(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 769 241**

(51) Int. CI.:

C12Q 1/6827 (2008.01)
G16B 20/10 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **04.09.2013 E 18207391 (6)**

(97) Fecha y número de publicación de la concesión europea: **06.11.2019 EP 3470533**

(54) Título: **Sistemas y métodos para detectar variación en el número de copias**

(30) Prioridad:

04.09.2012 US 201261696734 P
15.03.2013 US 201361793997 P
21.09.2012 US 201261704400 P
13.07.2013 US 201361845987 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
25.06.2020

(73) Titular/es:

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

(72) Inventor/es:

TALASAZ, AMIRALI y
ELTOUKHY, HELMY

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Sistemas y métodos para detectar variación en el número de copias

ANTECEDENTES DE LA INVENCIÓN

La detección y cuantificación de polinucleótidos es importante para la biología molecular y aplicaciones médicas como los diagnósticos. Las pruebas genéticas son particularmente útiles para una serie de métodos de diagnóstico. Por ejemplo, los trastornos causados por alteraciones genéticas raras (por ejemplo, variantes de secuencia) o cambios en los marcadores epigenéticos, como el cáncer y la aneuploidía parcial o completa, pueden detectarse o caracterizarse con más precisión con información de la secuencia del ADN.

La detección temprana y la monitorización de enfermedades genéticas, como el cáncer, es a menudo útil y necesaria para el tratamiento o gestión exitosa de la enfermedad. Un enfoque puede incluir la monitorización de una muestra derivada de ácidos nucleicos libres de células, una población de polinucleótidos que pueden encontrarse en diferentes tipos de fluidos corporales. En algunos casos, la enfermedad puede caracterizarse o detectarse en base a la detección de aberraciones genéticas, como un cambio en la variación del número de copias y/o la variación de la secuencia de una o más secuencias de ácidos nucleicos, o el desarrollo de otras ciertas alteraciones genéticas raras. El ADN libre de células ("ADNcf") se ha conocido en la técnica durante décadas, y puede contener aberraciones genéticas asociadas con una enfermedad en particular. Con mejoras en la secuenciación y técnicas para manipular ácidos nucleicos, hay una necesidad en la técnica de métodos y sistemas mejorados para usar ADN libre de células para detectar y monitorizar enfermedades.

SUMARIO DE LA INVENCIÓN

La invención proporciona un método para determinar la variación en el número de copias en una muestra que incluye polinucleótidos libres de células, el método comprendiendo:

a. proporcionar por lo menos dos conjuntos de polinucleótidos libres de células, que mapean a diferentes posiciones mapeables en una secuencia de referencia en un genoma, y, para los conjuntos de polinucleótidos libres de células;

i. marcar de manera no única los polinucleótidos libres de células con un conjunto de códigos de barras moleculares;
ii. amplificar los polinucleótidos libres de células para producir polinucleótidos amplificados;
iii. secuenciar un subconjunto del conjunto de polinucleótidos amplificados, para producir un conjunto de lecturas de secuenciación;
iv. agrupar el conjunto de lecturas de secuenciación secuenciadas de polinucleótidos amplificados en familias que se corresponden a lecturas de secuenciación de polinucleótidos amplificados del mismo polinucleótido libre de células;
v. inferir una medida cuantitativa a partir de las familias en los conjuntos; y

b. determinar la variación del número de copias en base a la medida cuantitativa de familias en los conjuntos.

La invención proporciona además un medio legible por ordenador que comprende código ejecutable por máquina no transitorio que, tras la ejecución por un procesador informático, implementa un método, el método comprendiendo:

a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia derivan de polinucleótidos de progenie amplificados a partir de polinucleótidos libres de células originales marcadas de manera no única;
b. agrupar las lecturas de secuenciación secuenciadas a partir de los polinucleótidos de progenie en familias que comprenden lecturas de secuenciación de polinucleótidos de progenie amplificados a partir del mismo polinucleótido libre de células original marcado;
c. inferir una medida cuantitativa de familias en los polinucleótidos libres de células originales marcadas de manera no única; y
d. determinar la variación del número de copias comparando la medida cuantitativa de familias en los polinucleótidos libres de células originales marcadas de manera no única.

