(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 840 003**

(51) Int. CI.:

C12Q 1/68 (2008.01)
C12Q 1/6806 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **29.09.2017 PCT/US2017/054607**

(87) Fecha y número de publicación internacional: **05.04.2018 WO18064629**

(96) Fecha de presentación y número de la solicitud europea: **29.09.2017 E 17857586 (6)**

(97) Fecha y número de publicación de la concesión europea: **04.11.2020 EP 3461274**

(54) Título: **Métodos para análisis multi-resolución de ácidos nucleicos libres de células**

(30) Prioridad:

30.09.2016 US 201662402940 P
07.03.2017 US 201762468201 P
24.04.2017 US 201762489391 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
06.07.2021

(73) Titular/es:

GUARDANT HEALTH, INC. (100.0%)
505 Penobscot Drive
Redwood City, CA 94063, US

(72) Inventor/es:

CHUDOVA, DARYA;
ELTOUKHY, HELMY;
MORTIMER, STEFANIE ANN, WARD;
ABDUEVA, DIANA y
SIKORA, MARCIN

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos para análisis multi-resolución de ácidos nucleicos libres de células

REFERENCIA CRUZADA

Esta solicitud reivindica la prioridad de la Solicitud Provisional de Estados Unidos Nº 62/402.940, presentada el 30 de septiembre de 2016, la Solicitud Provisional de Estados Unidos Nº 62/468.201, presentada el 7 de marzo de 2017, y la Solicitud Provisional de Estados Unidos Nº 62/489.391, presentada en abril 24, 2017

ANTECEDENTES

El análisis de ácidos nucleicos libres de células (por ejemplo, ácido desoxirribonucleico o ácido ribonucleico) para variantes genéticas derivadas de tumores es un paso crítico en un proceso de análisis típico para aplicaciones de detección, evaluación y monitorización del cáncer. La mayoría de los métodos actuales de ensayos de diagnóstico de cáncer de ácidos nucleicos libres de células se enfocan en la detección de variantes somáticas relacionadas con el tumor, incluyendo las variantes de un solo nucleótido (SNV), variaciones en el número de copias (CNV), fusiones e inserciones/deleciones (indeles), que son todos los objetivos principales de la biopsia líquida. Un enfoque de análisis típico puede comprender enriquecer una muestra de ácidos nucleicos para regiones objetivo de un genoma, seguido por secuenciación de ácidos nucleicos enriquecidos y análisis de datos de lectura de secuencia para variantes genéticas de interés. Estos ácidos nucleicos pueden enriquecerse usando una mezcla señuelo seleccionada para un ensayo particular de acuerdo con las restricciones de ensayo, incluyendo la carga de secuenciación limitada y la utilidad asociadas con cada región genómica de interés.

SUMARIO

En un aspecto, la presente divulgación proporciona un panel de conjuntos de señuelos que comprende uno o más conjuntos de señuelos que enriquecen selectivamente una o más regiones asociadas a nucleosomas de un genoma, dichas regiones asociadas a nucleosomas comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de una célula o un tipo de tejido de origen o un estado patológico.

En algunas realizaciones, cada una de las una o más regiones asociadas a nucleosomas de un panel de conjunto de señuelos comprende por lo menos una de: (i) variación estructural significativa, que comprende una variación en el posicionamiento nucleosómico, dicha variación estructural seleccionada del grupo que consiste de: una inserción, una deleción, una translocación, un reordenamiento genético, estado de metilación, un microsatélite, una variación del número de copias, una variación estructural relacionada con el número de copias o cualquier otra variación que indique diferenciación; y (ii) inestabilidad, que comprende una o más fluctuaciones o picos significativos en un mapa de partición del genoma que indica una o más localizaciones de alteraciones del mapa nucleosómico en un genoma.

En algunas realizaciones, el uno o más conjuntos de señuelos de un panel de conjuntos de señuelos están configurados para capturar regiones asociadas a nucleosomas del genoma en base a una función de una pluralidad de perfiles de ocupación de nucleosomas de referencia (i) asociados con una o más estados patológicos y uno o más estados no patológicos; (ii) asociado con una mutación somática conocida, como SNV, CNV, indel o reordenación; y/o (iii) asociado con patrones de expresión diferencial. En una realización, uno o más conjuntos de señuelos de un panel de conjuntos de señuelos se enriquecen selectivamente para una o más regiones asociadas a nucleosomas en una muestra de ácido desoxirribonucleico libre de células (ADNcf).

En otro aspecto, la presente divulgación proporciona un método para enriquecer una muestra de ácidos nucleicos para regiones asociadas a nucleosomas de un genoma que comprende (a) poner una muestra de ácidos nucleicos en contacto con un panel de conjuntos de señuelos, dicho panel de conjuntos de señuelos comprendiendo una o más conjuntos de señuelos que enriquecen selectivamente una o más regiones asociadas a nucleosomas de un genoma; y (b) enriquecer la muestra de ácidos nucleicos para una o más regiones asociadas a nucleosomas de un genoma.

En algunas realizaciones, uno o más conjuntos de señuelos en un panel de conjuntos de señuelos están configurados para capturar regiones del genoma asociadas a nucleosomas en base a una función de una pluralidad de perfiles de ocupación nucleosómica de referencia asociados con uno o más estados patológicos y uno más estados no patológicos. En una realización, el uno o más conjuntos de señuelos en un panel de conjuntos de señuelos se enriquecen selectivamente para la una o más regiones asociadas a nucleosomas en una muestra de ADNcf. En una realización, el método para enriquecer una muestra de ácidos nucleicos para regiones asociadas a nucleosomas de un genoma comprende además secuenciar los ácidos nucleicos enriquecidos para producir lecturas de secuencias de las regiones asociadas a nucleosomas de un genoma.

En otro aspecto, la presente divulgación proporciona un método para generar un conjunto de señuelos que comprende (a) identificar una o más regiones de un genoma, dichas regiones asociadas con un perfil de nucleosoma, y (b) seleccionar un conjunto de señuelos para capturar selectivamente dichas regiones. En una realización, un conjunto de señuelos en un panel de conjuntos de señuelos se enriquece selectivamente para una o más regiones asociadas a nucleosomas en una muestra de ácido desoxirribonucleico libre de células.

En otro aspecto, la presente divulgación proporciona un panel de señuelos que comprende un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas de una muestra de ácidos nucleicos que comprende una cantidad predeterminada de ADN, que se proporciona en una primera proporción de concentración que es menor que un punto de saturación del primer conjunto de señuelos; y un segundo conjunto de señuelos que se hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, que se proporciona en una segunda proporción de concentración que está asociada con un punto de saturación del segundo conjunto de señuelos. En una realización, el primer conjunto de regiones genómicas comprende una o más regiones genómicas de la estructura principal y el segundo conjunto de regiones genómicas comprende una o más regiones genómicas de punto crítico.

En otro aspecto, la presente divulgación proporciona un método para enriquecer múltiples regiones genómicas que comprende poner una cantidad predeterminada de una muestra de ácidos nucleicos en contacto con un panel de señuelos que comprende (i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas de la muestra de ácidos nucleicos, proporcionada a una primera proporción de concentración que es menor que un punto de saturación del primer conjunto de señuelos, y (ii) un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, proporcionada a una segunda proporción de concentración que está asociada con un punto de saturación del segundo conjunto de señuelos; y enriquecer la muestra de ácidos nucleicos para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas.

En algunas realizaciones, el método comprende además secuenciar los ácidos nucleicos enriquecidos para producir lecturas de secuencia del primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas.

En algunas realizaciones, el punto de saturación de un conjunto de señuelos se determina mediante (a) para cada uno de los señuelos en el conjunto de señuelos, generar una curva de titulación que comprende (i) medir la eficiencia de captura del señuelo en función de la concentración del señuelo, y (ii) identificar un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo; y (b) seleccionar un punto de saturación que sea sustancialmente más grande que todos los puntos de saturación asociados con los señuelos en el conjunto de señuelos, determinando de este modo el punto de saturación del conjunto de señuelos.

En algunas realizaciones, la eficacia de captura de un señuelo se determina (a) proporcionando una pluralidad de muestras de ácidos nucleicos obtenidas de una pluralidad de sujetos en una cohorte; (b) hibridando el señuelo con cada una de las muestras de ácidos nucleicos, en cada una de una pluralidad de concentraciones del señuelo; (c) enriqueciendo con el señuelo una pluralidad de regiones genómicas de las muestras de ácidos nucleicos, en cada una de la pluralidad de concentraciones del señuelo; y (d) midiendo el número de moléculas de ácido nucleico únicas o moléculas de ácido nucleico con la representación de ambas cadenas de una molécula de ácido nucleico de cadena doble original que representa la eficacia de captura en cada una de la pluralidad de concentraciones del señuelo.

En algunas realizaciones, un punto de inflexión es una primera concentración del señuelo de tal manera que la eficiencia de captura observada no aumenta significativamente a concentraciones del señuelo mayores que la primera concentración. Un punto de inflexión puede ser una primera concentración del señuelo de tal manera que un aumento observado entre (1) la eficiencia de captura a una concentración de señuelo del doble de la primera concentración en comparación con (2) la eficiencia de captura a la primera concentración de señuelo, es menor de aproximadamente un 1%, menor de aproximadamente un 2%, menor de aproximadamente un 3%, menor de aproximadamente un 4%, menor de aproximadamente un 5%, menor de aproximadamente un 6%, menor de aproximadamente un 7%, menor de aproximadamente un 8%, menor de aproximadamente un 9%, menor de aproximadamente un 10%, menor de aproximadamente un 12%, menor de aproximadamente un 14%, menor de aproximadamente un 16%, menor de aproximadamente un 18% o menor de aproximadamente un 20%.

En algunas realizaciones, la muestra de ácidos nucleicos comprende una muestra de ácidos nucleicos libre de células. En una realización, un método para enriquecer múltiples regiones genómicas comprende además secuenciar la muestra de ácidos nucleicos enriquecida para producir una pluralidad de lecturas de secuencia. En una realización, un método para enriquecer múltiples regiones genómicas comprende además producir una salida que comprende una secuencia de ácido nucleico representativa de la muestra de ácidos nucleicos.

En otro aspecto, la presente divulgación proporciona un panel de señuelos que comprende un primer

conjunto que captura selectivamente las regiones de la estructura principal de un genoma, dichas regiones de la estructura principal están asociadas con una función de clasificación de carga de secuenciación y utilidad, en donde la función de clasificación de cada región de la estructura principal tiene un valor menor que un valor umbral predeterminado; y un segundo conjunto de señuelos que captura selectivamente regiones de punto crítico de un genoma, dichas regiones de punto crítico están asociadas con una función de clasificación de carga de secuenciación y utilidad, en donde la función de clasificación de cada región de punto crítico tiene un valor mayor o igual al valor umbral predeterminado.

En algunas realizaciones, las regiones de punto crítico comprenden una o más regiones informativas de nucleosomas, dichas regiones informativas de nucleosomas comprendiendo una región de máxima diferenciación de nucleosomas. En una realización, el panel de señuelos comprende además un segundo conjunto de señuelos que captura selectivamente las regiones informativas de la enfermedad. En una realización, los señuelos en el primer conjunto de señuelos están a una primera concentración relativa al panel de señuelos, y los señuelos en el segundo conjunto de señuelos están a una segunda concentración relativa al panel de señuelos.

En otro aspecto, la presente divulgación proporciona un método para generar un conjunto de señuelos que comprende identificar una o más regiones genómicas de la estructura principal de interés, en donde la identificación de una o más regiones genómicas de la estructura principal comprende maximizar una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal; identificar una o más regiones genómicas de punto crítico de interés; crear un primer conjunto de señuelos que capture selectivamente las regiones genómicas de interés de la estructura principal; y crear un segundo conjunto de señuelos que capture selectivamente las regiones genómicas de interés de los puntos críticos, en donde el segundo conjunto de señuelos tiene una mayor eficacia de captura que el primer conjunto de señuelos.

En algunos aspectos, el uno o más puntos críticos se seleccionan usando uno o más de los siguientes: (i) maximizando una función de clasificación de la carga de secuenciación y la utilidad asociada con cada una de las regiones genómicas de puntos críticos, (ii) realizando perfiles de nucleosomas a través de una o más regiones genómicas de interés, (iii) mutaciones impulsoras de cáncer predeterminadas o prevalencia en una cohorte de pacientes relevante, y (iv) mutaciones impulsoras de cáncer identificadas empíricamente.

En algunos aspectos, identificar uno o más puntos críticos de interés comprende usar un procesador informático programado para clasificar un conjunto de regiones genómicas de puntos críticos en base a una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de puntos críticos. En algunas realizaciones, la identificación de una o más regiones genómicas de la estructura principal de interés comprende clasificar un conjunto de regiones genómicas de la estructura principal en base a una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal de interés. En algunas realizaciones, la identificación de una o más regiones genómicas de puntos críticos de interés comprende utilizar un conjunto de valores de frecuencia de alelos menores (MAF) determinados empíricamente o la clonalidad de una variante medida por su MAF con respecto a la mutación clonal o impulsora supuesta más alta en una muestra.

En algunos aspectos, la carga de secuenciación de una región genómica se calcula multiplicando entre sí uno o más de (i) el tamaño de la región genómica en pares de bases, (ii) la fracción relativa de lecturas gastadas en la secuenciación de fragmentos que mapean la región genómica, (iii) cobertura relativa como resultado del sesgo de secuencia de la región genómica, (iv) cobertura relativa como resultado del sesgo de amplificación de la región genómica, y (v) cobertura relativa como resultado del sesgo de captura de la región genómica.

En algunos aspectos, la utilidad de una región genómica se calcula multiplicando entre sí una o más de (i) la frecuencia de una o más mutaciones procesables en la región genómica, (ii) la frecuencia de una o más mutaciones asociadas con frecuencias de alelos menores (MAF) por encima de la media en la región genómica, (iii) fracción de pacientes en una cohorte que alberga una mutación somática dentro de la región genómica, (iv) suma de MAF para variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, y (v) proporción de (1) MAF para variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, a (2) MAF máximo para un paciente dado en la cohorte.

En algunos aspectos, las mutaciones procesables comprenden una o más de (i) mutaciones susceptibles de ser modificadas por fármacos, (ii) mutaciones para monitorización terapéutico, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v) mutaciones específicas del tipo celular, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico. En una realización, las mutaciones asociadas con frecuencias de alelos menores más altas comprenden una o más mutaciones impulsoras o se conocen a partir de datos externos o fuentes de anotaciones.

En otro aspecto, la presente divulgación proporciona un panel de señuelos que comprende una pluralidad de conjuntos de señuelos, cada conjunto de señuelos (i) comprendiendo uno o más señuelos que capturan

selectivamente una o más regiones genómicas con utilidad en el mismo cuantil en la pluralidad de señuelos, y (ii) tener una concentración relativa diferente de cada uno de los otros conjuntos de señuelos con utilidad en un cuantil diferente en la pluralidad de señuelos.

En otro aspecto, la presente divulgación proporciona un método para seleccionar un conjunto de bloques de panel que comprende (a) para cada bloque de panel, (i) calcular una utilidad del bloque de panel, (ii) calcular una carga secuencial del panel bloque, y (iii) calcular una función de clasificación del bloque del panel; y (b) realizar un proceso de optimización para seleccionar un conjunto de bloques de panel que maximiza los valores de la función de clasificación total de los bloques de panel seleccionados.

En algunos aspectos, una función de clasificación de un bloque de panel se calcula como la utilidad de un bloque de panel dividida por la carga de secuenciación de un bloque de panel. En algunas realizaciones, el proceso de optimización combinatoria comprende un algoritmo voraz.

En otro aspecto, la presente divulgación proporciona un método que comprende (a) proporcionar una pluralidad de mezclas de señuelos, en donde cada mezcla de señuelos comprende un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas y un segundo conjunto de señuelos que se hibrida selectivamente con un segundo conjunto de regiones genómicas, y en donde las mezclas de señuelos comprenden el primer conjunto de señuelos a diferentes concentraciones y el segundo conjunto de señuelos a las mismas concentraciones; (b) poner en contacto cada mezcla de señuelos con una muestra de ácidos nucleicos para capturar el ácido nucleico de la muestra con los conjuntos de señuelos, en donde las muestras de ácidos nucleicos tienen una concentración de ácido nucleico alrededor del punto de saturación del segundo conjunto de señuelos; (c) secuenciar los ácidos nucleicos capturados con cada mezcla de señuelos para producir conjuntos de lecturas de secuencia; (d) determinar el número relativo de lecturas de secuencia para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas para cada mezcla de señuelos; y (e) identificar por lo menos una mezcla de señuelos que proporcione profundidades de lectura para el segundo conjunto de regiones genómicas y, opcionalmente, el primer conjunto de regiones genómicas, en cantidades predeterminadas.

En otro aspecto, la presente divulgación proporciona un método para mejorar la precisión de la detección de una inserción o deleción (indel) de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células(ADNcf) en una muestra corporal de un sujeto, dicha pluralidad de lecturas de secuencia se generan mediante secuenciación de ácidos nucleicos, que comprende (a) para cada una de la pluralidad de lecturas de secuencia asociadas con las moléculas de ADN libre de células, proporcionar: una esperanza predeterminada de que se detecte un indel en una o más lecturas de secuencia de la pluralidad de lecturas de secuencia; una esperanza predeterminada de que un indel detectado sea un indel verdadero presente en una determinada molécula de ADN libre de células de las moléculas de ADN libre de células, dado que se ha detectado un indel en una o más de las lecturas de la secuencia; y una esperanza predeterminada de que un indel detectado se introduzca por error no biológico, dado que se ha detectado un indel en una o más de las lecturas de secuencia; (b) proporcionar medidas cuantitativas de uno o más parámetros del modelo característicos de las lecturas de secuencia generadas por secuenciación de ácidos nucleicos; (c) detectar uno o más indeles candidatos en la pluralidad de lecturas de secuencia asociadas con las moléculas de ADN libre de células; y (d) para cada indel candidato, realizar una prueba de hipótesis usando uno o más de los parámetros del modelo para clasificar dicho indel candidato como un indel verdadero o un indel introducido, mejorando de este modo la precisión de detección de un indel.

En otro aspecto, la presente divulgación proporciona un kit que comprende (a) una muestra que comprende una cantidad predeterminada de ADN; y (b) un panel de conjuntos de señuelos que comprende (i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas de una muestra de ácidos nucleicos que comprende una cantidad predeterminada de ADN, proporcionado a una primera proporción de concentración que es menor que un punto de saturación del primer conjunto de señuelos y (ii) un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, proporcionado a una segunda proporción de concentración que está asociada con un punto de saturación del segundo conjunto de señuelos.

En algunos aspectos, el método para mejorar la precisión de la detección de una inserción o deleción (indel) de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto comprende además enriquecer una o más loci del ADN libre de células en la muestra corporal antes del paso (a), produciendo de este modo polinucleótidos enriquecidos.

En algunos aspectos, el método comprende además amplificar los polinucleótidos enriquecidos para producir familias de amplicones, en donde cada familia comprende amplicones que se originan a partir de una sola cadena de moléculas de ADN libre de células. En algunos aspectos, el error no biológico comprende un error en la secuenciación en una pluralidad de ubicaciones de bases genómicas. En algunas realizaciones, el error no biológico comprende un error en la amplificación en una pluralidad de localizaciones de base genómicas.

En algunos aspectos, los parámetros del modelo comprenden uno o más de (por ejemplo, uno o más de, dos o más de, tres o más de, o cuatro de) (i) para cada uno del uno o más alelos variantes, una frecuencia del alelo variante (α) y una frecuencia de alelos que no son de referencia distintos del alelo variante (α'); (ii) una frecuencia de un error indel en toda la cadena delantera de una familia de cadenas (β_1), en donde una familia comprende una colección de amplicones que se originan a partir de una única cadena de moléculas de ADN libre de células; (iii) una frecuencia de un error de indel en toda la cadena inversa de una familia de cadenas (β_2); y (iv) una frecuencia de un error de indel en una lectura de secuencia (γ).

En algunos aspectos, el paso de realizar una prueba de hipótesis comprende realizar un algoritmo de maximización de múltiples parámetros. En algunos aspectos, el algoritmo de maximización de múltiples parámetros comprende un algoritmo de Nelder-Mead. En una realización, la clasificación de un indel candidato como un indel verdadero o un indel introducido comprende (a) maximizar una función de probabilidad multiparamétrica, (b) clasificar un indel candidato como un indel verdadero si el valor de la función de probabilidad máxima es mayor que un valor umbral predeterminado, y (c) clasificar un indel candidato como indel introducido si el valor de la función de probabilidad máxima es menor o igual a un valor umbral predeterminado.

En otro aspecto, la presente divulgación proporciona un medio legible por ordenador no transitorio que comprende un código ejecutable por máquina que, tras la ejecución por uno o más procesadores informáticos, implementa un método para generar un conjunto de señuelos que comprende identificar una o más regiones genómicas de la estructura principal de interés, en donde la identificación de una o más regiones genómicas de la estructura principal comprende maximizar una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal; identificar una o más regiones genómicas de puntos críticos de interés; crear un primer conjunto de señuelos que capture selectivamente las regiones genómicas de interés de la estructura principal; y crear un segundo conjunto de señuelos que capture selectivamente las regiones genómicas de interés de los puntos críticos, en donde el segundo conjunto de señuelos tiene una mayor eficacia de captura que el primer conjunto de señuelos.

En otro aspecto, la presente divulgación proporciona un medio legible por ordenador no transitorio que comprende un código ejecutable por máquina que, tras la ejecución por uno o más procesadores informáticos, implementa un método para seleccionar un conjunto de bloques de panel que comprende (a) para cada bloque de panel, (i) calcular una utilidad del bloque de panel, (ii) calcular una carga de secuenciación del bloque de panel, y (iii) calcular una función de clasificación del bloque de panel; y (b) realizar un proceso de optimización para seleccionar un conjunto de bloques de panel que maximiza los valores de la función de clasificación total del bloque de panel seleccionado.

