
Número de valores
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20
0 2 4 6 8 10 12 14 16 18 20
Desviación estándar de AF (%)

FIG. 8A

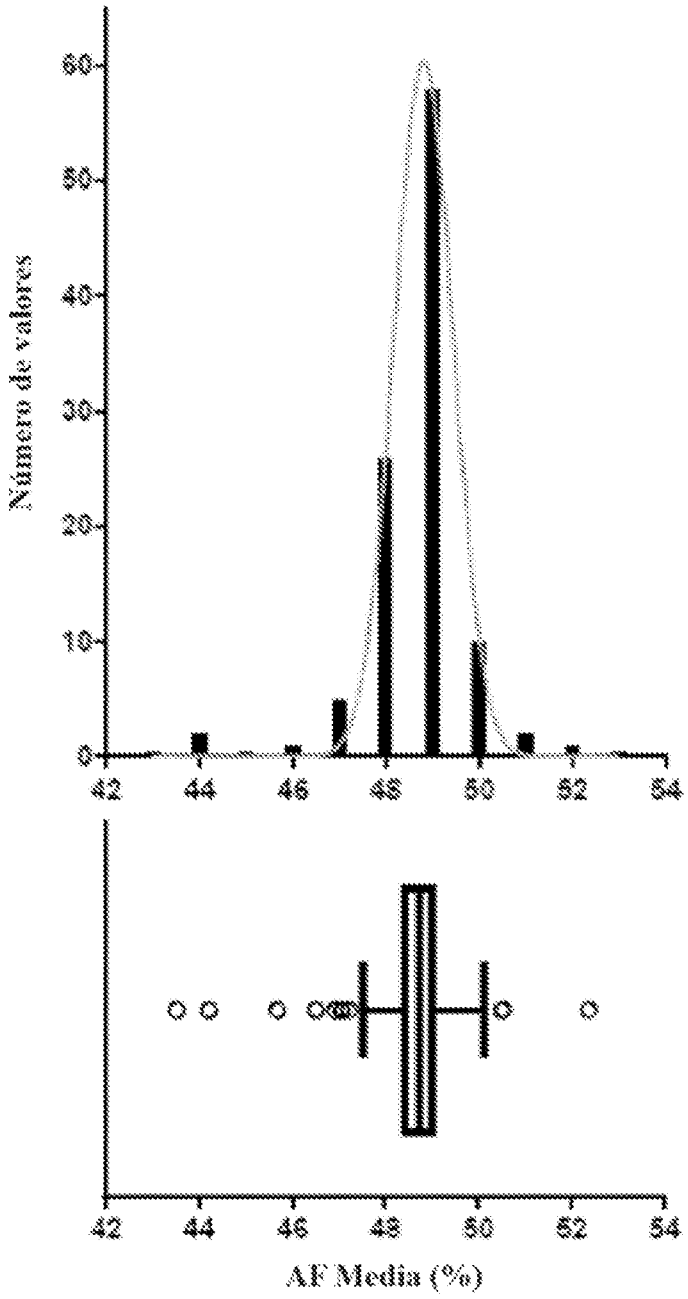

FIG. 8B

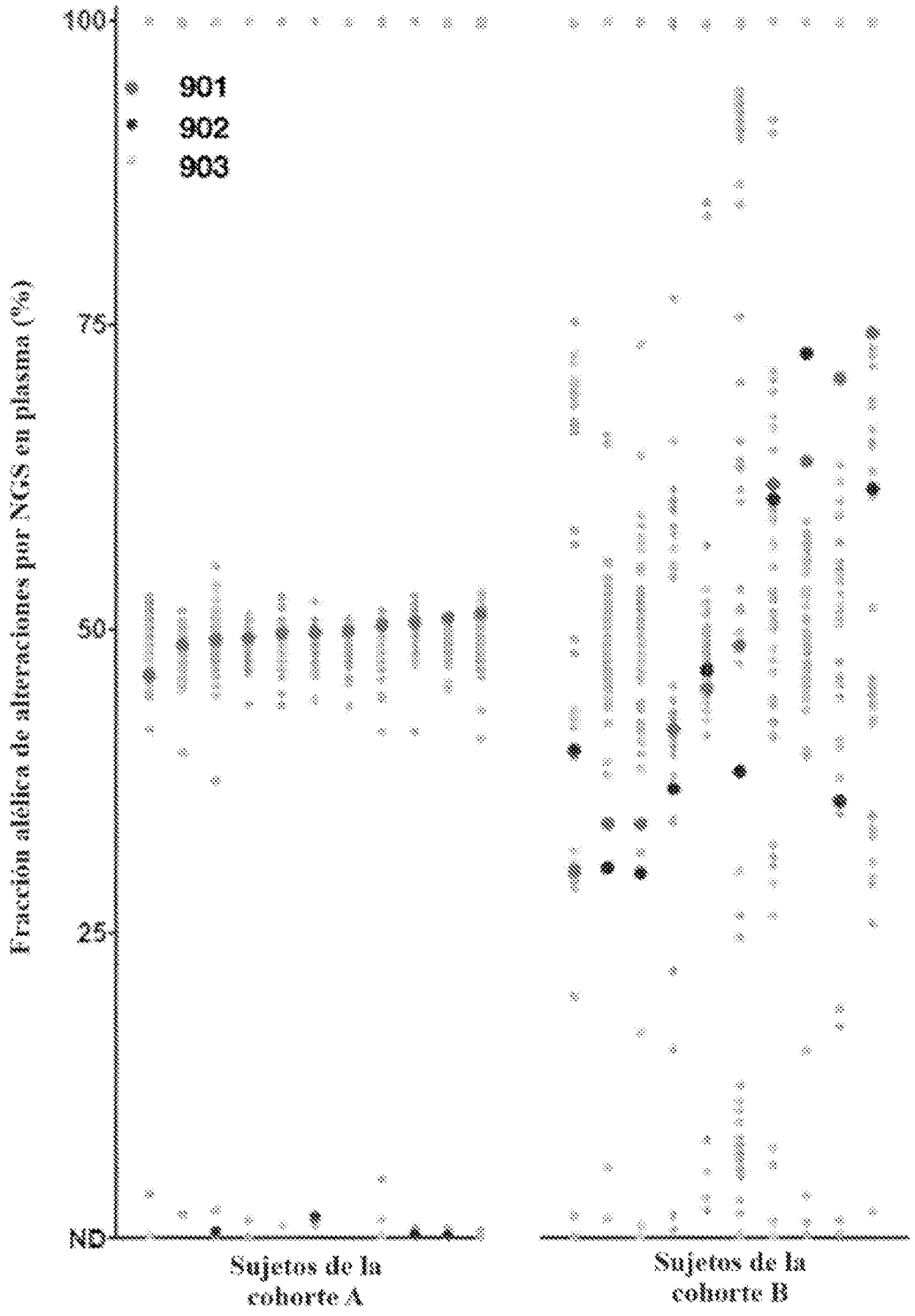
100
75
50
25
ND
Fracción alélica de alteraciones por NGS en plasma (%)
901
902
903
Sujetos de la cohorte A
Sujetos de la cohorte B

FIG. 9