La divulgación proporciona un método para detectar la variación del número de copias que comprende: a) secuenciar polinucleótidos extracelulares a partir de una muestra corporal de un sujeto, en donde cada uno de los polinucleótidos extracelulares está opcionalmente unido a códigos de barras únicos; b) filtrar las lecturas que no alcanzan un umbral establecido; c) mapear las lecturas de secuencia obtenidas del paso (a) a una secuencia de referencia; d) cuantificar/contar lecturas mapeadas en dos o más regiones predefinidas de la secuencia de referencia; e) determinar una variación del número de copias en una o más de las regiones predefinidas por (i)

normalizando el número de lecturas en las regiones predefinidas entre sí y/o el número de códigos de barras únicos en las regiones predefinidas entre sí; y (ii) comparar los números normalizados obtenidos en el paso (i) con los números normalizados obtenidos de una muestra de control.

La divulgación también proporciona un método para detectar una mutación rara en una muestra libre de células o sustancialmente libre de células obtenida de un sujeto que comprende: a) secuenciar polinucleótidos extracelulares a partir de una muestra corporal de un sujeto, en donde cada uno de los polinucleótidos extracelulares genera una pluralidad de lecturas de secuenciación; b) secuenciar polinucleótidos extracelulares a partir de una muestra corporal de un sujeto, en donde cada uno de los polinucleótidos extracelulares genera una pluralidad de lecturas de secuenciación; secuenciar polinucleótidos extracelulares a partir de una muestra corporal de un sujeto, en donde cada uno de los polinucleótidos extracelulares genera una pluralidad de lecturas de secuenciación; c) filtrar las lecturas que no alcanzan un umbral establecido; d) mapear lecturas de secuencia derivadas de la secuenciación en una secuencia de referencia; e) identificar un subconjunto de lecturas de secuencia mapeadas que se alinean con una variante de la secuencia de referencia en cada posición de base mapeable; f) para cada posición de base mapeable, calcular una proporción de (a) una cantidad de lecturas de secuencia mapeadas que incluyen una variante en comparación con la secuencia de referencia, con (b) una cantidad de lecturas de secuencia totales para cada posición de base mapeable; g) normalizar las proporciones o la frecuencia de varianza para cada posición de base mapeable y determinar las potenciales variantes o mutaciones raras; h) y comparar el número resultante para cada una de las regiones con potenciales variantes o mutaciones raras con números derivados similarmente de una muestra de referencia.

Adicionalmente, la divulgación también proporciona un método para caracterizar la heterogeneidad de una condición anormal en un sujeto, el método comprendiendo generar un perfil genético de polinucleótidos extracelulares en el sujeto, en donde el perfil genético comprende una pluralidad de datos resultantes de la variación del número de copias y/u otros análisis de mutaciones raras (por ejemplo, alteración genética).

En algunas realizaciones, la prevalencia/concentración de cada variante rara identificada en el sujeto se informa y cuantifica simultáneamente. En otras realizaciones, se informa de una puntuación de confianza, con respecto a la prevalencia/concentraciones de variantes raras en el sujeto.

En algunas realizaciones, los polinucleótidos extracelulares comprenden ADN. En otras realizaciones, los polinucleótidos extracelulares comprenden ARN. Los polinucleótidos pueden ser fragmentos o fragmentarse después del aislamiento. Adicionalmente, la divulgación proporciona un método para la circulación de aislamiento y extracción de ácido nucleico.

En algunas realizaciones, los polinucleótidos extracelulares se aíslan de una muestra corporal que puede seleccionarse de un grupo que consiste de sangre, plasma, suero, orina, saliva, excreciones mucosales, esputo, heces y lágrimas.

En algunas realizaciones, los métodos de la divulgación también comprenden un paso de determinar el porcentaje de secuencias que tienen variación en el número de copias u otra alteración genética rara (por ejemplo, variantes de secuencia) en dicha muestra corporal.

En algunas realizaciones, el porcentaje de secuencias que tienen variación en el número de copias en dicha muestra corporal se determina calculando el porcentaje de regiones predefinidas con una cantidad de polinucleótidos por encima o por debajo de un umbral predeterminado.