En otro aspecto, la presente divulgación proporciona un medio legible por ordenador no transitorio que comprende un código ejecutable por máquina que, tras la ejecución por uno o más procesadores informáticos, implementa un método para mejorar la precisión de detectar una inserción o deleción (indel) a partir de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto, dicha pluralidad de lecturas de secuencia siendo generadas por secuenciación de ácidos nucleicos, comprende (a) para cada una de la pluralidad de lecturas de secuencia asociadas con moléculas de ADN libre de células, proporcionar una esperanza predeterminada de un indel que se está detectando en una o más lecturas de secuencia de la pluralidad de lecturas de secuencia; una esperanza predetermina de que un indel detectado sea un indel verdadero presente en una determinada molécula de ADN libre de células de las moléculas de ADN libre de células, dado que se ha detectado un indel en una o más lecturas de la secuencia; y una esperanza predeterminada de que un indel detectado se introduzca por error no biológico, dado que se ha detectado un indel en una o más de las lecturas de la secuencia; (b) proporcionar medidas cuantitativas de uno o más parámetros del modelo característicos de las lecturas de secuencia generadas por secuenciación de ácidos nucleicos; (c) detectar uno o más indeles candidatos en la pluralidad de lecturas de secuencia asociadas con las moléculas de ADN libre de células; y (d) para cada indel candidato, realizar una prueba de hipótesis usando uno o más de los parámetros del modelo para clasificar dicho indel candidato como un indel verdadero o un indel introducido, mejorando de este modo la precisión de la detección de un indel.

En otro aspecto, la presente divulgación proporciona un método para enriquecer múltiples regiones genómicas, que comprende: (a) poner una cantidad predeterminada de ácido nucleico de una muestra en contacto con una mezcla de señuelos que comprende (i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas del ácido nucleico de la muestra, dicho primer conjunto de señuelos se proporciona a una primera concentración que es menor que un punto de saturación del primer conjunto de señuelos, y (ii) un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, dicho segundo conjunto de señuelos se proporciona a una segunda concentración que está asociada con un punto de saturación del segundo conjunto de señuelos; y (b) enriquecer la muestra de ácidos nucleicos para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas.

En algunas realizaciones, el segundo conjunto de señuelos tiene un punto de saturación que es mayor que

sustancialmente todos los puntos de saturación asociados con los señuelos en el segundo conjunto de señuelos cuando un señuelo del segundo conjunto de señuelos se somete a una curva de titulación generada i) midiendo la eficiencia de captura de un señuelo del segundo señuelo puesto como una función de la concentración del señuelo, y (ii) identificando un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo. En algunas realizaciones, el punto de saturación se selecciona de tal manera que una eficiencia de captura observada aumenta en menos del 20% a una concentración del señuelo dos veces mayor que la de la primera concentración.

En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumenta en menos del 10% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumenta en menos del 5% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumenta menos del 2% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumenta menos del 1% a una concentración del señuelo dos veces mayor que la de la primera concentración.

En algunas realizaciones, el primer conjunto de señuelos o el segundo conjunto de señuelos enriquecen selectivamente una o más regiones asociadas al nucleosoma de un genoma, dichas regiones asociadas al nucleosoma comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de un tipo de célula o tejido de origen o estado de enfermedad. En algunas realizaciones, la muestra de ácidos nucleicos comprende una muestra de ácidos nucleicos libre de células. En algunas realizaciones, el método comprende además: (c) secuenciar la muestra de ácidos nucleicos enriquecida para producir una pluralidad de lecturas de secuencia. En algunas realizaciones, el método comprende además: (d) producir una salida que comprende una secuencia de ácido nucleico representativa de la muestra de ácidos nucleicos.

En otro aspecto, la presente divulgación proporciona un método para generar un conjunto de señuelos que comprende: (a) identificar una o más regiones genómicas de la estructura principal predeterminadas, en donde la identificación de una o más regiones genómicas de la estructura principal comprende maximizar una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal; (b) identificar una o más regiones genómicas de puntos críticos predeterminadas, en donde el uno o más puntos críticos se seleccionan usando uno o más de los siguientes: (i) maximizar una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada uno de las regiones genómicas de puntos críticos, (ii) hacer perfiles de nucleosomas a través de una o más regiones genómicas predeterminadas, (iii) mutaciones impulsoras de cáncer predeterminadas o prevalencia en una cohorte de pacientes relevante, y (iv) mutaciones impulsoras del cáncer identificadas empíricamente; (c) crear un primer conjunto de señuelos que capture selectivamente las regiones genómicas de la estructura principal predeterminadas; y (d) crear un segundo conjunto de señuelos que capture selectivamente las regiones genómicas de puntos críticos predeterminadas, en donde el segundo conjunto de señuelos tiene una mayor eficacia de captura que el primer conjunto de señuelos. En algunas realizaciones, una región predeterminada (por ejemplo, una región de la estructura principal predeterminada o una región de punto crítico predeterminada) es una región de interés (por ejemplo, una región de estructura principal de interés o una región de punto crítico de interés, respectivamente).

En algunos aspectos, la identificación de uno o más punto críticos predeterminados comprende usar un procesador informático programado para clasificar un conjunto de regiones genómicas de puntos críticos en base a una función de clasificación de carga de secuenciación y utilidad asociadas con cada una de las regiones genómicas de puntos críticos. En algunas realizaciones, la identificación de una o más regiones genómicas de la estructura principal predeterminadas comprende: (i) clasificar un conjunto de regiones genómicas de la estructura principal en base a una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal predeterminadas; (ii) utilizar un conjunto de valores de frecuencia de alelos menores (MAF) determinados empíricamente o la clonalidad de una variante medida por su MAF con respecto a la mayor mutación impulsora o clonal presunta en una muestra; o (iii) una combinación de (i) y (ii).

En algunos aspectos, la carga de secuenciación de una región genómica se calcula multiplicando entre sí uno o más de: (i) el tamaño de la región genómica en pares de bases, (ii) la fracción relativa de lecturas gastadas en la secuenciación de fragmentos que mapean el región genómica, (iii) cobertura relativa como resultado del sesgo de secuencia de la región genómica, (iv) cobertura relativa como resultado del sesgo de amplificación de la región genómica, y (v) cobertura relativa como resultado del sesgo de captura de la región genómica. En algunas realizaciones, la utilidad de una región genómica se calcula multiplicando entre sí una o más de: (i) frecuencia de una o más mutaciones procesables en la región genómica, (ii) frecuencia de una o más mutaciones asociadas con frecuencias alélicas menores (MAF) por encima de la media en la región genómica, (iii) fracción de pacientes en una cohorte que alberga una mutación somática dentro de la región genómica, (iv) suma de MAF para variantes en pacientes en una cohorte, dichos pacientes teniendo una mutación somática dentro de la región genómica, y (v)

proporción de (1) MAF para variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, a (2) MAF máximo para un paciente dado en la cohorte.

En algunos aspectos, las mutaciones procesables comprenden una o más de: (i) mutaciones modificables con fármacos, (ii) mutaciones para monitorización terapéutica, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v) mutaciones específicas de tipo celular, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico. En algunos aspectos, las mutaciones asociadas con frecuencias de alelos menores más altas comprenden una o más mutaciones impulsoras o se conocen a partir de datos externos o fuentes de anotaciones.

En otro aspecto, la presente divulgación proporciona un método que comprende: (a) proporcionar una pluralidad de mezclas de señuelos, en donde cada mezcla de señuelos comprende un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas y un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas, y en donde las mezclas de señuelos comprenden el primer conjunto de señuelos a diferentes concentraciones y el segundo conjunto de señuelos a las mismas concentraciones; (b) poner en contacto cada mezcla de señuelos con una muestra de ácidos nucleicos para capturar el ácido nucleico de la muestra con los conjuntos de señuelos, en donde el segundo conjunto de señuelos en cada mezcla se proporciona a una concentración que está en o por encima del punto de saturación del segundo conjunto de señuelos, en donde el ácido nucleico de la muestra es capturado por los conjuntos de señuelos; (c) secuenciar una porción de los ácidos nucleicos capturados con cada mezcla de señuelos para producir conjuntos de lecturas de secuencia dentro de un número asignado de lecturas de secuencia; (d) determinar la profundidad de lectura de las lecturas de secuencia para el primer conjunto de señuelos y el segundo conjunto de señuelos para cada mezcla de señuelos; y (e) identificar por lo menos una mezcla de señuelos que proporcione profundidades de lectura para el segundo conjunto de regiones genómicas; en donde la profundidad de lectura para el segundo conjunto de regiones genómicas proporciona una sensibilidad de detección de por lo menos un 0,0001%.

En algunas realizaciones, el segundo conjunto de señuelos tiene un punto de saturación cuando se somete a titulación, dicha titulación comprende: generar una curva de titulación que comprende: (i) medir la eficiencia de captura del segundo conjunto de señuelos en función de la concentración del señuelos y (ii) identificar un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el segundo conjunto de señuelos.

En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumente en menos del 20% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumente en menos del 10% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumente en menos del 5% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumente menos del 2% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el punto de saturación se selecciona de tal manera que la eficiencia de captura observada aumente menos del 1% a una concentración del señuelo dos veces mayor que la de la primera concentración.

En algunos aspectos, el primer conjunto de señuelos o el segundo conjunto de señuelos enriquecen selectivamente una o más regiones asociadas a nucleosomas de un genoma, dichas regiones asociadas a nucleosomas comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de un tipo de origen o enfermedad de una célula o tejido. En algunas realizaciones, el primer conjunto de regiones genómicas o las segundas regiones genómicas comprenden una o más mutaciones procesables, en donde la una o más mutaciones procesables comprenden una o más de: (i) mutaciones modificables con fármacos, (ii) mutaciones para monitorización terapéutica, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v) mutaciones específicas del tipo de célula, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico.

En algunos aspectos, la primera y segunda regiones genómicas comprenden por lo menos una porción de cada uno de por lo menos 5 genes seleccionados de la Tabla 3. En algunas realizaciones, la primera y la segunda regiones genómicas tienen un tamaño entre aproximadamente 25 kilobases y 1.000 kilobases y una profundidad de lectura de entre 1.000 recuentos/base y 50.000 recuentos/base.

En un aspecto, la presente divulgación proporciona un método para enriquecer múltiples regiones genómicas, que comprende: (a) poner una cantidad predeterminada de ácido nucleico de una muestra en contacto con una mezcla de señuelos que comprende: (i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas del ácido nucleico de la muestra, dicho primer conjunto de señuelos se proporciona a una primera concentración que es menor que un punto de saturación del primer conjunto de señuelos, y (ii) un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas

del ácido nucleico de la muestra, dicho segundo conjunto de señuelos se proporciona a una segunda concentración que está en o por encima del punto de saturación del segundo conjunto de señuelos; y (b) enriquecer el ácido nucleico de la muestra para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas, produciendo de este modo un ácido nucleico enriquecido.

En algunas realizaciones, el segundo conjunto de señuelos tiene un punto de saturación que es mayor que sustancialmente todos los puntos de saturación asociados con los señuelos en el segundo conjunto de señuelos cuando un señuelo del segundo conjunto de señuelos se somete a una curva de titulación generada (i) midiendo la eficiencia de captura de un señuelo del segundo conjunto de señuelos en función de la concentración del señuelo, y (ii) identificando un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo. En algunas realizaciones, el punto de saturación del primer conjunto de señuelos se selecciona de tal manera que la eficiencia de captura observada aumente en menos del 10% a una concentración del señuelo dos veces mayor que la de la primera concentración. En algunas realizaciones, el primer conjunto de señuelos o el segundo conjunto de señuelos se enriquecen selectivamente para una o más regiones asociadas a nucleosomas de un genoma, las regiones asociadas a nucleosomas comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de un tipo de célula o tejido de origen o estado patológico. En algunas realizaciones, el método comprende además (c) secuenciar el ácido nucleico enriquecido para producir una pluralidad de lecturas de secuencia. En algunas realizaciones, el método comprende además (d) producir una salida que comprende secuencias de ácido nucleico representativas del ácido nucleico de la muestra.

En un aspecto, la presente divulgación proporciona un método que comprende: (a) proporcionar una pluralidad de mezclas de señuelos, en donde cada una de la pluralidad de mezclas de señuelos comprende un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas y un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas, en donde el primer conjunto de señuelos está a diferentes concentraciones en la pluralidad de mezclas de señuelos y el segundo conjunto de señuelos está a la misma concentración en la pluralidad de mezclas de señuelos; (b) poner en contacto cada una de la pluralidad de mezclas de señuelos con una muestra de ácidos nucleicos para capturar ácidos nucleicos de la muestra de ácidos nucleicos con el primer conjunto de señuelos y el segundo conjunto de señuelos, en donde el segundo conjunto de señuelos en cada mezcla de señuelos se proporciona a una primera concentración que está en o por encima de un punto de saturación del segundo conjunto de señuelos, en donde los ácidos nucleicos de la muestra de ácidos nucleicos son capturados por el primer conjunto de señuelos y el segundo conjunto de señuelos; (c) secuenciar una porción de los ácidos nucleicos capturados con cada mezcla de señuelos para producir conjuntos de lecturas de secuencia dentro de un número asignado de lecturas de secuencia; (d) determinar la profundidad de lectura de las lecturas de secuencia para el primer conjunto de señuelos y el segundo conjunto de señuelos para cada mezcla de señuelos; y (e) identificar por lo menos una mezcla de señuelos que proporcione profundidades de lectura para el segundo conjunto de regiones genómicas; en donde las profundidades de lectura para el segundo conjunto de regiones genómicas proporciona una sensibilidad de detección de una variante genética de por lo menos el 0,0001% de frecuencia de alelos menores (MAF). En algunas realizaciones, los pasos (d) y/o (e) son opcionales.

En algunas realizaciones, el segundo conjunto de señuelos tiene un punto de saturación cuando se somete a titulación, dicha titulación comprende generar una curva de titulación que comprende: (i) medir la eficiencia de captura del segundo conjunto de señuelos en función de la concentración de los señuelos; y (ii) identificar un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el segundo conjunto de señuelos. En algunas realizaciones, el punto de saturación se selecciona de tal manera que una eficiencia de captura observada aumenta en menos del 10% a una concentración del conjunto de señuelos veces mayor que la de la primera concentración. En algunas realizaciones, el primer conjunto de señuelos o el segundo conjunto de señuelos enriquecen selectivamente una o más regiones asociadas a nucleosomas de un genoma, las regiones asociadas a nucleosomas comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de un tipo de célula o tejido de origen o estado patológico. En algunas realizaciones, el primer conjunto de regiones genómicas comprende una o más mutaciones procesables, en donde la una o más mutaciones procesables comprenden una o más de: (i) mutaciones modificables con fármacos, (ii) mutaciones para monitorización terapéutica, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v) mutaciones específicas del tipo celular, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico. En algunas realizaciones, las primeras regiones genómicas comprenden por lo menos una porción de cada uno de por lo menos 5 genes seleccionados de la Tabla 1. En algunas realizaciones, las primeras regiones genómicas tienen un tamaño entre aproximadamente 25 kilobases y 1000 kilobases y una profundidad de lectura de entre 1.000 recuentos/base y 50.000 recuentos/base. En algunas realizaciones, el punto de saturación del segundo conjunto de señuelos se selecciona de tal manera que la eficiencia de captura observada aumenta en menos del 10% a una concentración del señuelo dos veces mayor que la de la segunda concentración. En algunas realizaciones, el segundo conjunto de regiones genómicas comprende una o más mutaciones procesables, en donde la una o más mutaciones procesables comprenden una o más de: (i) mutaciones modificables con fármacos, (ii) mutaciones para monitorización terapéutica, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v)

mutaciones específicas del tipo celular, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico. En algunas realizaciones, las segundas regiones genómicas comprenden por lo menos una porción de cada uno de por lo menos 5 genes seleccionados de la Tabla 1. En algunas realizaciones, las segundas regiones genómicas tienen un tamaño entre aproximadamente 25 kilobases y 1000 kilobases y una profundidad de lectura de entre 1000 recuentos/base y 50.000 recuentos/base.

Los aspectos y ventajas adicionales de la presente divulgación resultarán fácilmente evidentes para los expertos en esta técnica a partir de la siguiente descripción detallada, en la que sólo se muestran y describen realizaciones ilustrativas de la presente divulgación. Como se comprenderá, la presente divulgación es capaz de otras realizaciones diferentes, y sus varios detalles son susceptibles de modificaciones en varios aspectos obvios, todo sin apartarse de la divulgación. Por consiguiente, los dibujos y la descripción deben considerarse de naturaleza ilustrativa y no restrictiva. La US2012/208706 A1 (DOWNING ET AL.) y la US2016/28161 A1 (JAROSZ MIRNA [US] ET AL) describen el uso de conjuntos de señuelos para enriquecer múltiples regiones genómicas.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Las características novedosas de la divulgación se exponen con particularidad en las reivindicaciones adjuntas. Se obtendrá una mejor comprensión de las características y ventajas de la presente divulgación haciendo referencia a la siguiente descripción detallada que expone realizaciones ilustrativas, en las que se utilizan los principios de la divulgación, y los dibujos acompañantes (también "Figura" y "FIG." en la presente), de los cuales:

La FIG. 1 ilustra cómo pueden generarse una pluralidad de lecturas para cada locus enriquecido a partir de una muestra de ácido nucleico libre de células.

La FIG. 2 ilustra un ejemplo de inserción con el apoyo de una familia grande. La FIG. 3 ilustra un ejemplo de familias de lecturas pequeñas (que pueden parecer proporcionar evidencia de una variante real) y familias de lecturas grandes (que pueden indicar un probable error aleatorio derivado de la PCR o la secuenciación).

La FIG. 4 ilustra los varios parámetros que pueden usarse en una prueba de hipótesis y cómo cada parámetro puede estar relacionado con una probabilidad particular, por ejemplo, de una familia de lecturas que coinciden con una referencia, de lecturas de una cadena que coinciden con una referencia, y de una lectura que coincide con una referencia

La FIG. 5 ilustra un ejemplo de un sistema informático que puede programarse o configurarse de otro modo para implementar los métodos de la presente divulgación.

La FIG. 6 ilustra una curva de saturación ejemplar que muestra un recuento de moléculas únicas en el eje y como una función de la cantidad de ADNcf de entrada en el eje x.

DESCRIPCIÓN DETALLADA

Aunque en la presente se han mostrado y descrito varias realizaciones de la invención, será obvio para los expertos en la técnica que tales realizaciones se proporcionan a modo de ejemplo solamente. A los expertos en la técnica se les pueden ocurrir numerosas variaciones, cambios y sustituciones sin apartarse de la invención. Debe entenderse que pueden emplearse varias alternativas a las realizaciones de la invención descritas en la presente.

El término "variante genética", como se usa en la presente, se refiere de manera general a una alteración, variante o polimorfismo en una muestra de ácidos nucleicos o genoma de un sujeto. Tal alteración, variante o polimorfismo puede ser con respecto a un genoma de referencia, que puede ser un genoma de referencia del sujeto u otro individuo. Los polimorfismos de un solo nucleótido (SNP) son una forma de polimorfismos. En algunos ejemplos, uno o más polimorfismos comprenden una o más variaciones de un solo nucleótido (SNV), inserciones, deleciones, repeticiones, pequeñas inserciones, pequeñas deleciones, pequeñas repeticiones, uniones de variantes estructurales, repeticiones en tándem de longitud variable y/o secuencias flanqueantes. Las variaciones del número de copias (CNV), las transversiones y otros reordenamientos también son formas de variación genética. Una alteración genómica puede ser un cambio de base, inserción, deleción, repetición, variación del número de copias, o transversión.

El término "polinucleótido" o "ácido polinucleico" como se usa en la presente, se refiere de manera general a una molécula que comprende una o más subunidades de ácido nucleico (una "molécula de ácido nucleico"). Un polinucleótido puede incluir una o más subunidades seleccionadas entre adenosina (A), citosina (C), guanina (G), timina (T) y uracilo (U), o variantes de los mismos. Un nucleótido puede incluir A, C, G, T o U, o variantes de los mismos. Un nucleótido puede incluir cualquier subunidad que pueda incorporarse en una cadena de ácido nucleico en crecimiento. Dicha subunidad puede ser una A, C, G, T o U, o cualquier otra subunidad que sea específica de una o más A, C, G, T o U complementarias, o complementarias a una purina (es decir, A o G, o variante de la misma) o una pirimidina (es decir, C, T o U, o variante de la misma). La identificación de una subunidad puede permitir que se resuelvan bases de ácidos nucleicos individuales o grupos de bases (por ejemplo, AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA o contrapartidas de uracilo de los mismos). En algunos ejemplos, un polinucleótido es ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN), o derivados de los mismos. Un polinucleótido puede ser de cadena sencilla o de cadena doble.

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 puede ser un ácido nucleico libre de células. Como se usa en la presente, los términos ácido nucleico libre de células y ácido nucleico extracelular pueden usarse indistintamente. Un polinucleótido puede ser ADN libre de células (ADNcf). Por ejemplo, el polinucleótido puede ser ADN circulante. El ADN circulante puede comprender ADN tumoral circulante (ADNcf). Los ácidos nucleicos extracelulares o libres de células pueden derivarse de cualquier fluido corporal, incluyendo pero no limitado a, sangre completa, plaquetas, suero, plasma, líquido sinovial, líquido linfático, líquido ascítico, líquido intersticial o extracelular, el líquido en los espacios entre células, líquido crevicular gingival, médula ósea, líquido cefalorraquídeo, saliva, mucosidad, esputo, semen, sudor, orina, líquido o lavado cervical, líquido o lavado vaginal, lavado o glándula mamaria y/o cualquier combinación de los mismos. En algunas realizaciones, los ácidos nucleicos extracelulares o libres de células pueden derivarse del plasma. En algunas realizaciones, puede procesarse un fluido corporal que contiene células para eliminar las células para purificar y/o extraer ácidos nucleicos extracelulares o libres de células. Un polinucleótido puede ser de cadena sencilla o de cadena doble. Alternativamente, un polinucleótido puede comprender una combinación de una porción de cadena doble y una porción de cadena sencilla.

Los polinucleótidos no tienen que estar libres de células. En algunos casos, los polinucleótidos pueden aislarse de una muestra. Una muestra puede ser una composición que comprende un analito. Por ejemplo, una muestra puede ser cualquier muestra biológica aislada de un sujeto que incluye, sin limitación, fluido corporal, sangre completa, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejido, líquido sinovial, líquido linfático, líquido ascítico, líquido intersticial o extracelular, el líquido en los espacios entre las células, incluyendo el líquido crevicular gingival, la médula ósea, el líquido cefalorraquídeo, la saliva, la mucosa, el esputo, el semen, el sudor, la orina o cualquier otro líquido corporal, y/o cualquier combinación de los mismos. 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 de tumor, que puede obtenerse de un sujeto mediante varios enfoques, que incluyen pero no están limitados a, punción venosa, excreción, eyaculación, masaje, biopsia, aspiración con aguja, lavado, raspado, incisión quirúrgica o intervención u otros enfoques. En algunas realizaciones, una muestra es una muestra de ácido nucleico, por ejemplo, una muestra de ácido nucleico purificada. En algunas realizaciones, una muestra de ácidos nucleicos comprende ADN libre de células (ADNcf). Un analito en una muestra puede estar en varias etapas de pureza. Por ejemplo, puede tomarse una muestra bruta directamente de un sujeto que puede contener el analito en un estado no purificado. Una muestra también puede enriquecerse para un analito. Un analito también puede estar presente en la muestra en forma aislada o sustancialmente aislada.

Los polinucleótidos pueden comprender secuencias asociadas con cáncer, como leucemia linfoblástica aguda (LLA), leucemia mieloide aguda (LMA), carcinoma adrenocortical, sarcoma de Kaposi, cáncer anal, carcinoma de células basales, cáncer de vías biliares, cáncer de vejiga, cáncer de huesos, osteosarcoma, histiocitoma fibroso maligno, glioma de tronco encefálico, cáncer de cerebro, craneofaringioma, ependimoblastoma, ependimoma, meduloblastoma, meduloeptitelioma, tumor de parénquima pineal, cáncer de mama, tumor bronquial, linfoma de Burkitt, linfoma no de Hodgkin, tumor carcinoide, cáncer cervical, cordoma, leucemia linfocítica crónica (LLC), leucemia mielógena crónica (LMC), cáncer de colon, cáncer colorrectal, linfoma de células T cutáneo, carcinoma ductal in situ, cáncer de endometrio, cáncer de esófago, sarcoma de Ewing, cáncer de ojo, melanoma intraocular, retinoblastoma, histiocitoma fibroso, cáncer de vesícula biliar, cáncer gástrico, glioma, leucemia de células pilosas, cáncer de cabeza y cuello, cáncer de corazón, cáncer hepatocelular (de hígado), linfoma de Hodgkin, cáncer de hipofaringe, cáncer de riñón, cáncer de laringe, cáncer de labios, cáncer de cavidad oral, cáncer de pulmón, carcinoma de células no pequeñas, carcinoma de células pequeñas, melanoma, cáncer de boca, síndromes mielodisplásicos, mieloma múltiple, meduloblastoma, cáncer de cavidad nasal, cáncer de seno paranasal, neuroblastoma, cáncer nasofaríngeo, cáncer oral, cáncer orofaríngeo, osteosarcoma, cáncer de ovario, cáncer de páncreas, papilomatosis, paraganglioma, cáncer de paratiroides, cáncer de pene, cáncer de faringe, tumor pituitario, neoplasia de células plasmáticas, cáncer de próstata, cáncer de recto, cáncer de células renales, rabdomiosarcoma, cáncer de glándulas salivales, síndrome de Sezary, cáncer de piel, no melanoma, cáncer de intestino delgado, sarcoma de tejido blando, carcinoma de células escamosas, cáncer de testículo, cáncer de garganta, timoma, cáncer de tiroides, cáncer de uretra, cáncer de útero, sarcoma de útero, cáncer de vagina, cáncer de vulva, macroglobulinemia de Waldenstrom y/o tumor de Wilms.

Una muestra puede comprender varias cantidades de ácido nucleico que contiene equivalentes de genoma. Por ejemplo, una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 (10^4) equivalentes de genoma humano haploide y, en el caso de ADNcf aproximadamente 200 billones (2×10^{11}) moléculas de polinucleótidos individuales. De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes de genoma humano haploide y, en el caso de ADNcf, aproximadamente 600 billones de moléculas individuales.

Una muestra puede comprender ácidos nucleicos de diferentes fuentes. Por ejemplo, una muestra puede comprender ADN de línea germinal o ADN somático. Una muestra puede comprender ácidos nucleicos portadores

de mutaciones. Por ejemplo, una muestra puede comprender ADN que porta mutaciones de la línea germinal y/o mutaciones somáticas. Una muestra también puede comprender ADN que porta mutaciones asociadas con el cáncer (por ejemplo, mutaciones somáticas asociadas con el cáncer).

El término "sujeto", como se usa en la presente, se refiere de manera general a un animal, como una especie de mamífero (por ejemplo, un humano) o una especie de ave (por ejemplo, un pájaro), u otro organismo, como una planta. Más específicamente, el sujeto puede ser un vertebrado, un mamífero, un ratón, un primate, un simio o un humano. Los animales incluyen, pero no están limitados a, animales de granja, animales deportivos y mascotas. Un sujeto puede ser un individuo sano, un individuo que tiene o se sospecha que tiene una enfermedad o predisposición a la enfermedad, o un individuo con necesidad de terapia o que se sospecha que necesita terapia. Un sujeto puede ser un paciente.

El término "genoma", como se usa en la presente, se refiere de manera general a la totalidad de la información hereditaria de un organismo. Un genoma puede estar codificado en ADN o en ARN. Un genoma puede comprender regiones codificantes que codifican proteínas, así como regiones no codificantes. Un genoma puede incluir la secuencia de todos los cromosomas juntos en un organismo. Por ejemplo, el genoma humano tiene un total de 46 cromosomas. La secuencia de todos estos juntos puede constituir un genoma humano. Un genoma puede comprender un genoma diploide o haploide.

El término "señuelo", como se usa en la presente, se refiere de manera general a un oligonucleótido específico del objetivo (por ejemplo, una sonda de captura) diseñado y usado para capturar regiones genómicas específicas de interés (por ejemplo, objetivos o regiones genómicas de interés predeterminadas). El señuelo puede capturar sus objetivos previstos hibridando selectivamente con ácidos nucleicos complementarios.

El término "panel de señuelos" o "panel de conjuntos de señuelos", como se usa en la presente, se refiere de manera general a un conjunto de señuelos dirigidos hacia un conjunto seleccionado de regiones genómicas de interés. Puede hacerse referencia a un panel de señuelos o un panel de conjuntos de señuelos como mezcla de señuelos. El panel de señuelos puede capturar sus objetivos previstos en un solo paso de hibridación selectiva.

El término "precisión" de detectar una variante genética (por ejemplo, un indel), como se usa en la presente, se refiere de manera general al porcentaje de variantes genéticas candidatas (por ejemplo, detectadas) detectadas mediante el análisis de una o más lecturas de secuencia que son identificada como una variante genética verdadera atribuible al origen biológico (por ejemplo, no atribuible a un error introducido, como el derivado del error de secuenciación o amplificación). El término "tasa de error" de detección de una variante genética (por ejemplo, un indel), como se usa en la presente, se refiere de manera general al porcentaje de variantes genéticas candidatas (por ejemplo, detectadas) detectadas a través del análisis de una o más lecturas de secuencia que se identifican como una variante genética introducida atribuible a un origen no biológico (por ejemplo, error de secuenciación o amplificación). Por ejemplo, si el análisis de una o más lecturas de secuencia identifica 100 variantes genéticas candidatas, de las cuales 90 son atribuibles a un origen biológico y 10 se atribuyen a un origen no biológico, entonces este análisis tiene una precisión de detección de la variante genética del 90% y una tasa de error del 10%.

El término "aproximadamente" y sus equivalentes gramaticales en relación a un valor numérico de referencia pueden incluir un intervalo de valores hasta más o menos el 10% de ese valor. Por ejemplo, la cantidad "aproximadamente 10" puede incluir cantidades de 9 a 11. En otras realizaciones, el término "aproximadamente" en relación a un valor numérico de referencia puede incluir un intervalo de valores más o menos el 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2% o 1% de ese valor.

El término "por lo menos" y sus equivalentes gramaticales en relación a un valor numérico de referencia pueden incluir el valor numérico de referencia y más de ese valor. Por ejemplo, la cantidad "por lo menos 10" puede incluir el valor 10 y cualquier valor numérico superior a 10, como 11, 100 y 1.000.

El término "como máximo" y sus equivalentes gramaticales en relación a un valor numérico de referencia pueden incluir el valor numérico de referencia y menos de ese valor. Por ejemplo, la cantidad "como máximo 10" puede incluir el valor 10 y cualquier valor numérico menor de 10, como 9, 8, 5, 1, 0,5 y 0,1.

Los términos "procesamiento", "cálculo" y "comparación" pueden usarse indistintamente. El término puede referirse a determinar una diferencia, por ejemplo, una diferencia en número o secuencia. Por ejemplo, pueden procesarse valores o secuencias de expresión génica, variación del número de copias (CNV), indel y/o variantes de un solo nucleótido (SNV).

La presente divulgación proporciona métodos y sistemas para el análisis de resolución múltiple de ácidos nucleicos libres de células (por ejemplo, ácido desoxirribonucleico (ADN)), en los que las regiones genómicas objetivo de interés pueden enriquecerse con sondas de captura ("señuelos") seleccionados para uno o más paneles de conjuntos de señuelos que usan un esquema de colocación en mosaico y captura diferencial. Un esquema de colocación en mosaico y captura diferencial usa conjuntos de señuelos de diferentes concentraciones relativas para

colocar en mosaico diferencialmente (por ejemplo, a diferentes "resoluciones") a través de las regiones genómicas asociadas con los señuelos, sujeto a un conjunto de restricciones (por ejemplo, restricciones del secuenciador como la carga de secuenciación, la utilidad de cada señuelo, etc.), y capturarlos al nivel deseado para secuenciarlos en sentido descendente. Estas regiones genómicas objetivo de interés pueden incluir variantes de un solo nucleótido (SNV) e indeles (es decir, inserciones o eliminaciones). Las regiones genómicas objetivo de interés pueden comprender regiones genómicas de interés de la estructura principal ("regiones de la estructura principal") o regiones genómicas de interés de puntos críticos ("regiones de punto crítico" o "puntos críticos"). Mientras que "puntos críticos" pueden referirse a loci particulares asociados con variantes de secuencia, las regiones de "estructura principal" pueden referirse a regiones genómicas más grandes, cada una de las cuales puede tener una o más variantes de secuencia potenciales. Por ejemplo, una región de la estructura principal puede ser una región que contiene una o más mutaciones asociadas con el cáncer, mientras que un punto crítico puede ser un locus con una mutación particular asociada con el cáncer recurrente. Tanto las regiones genómicas de la estructura principal como de puntos críticos de interés pueden comprender genes marcadores relevantes para el tumor que se incluyen comúnmente en los ensayos de biopsia líquida (por ejemplo, BRAF, BRCA, EGFR, KRAS, PIK3CA, ROS1, TP53 y otros), para los cuales puede esperarse que se vean una o más variantes en sujetos con cáncer.

Entre el conjunto de genes marcadores relevantes para el tumor que pueden seleccionarse para su inclusión en un panel de conjuntos de señuelos, las regiones genómicas de puntos críticos de interés pueden seleccionarse para estar representadas por una proporción más alta de lecturas de secuencia en comparación con las regiones genómicas de la estructura principal de interés en el protocolo experimental. Este protocolo experimental puede comprender pasos que incluyen aislamiento, amplificación, captura, secuenciación y análisis de datos. La selección de regiones como regiones de punto crítico o regiones de la estructura principal puede estar impulsada por consideraciones como la eficiencia de captura, la carga de secuenciación y/o la utilidad asociada con cada una de las regiones y su señuelo correspondiente. La utilidad puede evaluarse por la relevancia clínica (por ejemplo, "valor clínico") de un marcador genómico de interés (por ejemplo, un marcador tumoral) hacia un ensayo de biopsia líquida, por ejemplo, mutaciones impulsoras de cáncer predeterminadas, regiones genómicas con prevalencia en una cohorte de pacientes relevante, mutaciones impulsoras de cáncer identificadas empíricamente o regiones genómicas asociadas a nucleosomas. Por ejemplo, la utilidad puede medirse mediante una métrica representativa del rendimiento esperado de variantes genéticas procesables y/o asociadas a enfermedades en la detección o contribución para determinar el tejido de origen o el estado patológico de una muestra. La utilidad puede ser una función de valor clínico que aumenta monótonamente.

Como cada ciclo de secuenciación de una muestra dada de ácidos nucleicos libres de células está típicamente limitado por un cierto número total de lecturas, un enfoque de análisis de resolución múltiple para generar un panel de conjuntos de señuelos que enriquece preferentemente las "regiones de punto crítico" en comparación con las regiones de la estructura principal, permitirá el uso eficiente de las lecturas de secuenciación para la detección de variantes genéticas para aplicaciones de detección y evaluación del cáncer, al enfocar la secuenciación a profundidades de lectura más altas para las regiones de punto crítico sobre las regiones de la estructura principal. El uso de este enfoque puede permitir la mejora de un ensayo de muestra, dada una carga de secuenciación limitada o restringida (por ejemplo, número de lecturas secuenciadas por muestra analizada), de tal manera que pueda detectarse un mayor número de variantes genéticas clínicamente procesables por ensayo de muestra en comparación con un ensayo de muestra no optimizado.

La presente divulgación proporciona métodos para mejorar la precisión de la detección de una inserción o deleción (indel) de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto, dicha pluralidad de lecturas de secuencia se generan mediante secuenciación de ácidos nucleicos. Para cada una de la pluralidad de lecturas de secuencia asociadas con moléculas de ADNcf puede identificarse un indel candidato. Cada indel candidato puede clasificarse como un indel verdadero o un indel introducido, usando una combinación de esperanzas predeterminadas de (i) que se detecte un indel en una o más lecturas de secuencia de la pluralidad de lecturas de secuencia, (ii) que un indel detectado sea un indel verdadero presente en una molécula de ADNcf dada de las moléculas de ADN libre de células, dado que se ha detectado un indel en una o más de las lecturas de secuencia, y/o (iii) que un indel detectado sea introducido por error no biológico, dado que se ha detectado un indel en una o más lecturas de la secuencia, junto con uno o más parámetros del modelo para realizar una prueba de hipótesis. Este enfoque puede reducir el error y mejorar la precisión de la detección de un indel a partir de datos de lectura de secuencia.

Introducción

Una realización del análisis de resolución múltiple procede como sigue. Las regiones de un genoma se seleccionan para secuenciación. A estas regiones puede hacerse referencia colectivamente como panel o un bloque de paneles. El panel se divide en un primer conjunto de regiones genómicas y un segundo conjunto de regiones genómicas. Al primer conjunto de regiones genómicas puede hacerse referencia como región de la estructura principal, mientras que al segundo conjunto puede hacerse referencia como regiones de punto crítico. Estas regiones pueden dividirse entre genes o dentro de genes o fuera de genes según lo desee el practicante médico. Por ejemplo, un exón de un gen puede dividirse en porciones asignadas a la región del puntos críticos y porciones

asignadas a la región de la estructura principal.

Se preparan un primer conjunto de señuelos y un segundo conjunto de señuelos que hibridan selectivamente con las primeras regiones genómicas y las segundas regiones genómicas, respectivamente. Usando los métodos descritos en la presente, por ejemplo, preparación de curvas de titulación, se determinan las concentraciones del conjunto de señuelos que, para una muestra de prueba que tiene una cantidad predeterminada de ADN, captura el ADN en la muestra en un punto de saturación (para el conjunto de señuelos dirigido a las regiones del punto crítico) y por debajo del punto de saturación (para el conjunto de señuelos dirigido a las regiones de la estructura principal). La captura de moléculas de ADN de una muestra en el punto de saturación contribuye a detectar variantes genéticas en el nivel más alto de sensibilidad ya que es más probable que se capturen las variantes genéticas de moléculas.

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 en las técnicas de secuenciación, y consideraciones prácticas como el tiempo y el gasto. Por tanto, un "presupuesto de lectura" es una forma de conceptualizar la cantidad de información genética que puede extraerse de una muestra. Puede seleccionarse un presupuesto de lectura por muestra que identifique el número total de lecturas base que se asignarán a una muestra de prueba que comprende una cantidad predeterminada de ADN en un experimento de secuenciación. El presupuesto de lectura puede basarse en las lecturas totales producidas, por ejemplo, incluyendo las lecturas redundantes producidas mediante 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 lectura de 3.000.000.000 de lecturas puede asignarse como 150.000 bases a una profundidad de lectura media 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 estructura principal del panel, a una primera profundidad de lectura media y bases en la región del punto crítico del panel, a una profundidad de lectura más profunda.

A modo de ejemplo no limitativo, si un presupuesto de lectura consiste de 100.000 recuentos de lectura para una muestra dada, esos 100.000 recuentos de lectura se dividirán entre las lecturas de las regiones de la estructura principal y las lecturas de las regiones de puntos de acceso. La asignación de un gran número de esas lecturas (por ejemplo, 90.000 lecturas) a las regiones de la estructura principal dará como resultado que una pequeña cantidad de lecturas (por ejemplo, las 10.000 lecturas restantes) se asignen a las regiones de punto crítico. Por el contrario, la asignación de un gran número de lecturas (por ejemplo, 90.000 lecturas) a regiones de punto crítico dará como resultado que una pequeña cantidad de lecturas (por ejemplo, las 10.000 lecturas restantes) se asignen a regiones de la estructura principal. Por tanto, un trabajador calificado puede asignar un presupuesto de lectura para proporcionar los niveles deseados de sensibilidad y especificidad. En ciertas realizaciones, el presupuesto de lectura 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, o entre 1.000.000.000 lecturas y 5.000.000.000 lecturas a través de, por ejemplo, 20.000 bases a 100.000 bases.

El primer y el segundo niveles de sensibilidad se seleccionan para la detección de variantes genéticas en las regiones de la estructura principal y del punto crítico, respectivamente. La sensibilidad, como se usa en la presente, se refiere al límite de detección de una variante genética en función de la frecuencia en una muestra. Por ejemplo, la sensibilidad puede ser por lo menos del 1%, por lo menos del 0,1%, por lo menos del 0,01%, por lo menos del 0,001%, por lo menos del 0,0001% o por lo menos del 0,00001%, lo que significa que una secuencia dada puede detectarse en una muestra a una frecuencia de por lo menos el 1%, por lo menos el 0,1%, por lo menos el 0,01%, por lo menos el 0,001%, por lo menos el 0,0001% o por lo menos el 0,00001%, respectivamente. Es decir, las variantes genéticas presentes en la muestra en los niveles son detectables mediante secuenciación. Típicamente, la sensibilidad seleccionada para las regiones de punto crítico será mayor que la sensibilidad seleccionada para las regiones de la estructura principal. Por ejemplo, el nivel de sensibilidad para las regiones de punto crítico puede seleccionarse a por lo menos el 0,001%, mientras que el nivel de sensibilidad para las regiones de fondo puede seleccionarse a por lo menos el 0,1% o por lo menos el 1%.

Las concentraciones relativas de conjuntos de señuelos dirigidas a regiones de fondo y regiones de punto crítico pueden seleccionarse para optimizar las lecturas en un experimento de secuenciación con respecto al presupuesto de lectura seleccionado y las sensibilidades seleccionadas para las regiones de estructura principal y puntos críticos para una muestra seleccionada. Entonces, por ejemplo, dada una muestra de prueba que contiene una cantidad predeterminada de ADN, y un conjunto de señuelos de punto crítico que captura el ADN de las regiones de punto crítico en saturación, se selecciona una cantidad de conjunto de señuelos de la estructura

principal que está por debajo de la saturación para la muestra de tal manera que en un experimento de secuenciación que produce lecturas dentro del presupuesto de lectura seleccionado, el conjunto de lecturas resultante detecta variantes genéticas en las regiones de punto crítico y en las regiones de la estructura principal a los niveles de sensibilidad preseleccionados.

Las cantidades relativas de los conjuntos de señuelos son una función de varios factores. Uno de estos factores es la proporción relativa del panel asignado a las regiones del punto crítico y a las regiones de estructura principal, respectivamente. Cuanto mayor sea el porcentaje relativo de regiones de punto crítico en el panel, menor será el número de lecturas y el presupuesto que puede asignarse a la región de estructura principal. Otro factor es la sensibilidad de detección seleccionada para las regiones de punto crítico. Para una muestra dada, cuanto mayor sea la sensibilidad necesaria para las regiones de punto crítico, menor será la sensibilidad para la región de la estructura principal. Otro factor es el presupuesto de lectura. Para una sensibilidad para las regiones de punto crítico, cuanto menor sea el presupuesto de lectura, menor será la sensibilidad posible para la región de la estructura principal. Otro factor es el tamaño del panel general. Para cualquier presupuesto de lectura dado, cuanto mayor sea el panel, más sensibilidad de las regiones de la estructura principal deberá sacrificarse para lograr la sensibilidad deseada en las regiones de punto crítico.