En algunas realizaciones, se extraen fluidos corporales de un sujeto que se sospecha tiene una condición anormal que puede seleccionarse del grupo que consiste de, mutaciones, mutaciones raras, variantes de un único nucleótido, indeles, variaciones en el número de copias, transversiones, translocaciones, inversión, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones de genes, fusiones de cromosomas, truncamientos de genes, amplificación de genes, duplicaciones de genes, lesiones cromosómicas, lesiones de ADN, cambios anormales en las modificaciones químicas del ácido nucleico, cambios anormales en los patrones epigenéticos, cambios anormales en la infección por metilación de ácidos nucleicos y cáncer.

En algunas realizaciones, el sujeto puede ser una mujer embarazada en la que la condición anormal puede ser una anomalía fetal seleccionada del grupo que consiste de, variantes de un único nucleótido, indeles, variaciones en el número de copias, transversiones, translocaciones, inversión, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones de genes, fusiones de cromosomas, truncamientos de genes, amplificación de genes, duplicaciones de genes, lesiones cromosómicas, lesiones de ADN, cambios anormales en las modificaciones químicas del ácido nucleico, cambios anormales en los patrones epigenéticos, cambios anormales en la infección por metilación de ácidos nucleicos y cáncer.

En algunas realizaciones, el método comprende unir uno o más códigos de barras a los polinucleótidos

extracelulares o fragmentos de los mismos antes de la secuenciación, en el que los códigos de barras son únicos. En otras realizaciones los códigos de barras unidos a los polinucleótidos extracelulares o fragmentos de los mismos antes de la secuenciación no son únicos.

En algunas realizaciones, los métodos de la divulgación pueden comprender enriquecer selectivamente regiones del genoma del sujeto antes de la secuenciación. En otras realizaciones los métodos de la divulgación comprenden enriquecer selectivamente regiones del genoma del sujeto antes de la secuenciación. En otras realizaciones, los métodos de la divulgación comprenden enriquecer no selectivamente regiones del genoma del sujeto antes de la secuenciación.

Además, los métodos de la divulgación comprenden unir uno o más códigos de barras a los polinucleótidos extracelulares o fragmentos de los mismos antes de cualquier paso de amplificación o enriquecimiento.

En algunas realizaciones, el código de barras es un polinucleótido, que puede comprender además una secuencia aleatoria o un conjunto fijo o semi-aleatorio de oligonucleótidos que en combinación con la diversidad de moléculas secuenciadas de una región seleccionada permite la identificación de moléculas únicas y es por lo menos de 3, 5, 10, 15, 20 25, 30, 35, 40, 45, o 50mer pares de bases de longitud.

En algunas realizaciones, los polinucleótidos extracelulares o fragmentos de los mismos pueden amplificarse. En algunas realizaciones, la amplificación comprende la amplificación global o la amplificación del genoma completo.

En algunas realizaciones, las lecturas de secuencia de identidad única pueden detectarse en base a la información de secuencia en las regiones de comienzo (inicio) y final (parada) de la lectura de secuencia y la longitud de la lectura de secuencia. En otras realizaciones, las moléculas de secuencias de identidad única se detectan en base a la información de secuencia en las regiones de comienzo (inicio) y final (parada) de la lectura de secuencia, la longitud de la lectura de secuencia y la unión de un código de barras.

En algunas realizaciones, la amplificación comprende amplificación selectiva, amplificación no selectiva, amplificación por supresión o enriquecimiento sustractivo.

En algunas realizaciones, los métodos de la divulgación comprenden eliminar un subconjunto de las lecturas de un análisis adicional antes de cuantificar o enumerar las lecturas.

En algunas realizaciones, el método puede comprender filtrar las lecturas con una puntuación de precisión o calidad menor que un umbral, por ejemplo, 90%, 99%, 99,9%, o 99,99% y/o puntuación de mapeo menor que un umbral, por ejemplo, 90%, 99%, 99,9% o 99,99%. En otras realizaciones, los métodos de la divulgación comprenden filtrar lecturas con una puntuación de calidad menor que un umbral establecido.

En algunas realizaciones, las regiones predefinidas son de tamaño uniforme o sustancialmente uniforme, aproximadamente de 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb o 100kb de tamaño. En algunas realizaciones, se analizan por lo menos 50, 100, 200, 500, 1000, 2000, 5000, 10.000, 20.000 o 50.000 regiones.

En algunas realizaciones, se produce una variante genética, una mutación rara o una variación del número de copias en una región del genoma seleccionada del grupo que consiste de fusiones de genes, duplicaciones de genes, deleciones de genes, translocaciones de genes, regiones de microsatélites, fragmentos de genes o combinaciones de los mismos. En otras realizaciones, se produce una variante genética, mutación rara o variación en el número de copias en una región del genoma seleccionada del grupo que consiste de genes, oncogenes, genes supresores de tumores, promotores, elementos de secuencias reguladoras o combinaciones de los mismos. En algunas realizaciones, la variante es una variante de nucleótido, sustitución de una única base, o indel pequeño, transversión, translocación, inversión, deleción, truncamiento o truncamiento del gen de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 o 20 nucleótidos de longitud.

En algunas realizaciones, el método comprende corregir/normalizar/ajustar la cantidad de lecturas mapeadas usando los códigos de barras o propiedades únicas de lecturas individuales.

En algunas realizaciones, la enumeración de las lecturas se realiza mediante la enumeración de códigos de barras únicos en cada una de las regiones predefinidas y normalizando esos números en por lo menos un subconjunto de regiones predefinidas que se secuenciaron. En algunas realizaciones, las muestras a intervalos temporales sucesivos del mismo sujeto se analizan y comparan con resultados de muestras anteriores. El método de la divulgación puede comprender además determinar la frecuencia de variación del número de copias parcial, pérdida de heterocigosidad, análisis de expresión génica, análisis epigenético y análisis de hipermetilación después de amplificar los polinucleótidos extracelulares unidos al código de barras.

En algunas realizaciones, el análisis de la variación del número de copias y de mutaciones raras se

determinan en una muestra libre de células o sustancialmente libre de células obtenida de un sujeto usando secuenciación multiplex, que comprende realizar más de 10.000 reacciones de secuenciación; secuenciar simultáneamente por lo menos 10.000 lecturas diferentes; o realizar análisis de datos en por lo menos 10.000 lecturas diferentes en todo el genoma. El método puede comprender una secuenciación multiplex que comprende realizar un análisis de datos en por lo menos 10.000 lecturas diferentes en todo el genoma. El método puede comprender además enumerar las lecturas secuenciadas que son identificables de manera única.

En algunas realizaciones, los métodos de la divulgación comprenden normalizar y la detección se realiza usando uno o más de Markov oculto, programación dinámica, máquina de vectores de soporte, red bayesiana, decodificación de entramados, decodificación de Viterbi, maximización de la esperanza, filtrado de Kalman, o metodologías de redes neuronales.

En algunas realizaciones, los métodos de la divulgación comprenden monitorizar la progresión de la enfermedad, monitorizar la enfermedad residual, monitorizar la terapia, diagnosticar una afección, pronosticar una afección o seleccionar una terapia en base a las variantes descubiertas.

En algunas realizaciones, una terapia se modifica en base al análisis de la muestra más reciente. Además, los métodos de la divulgación comprenden inferir el perfil genético de un tumor, infección u otra anomalía tisular. En algunas realizaciones, se monitoriza el crecimiento, la remisión o la evolución de un tumor, una infección u otra anomalía del tejido. En algunas realizaciones, se analiza y se monitoriza el sistema inmune del sujeto en casos individuales o a lo largo del tiempo.

En algunas realizaciones, los métodos de la divulgación comprenden la identificación de una variante que se sigue a través de una prueba de imagen (por ejemplo, CT, PET-CT, MRI, rayos X, ultrasonido) para la localización de la anomalía del tejido sospechosa de provocar la variante identificada.

En algunas realizaciones, los métodos de la divulgación comprenden el uso de datos genéticos obtenidos de una biopsia de tejido o tumor del mismo paciente. En algunas realizaciones, se infiere la filogenética de un tumor, infección u otra anomalía del tejido.