Será evidente que para cualquier presupuesto de lectura dado, aumentar el porcentaje de lecturas asignadas a las regiones de la estructura principal disminuirá la sensibilidad de detección en las regiones de punto crítico. Al contrario, aumentar la sensibilidad de detección en las regiones del punto crítico, aumentando la cantidad del presupuesto de lectura asignado a las regiones del punto crítico, disminuye la detección de las regiones de la estructura principal. En consecuencia, los niveles de sensibilidad relativos de las regiones de punto crítico pueden ser lo suficientemente altos para lograr los niveles de detección objetivo, mientras que el nivel de sensibilidad en las regiones de la estructura principal no es tan bajo como para que se pierdan niveles significativos de variantes genéticas. El médico selecciona estos niveles relativos para lograr los resultados deseados. En algunas realizaciones, un trabajador experto usará una mezcla de señuelos calculada para capturar todas (o sustancialmente todas) las regiones de punto crítico en una muestra y una porción de las regiones de estructura principal, de tal manera que la profundidad de lectura de las regiones capturadas proporcionará las sensibilidades de punto crítico y estructura principal deseadas.

Regiones genómicas asociadas a nucleosomas

En un aspecto, un panel de conjuntos de señuelos puede comprender uno o más conjuntos de señuelos que se enriquecen selectivamente para una o más regiones asociadas a nucleosomas de un genoma. Las regiones asociadas a nucleosomas pueden comprender regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial. La ocupación nucleosómica diferencial puede ser característica de un tipo de célula o tejido de origen o estado patológico. El análisis de la ocupación nucleosómica diferencial puede realizarse usando uno o más perfiles de ocupación nucleosómica de un tipo de célula o de tejido dado. Ejemplos de técnicas de elaboración de perfiles de ocupación nucleosómica incluyen Statham et al, Genomics Data, Volumen 3, marzo de 2015, páginas 94-96 (2015). Los ácidos nucleicos libres de células de una muestra obtenida de un sujeto pueden eliminarse principalmente mediante una combinación de procesos apoptóticos y necróticos en células, tejidos y órganos. Como resultado de la ocupación nucleosómica variable y la protección contra la escisión del ADN en ciertas localizaciones de un genoma, los patrones o perfiles nucleosómicos asociados con procesos apoptóticos y procesos necróticos pueden ser evidentes al analizar fragmentos de ácido nucleico libre de células para las regiones de un genoma asociadas a nucleosomas.

La detección de tales patrones asociados a nucleosomas puede usarse, independientemente o junto con variantes somáticas detectadas, para monitorizar una afección en un sujeto. Por ejemplo, a medida que un tumor se expande, la proporción de necrosis a apoptosis en el microambiente del tumor puede cambiar. Tales cambios en la necrosis y/o apoptosis pueden detectarse enriqueciendo selectivamente una muestra de ácidos nucleicos libres de células para una o más regiones asociadas a nucleosomas. Como otro ejemplo, puede observarse una distribución de las longitudes de los fragmentos debida a la protección nucleosómica diferencial entre diferentes tipos de células, o entre células tumorales y no tumorales. El análisis de las regiones asociadas a nucleosomas para la distribución de la longitud de los fragmentos puede ser clínicamente relevante para las aplicaciones de detección y evaluación del cáncer. Este análisis puede comprender el enriquecimiento selectivo de las regiones asociadas a nucleosomas, luego la secuenciación de las regiones enriquecidas para producir una pluralidad de lecturas de secuencia representativas de la muestra de ácido nucleico, y analizar las lecturas de secuencia para variantes genéticas y perfiles de nucleosomas de interés.

Una vez que se han identificado las regiones asociadas a nucleosomas, pueden usarse para el diseño de paneles modulares. Ver a continuación. Tal diseño de paneles modulares puede permitir diseños de un conjunto de sondas o señuelos que enriquecen selectivamente regiones del genoma que son relevantes para la elaboración de perfiles nucleosómicos. Al incorporar esta "conciencia nucleosómica", pueden obtenerse datos de secuencia de muchos individuos para optimizar el procedimiento de diseño del panel, por ejemplo, la determinación de qué localizaciones genómicas apuntar y la concentración óptima de sondas para estas localizaciones genómicas.

Al incorporar el conocimiento tanto de las variaciones somáticas como de las variaciones estructurales e inestabilidad, pueden diseñarse paneles de sondas, señuelos o cebadores para dirigirse a porciones específicas del genoma ("puntos críticos") con patrones o grupos conocidos de variación estructural o inestabilidad. Por ejemplo, el análisis estadístico de datos de secuencia revela una serie de eventos somáticos acumulados y variaciones estructurales, y por lo tanto permite estudios de evolución clonal. El análisis de datos revela importantes conocimientos biológicos, incluyendo la cobertura diferencial entre cohortes, patrones que indican la presencia de ciertos subconjuntos de tumores, eventos estructurales extraños en muestras con alta carga de mutación somática y cobertura diferencial atribuida de células sanguíneas frente a células tumorales.

Una región genómica localizada se refiere a una región corta del genoma que puede variar en longitud de, o de aproximadamente, 2 a 200 pares de bases, de 2 a 190 pares de bases, de 2 a 180 pares de bases, de 2 a 170 pares de bases, de 2 a 160 pares de bases, de 2 a 150 pares de bases, de 2 a 140 pares de bases, de 2 a 130 pares de bases, de 2 a 120 pares de bases, de 2 a 110 pares de bases, de 2 a 100 pares de bases pares, de 2 a 90 pares de bases, de 2 a 80 pares de bases, de 2 a 70 pares de bases, de 2 a 60 pares de bases, de 2 a 50 pares de bases, de 2 a 40 pares de bases, de 2 a 30 pares de bases, de 2 a 20 pares de bases, de 2 a 10 pares de bases y/o de 2 a 5 pares de bases. Cada región genómica localizada puede contener un patrón o grupo de variación estructural significativa o inestabilidad. Pueden proporcionarse mapas de partición del genoma para identificar regiones genómicas localizadas relevantes. Una región genómica localizada puede contener un patrón o grupo de variación estructural significativa o inestabilidad estructural. Un grupo puede ser una región de punto crítico dentro de una región genómica localizada. La región del punto crítico puede contener una o más fluctuaciones o picos significativos. Puede seleccionarse una variación estructural del grupo que consiste de: una inserción, una deleción, una translocación, un reordenamiento genético, estado de metilación, un microsatélite, una variación del número de copias, una variación estructural relacionada con el número de copias o cualquier otra variación que indica diferenciación. Una variación estructural puede provocar una variación en el posicionamiento nucleosómico.

Un mapa de partición del genoma puede obtenerse: (a) proporcionando muestras de ADN o ARN libre de células de dos o más sujetos en una cohorte, (b) obteniendo una pluralidad de lecturas de secuencia de cada una de las muestras de ADN o ARN libre de células, y (c) analizando la pluralidad de lecturas de secuencia para identificar una o más regiones genómicas localizadas, cada una de las cuales contiene un patrón o grupo de variación estructural o inestabilidad significativas. El análisis estadístico puede realizarse sobre la información de la secuencia para asociar un conjunto de lecturas de secuencia con uno o más perfiles de ocupación nucleosómica que representan distintas cohortes (por ejemplo, un grupo de sujetos con una característica común como un estado patológico o un estado no patológico).

El análisis estadístico puede comprender proporcionar uno o más mapas de partición del genoma que enumeran los intervalos genómicos relevantes representativos de los genes de interés para un análisis adicional. El análisis estadístico puede comprender además seleccionar un conjunto de una o más regiones genómicas localizadas en base a los mapas de partición del genoma. El análisis estadístico puede comprender además analizar una o más regiones genómicas localizadas en el conjunto para obtener un conjunto de una o más alteraciones del mapa nucleosómico. El análisis estadístico puede comprender uno o más de (por ejemplo, uno o más, dos o más, o tres de): reconocimiento de patrones, aprendizaje profundo y aprendizaje no supervisado.

Una alteración del mapa nucleosómico es un valor medido que caracteriza una región genómica localizada dada en términos de información biológicamente relevante. Una alteración del mapa nucleosómico puede estar asociada con una mutación impulsora elegida del grupo que consiste de: tipo salvaje, variante somática, variante de la línea germinal y metilación del ADN.

Pueden usarse una o más alteraciones del mapa nucleosómico para clasificar un conjunto de lecturas de secuencia como asociadas con uno o más perfiles de ocupación nucleosómica que representan distintas cohortes. Estos perfiles de ocupación nucleosómica pueden estar asociados con una o más evaluaciones. Una evaluación puede considerarse como parte de una intervención terapéutica (por ejemplo, opciones de tratamiento, selección de tratamiento, evaluación adicional mediante biopsia y/o imagenología).

Puede seleccionarse una evaluación del grupo que consiste de: indicación, tipo de tumor, gravedad del tumor, agresividad del tumor, resistencia del tumor al tratamiento, y clonalidad del tumor. Puede determinarse una evaluación de la clonalidad del tumor observando la heterogeneidad en la alteración del mapa nucleosómico a través de moléculas de ADN libres de células en una muestra. Se determina una evaluación de las contribuciones relativas de cada uno de dos o más clones.

Cada una de las una o más regiones asociadas a nucleosomas de un panel de conjuntos de señuelos puede comprender por lo menos una de: (i) variación estructural significativa, que comprende una variación en el posicionamiento nucleosómico, dicha variación estructural seleccionada del grupo que consiste de: una inserción, una deleción, una translocación, un reordenamiento genético, estado de metilación, un microsatélite, una variación del número de copias, una variación estructural relacionada con el número de copias o cualquier otra variación que

indique diferenciación; y (ii) inestabilidad, que comprende una o más fluctuaciones o picos significativos en un mapa de partición del genoma que indica una o más localizaciones de alteraciones del mapa nucleosómico en un genoma. El uno o más conjuntos de señuelos de un panel de conjuntos de señuelos pueden configurarse para capturar regiones del genoma asociadas a nucleosomas en función de una pluralidad de perfiles de ocupación de nucleosomas de referencia asociados con uno o más estados patológicos y uno o más estados no patológicos.

El uno o más conjuntos de señuelos de un panel de conjuntos de señuelos pueden enriquecerse selectivamente para una o más regiones asociadas a nucleosomas en una muestra de ácido desoxirribonucleico libre de células (ADNcf). Por ejemplo, el conjunto de señuelos puede enriquecerse selectivamente para una o más regiones asociadas a nucleosomas poniendo una muestra nucleica en contacto con el conjunto de señuelos y permitiendo que el conjunto de señuelos hibride selectivamente con el conjunto de regiones genómicas asociadas a nucleosomas asociadas con el conjunto de señuelos.

En un aspecto, un método para enriquecer una muestra de ácidos nucleicos para regiones asociadas a nucleosomas de un genoma puede comprender (a) poner una muestra de ácidos nucleicos en contacto con un panel de conjuntos de señuelos, dicho panel de conjuntos de señuelos comprendiendo uno o más conjuntos de señuelos que se enriquecen selectivamente para una o más regiones asociadas a nucleosomas de un genoma; y (b) enriquecer la muestra de ácidos nucleicos para una o más regiones asociadas a nucleosomas de un genoma. Los uno o más conjuntos de señuelos en un panel de conjuntos de señuelos pueden configurarse para capturar regiones del genoma asociadas a nucleosomas en base a una función de una pluralidad de perfiles de ocupación de nucleosomas de referencia asociados con uno o más estados patológicos y uno o más estados no patológicos. La pluralidad de perfiles de ocupación nucleosómica de referencia puede servir como un "mapa" para el que el análisis puede revelar patrones o grupos de regiones y/o localizaciones genómicas que pueden ser objetivo para la captura para la detección de variantes asociadas a nucleosomas.

El uno o más conjuntos de señuelos en un panel de conjuntos de señuelos pueden enriquecerse selectivamente para la una o más regiones asociadas a nucleosomas en una muestra de ácido desoxirribonucleico libre de células (ADNcf). El método para enriquecer una muestra de ácidos nucleicos para regiones asociadas a nucleosomas de un genoma puede comprender además secuenciar los ácidos nucleicos enriquecidos para producir lecturas de secuencias de las regiones asociadas a nucleosomas de un genoma. Estas lecturas de secuencia pueden alinearse con un genoma de referencia y analizarse en busca de variantes asociadas a nucleosomas y/o genéticas (por ejemplo, SNV y/o indeles).

En un aspecto, un método para generar un conjunto de señuelos puede comprender (a) identificar una o más regiones de un genoma, dichas regiones asociadas con un perfil de nucleosoma, y (b) seleccionar un conjunto de señuelos para capturar selectivamente dichas regiones. Un conjunto de señuelos en un panel de conjuntos de señuelos puede enriquecerse selectivamente para una o más regiones genómicas asociadas a nucleosomas en una muestra de ácido desoxirribonucleico libre de células (ADNcf). Por ejemplo, el conjunto de señuelos puede enriquecerse selectivamente para una o más regiones asociadas a nucleosomas poniendo una muestra nucleica en contacto con el conjunto de señuelos y permitiendo que el conjunto de señuelos hibride selectivamente con el conjunto de regiones genómicas asociadas a nucleosomas asociadas con el conjunto de señuelos.

Paneles de señuelos para el enriquecimiento de múltiples regiones genómicas

En un aspecto, un panel de señuelos puede comprender un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas de una muestra de ácidos nucleicos que comprende una cantidad predeterminada de ADN, en donde el primer conjunto de señuelos puede proporcionarse a una primera proporción de concentración que es menor que un punto de saturación del primer conjunto de señuelos; y un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, en donde el segundo conjunto de señuelos puede proporcionarse a una segunda proporción de concentración que está asociada con un punto de saturación del segundo conjunto de señuelos. Como se usa en la presente, una concentración asociada con un punto de saturación puede estar en o por encima del punto de saturación. En algunas realizaciones, una concentración asociada con un punto de saturación está en o por encima de un punto que está un 10% por debajo del punto de saturación. El primer conjunto de regiones genómicas puede comprender una o más regiones genómicas de la estructura principal. El segundo conjunto de regiones genómicas puede comprender una o más regiones genómicas del punto crítico. La cantidad predeterminada de ADN puede ser aproximadamente 200 ng, aproximadamente 150 ng, aproximadamente 125 ng, aproximadamente 100 ng, aproximadamente 75 ng, aproximadamente 50 ng, aproximadamente 25 ng, aproximadamente 10 ng, aproximadamente 5 ng y/o aproximadamente 1 ng.

En un aspecto, un método para enriquecer múltiples regiones genómicas puede comprender poner en contacto una cantidad predeterminada de una muestra de ácidos nucleicos con un panel de señuelos que comprende (i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas de la muestra de ácidos nucleicos, que puede proporcionarse a una primera proporción de concentración que es menor que el punto de saturación del primer conjunto de señuelos, y (ii) un segundo conjunto de señuelos

que hibrida selectivamente con un segundo conjunto de regiones genómicas de la muestra de ácidos nucleicos, que puede proporcionarse a una segunda proporción de concentración asociada con un punto de saturación del segundo conjunto de señuelos; y enriquecer la muestra de ácidos nucleicos para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas.

El enriquecimiento puede comprender los pasos siguientes: (a) poner en contacto la muestra de ácidos nucleicos con un conjunto de señuelos; (b) capturar los ácidos nucleicos de la muestra hibridándolos con las sondas en el conjunto de señuelos; y (c) separar los ácidos nucleicos capturados de los ácidos nucleicos no capturados.

Usando este enfoque, la captura del segundo conjunto de regiones genómicas a un punto de saturación de su conjunto de señuelos puede producir una detección de alta sensibilidad de variantes del segundo conjunto de regiones genómicas (por ejemplo, regiones de punto crítico), mientras que la captura del primer conjunto de regiones genómicas por debajo del punto de saturación de su conjunto de señuelos puede ser deseable para el primer conjunto de regiones genómicas (por ejemplo, regiones de la estructura principal). La flexibilidad de este método para ajustar la captura de diferentes conjuntos de señuelos en o por debajo de sus niveles de saturación puede aprovecharse para seleccionar estratégicamente regiones genómicas de interés para los paneles de conjuntos de señuelos de la estructura principal o el punto crítico, dadas las características de cada región genómica, como la carga de secuenciación y la utilidad..

El método puede comprender además secuenciar los ácidos nucleicos enriquecidos para producir una pluralidad de lecturas de secuencia del primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas. Estas lecturas de secuencia pueden analizarse en busca de variantes genéticas relevantes para el cáncer (por ejemplo, SNV e indeles) para aplicaciones de detección y evaluación del cáncer.

El experto en la técnica apreciará que el punto de saturación se refiere a la saturación de la cinética de unión. En esencia, a medida que aumenta la concentración de un señuelo (o conjunto de señuelos), también aumentará la cantidad de objetivo que se une al señuelo (o conjunto de señuelos). Sin embargo, la cantidad de objetivo en una muestra dada será fija y, por tanto, en un cierto punto, todo el objetivo de la muestra estará unido al señuelo (o conjunto de señuelos). Por lo tanto, a medida que las concentraciones de señuelo aumentan más allá de este punto, la cantidad de objetivo unido no aumentará sustancialmente ya que el sistema se acercará al equilibrio de unión (las tasas a las que las moléculas de señuelo se unen y liberan las moléculas objetivo comenzarán a converger).

El punto de saturación se refiere a una concentración o cantidad de señuelo en cuyo punto aumentar esa concentración o cantidad no aumenta sustancialmente la cantidad de material objetivo capturado de una muestra, por ejemplo, ese punto en el que los aumentos en la concentración de señuelo producen aumentos cada vez menores en la cantidad total de material objetivo capturado. En algunas realizaciones, el punto en el que aumentar la concentración o la cantidad de señuelo no aumenta sustancialmente la cantidad de material objetivo capturado de una muestra es el punto en el que aumentar la concentración o cantidad de señuelo no produce un aumento en la cantidad de objetivo capturado de la muestra. El punto de saturación puede ser un punto de inflexión en una curva de saturación que mide la cantidad de ácido nucleico objetivo capturado con concentraciones crecientes del conjunto de señuelos. Por ejemplo, el punto de saturación puede ser el punto en el que un aumento del 100% en la concentración del señuelo (por ejemplo, 2 X o dos veces la concentración) aumenta la cantidad de objetivo capturado en menos del 20%, menos del 19%, menos del 18 %>, menos del 17%, menos del 16%, menos del 15%, menos del 14%, menos del 13%, menos del 12%, menos del 11%, menos del 10%, menos del 9%, menos del 8%, menos del 7%, menos del 6%, menos del 5%, menos del 4%, menos del 3%, menos del 2% o menos del 1%. En algunas realizaciones, un aumento del 50% en la concentración del señuelo (por ejemplo, 1,5 X o una vez y media la concentración) aumenta la cantidad de objetivo capturado en menos del 20%, menos del 19%, menos del 18%, menos del 17%, menos del 16%, menos del 15%, menos del 14%, menos del 13%, menos del 12%, menos del 11%, menos del 10%, menos del 9%, menos del 8%, menos del 7%, menos del 6%, menos del 5%, menos del 4%, menos del 3%, menos del 2% o menos del 1%. En algunas realizaciones, un aumento del 20% en la concentración de señuelo (por ejemplo, 1,2 X) aumenta la cantidad de objetivo capturado en menos del 20%, menos del 19%, menos del 18%, menos del 17%, menos del 16%, menos del 15%, menos del 14%, menos del 13%, menos del 12%, menos del 11%, menos del 10%, menos del 9%, menos del 8%, menos del 7%, menos del 6%, menos del 5%, menos del 4%, menos del 3%, menos del 2% o menos del 1%. En algunas realizaciones, un aumento del 10% en la concentración de señuelo (por ejemplo, 1,1 X) aumenta la cantidad de objetivo capturado en menos del 20%, menos del 19%, menos del 18%, menos del 17%, menos del 16%, menos del 15%, menos del 14%, menos del 13%, menos del 12%, menos del 11%, menos del 10%, menos del 9%, menos del 8%, menos del 7%, menos del 6%, menos del 5%, menos del 4%, menos del 3%, menos del 2% o menos del 1%.

Como otro ejemplo, el punto de saturación puede ser el punto en el que un aumento del 100% en la concentración del señuelo (por ejemplo, 2 X o dos veces la concentración) aumenta la cantidad de objetivo capturado como máximo en un 20%. El punto de saturación puede ser el punto en el que un aumento del 50% en la concentración del señuelo (por ejemplo, 1,5 X o el doble de la concentración) aumenta la cantidad de objetivo capturado como máximo en un 20%. El punto de saturación puede ser el punto en el que un aumento del 20% en la

concentración del señuelo (por ejemplo, 1,2 X o dos veces la concentración) aumenta la cantidad de objetivo capturado como máximo en un 20%. El punto de saturación puede ser el punto en donde un aumento del 10% en la concentración del señuelo (por ejemplo, 1,1 X o dos veces la concentración) aumenta la cantidad de objetivo capturado como máximo en un 20%.

Puede generarse una curva de saturación, por ejemplo, titulando diferentes cantidades de ácidos nucleicos objetivo frente a una cantidad fija o variable de señuelos (por ejemplo, señuelos fijados en una micromatriz) para medir la cantidad de ácido nucleico objetivo (incluyendo, por ejemplo, el número de moléculas únicas) unidas a los señuelos. También puede generarse una curva de saturación, por ejemplo, titulando diferentes cantidades de señuelos contra una cantidad fija o variable de ácidos nucleicos objetivo para medir la cantidad de ácido nucleico objetivo (incluyendo, por ejemplo, el número de moléculas únicas) unido a los señuelos. En algunas realizaciones, puede generarse una curva de saturación usando un subconjunto de lecturas de secuencia como una medida del ácido nucleico objetivo (por ejemplo, recuento de moléculas únicas) capturado. Por ejemplo, las lecturas de secuencia pueden clasificarse como que tienen soporte de cadena sencilla (cuando todas las lecturas dentro de un grupo de lecturas únicas son de la misma cadena de ácido nucleico original de un ácido nucleico de cadena doble como el ADN) o soporte de cadena doble (cuando las lecturas dentro de un grupo de lecturas únicas son de ambas cadenas de ácido nucleico originales de un ácido nucleico de cadena doble como el ADN). En las realizaciones que seleccionan el soporte de cadena doble, el experto en la técnica entenderá que debe contar sólo las moléculas únicas capturadas para las que se observan ambas cadenas. El soporte de doble cadena puede determinarse, por ejemplo, marcando diferencialmente cada una de las dos cadenas diferentes de un ácido nucleico de tal manera que las lecturas de cada cadena puedan contarse por separado. En algunas realizaciones, un ácido nucleico objetivo con soporte de cadena doble requerirá una cantidad más alta de señuelo para alcanzar la saturación para esa objetivo que la que se requeriría para un señuelo con soporte de cadena sencilla.

La **FIG. 6** representa una curva de saturación ejemplar que muestra el recuento de moléculas únicas en el eje y en función de la cantidad de señuelo de entrada en el eje x. En cada cantidad de entrada (mostrada como una serie de volúmenes de una solución de señuelo), se tituló la cantidad de señuelo para generar la curva. En la Tabla 1 y la Tabla 2 siguientes se muestran diseños de curvas de titulación experimentales ejemplares. El número de lecturas de secuencias únicas frente a la cantidad de señuelo de entrada puede usarse para generar una curva de titulación como se muestra en la **FIG. 6**.

Tabla 1: Diseño de la curva de titulación

	Cantidad de objetivo de entrada (0, 5, 15 o 30 ng)							
Cantidad de señuelo (estructura principal o punto crítico; µl)	**Vol. A**	**Vol. B**	**Vol. C**	**Vol. D**	**Vol. E**	**Vol. F**	**Vol. G**	**Vol. H**
Estructura principal 1 (ng de ácido nucleico objetivo de entrada)	0	5	5	0	5	0	5	5
Estructura principal 2 (ng de ácido nucleico objetivo de entrada)	30	30	30	0	30	0	30	30
Punto crítico 1 (ng de ácido nucleico objetivo de entrada)	0	0	0	0	0	5	0	0
Punto crítico 2 (ng de ácido nucleico objetivo de entrada)	0	0	0	0	0	15	0	0
Punto crítico 3 (ng de ácido nucleico objetivo de entrada)	0	0	0	0	0	30	0	0
Estructura principal 3 (ng de ácido nucleico objetivo de entrada)	5	5	5	0	5	0	5	5
Estructura principal 4 (ng de ácido nucleico objetivo de entrada)	0	0	15	0	15	0	15	15
Estructura principal 5 (ng de ácido nucleico objetivo de entrada)	30	30	30	0	30	0	30	30
Punto crítico 4 (ng de ácido nucleico objetivo de entrada)	5	5	0	5	0	5	0	0
Punto crítico 5 (ng de ácido nucleico objetivo de entrada)	0	0	0	0	0	15	0	0
Punto crítico 6 (ng de ácido nucleico objetivo de entrada)	30	30	0	30	0	30	0	0

Tabla 2: Diseño de la curva de titulación. Hibridación realizada a 65º C.

Condición #	Señuelo de punto crítico (µl)	Señuelo de estructura principal (µl)	Cantidad de ácido nucleico objetivo de entrada (ng)
1	A	B	5
2	A	B	5
3	A	B	5
4	A	B	5
5	A2	B1	5
6	A2	B1	5
7	A2	B2	5
8	A2	B2	5
9	A	B1	15
10	A	B1	15
11	A	B2	15
12	A	B2	15
13	A2	B1	15
14	A2	B1	15
15	A2	B2	15
16	A2	B2	15
17	A2	B2	30
18	A2	B2	30

Usando una curva de titulación como la de la FIG. 6, un experto en la técnica puede calcular un punto de saturación. Por ejemplo, mirando el Vol. 0.8X, el recuento de moléculas únicas es de aproximadamente 2700. A 2 X la cantidad de señuelo (Vol. 1.6X), el recuento de moléculas únicas es aproximadamente 3200, una diferencia de 500. Por tanto, duplicar la cantidad de señuelo da como resultado un aumento en la captura de aproximadamente el 18,5%. Por el contrario, en el Vol. 2X, el recuento de moléculas únicas es de aproximadamente 3250, y a 1 µl, el recuento de moléculas únicas es de aproximadamente 3500, una diferencia de 250. Duplicar la cantidad de señuelo aquí da como resultado un aumento en la captura de sólo alrededor del 7,7%. Por consiguiente, una persona experta en la técnica que busque usar un punto de saturación en el que un aumento del 100% en la concentración de señuelo para aumentar la cantidad de objetivo capturado en menos del 8% podría usar por lo tanto el Vol. 2X de señuelo como el punto de saturación.

En el punto de saturación, el conjunto de señuelos puede capturar cualquiera de por lo menos el 40%, por lo menos el 50%, por lo menos el 60%, por lo menos el 70%, por lo menos el 80%, por lo menos el 85%, por lo menos el 86%, por lo menos el 87%, por lo menos el 88%, por lo menos el 89%, por lo menos el 90%, por lo menos el 91%, por lo menos el 92%, por lo menos el 93%, por lo menos el 94%, por lo menos el 95%, por lo menos el 96%, por lo menos el 97%, por lo menos el 98% y/o por lo menos el 99% de una secuencia objetivo en una muestra. El punto de saturación puede referirse al punto de saturación de un conjunto de señuelos o de un señuelo en particular, según el contexto en el que se use el término.

El punto de saturación de un conjunto de señuelos puede determinarse mediante el método siguiente: (a) para cada uno de los señuelos en el conjunto de señuelos, generar una curva de titulación que comprende (i) medir la eficiencia de captura del señuelo en una cantidad dada de la muestra de entrada (por ejemplo, muestra de prueba) en función de la concentración del señuelo, y (ii) identificar un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo; y (b) seleccionar un punto de saturación que sea mayor que sustancialmente todos los puntos de saturación asociados con los señuelos en el conjunto de señuelos, determinando de este modo el punto de saturación del conjunto de señuelos. La selección de un punto de saturación puede verse influenciada por la eficiencia de captura de un señuelo y los costos asociados, de tal manera que la concentración en el punto de saturación puede ser lo suficientemente alta para lograr una eficiencia de captura deseada, a la vez que es los suficientemente baja para asegurar costes de reactivos de ensayo

razonables.

La eficacia de captura de un señuelo puede determinarse (a) proporcionando una pluralidad de muestras de ácidos nucleicos obtenidas de una pluralidad de sujetos en una cohorte; (b) hibridando el señuelo con cada una de las muestras de ácidos nucleicos, a cada una de una pluralidad de concentraciones del señuelo; (c) enriqueciendo con el señuelo una pluralidad de regiones genómicas de las muestras de ácidos nucleicos, a cada una de la pluralidad de concentraciones del señuelo; y (d) midiendo el número de moléculas de ácido nucleico únicas o moléculas de ácido nucleico con la representación de ambas cadenas de una molécula de ácido nucleico de doble cadena original que representa la eficacia de captura en cada una de la pluralidad de concentraciones del señuelo. Típicamente, la eficiencia de captura de un señuelo (por ejemplo, el porcentaje de moléculas que contienen la región genómica objetivo del señuelo que se capturan de una muestra que comprende tales moléculas) aumenta rápidamente con la concentración hasta que se alcanza un punto de inflexión, después de lo cual el porcentaje de moléculas capturada aumenta mucho más lentamente.

Un punto de inflexión puede ser una primera concentración de un señuelo de tal manera que la eficiencia de captura observada no aumente significativamente a concentraciones del señuelo mayores que la primera concentración. Un punto de inflexión puede ser una primera concentración del señuelo de tal manera que un aumento observado entre (1) la eficiencia de captura a una concentración de señuelo del doble de la primera concentración en comparación con (2) la eficiencia de captura a la primera concentración de señuelo, es menor de aproximadamente el 1%, menor de aproximadamente el 2%, menor de aproximadamente el 3%, menor de aproximadamente el 4%, menor de aproximadamente el 5%, menor de aproximadamente el 6%, menor de aproximadamente el 7%, menor de aproximadamente el 8%, menor de aproximadamente el 9%, menor de aproximadamente el 10%, menor de aproximadamente el 12%, menor de aproximadamente el 14%, menor de aproximadamente el 16%, menor de aproximadamente el 18%, menor de aproximadamente el 20%, menor de aproximadamente el 30%, menor de aproximadamente el 40%, o menor de aproximadamente el 50%. Dicho punto de inflexión identificado puede considerarse un punto de saturación asociado con un señuelo. Puede usarse un señuelo a una concentración de un punto de saturación en un ensayo para permitir la captura óptima de una región genómica objetivo y, por lo tanto, la sensibilidad de detectar variantes genéticas de la región genómica objetivo. En algunas realizaciones, el punto de saturación asociado con un conjunto de señuelos es el punto de saturación del señuelo más débil en ese conjunto de señuelos. Por ejemplo, el conjunto de señuelos tiene un punto de saturación que es mayor que sustancialmente todos los puntos de saturación asociados con los señuelos en el conjunto de señuelos cuando un señuelo del conjunto de señuelos se somete a una curva de titulación generada (i) midiendo la eficiencia de captura de un señuelo del conjunto de señuelos puesto en función de la concentración del señuelo, y (ii) identificando un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo. Cuando cada señuelo en el conjunto de señuelos está a una primera concentración que está por lo menos en su punto de saturación, el conjunto de señuelos habrá capturado las secuencias objetivo de tal manera que la eficiencia de captura observada de las secuencias objetivo aumente en menos del 20% a una concentración de los señuelos del doble que la de la primera concentración.

La muestra de ácidos nucleicos puede ser una muestra de ácidos nucleicos libres de células (por ejemplo, ADNcf). Un método para enriquecer múltiples regiones genómicas puede comprender además secuenciar la muestra de ácidos nucleicos enriquecida para producir una pluralidad de lecturas de secuencia. Un método para enriquecer múltiples regiones genómicas puede comprender además producir una salida que comprenda una secuencia de ácidos nucleicos representativa de la muestra de ácidos nucleicos. Esta secuencia de ácidos nucleicos puede luego alinearse con un genoma de referencia y analizarse en busca de variantes genéticas relevantes para el cáncer mediante enfoques bioinformáticos.

Una molécula original puede producir lecturas de secuencia redundantes, por ejemplo, después de la amplificación y secuenciación de amplicones, o mediante secuenciación repetida de la misma molécula. Las lecturas de secuencia redundantes de una molécula original pueden colapsarse en una secuencia de consenso (por ejemplo, una "secuencia única") que representa la secuencia de la molécula original. Esto puede hacerse generando una secuencia de consenso para la molécula completa, para parte de la molécula o en una posición de un solo nucleótido en la molécula (nucleótido de consenso). Como se usa en la presente, "polinucleótido secuenciado" se refiere o a lecturas de secuencia generadas a partir de amplicones de una molécula original, o una secuencia de consenso de una molécula original derivada de dichos amplicones. Las lecturas únicas son lecturas que son diferentes de cualquier otra lectura. Las lecturas pueden ser únicas en base a la secuencia de una molécula original, o en base a la secuencia de una molécula original más una o más secuencias de códigos de barras unidas a una molécula original. Por ejemplo, dos moléculas originales idénticas aún pueden producir lecturas únicas si sus códigos de barras son diferentes. De igual manera, dos moléculas originales diferentes producirán lecturas únicas incluso si sus códigos de barras son iguales. Las secuencias de consenso pueden ser secuencias únicas cuando se generan agrupando lecturas únicas.

En un aspecto, un panel de señuelos puede comprender un primer conjunto que captura selectivamente regiones de la estructura principal de un genoma, dichas regiones de la estructura principal asociadas con una función de clasificación de carga de secuenciación y utilidad, en donde la función de clasificación de cada región de

la estructura principal tiene un valor menor que un valor umbral predeterminado; y un segundo conjunto de señuelos que captura selectivamente regiones de punto crítico de un genoma, dichas regiones de punto crítico asociadas con una función de clasificación de carga de secuenciación y utilidad, en donde la función de clasificación de cada región de punto crítico tiene un valor mayor o igual al valor umbral predeterminado. Este enfoque puede usar por lo menos dos conjuntos de señuelos correspondientes a las regiones de la estructura principal y del punto crítico.

Las regiones de punto crítico pueden ser relativamente más importantes que las regiones de la estructura principal para capturar y analizar en una muestra de ácidos nucleicos libre de células dada debido a su utilidad relativamente alta y/o carga de secuenciación relativamente baja. La selección de una región dada como región de punto crítico o región de la estructura principal depende de su valor de función de clasificación, que se calcula como una función de la carga de secuenciación y la utilidad. Puede calcularse un valor de función de clasificación como la utilidad de una región genómica dividida por la carga de secuenciación de una región genómica.

Las regiones de la estructura principal o del punto crítico pueden comprender una o más regiones informativas de nucleosomas. Las regiones informativas de nucleosomas pueden comprender una región de máxima diferenciación de nucleosomas. El panel de señuelo puede comprender además un segundo conjunto de señuelos que captura selectivamente las regiones informativas de la enfermedad. Los señuelos en el primer conjunto de señuelos pueden estar a una primera concentración (por ejemplo, una primera concentración con respecto al panel de señuelos), y los señuelos en el segundo conjunto de señuelos pueden estar a una segunda concentración (por ejemplo, una segunda concentración con respecto al panel de señuelos).

En un aspecto, un método para generar un conjunto de señuelos puede comprender identificar una o más regiones genómicas de la estructura principal de interés, en donde la identificación de la una o más regiones genómicas de la estructura principal puede comprender maximizar una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal; identificar una o más regiones genómicas de punto crítico de interés; crear un primer conjunto de señuelos que capture selectivamente las regiones genómicas de la estructura principal de interés; y crear un segundo conjunto de señuelos que captura selectivamente las regiones genómicas de punto crítico de interés. El segundo conjunto de señuelos puede tener una mayor eficiencia de captura que el primer conjunto de señuelos.

El uno o más puntos críticos pueden seleccionarse usando uno o más de (por ejemplo, uno o más, dos o más, tres o más, o cuatro) de lo siguiente: (i) maximizar una función de clasificación de la carga de secuenciación y utilidad asociadas con cada una de las regiones genómicas de puntos críticos, (ii) elaborar perfiles de nucleosomas a través de una o más regiones genómicas de interés, (iii) mutaciones impulsoras de cáncer predeterminadas o prevalencia en una cohorte de pacientes relevante, y (iv) mutaciones impulsoras de cáncer identificadas empíricamente.

La identificación de uno o más puntos críticos de interés puede comprender usar un procesador informático programado para clasificar un conjunto de regiones genómicas de punto crítico en base a una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de punto crítico. La identificación de una o más regiones genómicas de la estructura principal de interés puede comprender clasificar un conjunto de regiones genómicas de la estructura principal en base a una función de clasificación de la carga de secuenciación y la utilidad asociadas con cada una de las regiones genómicas de la estructura principal de interés. La identificación de una o más regiones genómicas de punto crítico de interés puede comprender utilizar un conjunto de valores de frecuencia de alelos menores (MAF) determinados empíricamente o la clonalidad de una variante medida por su MAF en relación con la mayor mutación impulsora o clonal presunta en una muestra obtenida de uno o más sujetos en una cohorte de interés. Las regiones genómicas que tienen valores de MAF relativamente altos en una cohorte de interés pueden ser puntos críticos adecuados porque pueden indicar evaluaciones relevantes para el cáncer como detección, tipo de células o tejido de origen, carga tumoral, y/o eficacia del tratamiento.

La carga de secuenciación de una región genómica puede calcularse multiplicando entre sí uno o más de (por ejemplo, uno o más, dos o más, tres o más, cuatro o más, o cinco de) (i) el tamaño de la región genómica en pares de bases, (ii) fracción relativa de lecturas gastadas en fragmentos de secuenciación que mapean la región genómica, (iii) cobertura relativa como resultado del sesgo de secuencia de la región genómica, (iv) cobertura relativa como resultado del sesgo de amplificación del región genómica y (v) cobertura relativa como resultado del sesgo de captura de la región genómica. Este indicador puede calcularse para cada región genómica en un conjunto de paneles de señuelo para identificar los "costes" asociados con la generación de lecturas de secuencia asociadas con la región genómica de una muestra de ácidos nucleicos.

La carga de secuenciación de una región genómica es linealmente proporcional al tamaño de la región genómica en pares de bases. La fracción relativa de lecturas gastadas en fragmentos de secuenciación que mapean la región genómica también influye en la carga de secuenciación de la región genómica, ya que algunas regiones genómicas pueden ser especialmente difíciles de secuenciar de manera fiable (por ejemplo, debido al alto contenido de GC o la presencia de secuencias altamente repetitivas) y, por lo tanto, puede requerir una mayor profundidad de

secuenciación para el análisis a la resolución deseada del señuelo. De manera similar, la cobertura relativa como resultado del sesgo de secuencia, el sesgo de amplificación y/o el sesgo de captura de la región genómica también puede afectar la carga de secuenciación de la región genómica. La carga de secuenciación total de un ciclo de secuenciación de un ensayo dado puede entonces calcularse sumando todas las cargas de secuenciación de los señuelos (incluyendo regiones de punto crítico y de la estructura principal) en el conjunto de paneles de señuelos seleccionados del ensayo.

En algunos ejemplos, la utilidad de una región genómica puede calcularse multiplicando entre sí uno o más de (por ejemplo, uno o más, dos o más, tres o más, cuatro o más, cinco o más, seis o más, o siete) de los siguientes factores de utilidad: (i) presencia de una o más mutaciones procesables en la región genómica, (ii) frecuencia de una o más mutaciones procesables en la región genómica, (iii) presencia de una o más mutaciones asociadas con las frecuencias de alelos menores (MAF) por encima de la media en la región genómica, (iv) frecuencia de una o más mutaciones asociadas con MAF por encima de la media en la región genómica, (v) fracción de pacientes en una cohorte que alberga una mutación somática dentro de la región genómica, (vi) suma de MAF para variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, y (vii) proporción de (1) MAF para variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, con (2) MAF máximo para un paciente dado en la cohorte.

El objetivo de calcular la utilidad de una región genómica puede ser ayudar a evaluar su importancia relativa para la inclusión en un panel de conjuntos de señuelos. Por ejemplo, la presencia y/o frecuencia de una o más mutaciones procesables en la región genómica afectan la utilidad de una región genómica para su inclusión en un panel de conjuntos de señuelos, ya que las regiones genómicas que contienen mutaciones altamente frecuentes son buenos marcadores (por ejemplo, Indicadores) de estados patológicos, incluyendo el cáncer. De manera similar, la selección de regiones genómicas con presencia y/o frecuencia de mutaciones asociadas con MAF por encima de la media permitirá una detección altamente sensible de estas mutaciones en un ensayo de biopsia líquida.

La fracción de pacientes en una cohorte que alberga una mutación somática dentro de la región genómica puede indicar mutaciones impulsoras que son adecuadas como marcador para la enfermedad de la cohorte (por ejemplo, mama, colorrectal, pancreático, próstata, melanoma, pulmón o hígado). Para maximizar las posibilidades de detectar el MAF más alto o la variante impulsora, puede usarse la suma de MAF para las variantes en pacientes en una cohorte, dichos pacientes albergando una mutación somática dentro de la región genómica, como factor de utilidad. Para dar el peso máximo a las mutaciones impulsoras, puede usarse la proporción de (1) MAF para variantes en pacientes en una cohorte, dichos pacientes que albergan una mutación somática dentro de la región genómica, con (2) MAF máximo para un paciente dado en la cohorte como factor de utilidad. Las mutaciones asociadas con frecuencias de alelos menores más altas pueden comprender una o más mutaciones impulsoras o son conocidas a partir de datos externos o fuentes de anotación.

Las mutaciones procesables pueden comprender mutaciones cuya presencia detectada puede influir o determinar decisiones clínicas (por ejemplo, diagnóstico, monitorización del cáncer, monitorización de la terapia, evaluación de la eficacia de la terapia). Las mutaciones procesables pueden comprender una o más de (por ejemplo, una o más, dos o más, tres o más, cuatro o más, cinco o más, seis o más, o siete) (i) mutaciones modificables por fármacos, (ii) mutaciones para monitorización terapéutica, (iii) mutaciones específicas de la enfermedad, (iv) mutaciones específicas del tejido, (v) mutaciones específicas del tipo celular, (vi) mutaciones de resistencia y (vii) mutaciones de diagnóstico.

Las mutaciones modificables con fármacos pueden incluir aquellas mutaciones cuya presencia detectada en una muestra de ácidos nucleicos de un sujeto puede indicar que el sujeto es un candidato apropiado para el tratamiento con un determinado fármaco asociado con la mutación (por ejemplo, la detección de la mutación EGFR L858R puede indicar la necesidad de tratar con un tratamiento de inhibidores de la tirosina quinasa (TKI)). Las mutaciones para la monitorización terapéutica incluyen aquellas mutaciones cuya presencia detectada o nivel aumentado en una muestra de ácidos nucleicos de un sujeto puede indicar que el cáncer del sujeto está respondiendo a un curso de tratamiento. Las mutaciones de resistencia incluyen aquellas mutaciones cuya presencia detectada o nivel aumentado en una muestra de ácidos nucleicos de un sujeto puede indicar que el cáncer del sujeto se ha vuelto resistente a un curso de tratamiento (por ejemplo, la aparición de la mutación EGFR T790M puede indicar la aparición de resistencia). Las mutaciones pueden ser específicas de una enfermedad (por ejemplo, tipo de tumor), tipo de tejido o tipo de célula, cuya detección puede indicar cáncer, inflamación u otro estado patológico en un órgano, tejido o tipo de célula particular.