En algunas realizaciones, los métodos de la divulgación comprenden realizar no-tipificación basada en la población e identificación de regiones de baja confianza. En algunas realizaciones, la obtención de datos de mediciones para la cobertura de secuencia comprende medir la profundidad de cobertura de secuencia en cada posición del genoma. En algunas realizaciones, corregir los datos de mediciones para el sesgo de cobertura de secuencia comprende calcular la cobertura promediada en ventanas. En algunas realizaciones, corregir los datos de mediciones para el sesgo de cobertura de secuencia comprende realizar ajustes para tener en cuenta el sesgo de GC en la construcción de la biblioteca y el proceso de secuenciación. En algunas realizaciones, corregir los datos de mediciones para el sesgo de cobertura de secuencia comprende realizar ajustes en base al factor de ponderación adicional asociado con los mapeos individuales para compensar el sesgo.

En algunas realizaciones, los métodos de la divulgación comprenden polinucleótidos extracelulares derivados de un origen celular enfermo. En algunas realizaciones, el polinucleótido extracelular se deriva de un origen celular sano.

La divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los siguientes pasos: seleccionar regiones predefinidas en un genoma; enumerar el número de lecturas de secuencia en las regiones predefinidas; normalizar el número de lecturas de secuencia en las regiones predefinidas; y determinar el porcentaje de variación del número de copias en las regiones predefinidas. En algunas realizaciones, se analiza la totalidad del genoma o por lo menos el 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% o 90% del genoma. En algunas realizaciones, el medio legible por ordenador proporciona datos sobre el porcentaje de ADN o ARN de cáncer en plasma o suero al usuario final.

En algunas realizaciones, se analiza la cantidad de variación genética, como polimorfismos o variantes causales. En algunas realizaciones, se detecta la presencia o ausencia de alteraciones genéticas.

La divulgación también proporciona un método para detectar una mutación rara en una muestra libre de células o sustancialmente libre de células obtenida de un sujeto que comprende: a) secuenciar polinucleótidos extracelulares a partir de una muestra corporal de un sujeto, en donde cada uno de los polinucleótidos extracelulares genera un pluralidad de lecturas de secuenciación; b) filtrar las lecturas que no alcanzan un umbral establecido; c) mapear las lecturas de secuencia derivadas de la secuenciación en una secuencia de referencia; d) identificar un subconjunto de lecturas de secuencia mapeadas que se alinean con una variante de la secuencia de referencia en cada posición de base mapeable; e) para cada posición de base mapeable, calcular una proporción de (a) una cantidad de lecturas de secuencia mapeadas que incluyen una variante en comparación con la secuencia de referencia, con (b) una cantidad de lecturas de secuencia totales para cada posición de base mapeable; f) normalizar las proporciones o la frecuencia de la varianza para cada posición de base mapeable y determinar las potenciales

variantes raras u otras alteraciones genéticas; y g) comparar el número resultante para cada una de las regiones.

Esta divulgación también proporciona un método que comprende: a. proporcionar por lo menos un conjunto de polinucleótidos principales marcados, y para cada conjunto de polinucleótidos principales marcados; b. amplificar los polinucleótidos originales marcados en el conjunto para producir un conjunto correspondiente de polinucleótidos de progenie amplificados; c. secuenciar un subconjunto (incluyendo un subconjunto apropiado) del conjunto de polinucleótidos de progenie amplificados, para producir un conjunto de lecturas de secuenciación; y d. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiente a un polinucleótido único entre el conjunto de polinucleótidos originales marcados. En ciertas realizaciones, el método comprende además: e. analizar el conjunto de secuencias de consenso para cada conjunto de moléculas originales marcadas.

En algunas realizaciones, cada polinucleótido en un conjunto es mapeable a una secuencia de referencia.

En algunas realizaciones, el método comprende proporcionar una pluralidad de conjuntos de polinucleótidos originales marcados, en donde cada conjunto es mapeable a una secuencia de referencia diferente.

En algunas realizaciones, el método comprende además convertir material genético de partida inicial en los polinucleótidos originales marcados.

En algunas realizaciones, el material genético de partida inicial comprende no más de 100 ng de polinucleótidos.

En algunas realizaciones, el método comprende hacer cuellos de botella con el material genético de partida inicial antes de la conversión.