En la Tabla 3 y la Tabla 4 pueden encontrarse listados ejemplares de localizaciones genómicas de interés. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, por lo menos 75, por lo menos 80, por lo menos 85, por lo menos 90, por lo menos 95, o 97 de los genes de la Tabla 3. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 5, por lo menos 10, por lo menos 15,

por lo menos por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65 o 70 de los SNV de la Tabla 3. En algunos realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos por lo menos 10, por lo menos 11, en l este 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17 o 18 de las CNV de la Tabla 3. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5 o 6 de las fusiones de la Tabla 3. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos una porción de por lo menos 1, por lo menos 2 o 3 de los indeles de la Tabla 3. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos una porción de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, por lo menos 75, por lo menos 80, por lo menos 85, por lo menos 90, por lo menos 95, por lo menos 100, por lo menos 105, por lo menos 110 o 115 de los genes de la Tabla 4. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, o 73 de los SNV de la Tabla 4. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17 o 18 de las CNV de la Tabla 4. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5 o 6 de las fusiones de la Tabla 4. En algunas realizaciones, Las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, a por lo menos 15, por lo menos 16, por lo menos 17 o 18 de los indeles de la Tabla 4. Cada una de estas localizaciones genómicas de interés puede identificarse como una región de la estructura principal o una región de punto crítico para un panel de conjuntos de señuelos dado. En la Tabla 5 puede encontrarse una lista ejemplar de las localizaciones genómicas de interés de los puntos críticos. En algunas realizaciones, las regiones genómicas usadas en los métodos de la presente divulgación comprenden por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19 o por lo menos 20 de los genes de la Tabla 5. Cada región genómica de punto crítico se enumera con varias características, incluyendo el gen asociado, el cromosoma en el que reside, la posición de inicio y de parada 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, tipo de mutación) que una región genómica de interés dada puede buscar capturar.

Tabla 3

Mutaciones de Punto (SNVs)						Amplificaciones (CNVs)		Fusiones	Indeles
AKT1	ALK	APC	AR	ARAF	ARID1A	AR	BRAF	ALK	
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	CCND1	CCND2	FGFR2	EGFR (exones 19 & 20)
CCNE1	CDH1	CDK4	CDK6	CDKN2A	CDKN2B	CCNE1	CDK4	FGFR3	
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	CDK6	EGFR	NTRK1	
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	ERBB2	FGFR1	RET	ERBB2 (exones 19 & 20)
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	FGFR2	KIT	ROS1	
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	KRAS	MET		
MLH 1	MPL	MYC	NF1	NFE2L2	NOTCH 1	MYC	PDGFRA		MET (falta exón 14)
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	PIK3CA	RAF1		
PTPN11	RAF1	RB1	RET	RHEB	RHOA				
RIT1	ROS1	SMAD4	SMO	SRC	STK11				
TERT	TP53	TSC1	VHL						

ES 2 840 003 T3

Tabla 4

Mutaciones de Punto (SNVs)						Amplificaciones (CN- Vs)		Fusiones	Indeles
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	CDKN2 A	DDR2	CCNE1	CDK4	FGFR3	
CTNNB 1	EGFR	ERBB2	ESR1	EZH2	FBXW7	CDK6	EGFR	NTRK1	ERBB2 (exones 19 & 20)
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	ERBB2	FGFR1	RET	
GNAS	HNF1 A	HRAS	IDH1	IDH2	JAK2	FGFR2	KIT	ROS1	
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	KRAS	MET		MET (falta exón 14)
MLH 1	MPL	MYC	NF1	NFE2L2	NOTCH 1	MYC	PDGFR A		
NPM1	NRAS	NTRK1	PDGFR A	PIK3CA	PTEN	PIK3C A	RAF1		
PTPN11	RAF1	RB1	RET	RHEB	RHOA				ATM
RIT1	ROS1	SMAD 4	SMO	MAPK1	STK11				
TERT	TP53	TSC1	VHL	MAPK3	MTOR				
NTRK3									APC
									ARID1A
									BRCA1
									BRCA2
									CDH1
									CDKN2 A
									GATA3
									KIT
									MLH 1
									MTOR
									NF1
									PDGFRA
									PTEN
									RB1
									SMAD4
									STK11
									TP53
									TSC1
									VHL

Tabla 5

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones cubiertos	Característica crítica
ALK	chr2	29446405	29446655	250	intrón 19	Fusión
ALK	chr2	29446062	29446197	135	intrón 20	Fusión

(continuación)

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones cubiertos	Característica crítica
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 / Indeles
GATA3	chr10	8115692	8116002	310	6	SS / Indeles
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	
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

(continuación)

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones cubiertos	Característica crítica
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
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 objetivo total)		
				16330 (cobertura de sonda total)		

En un aspecto, un panel de señuelos puede comprender una pluralidad de conjuntos de señuelos, cada conjunto de señuelos (i) comprendiendo uno o más señuelos que capturan selectivamente una o más regiones genómicas con utilidad en el mismo cuantil a través de la pluralidad de señuelos, y (ii) teniendo una concentración relativa diferente de cada uno de los otros conjuntos de señuelos con utilidad en un cuantil diferente en la pluralidad de señuelos. Los cuantiles pueden ser, por ejemplo, dos mitades, tres tercios, cuatro cuartos, etc. Por ejemplo, un panel de señuelos puede comprender tres conjuntos de señuelos, cada conjunto de señuelos comprendiendo señuelos que capturan selectivamente regiones genómicas con utilidad en el tercio superior, tercio medio, o tercio inferior de los valores de utilidad en la pluralidad de señuelos, cada uno de los tres conjuntos de señuelos teniendo una concentración relativa diferente.

Un panel de señuelos puede comprender una pluralidad de conjuntos de señuelos, cada conjunto de señuelos (i) comprendiendo uno o más señuelos que capturan selectivamente una o más regiones genómicas con carga de secuenciación en el mismo cuantil en la pluralidad de señuelos, y (ii) teniendo una concentración relativa diferente de cada uno de los otros conjuntos de señuelos con carga de secuenciación en un cuantil diferente a través de la pluralidad de señuelos. Un panel de señuelos puede comprender una pluralidad de conjuntos de señuelos, cada conjunto de señuelos (i) comprendiendo uno o más señuelos que capturan selectivamente una o más regiones

genómicas con un valor de función de clasificación (por ejemplo, utilidad dividida por carga de secuenciación) en el mismo cuantil en la pluralidad de señuelos, y (ii) teniendo una concentración relativa diferente de cada uno de los otros conjuntos de señuelos con un valor de función de clasificación en un cuantil diferente en la pluralidad de señuelos.

En un aspecto, un método para seleccionar un conjunto de bloques de panel puede comprender (a) para cada bloque de panel, (i) calcular una utilidad del bloque de panel, (ii) calcular una carga secuencial del bloque de panel, y (iii) calcular una función de clasificación del bloque de panel; y (b) realizar un proceso de optimización para seleccionar un conjunto de bloques de panel que maximiza los valores de la función de clasificación total de los bloques de panel seleccionados. Una función de clasificación de un bloque de panel puede calcularse como la utilidad de un bloque de panel dividida por la carga de secuenciación de un bloque de panel. El proceso de optimización combinatoria puede optimizar la suma total de los valores de la función de clasificación de todos los bloques de panel seleccionados para el conjunto de bloques de panel en un solo ensayo. Este enfoque puede permitir una selección de panel óptima dadas las restricciones en la carga de secuencia y la utilidad. El proceso de optimización combinatoria puede ser un algoritmo voraz. En un aspecto, un método puede comprender (a) proporcionar una pluralidad de mezclas de señuelos, en donde cada una de la pluralidad de mezclas de señuelos comprende un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas y un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas, en donde el primer conjunto de señuelos está a diferentes concentraciones en la pluralidad de mezclas de señuelos y el segundo conjunto de señuelos está a la misma concentración en la pluralidad de mezclas de señuelos; (b) poner en contacto cada una de la pluralidad de mezclas de señuelos con una muestra de ácidos nucleicos para capturar ácidos nucleicos de la muestra de ácidos nucleicos con el primer conjunto de señuelos y el segundo conjunto de señuelo, en donde los ácidos nucleicos de las muestras de ácidos nucleicos son capturados por el primero conjunto de señuelos y el segundo conjunto de señuelo; (c) secuenciar una parte de los ácidos nucleicos capturados con cada mezcla de señuelos para producir conjuntos de lecturas de secuencia dentro de un número asignado de lecturas de secuencia; (d) determinar la profundidad de lectura para el primer conjunto de señuelos y el segundo conjunto de señuelos para cada mezcla de señuelos; y (e) identificar por lo menos una mezcla de señuelos que proporcione profundidades de lectura para el segundo conjunto de regiones genómicas y, opcionalmente, el primer conjunto de regiones genómicas, en cantidades predeterminadas. En algunas realizaciones, las profundidades de lectura para el segundo conjunto de regiones genómicas proporcionan una sensibilidad de detección de una variante genética de por lo menos un 0,0001% de MAF. En algunas realizaciones, un primer conjunto de regiones genómicas y/o un segundo conjunto de regiones tienen un tamaño entre 25 kilobases a 1.000 kilobases. En algunas realizaciones, un primer conjunto de regiones genómicas y/o un segundo conjunto de regiones tienen una profundidad de lectura de entre 1.000 recuentos/base y 50.000 recuentos/base.

Precisión mejorada de detección de indeles

Se divulga un método para mejorar la precisión de la detección de una inserción o deleción (indel) de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto, dicha pluralidad de lecturas de secuencia se generan por secuenciación de ácidos nucleicos. Para cada una de la pluralidad de lecturas de secuencia asociadas con moléculas de ADNcf, puede identificarse un indel candidato. Cada indel candidato puede clasificarse como un indel verdadero o un indel introducido, usando una combinación de esperanzas predeterminadas de (i) que se detecte un indel en una o más lecturas de secuencia de la pluralidad de lecturas de secuencia, (ii) que un indel detectado sea un indel verdadero presente en una determinada molécula de ADN libre de células de las moléculas de ADN libre de células, dado que se ha detectado un indel en una o más de las lecturas de la secuencia, y/o (iii) que un indel detectado se haya introducido por error no biológico, dado que se ha detectado un indel en una o más de las lecturas de secuencia, junto con uno o más parámetros del modelo para realizar una prueba de hipótesis. Este enfoque puede reducir el error y mejorar la precisión de la detección de un indel a partir de los datos de lecturas de secuencia.

La **FIG. 1** ilustra cómo pueden generarse una pluralidad de lecturas para cada locus enriquecido a partir de una muestra de ácidos nucleicos libres de células. Cada molécula de ácido nucleico enriquecida (por ejemplo, molécula de ADN) se amplifica para producir una familia de amplicones. Luego, estos amplicones pueden secuenciarse tanto en cadenas directas como inversas para producir una pluralidad de datos de lecturas de secuencia. A partir de la pluralidad de datos de lecturas de secuencia, pueden detectarse y clasificarse los indeles candidatos como indeles verdaderos o indeles introducidos (por ejemplo, no biológicos).

Este algoritmo supone que para cualquier molécula de ADN dada para la cual se analiza una pluralidad de lecturas de secuencia para variantes que comprenden indeles, hay una esperanza predeterminada (por ejemplo, probabilidad) de que un indel esté presente o en la molécula original (por ejemplo, un indel biológico "verdadero") o introducido en algún punto de un protocolo que culmina un conjunto de lecturas de secuencia (por ejemplo, un indel no biológico introducido derivado de un error, incluyendo la amplificación o el error de secuenciación). El modelo puede tener como objetivo realizar una prueba de hipótesis que pregunte, dado un patrón de mapeo de lecturas para una posición de bases particular (por ejemplo, cubrir la posición de bases en algún lugar de la lectura), si el patrón observado es más indicativo de un indel en una secuencia presente al comienzo del protocolo (por ejemplo, un indel

biológico verdadero) o introducido durante el protocolo (un indel no biológico).

En un aspecto, un método para mejorar la precisión de la detección de una inserción o deleción (indel) de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto, dicha pluralidad de lecturas de secuencia se generan mediante secuenciación de ácidos nucleicos, puede comprender (a) para cada una de la pluralidad de lecturas de secuencia asociadas con las moléculas de ADN libre de células, proporcionar: una esperanza predeterminada de que se detecte un indel en una o más lecturas de secuencia de la pluralidad de lecturas de secuencia; una esperanza predeterminada de que un indel detectado sea un indel verdadero presente en una determinada molécula de ADN libre de células de las moléculas de ADN libre de células, dado que se ha detectado un indel en una o más lecturas de la secuencia; y una esperanza predeterminada de que un indel detectado se haya introducido por error no biológico, dado que se ha detectado un indel en una o más de las lecturas de la secuencia; (b) proporcionar medidas cuantitativas de uno o más parámetros del modelo característicos de las lecturas de secuencia generadas por secuenciación de ácidos nucleicos; (c) detectar uno o más indeles candidatos en la pluralidad de lecturas de secuencia asociadas con las moléculas de ADN libre de células; y (d) para cada indel candidato, realizar una prueba de hipótesis usando uno o más de los parámetros del modelo para clasificar dicho indel candidato como un indel verdadero o un indel introducido, mejorando de este modo la precisión de detección de un indel.

El método para mejorar la precisión de detección de una inserción o deleción (indel) a partir de una pluralidad de lecturas de secuencia derivadas de moléculas de ácido desoxirribonucleico libre de células (ADNcf) en una muestra corporal de un sujeto puede comprender además enriquecer uno o más loci del ADN libre de células en la muestra corporal antes del paso (a), produciendo de este modo polinucleótidos enriquecidos.

El método puede comprender además amplificar los polinucleótidos enriquecidos para producir familias de amplicones, en donde cada familia comprende amplicones que se originan a partir de una sola cadena de moléculas de ADN libre de células. El error no biológico puede comprender un error en la secuenciación en una pluralidad de localizaciones de bases genómicas. El error no biológico puede comprender un error en la amplificación en una pluralidad de localizaciones de bases genómicas.

La **FIG. 2** ilustra un ejemplo de familias pequeñas de lecturas (que pueden parecer proporcionar evidencia de una variante de indel verdadero) y familias grandes de lecturas (que pueden indicar un error introducido probablemente derivado de PCR o secuenciación. En general, puede esperarse que los indeles verdaderos se detecten o midan como familias pequeñas de lecturas, ya que no se espera que afecten biológicamente a un gran número de moléculas de ADN. Por el contrario, puede esperarse que los indeles introducidos se detecten o midan como familias más grandes de lecturas, lo que puede indicar un error introducido durante la PCR o la secuenciación Algunas lecturas no recortadas o erróneas pueden hacer que el algoritmo descalifique a la familia basándose en una prueba de hipótesis que clasifica un indel (por ejemplo, inserción o deleción) como introducido en lugar de biológico.

La **FIG. 3** ilustra un ejemplo de una inserción soportada por una gran familia tras alinear y comparar una pluralidad de lecturas de secuencia con un genoma de referencia. Como en el caso anterior en la FIG. 3, algunas lecturas no recortadas o erróneas pueden hacer que el algoritmo descalifique a la familia basándose en una prueba de hipótesis que clasifica un indel (por ejemplo, inserción o eliminación) como introducido en lugar de biológico.

Los parámetros del modelo pueden comprender uno o más (por ejemplo, uno o más, dos o más, tres o más, o cuatro) de (i) para cada uno del uno o más alelos variantes, una frecuencia del alelo variante (α) y una frecuencia de alelos no de referencia distintos del alelo variante (α'); (ii) una frecuencia de un error de indel en toda la cadena directa de una familia de cadenas (β_1), en donde una familia comprende una colección de amplicones que se originan a partir de una única cadena de moléculas de ADN libre de células; (iii) una frecuencia de un error de indel en toda la cadena inversa de una familia de cadenas (β_2); y (iv) una frecuencia de error de indel en una secuencia leída (γ).

La **FIG. 4** ilustra los varios parámetros que pueden usarse en una prueba de hipótesis y cómo cada parámetro puede estar relacionado con una probabilidad particular, por ejemplo, de una familia de lecturas que coinciden con una referencia, de unas lecturas de cadenas que coinciden con una referencia, y de una lectura que coincide con una referencia. La Fig. 2 también ilustra cómo puede realizarse una prueba de parámetros que contenga una función de máxima probabilidad. Si la prueba de parámetros es mayor que un umbral predeterminado cuando se realiza en un indel candidato, entonces el candidato puede clasificarse como un indel verdadero. Si la prueba de parámetro es menor o igual que un umbral predeterminado cuando se realiza en un indel candidato, entonces el candidato puede clasificarse como un indel introducido (por ejemplo, no biológico).

El paso de realizar una prueba de hipótesis puede comprender realizar un algoritmo de maximización de múltiples parámetros. El algoritmo de maximización de múltiples parámetros puede comprender un algoritmo de Nelder-Mead. La clasificación de un indel candidato como un indel verdadero o un indel introducido puede comprender (a) maximizar una función de probabilidad de múltiples parámetros, (b) clasificar un indel candidato como un indel verdadero si el valor de la función de máxima probabilidad es mayor que un umbral predeterminado

valor, y (c) clasificar un indel candidato como indel introducido si el valor de la función de máxima probabilidad es menor o igual que un valor de umbral predeterminado. La función de probabilidad de múltiples parámetros puede proporcionarse como:

$$\Pr\{\text{Lecturas} \mid \alpha, \alpha', \beta_1, \beta_2, \gamma\} = \prod_{Familias} \left(\alpha \cdot \left((1-\beta_1)(1-\gamma)^{R_1} \gamma^{V_1 + O_1} + \beta_1 \gamma^{R_1} (1-\gamma)^{V_1 + O_1} \right) \cdot \left((1-\beta_2)(1-\gamma)^{R_2} \gamma^{V_2 + O_2} + \beta_2 \gamma^{R_2} (1-\gamma)^{V_2 + O_2} \right) + \alpha' \cdot (\ldots) + (1-\alpha-\alpha') \cdot (\ldots) \right)$$

Una función de probabilidad de múltiples parámetros Pr{Lecturas | α, α', β_1, β_2, γ} pueden representar una probabilidad de una configuración observada de lecturas de acuerdo con el modelo ilustrado en la **Fig. 4** (y descrito en el párrafo [00112]). Una suposición del modelo puede ser que, dados ciertos valores de parámetros (por ejemplo, α, α', β_1, β_2 y γ), una configuración observada de lecturas dentro de una familia es estadísticamente independiente de una configuración observada de lecturas dentro de todos las otras familias. Por lo tanto, la probabilidad Pr{Lecturas | α, α', β_1, β_2, γ} pueden expresarse como un producto de Pr{lecturas en la familia f | α, α', β_1, β_2, γ} sobre todas las familias. Esta propia probabilidad por familia puede comprender una suma ponderada de por lo menos tres componentes, en donde cada componente corresponde a un posible tipo de familia: a) que tiene el alelo variante (con peso α), b) que tiene otro alelo variante no de referencia (con peso α', o c) que tiene el alelo de referencia (con peso 1-α-α'). Estos componentes sumados pueden ser probabilidades de configuración de lectura observada para el el tipo de familia respectivo Pr{lecturas en la familia f | α, α', β_1, β_2, γ, y la familia f teniendo el alelo variante}, Pr{lecturas en la familia f | α, α', β_1, β_2, γ, y la familia f teniendo otro alelo variante no de referencia}, y Pr{lecturas en la familia f | α, α', β_1, β_2, γ, y la familia f teniendo el alelo de referencia}.

Como el modelo postula que dentro de una familia cada cadena puede verse afectada por un error de indel independientemente de la otra cadena, la probabilidad de configuración de lectura observada para una familia que tiene el alelo variante Pr {lecturas en la familia f | α, α', β_1, β_2, γ, y la familia f teniendo el alelo variante} puede ser en sí misma un producto de la probabilidad de configuración observada de lecturas de la cadena directa y la probabilidad de configuración observada de lecturas de la cadena inversa. Cada una de estas probabilidades puede ser en sí misma una suma ponderada de por lo menos dos componentes, en donde cada componente corresponde a un resultado posible: X) el error de indel específico de la cadena afectó a esta cadena de la familia (con peso β_1 o β_2) e Y) el error de indel específico de la cadena no afectó a esta cadena de la familia (con peso 1-β_1 o 1-β_2).

Finalmente, dentro de una familia de tipo supuesto a), b) o c), y/o dentro de una cadena de tipo supuesto X) o Y), la probabilidad de una configuración de lectura específica puede ser un producto de probabilidades para lecturas individuales, ya que el modelo postula que estas lecturas tienen una probabilidad estadísticamente independiente de caer en una de las tres categorías: i) la lectura apoya el alelo variante, ii) la lectura apoya otro alelo variante no de referencia, o iii) la lectura apoya el alelo de referencia. Estas probabilidades se enumeran en la Tabla 6 a continuación.

Tabla 6

Familia	Error de cadena	i) lectura apoya variante	ii) lectura apoya otro	iii) lectura apoya referencia
a) alelo variante	presente	γ	1-γ	1-γ
	ausente	1-γ	γ	γ
b) otro alelo variante	presente	1-γ	γ	1-γ
	ausente	γ	1-γ	γ
c) alelo de referencia	presente	1-γ	1-γ	γ
	ausente	γ	γ	1-γ

Aunque en la presente se han mostrado y descrito realizaciones preferidas de la presente invención, resultará obvio para los expertos en la técnica que tales realizaciones se proporcionan a modo de ejemplo solamente. No se pretende que la invención esté limitada por los ejemplos específicos proporcionados dentro de la especificación. Aunque la invención se ha descrito con referencia a la especificación mencionada anteriormente, no se pretende que las descripciones e ilustraciones de las realizaciones en la presente se interpreten en un sentido limitativo.

Sistemas de control informático

La presente divulgación proporciona sistemas de control informático que están programados para

implementar métodos de la divulgación. En un aspecto, la presente divulgación proporciona un sistema que comprende un ordenador que comprende un procesador y una memoria de ordenador, en donde el ordenador está en comunicación con una red de comunicaciones, y en donde la memoria del ordenador comprende un código que, cuando es ejecutado por el procesador, (1) recibe los datos de secuencia en la memoria del ordenador desde la red de comunicaciones; (2) determina si una variante genética en los datos de secuencia representa un mutante; y (3) informa, a través de la red de comunicaciones, la determinación.

Una 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 incluye, sin limitación, banda ancha sobre líneas eléctricas (BPL), cable módem, línea de abonado digital (DSL), fibra, satélite e inalámbrica.