En algunas realizaciones, el método comprende convertir el material genético de partida inicial en polinucleótidos parentales marcados con una eficiencia de conversión de por lo menos el 10%, por lo menos el 20%, por lo menos el 30%, por lo menos el 40%, por lo menos el 50%, por lo menos el 60%, por lo menos el 80% o por lo menos el 90%.

En algunas realizaciones, la conversión comprende cualquiera de ligación de extremo romo, ligación de extremo pegajoso, sondas de inversión molecular, PCR, PCR basada en ligación, ligación de cadena sencilla y circularización de cadena sencilla.

En algunas realizaciones el material genético de partida inicial es ácido nucleico libre de células.

En algunas realizaciones, una pluralidad de secuencias de referencia son del mismo genoma.

En algunas realizaciones cada polinucleótido original marcado en el conjunto se marca de manera única.

En algunas realizaciones los marcadores son no únicos.

En algunas realizaciones la generación de secuencias de consenso se basa en la información del marcador y/o por lo menos uno de la información de secuencia en la región de comienzo (partida) de la lectura de secuencias, las regiones finales (de parada) de la lectura de secuencia y la longitud de la lectura de secuencias.

En algunas realizaciones, el método comprende secuenciar un subconjunto del conjunto de polinucleótidos de la progenie amplificada suficientes para producir lecturas de secuencia para por lo menos una progenie de cada uno de por lo menos el 20%, por lo menos el 30%, 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 90% por lo menos el 95%, por lo menos el 98%, por lo menos el 99%, por lo menos el 99,9% o por lo menos el 99,99% de polinucleótidos únicos en el conjunto de polinucleótidos parentales marcados.

En algunas realizaciones, la por lo menos una progenie es una pluralidad de progenies, por ejemplo, por lo menos 2, por lo menos 5 o por lo menos 10 progenies.

En algunas realizaciones, el número de lecturas de secuencia en el conjunto de lecturas de secuencia es mayor que el número de polinucleótidos parentales marcados únicos en el conjunto de polinucleótidos parentales marcados.

En algunas realizaciones, el subconjunto del conjunto de polinucleótidos de la progenie amplificada secuenciados es del tamaño suficiente para que cualquier secuencia de nucleótidos representada en el conjunto de polinucleótidos parentales marcados en un porcentaje que sea igual a la tasa error de secuenciación por base del porcentaje de la plataforma de secuenciación usado, tenga por lo menos un 50%, por lo menos un 60%, por lo

menos un 70%, por lo menos un 80%, por lo menos un 90%, por lo menos un 95%, por lo menos un 98%, por lo menos un 99%, por lo menos un 99,9% o por lo menos un 99,99% de probabilidad de estar representado entre el conjunto de secuencias de consenso.

En algunas realizaciones, el método comprende enriquecer el conjunto de polinucleótidos de la progenie amplificada para el mapeo de polinucleótidos en una o más secuencias de referencia seleccionadas mediante: (i) la amplificación selectiva de secuencias del material genético de partida inicial convertido en polinucleótidos parentales marcados; (ii) la amplificación selectiva de polinucleótidos parentales marcados; (iii) la captura de secuencia selectiva de polinucleótidos de la progenie amplificada; o (iv) la captura de secuencia selectiva del material genético de partida inicial.

En algunas realizaciones, el análisis comprende la normalización de una medida (por ejemplo, número) tomada de un conjunto de secuencias de consenso frente a una medida tomada de un conjunto de secuencias de consenso de una muestra de control.

En algunas realizaciones, el análisis comprende detectar mutaciones, mutaciones raras, variantes de un único nucleótido, indeles, variaciones en el número de copias, transversiones, translocaciones, inversión, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones de genes, fusiones de cromosomas, truncamientos de genes, amplificación de genes, duplicaciones de genes, lesiones de cromosomas, lesiones de ADN, cambios anormales en las modificaciones químicas de ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la infección por metilación de los ácidos nucleicos o cáncer.

En algunas realizaciones los polinucleótidos comprenden ADN, ARN, una combinación de dos o ADN más ADNC derivado de ARN.

En algunas realizaciones, un cierto subconjunto de polinucleótidos se selecciona para o se enriquece en base a la longitud del polinucleótido en pares de bases del conjunto inicial de polinucleótidos o de los polinucleótidos amplificados.