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

La **FIG. 5** muestra un sistema informático 501 que está programado o configurado de otro modo para implementar métodos para generar un conjunto de señuelos, para seleccionar un conjunto de bloques de panel, y para mejorar la precisión de la detección de un indel de una pluralidad de lecturas de secuencia derivadas de moléculas ADNcf. El sistema informático 501 puede regular varios aspectos de la presente divulgación, tales como, por ejemplo, métodos para generar un conjunto de señuelos, para seleccionar un conjunto de bloques de panel, o para mejorar la precisión de la detección de un indel a partir de una pluralidad de lecturas de secuencia derivadas de moléculas ADNcf. El sistema informático 501 puede ser un dispositivo electrónico de un usuario o un sistema informático que está localizado remoto con respecto al dispositivo electrónico. El dispositivo electrónico puede ser un dispositivo electrónico móvil.

El sistema informático 501 incluye una unidad de procesamiento central (CPU, también "procesador" y "procesador informático" en la presente) 505, que puede ser un procesador de un solo núcleo o de múltiples núcleos, o una pluralidad de procesadores para procesamiento paralelo. El sistema informático 501 también incluye memoria o localización de memoria 510 (por ejemplo, memoria de acceso aleatorio, memoria de solo lectura, memoria flash), unidad de almacenamiento electrónico 515 (por ejemplo, disco duro), interfaz de comunicación 520 (por ejemplo, adaptador de red) para comunicarse con uno o más de otros sistemas, y dispositivos periféricos 525 como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónicos. La memoria 510, la unidad de almacenamiento 515, la interfaz 520 y los dispositivos periféricos 525 están en comunicación con la CPU 505 a través de un bus de comunicación (líneas continuas), como una placa base. La unidad de almacenamiento 515 puede ser una unidad de almacenamiento de datos (o depósito de datos) para almacenar datos. El sistema informático 501 puede acoplarse operativamente a una red informática ("red") 530 con la ayuda de la interfaz de comunicación 520. La red 530 puede ser Internet, una internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. La red 530 en algunos casos es una red de telecomunicaciones y/o datos. La red 530 puede incluir uno o más servidores informáticos, que pueden permitir la informática distribuida, como la informática en la nube. La red 530, en algunos casos con la ayuda del sistema informático 501, puede implementar una red entre pares, que puede permitir que los dispositivos acoplados al sistema informático 501 se comporten como un cliente o un servidor.

La CPU 505 puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden incorporarse en un programa o software. Las instrucciones pueden almacenarse en una localización de memoria, como la memoria 510. Las instrucciones pueden dirigirse a la CPU 505, que posteriormente puede programar o configurar de otro modo la CPU 505 para implementar los métodos de la presente divulgación. Los ejemplos de operaciones realizadas por la CPU 505 pueden incluir buscar, decodificar, ejecutar y reescribir.

La CPU 505 puede ser parte de un circuito, como un circuito integrado. En el circuito pueden incluirse uno o más de otros componentes del sistema 501. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC).

La unidad de almacenamiento 515 puede almacenar archivos como controladores, bibliotecas y programas guardados. La unidad de almacenamiento 515 puede almacenar datos de usuario, por ejemplo, preferencias de usuario y programas de usuario. El sistema informático 501 en algunos casos puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático 501, como localizadas en un servidor remoto que está en comunicación con el sistema informático 501 a través de una intranet o Internet.

El sistema informático 501 puede comunicarse con uno o más sistemas informáticos remotos a través de la

red 530. Por ejemplo, el sistema informático 501 puede comunicarse con un sistema informático remoto de un usuario. Ejemplos de sistemas informáticos remotos incluyen ordenadores personales (por ejemplo, PC portátil), pizarras o tabletas PC (por ejemplo, Apple® iPad, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, Apple® iPhone, dispositivo con Android, Blackberry®) o asistentes digitales personales. El usuario puede acceder al sistema informático 501 a través de la red 530.

Los métodos descritos en la presente pueden implementarse mediante un código ejecutable por máquina (por ejemplo, procesador de ordenador) almacenado en una localización de almacenamiento electrónico del sistema informático 501 como, por ejemplo, en la memoria 510 o la unidad de almacenamiento electrónico 515. El código ejecutable o legible por máquina puede proporcionarse en forma de software. Durante el uso, el código puede ser ejecutado por el procesador 505. En algunos casos, el código puede ser recuperado de la unidad de almacenamiento 515 y almacenarse en la memoria 510 para que el procesador 505 tenga fácil acceso. En algunas situaciones, la unidad de almacenamiento electrónico 515 puede excluirse, y las instrucciones ejecutables por máquina se almacenan en la memoria 510.

El código puede precompilarse y configurarse para su uso con una máquina que tenga un procesador adaptado para ejecutar el código, o puede compilarse durante el tiempo de ejecución. El código puede proporcionarse en un lenguaje de programación que puede seleccionarse para permitir que el código se ejecute de una manera precompilada o compilada.

Los aspectos de los sistemas y métodos proporcionados en la presente, como el sistema informático 501, pueden incorporarse en programación. Varios aspectos de la tecnología pueden considerarse como "productos" o "artículos de fabricación" típicamente 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 solo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir parte o toda la memoria tangible de los ordenadores, procesadores o similares, o módulos asociados de los mismos, como varias memorias de semiconductores, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación del software. En ocasiones, todo o parte del software pueden comunicarse a través de Internet o de varias otras redes de telecomunicaciones. Tales 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 principal a la plataforma informática de un servidor de aplicaciones. Por tanto, otro tipo de medio que puede contener los elementos de software incluye ondas ópticas, eléctricas y electromagnéticas, como las que se usan en interfaces físicas entre dispositivos locales, a través de redes terrestres de cable y ópticas y a través de varios enlaces aéreos. Los elementos físicos que transportan tales ondas, como enlaces por cable o inalámbricos, enlaces ópticos o similares, también pueden considerarse medios que llevan el software. Como se usa en la presente, a menos que se limite a medios de "almacenamiento" no transitorios, tangibles, términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participa en proporcionar 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, entre otras, 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 en cualquier ordenador o similares, como los que pueden usarse para implementar las bases de datos, etc. mostrados 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; cable de cobre y fibra óptica, incluyendo 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 de luz como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarroja (IR). Por lo tanto, 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, cinta de papel para tarjetas perforadas, 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 desde el cual un ordenador pueda leer código de programación y/o datos. Muchas de estas formas de medios legibles por ordenador pueden estar implicadas en llevar una o más secuencias de una o más instrucciones a un procesador para su ejecución.

El sistema informático 501 puede incluir o estar en comunicación con una pantalla electrónica 535 que comprende una interfaz de usuario (UI) 540 para proporcionar, por ejemplo, parámetros de entrada para métodos para generar un conjunto de señuelos, para seleccionar un conjunto de bloques de panel., o para mejorar la precisión de la detección de un indel a partir de una pluralidad de lecturas de secuencia derivadas de ADNcf. Los ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en red.

Los métodos y sistemas de la presente divulgación pueden implementarse mediante uno o más algoritmos. Un algoritmo puede implementarse por medio de software tras la ejecución por la unidad central de procesamiento 505. El algoritmo puede, por ejemplo, generar un conjunto de señuelos, seleccionar un conjunto de bloques de panel o mejorar la precisión de la detección de una indel de una pluralidad de lecturas de secuencia derivadas de moléculas de ADNcf.

EJEMPLOS

Ejemplo 1: Evaluación del rendimiento analítico

La sensibilidad analítica (definida por el límite de detección y por el porcentaje de concordancia positiva) y la precisión se evaluaron a lo largo de la fracción alélica notificable y los intervalos de número de copias mediante múltiples estudios de dilución en serie de material artificial y muestras de pacientes caracterizados ortogonalmente. La especificidad analítica se evaluó calculando la tasa de falsos positivos en mezclas de muestras de donantes sanos precaracterizadas diluidas en serie en el intervalo inferior notificable hasta las fracciones alélicas por debajo del límite de detección. El valor predictivo positivo (PPV) se estimó como una función de la fracción alélica/número de copias de muestras de pacientes clínicos precaracterizados y se ajustó a la prevalencia usando una cohorte de 2585 muestras clínicas consecutivas. La confirmación ortogonal cualitativa y cuantitativa se realizó mediante ddPCR.

El rendimiento analítico se resume en la Tabla 7 a continuación. La especificidad analítica fue del 100% para variantes de un solo nucleótido (SNV), fusiones y alteraciones del número de copias (CNA) y del 96% (24/25) para indeles en 25 muestras definidas. En relación con otros métodos, este ensayo demostró aumentos del 20% al 50% en la recuperación de la molécula de fusión, dependiendo del contexto de la secuencia. El análisis retrospectivo in silico de 2585 muestras clínicas consecutivas demostró un aumento relativo de >15% en la detección de fusión procesable, un aumento del 6%-15% en la detección de indel procesable (excluyendo los indeles notificables recientemente) y un aumento del 3%-6% en la detección de SNV procesable.

Tabla 7

Alteraciones	Intervalo informable	95% de límite de detención	Fracción alélica / Número de copias	Sensibilidad analítica	Fracción alélica / Número de copias	PPV
SNVs	≥0.04%	0.25%	≥0.25%	>99.9%	≥0.25%	98.7%
			0.05-0.25%	63.8%	<0.25%	92.3%
Indeles	≥0.02%	0.2%	≥0.25%	>99.9%	≥0.25%	98.4%
			0.05-0.25%	67.8%	<0.25%	88.5%
Fusiones	≥0.04%	0.4%	≥0.3%	100%	cualquiera	100%
			<0.3%	83.0%		
CNAs	≥2.12 copias	2.24-2.93 copias	2.3 copias	95.0%	cualquiera	100%

Tabla 7: Características de rendimiento analítico basadas en la entrada de ADNcf estándar (30 ng). Se proporcionan estimaciones de sensibilidad analítica/límite de detección para variantes clínicamente procesables y pueden variar según el contexto de secuencia y la entrada de ADNcf. El valor predictivo positivo se estima en todo el espacio del panel notificable (el VPP fue del 100% para las variantes clínicamente procesables).

En resumen, el ensayo detectó de forma exhaustiva todas las variantes genómicas somáticas recomendadas por las guías de tumores sólidos en adultos con alta sensibilidad, precisión y especificidad.

Ejemplo 2: Titulación de punto crítico y de estructura principal

En este experimento, se determinaron la replicación de sondas apropiada y el punto de saturación para cada panel. Los paneles de punto crítico y estructura principal se diseñaron tanto para la replicación de sonda predeterminada como para la replicación de sonda optimizada. El panel de puntos críticos es de aproximadamente 12 kb y se dirige a regiones de objetivos genómicos que pueden ser indicativas de la respuesta al fármaco, un estado patológico (por ejemplo, cáncer) y/o un objetivo genómico incluido en la National Comprehensive Cancer Network ("NCCN"). El panel de la estructura principal es de aproximadamente 140 kb y cubre el resto del contenido del panel. El panel del punto crítico y de la estructura principal puede comprender cualquier localización genética en la Tabla 3. Se realizó un experimento de titulación para la cantidad de entrada del panel para cada uno de los cuatro paneles a 5 ng, 15 ng y 30 ng de ADNcf como se expone en la Tabla 1. La **FIG. 6** muestra la cantidad de entrada

frente al recuento de moléculas únicas para el panel genérico. El recuento de moléculas únicas saturadas a aproximadamente Vol. 3X para el señuelo de la estructura principal y aproximadamente Vol. 1.2X para el señuelo del punto crítico (datos no mostrados), sugiere que el panel de estructura principal optimizado era menos variable en comparación con el panel predeterminado.

Ejemplo 3: Captura selectiva de una región de punto crítico

En base al punto de saturación de cada panel en el Ejemplo 2, se determinó una concentración de señuelo de la estructura principal y una concentración de señuelo de punto crítico. Se generó Una mezcla de señuelo de estructura principal (por ejemplo, Vol. A) y señuelo de punto crítico (por ejemplo, Vol. B) y el recuento de moléculas para la mezcla de señuelos de punto crítico/estructura principal se comparó con el recuento de moléculas para un panel genérico. Los recuentos de moléculas del panel de punto crítico fueron más altos que los del panel de la estructura principal. La diferencia se hizo más notoria con una mayor cantidad de entrada de ADNcf, ya que el señuelo de la estructura principal se saturó mucho más rápido, por ejemplo, con una menor cantidad de entrada, en comparación con el señuelo del punto crítico. Se observó una tendencia similar con el recuento de cadena doble (datos no mostrados). El tamaño de la familia también fue más alto para el panel de punto crítico que para el panel de estructura principal (datos no mostrados). La diferencia en los tamaño de las familias puede indicar que el panel de punto crítico captura más que el panel de la estructura principal, a pesar de que el efecto estaba enmascarado con recuentos de moléculas. Por ejemplo, con tamaños de familias más grandes para 5 ng, es probable que se capturasen la mayoría de las moléculas únicas, por lo que no hubo una diferencia obvia entre el panel de punto crítico y el de la estructura principal. Con las diferencias de tamaños de las familias, es probable que el panel de punto crítico capturase más duplicados de PCR que el panel de la estructura principal.

En resumen, este experimento demuestra que las regiones de punto crítico pueden capturarse selectivamente con una cantidad de panel de punto crítico aumentada.

LISTADO DE SECUENCIAS

<110> GUARDANT HEALTH, INC.

<120> MÉTODOS PARA ANÁLISIS MULTI-RESOLUCIÓN DE ÁCIDOS NUCLEICOS LIBRES DE CÉLULAS

<130> 42534-733.601

<150> 62/489,391
<151> 2017-04-24

<150> 62/4468,201
<151> 2017-03-07

<150> 62/402,940
<151> 2016-09-30

<160> 11

<170> PatentIn versión 3.5

<210> 1
<211> 36
<212> ADN
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Oligonucleótido sintético

<400> 1
gctacccccc gcagcagcag cagcagcagc agcaac 36

<210> 2
<211> 12
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 2

Arg Tyr Pro Pro Gln Gln Gln Gln Gln Gln Gln Gln
1 5 10

<210> 3
<211> 12
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 3

Ala Thr Pro Arg Ser Ser Ser Ser Ser Ser Ser Asn
1 5 10

<210> 4
<211> 12
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 4

Leu Pro Pro Ala Ala Ala Ala Ala Ala Ala Ala Thr
1 5 10

<210> 5
<211> 13
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 5

Arg Tyr Pro Pro Gln Gln Gln Gln Gln Gln Gln Gln Arg
1 5 10

<210> 6
<211> 12
<212> ADN
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Oligonucleótido sintético

<400> 6
gcatacgtga tg 12

<210> 7
<211> 32
<212> ADN
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Oligonucleótido sintético

<400> 7

gtacccttgt ccccaggaag catacgtgat gg 32

<210> 8
<211> 9
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 8

Arg Thr Leu Val Pro Arg Lys His Thr
1 5

<210> 9
<211> 11
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 9

Val Pro Leu Ser Pro Gly Ser Ile Arg Asp Gly
1 5 10

<210> 10
<211> 10
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 10

Tyr Pro Cys Pro Gln Glu Ala Tyr Val Met
1 5 10

<210> 11
<211> 5
<212> PRT
<213> Secuencia artificial

<220>
<223> Descripción de Secuencia artificial: Péptido sintético

<400> 11

Glu Ala Tyr Val Met
1 5

REIVINDICACIONES

1. Un método para enriquecer múltiples regiones genómicas, que comprende:

(a) poner en contacto una cantidad predeterminada de ácido nucleico de una muestra con una mezcla de señuelos que comprende:

(i) un primer conjunto de señuelos que hibrida selectivamente con un primer conjunto de regiones genómicas del ácido nucleico de la muestra, dicho primer conjunto de señuelos proporcionado a una primera concentración que es menor que un punto de saturación del primer conjunto de señuelos, y
(ii) un segundo conjunto de señuelos que hibrida selectivamente con un segundo conjunto de regiones genómicas del ácido nucleico de la muestra, dicho segundo conjunto de señuelos proporcionado a una segunda concentración que es igual o está por encima de un punto de saturación del segundo conjunto de señuelos; y

(b) enriquecer el ácido nucleico de la muestra para el primer conjunto de regiones genómicas y el segundo conjunto de regiones genómicas, produciendo de este modo un ácido nucleico enriquecido,

en donde los señuelos son oligonucleótidos específicos del objetivo, y en donde la muestra comprende ADN libre de células.

2. El método de la reivindicación 1, en donde el segundo conjunto de señuelos tiene un punto de saturación que es mayor que sustancialmente todos los puntos de saturación asociados con los señuelos en el segundo conjunto de señuelos cuando un señuelo del segundo conjunto de señuelos se somete a una curva de titulación generada:

(i) midiendo la eficiencia de captura de un señuelo del segundo conjunto de señuelos en función de la concentración del señuelo, y
(ii) identificando un punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo.

3. El método de la reivindicación 1 o la reivindicación 2, en donde el punto de saturación del primer conjunto de señuelos se selecciona de tal manera que una eficiencia de captura observada aumento en menos del 10% a una concentración del señuelo de dos veces la de la primera concentración.

4. El método de cualquiera de las reivindicaciones 1 a 3, en donde el punto de saturación se selecciona de tal manera que una eficiencia de captura observada aumenta en menos del 10% a una concentración del señuelo de dos veces la de la segunda concentración.

5. El método de cualquiera de las reivindicaciones 1 a 4, en donde el primer conjunto de señuelos o el segundo conjunto de señuelos enriquecen selectivamente una o más regiones asociadas a nucleosomas de un genoma, las regiones asociadas a nucleosomas comprendiendo regiones genómicas que tienen una o más posiciones de bases genómicas con ocupación nucleosómica diferencial, en donde la ocupación nucleosómica diferencial es característica de un tipo de célula o tejido de origen o estado patológico.

6. El método de cualquiera de las reivindicaciones 1 a 5, que comprende además:

(c) secuenciar el ácido nucleico enriquecido para producir una pluralidad de lecturas de secuencia, comprendiendo además opcionalmente:
(d) producir una salida que comprende secuencias de ácido nucleico representativas del ácido nucleico de la muestra.

7. El método de la reivindicación 6, en donde las lecturas de secuencia se analizan para variantes genéticas relevantes para el cáncer.

8. El método de la reivindicación 6 o la reivindicación 7, en donde las lecturas de secuencia redundantes de un ácido nucleico original en la muestra se colapsan en una secuencia de consenso que representa la secuencia del ácido nucleico original.

9. El método de cualquiera de las reivindicaciones 6 a 8, en donde el primer conjunto de regiones genómicas y/o el segundo conjunto de regiones genómicas tienen una profundidad de lectura de entre 1.000 recuentos/base y 50.000 recuentos/base.

10. El método de cualquiera de las reivindicaciones 1 a 9, en donde el punto de saturación de un conjunto de señuelos se determina:

(a) para cada uno de los señuelos en el conjunto de señuelos, generando una curva de titulación que comprende:

(i) medir la eficiencia de captura del señuelo en función de la concentración del señuelo, y
(ii) identificar el punto de inflexión dentro de la curva de titulación, identificando de este modo un punto de saturación asociado con el señuelo; y

(b) seleccionando un punto de saturación que es mayor que sustancialmente todos los puntos de saturación asociados con señuelos en el conjunto de señuelos, determinando de este modo el punto de saturación del conjunto de señuelos.

11. El método de cualquiera de las reivindicaciones 2 a 4 o la reivindicación 10, en donde la eficiencia de captura de un señuelo se determina:

(a) proporcionando una pluralidad de muestras de ácidos nucleicos obtenidas de una pluralidad de sujetos en una cohorte;
(b) hibridando el señuelo con cada una de las muestras de ácidos nucleicos, en cada una de una pluralidad de concentraciones del señuelo;
(c) enriqueciendo con el señuelo una pluralidad de regiones genómicas de las muestras de ácidos nucleicos, en cada una de la pluralidad de concentraciones del señuelo; y
(d) midiendo el número de moléculas de ácido nucleico únicas o moléculas de ácido nucleico con la representación de ambas cadenas de una molécula de ácido nucleico de cadena doble original que representa la eficacia de captura en cada una de la pluralidad de concentraciones del señuelo.

12. El método de cualquiera de las reivindicaciones anteriores, en donde las regiones genómicas comprenden:

(a) por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, por lo menos 75, por lo menos 80, por lo menos 85, por lo menos 90, por lo menos 95, o 97 de los genes de la Tabla 3; y/o
(b) por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, por lo menos 75, por lo menos 80, por lo menos 85, por lo menos 90, por lo menos 95, por lo menos 100, por lo menos 105, por lo menos 110, o 115 de los genes de la Tabla 4.

13. El método de cualquiera de las reivindicaciones anteriores, en donde la cantidad predeterminada es de aproximadamente 200 ng, aproximadamente 150 ng, aproximadamente 125 ng, aproximadamente 100 ng, aproximadamente 75 ng, aproximadamente 50 ng, aproximadamente 25 ng, aproximadamente 10 ng, aproximadamente 5 ng o aproximadamente 1 ng de ADN.

14. El método de cualquiera de las reivindicaciones anteriores, en donde el ADN libre de células es ADN libre de células aislado de la sangre o suero.

15. El método de cualquiera de las reivindicaciones anteriores, en donde el ADN libre de células comprende ADN tumoral circulante.