En algunas realizaciones, el análisis comprende además la detección y la monitorización de una anomalía o enfermedad dentro de un individuo, como, infección y/o cáncer.

En algunas realizaciones, el método se realiza en combinación con realización de perfiles del repertorio inmune.

En algunas realizaciones, los polinucleótidos se extraen del grupo que consiste de sangre, plasma, suero, orina, saliva, excreciones de la mucosa, esputo, heces y lágrimas.

En algunas realizaciones, colapsar comprende detectar y/o corregir errores, incisiones o lesiones presentes en la cadena de sentido o antisentido de los polinucleótidos parentales marcados o polinucleótidos de la progenie amplificada.

Esta divulgación también proporciona un método que comprende detectar la variación genética en el material genético de partida inicial con una sensibilidad de por lo menos un 5%, por lo menos un 1%, por lo menos un 0,5%, por lo menos un 0,1% o por lo menos un 0,05%. En algunas realizaciones, el material genético de partida inicial se proporciona en una cantidad inferior a 100 ng de ácido nucleico, la variación genética es la variación del número de copia/heterocigosidad y la detección se realiza con resolución subcromosómica; por ejemplo, resolución de por lo menos 100 megabases, resolución de por lo menos 10 megabases, resolución de por lo menos 1 megabase, resolución de por lo menos 100 kilobases, resolución de por lo menos 10 kilobases o resolución de por lo menos 1 kilobase. En otra realización, el método comprende proporcionar una pluralidad de conjuntos de polinucleótidos originales marcados, en donde cada conjunto es mapeable a una secuencia de referencia diferente. En otra realización, la secuencia de referencia es el locus de un marcador tumoral, y el análisis comprende detectar el marcador tumoral en el conjunto de secuencias consenso. En otra realización, el marcador tumoral está presente en el conjunto de secuencias de consenso a una frecuencia menor que la tasa de error introducida en el paso de amplificación. En otra realización, el por lo menos un conjunto es una pluralidad de conjuntos, y las secuencias de referencia comprenden una pluralidad de secuencias de referencia, cada una de las cuales es el locus de un marcador tumoral. En otra realización, el análisis comprende detectar la variación en el número de copias de las secuencias de consenso entre por lo menos dos conjuntos de polinucleótidos originales. En otra realización, el análisis comprende detectar la presencia de variaciones de secuencia en comparación con las secuencias de referencia. En otra realización, el análisis comprende detectar la presencia de variaciones de secuencia en comparación con las secuencias de referencia y detectar la variación en el número de copias de secuencias de consenso entre por lo menos dos conjuntos de polinucleótidos originales. En otra realización, colapsar comprende: i. agrupar las lecturas de secuencias de los polinucleótidos de progenie amplificados en familias, cada familia amplificada del mismo polinucleótido original marcado; y ii. determinar una secuencia de consenso en base a las

lecturas de secuencia en una familia.

Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los pasos siguientes: a. proporcionar por lo menos un conjunto de polinucleótidos originales marcados, y para cada conjunto de polinucleótidos originales marcados; b. amplificar los polinucleótidos originales marcados en el conjunto para producir un conjunto correspondiente de polinucleótidos de progenie amplificados; c. secuenciar un subconjunto (incluyendo un subconjunto apropiado) del conjunto de polinucleótidos de progenie amplificados, para producir un conjunto de lecturas de secuenciación; y d. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiente a un polinucleótido único entre el conjunto de polinucleótidos originales marcados y, opcionalmente, e. analizar el conjunto de secuencias de consenso para cada conjunto de moléculas originales marcadas.

Esta divulgación también proporciona un método que comprende: a. proporcionar por lo menos un conjunto de polinucleótidos originales marcados, y para cada conjunto de polinucleótidos originales marcados; b. amplificar los polinucleótidos originales marcados en el conjunto para producir un conjunto correspondiente de polinucleótidos de la progenie amplificados; c. secuenciar un subconjunto (incluyendo un subconjunto apropiado) del conjunto de polinucleótidos de la progenie amplificados, para producir un conjunto de lecturas de secuenciación; d. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiendo a un polinucleótido único entre el conjunto de polinucleótidos originales marcados; y e. filtrar de entre las secuencias de consenso aquellas que no logran alcanzar un umbral de calidad. En una realización, el umbral de calidad considera una serie de lecturas de secuencia de polinucleótidos de progenie amplificados colapsados en una secuencia consenso. En otra realización, el umbral de calidad considera una serie de lecturas de secuencia de polinucleótidos de progenie amplificados colapsados en una secuencia de consenso. Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar el método mencionado anteriormente.