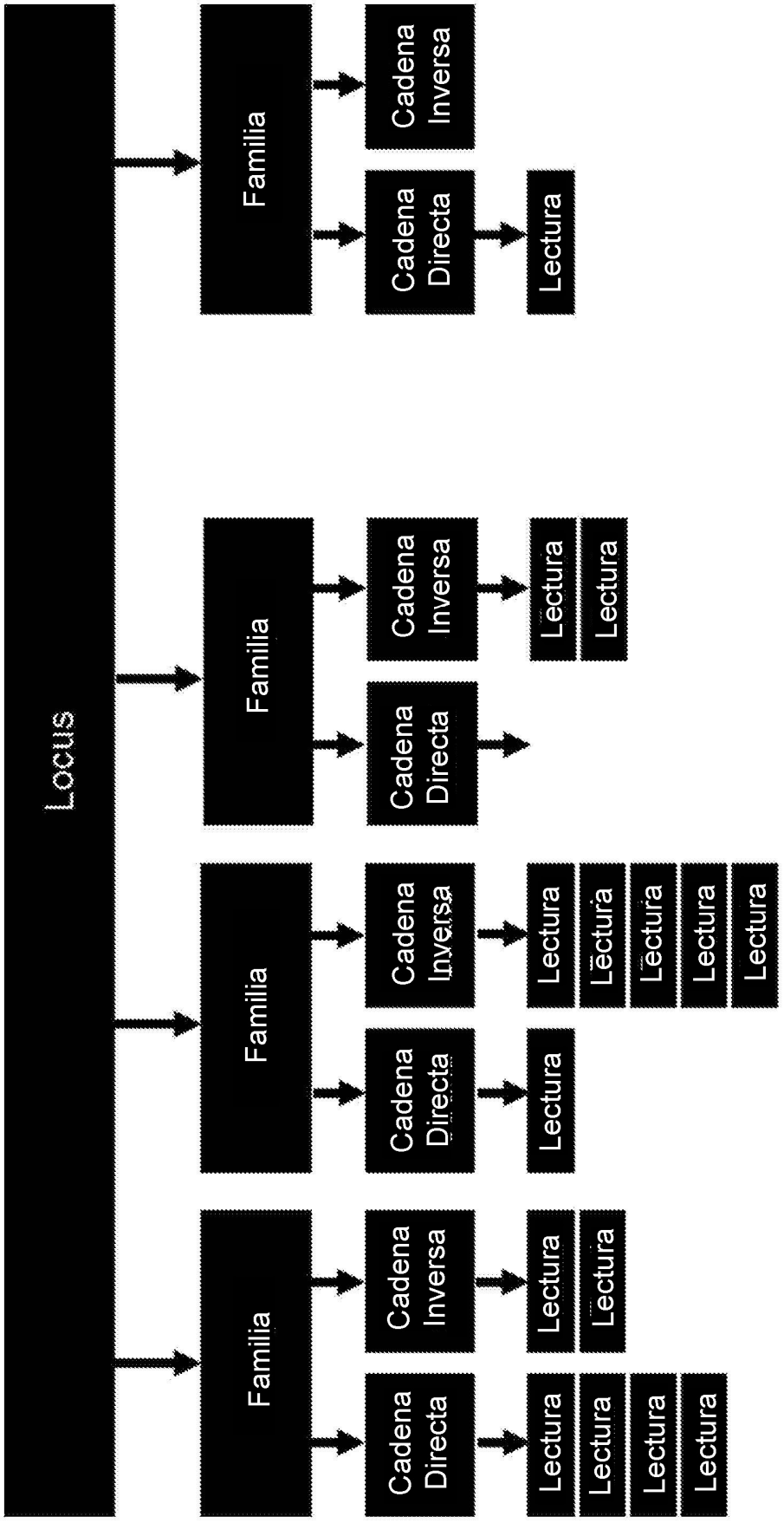

Locus
Familia
Cadena Inversa
Cadena Directa
Lectura
Familia
Cadena Inversa
Cadena Directa
Lectura
Lectura
Familia
Cadena Inversa
Cadena Directa
Lectura
Lectura
Lectura
Lectura
Lectura
Familia
Cadena Inversa
Cadena Directa
Lectura
Lectura
Lectura
Lectura
Lectura

FIG. 1

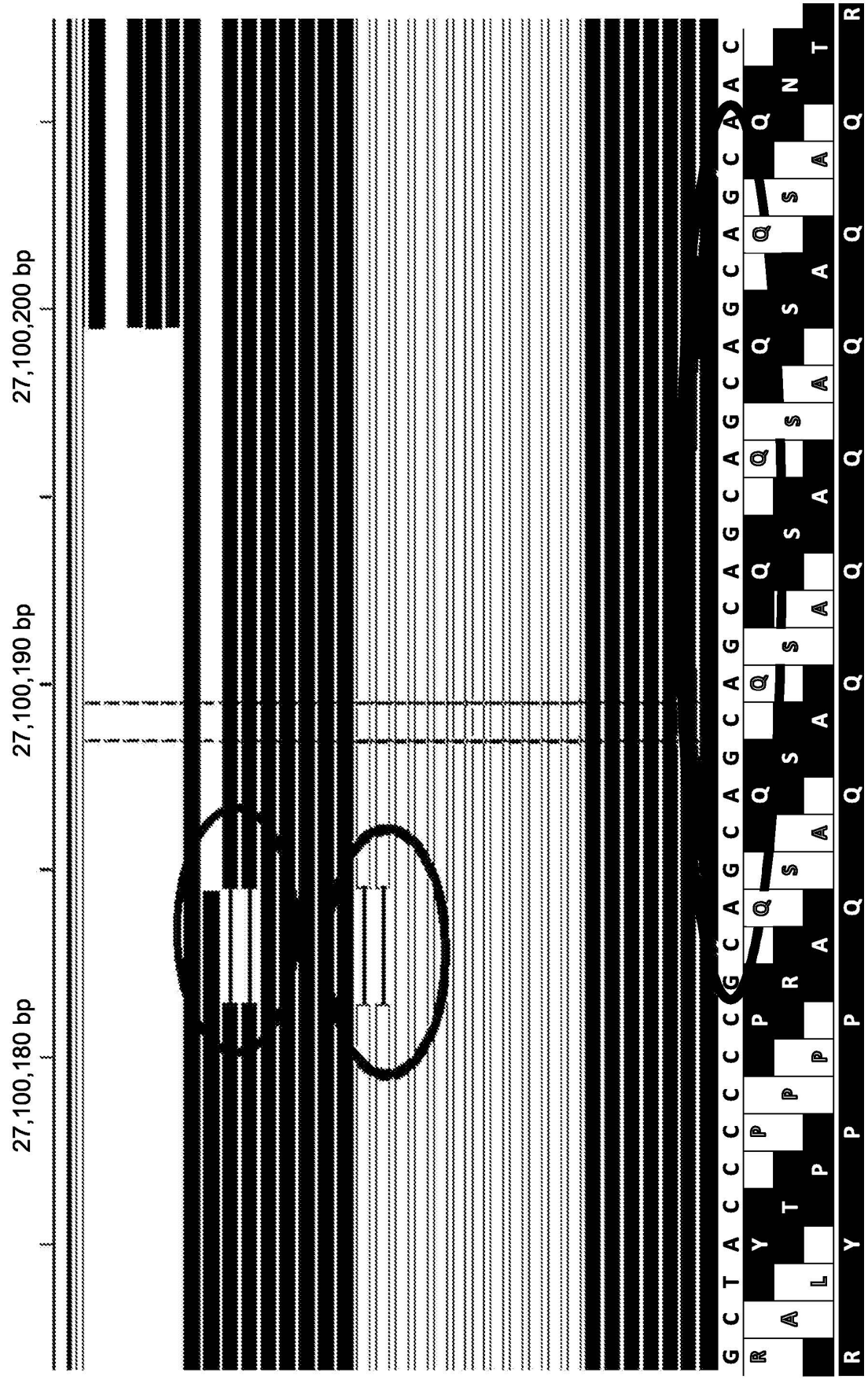


FIG. 2

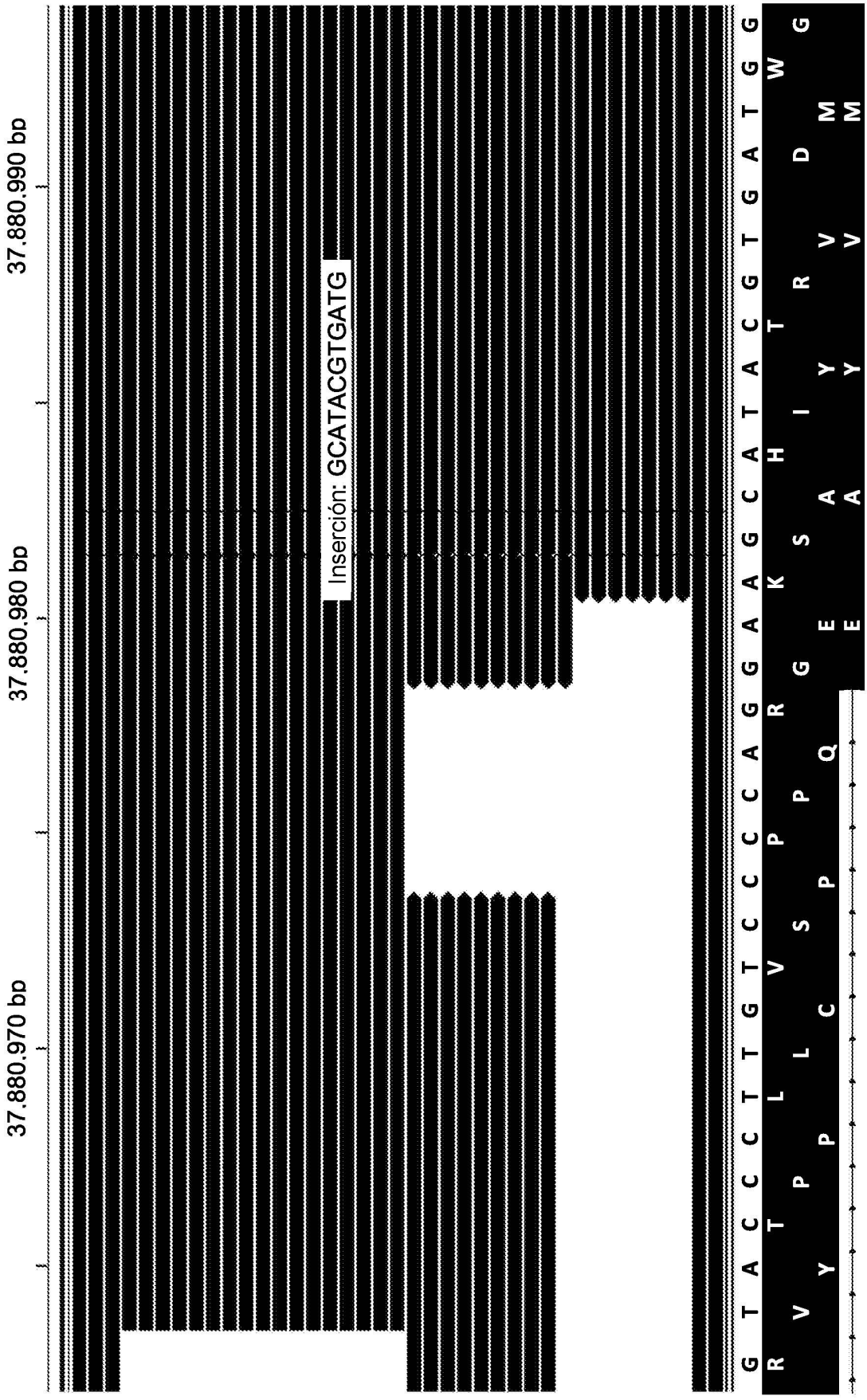
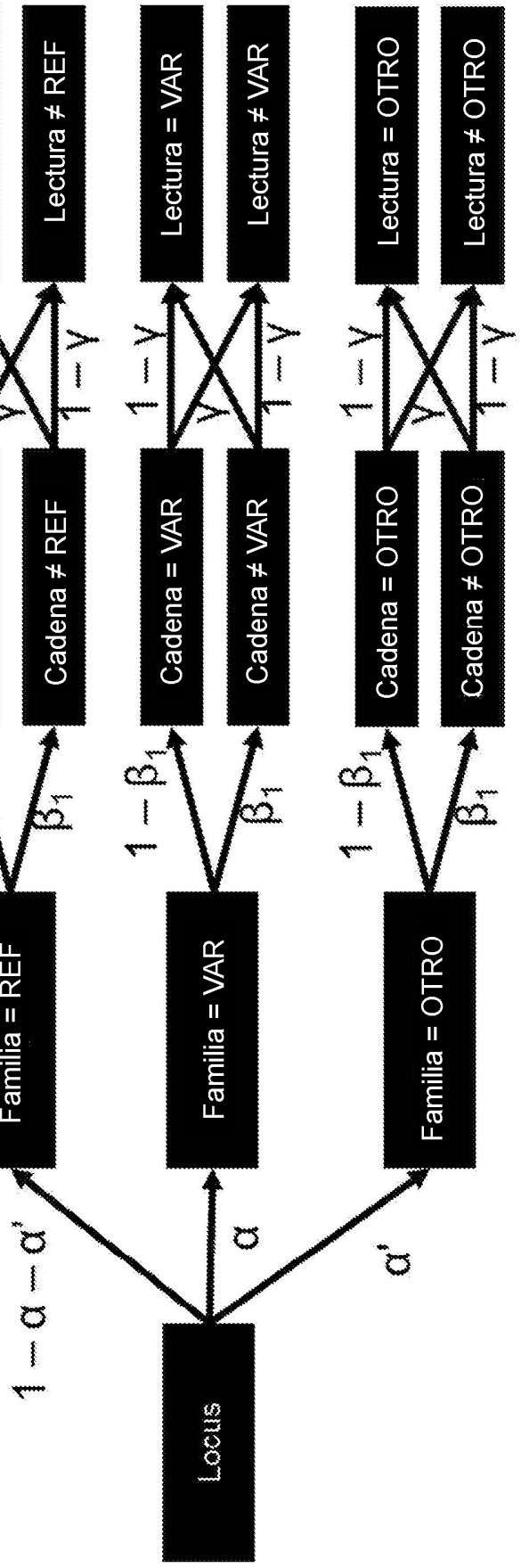

37.880.970 bp
37.880.980 bp
37.880.990 bp
Inserción: GCATACGTGATG
ERBB2

FIG. 3



Prueba de Hipótesis:

$$\log \frac{\max_{\alpha,\alpha',\beta_1,\beta_2,\gamma} \Pr\{\text{Lecturas} \mid \alpha,\alpha',\beta_1,\beta_2,\gamma\}}{\max_{\alpha,\beta_1,\beta_2,\gamma} \Pr\{\text{Lecturas} \mid \alpha=0,\alpha',\beta_1,\beta_2,\gamma\}} > Umbral$$

FIG. 4

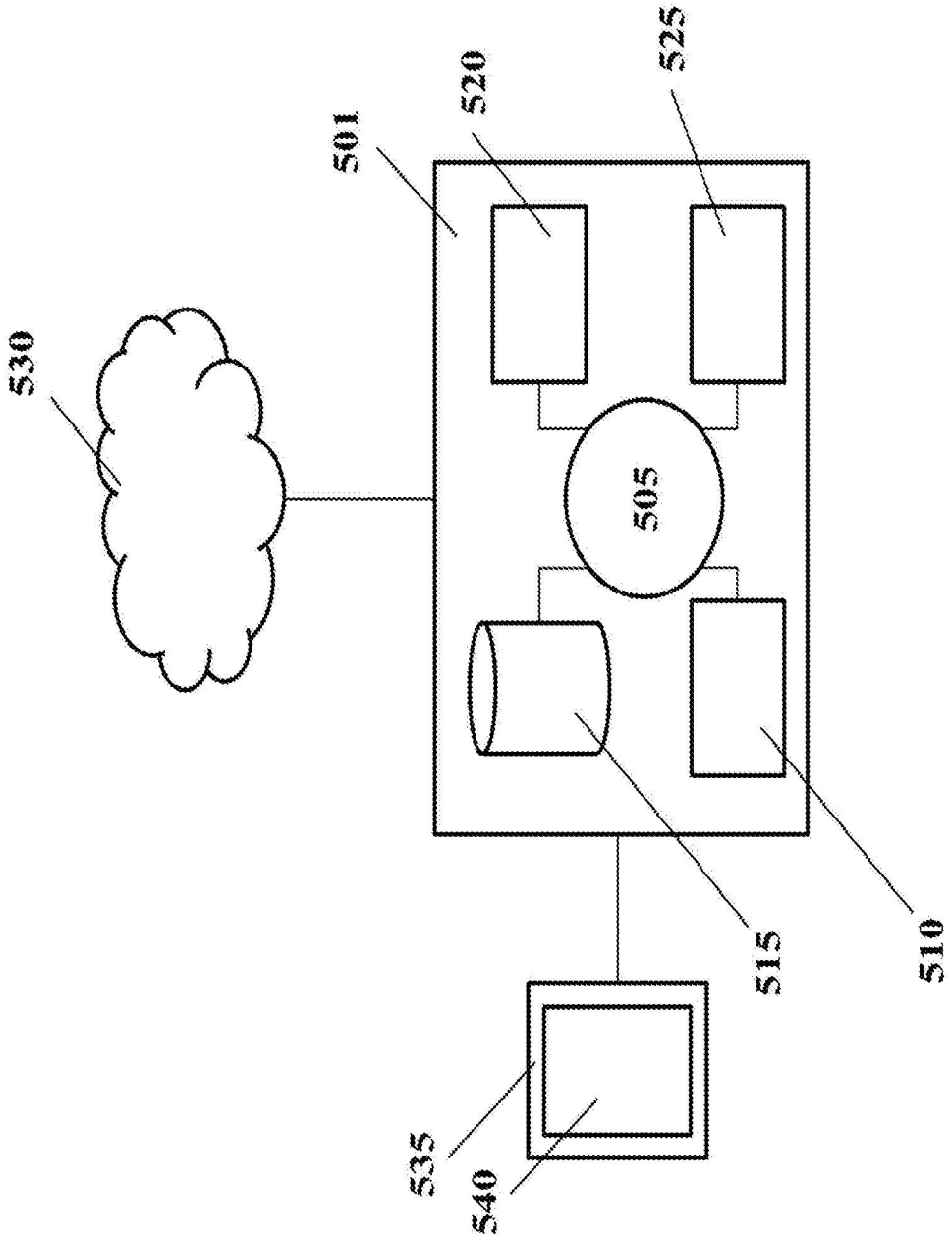


FIG. 5

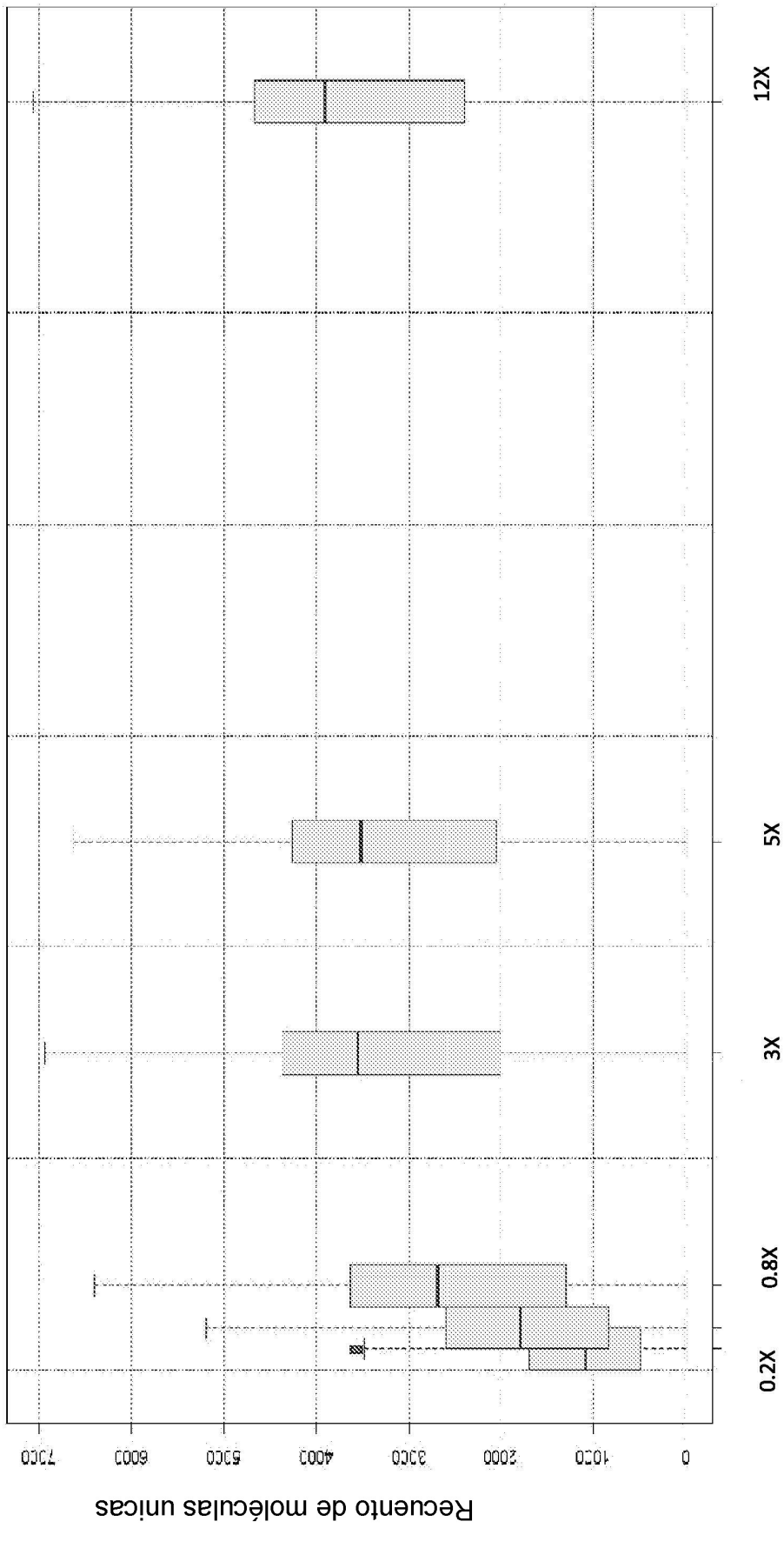

12X
5X
3X
0.8X
0.2X
0
1000
2000
3000
4000
5000
6000
7000
Recuento de moléculas únicas

FIG. 6