Esta divulgación también proporciona un método que comprende: a. proporcionar por lo menos un conjunto de polinucleótidos originales marcados, en donde cada conjunto mapea para una secuencia de referencia diferente en uno o más genomas, y, para cada conjunto de polinucleótidos originales marcados; i. amplificar los primeros polinucleótidos para producir un conjunto de polinucleótidos amplificados; ii) secuenciar un subconjunto del conjunto de polinucleótidos amplificados, para producir un conjunto de lecturas de secuenciación; y iii. colapsar las lecturas de secuencia: 1. agrupando las lecturas de secuencias secuenciadas de polinucleótidos de progenie amplificados en familias, cada familia amplificada a partir del mismo polinucleótido original marcado. En una realización, el colapso comprende además: 2. determinar una medida cuantitativa de lecturas de secuencia en cada familia. En otra realización, el método comprende además (incluyendo a): b. determinar una medida cuantitativa de familias únicas; y c. en base a (1) la medida cuantitativa de familias únicas y (2) la medida cuantitativa de lecturas de secuencia en cada grupo, inferir una medida de polinucleótidos originales marcados únicos en el conjunto. En otra realización, la inferencia se realiza usando modelos estadísticos o probabilísticos. En otra realización en la que el por lo menos un conjunto es una pluralidad de conjuntos. En otra realización, el método comprende además corregir el sesgo de la amplificación o de la representación entre los dos conjuntos. En otra realización, el método comprende además usar un control o un conjunto de muestras de control para corregir los sesgos de la amplificación o de la representación entre los dos conjuntos. En otra realización, el método comprende además determinar la variación en el número de copias entre los conjuntos. En otra realización, el método comprende además (incluyendo a, b, c): d. determinar una medida cuantitativa de formas polimórficas entre las familias; y e. en base a la medida cuantitativa determinada de formas polimórficas, inferir una medida cuantitativa de formas polimórficas en el número de polinucleótidos originales marcados únicos inferidos. En otra realización en la que las formas polimórficas incluyen, pero no están limitadas a: sustituciones, inserciones, deleciones, inversiones, cambios de microsatélites, transversiones, translocaciones, fusiones, metilación, hipermetilación, hidroximetilación, acetilación, variantes epigenéticas, variantes asociadas a la regulación o sitios de unión a proteínas. En otra realización en la que los conjuntos derivan de una muestra común, el método comprende además: a. inferir la variación en el número de copias para la pluralidad de conjuntos en base a una comparación del número inferido de polinucleótidos originales marcados en cada mapeo de conjunto a cada una de una pluralidad de secuencias de referencia. En otra realización, se infiere adicionalmente el número original de polinucleótidos en cada conjunto. Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los métodos mencionados anteriormente.

Esta divulgación también proporciona un método para determinar la variación en el número de copias en una muestra que incluye polinucleótidos, el método comprendiendo: a. proporcionar por lo menos dos conjuntos de primeros polinucleótidos, en donde cada conjunto mapea para una secuencia de referencia diferente en un genoma y, para cada conjunto de primeros polinucleótidos; i. amplificar los polinucleótidos para producir un conjunto de polinucleótidos amplificados; ii. secuenciar un subconjunto del conjunto de polinucleótidos amplificados, para producir un conjunto de lecturas de secuenciación; iii. agrupar las lecturas de secuencias secuenciadas a partir de polinucleótidos amplificados en familias, cada familia amplificada a partir del mismo primer polinucleótido del conjunto; iv. inferir una medida cuantitativa de familias en el conjunto; b. determinar la variación en el número de copias comparando la medida cuantitativa de familias en cada conjunto. Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los métodos mencionados anteriormente.

Esta divulgación también proporciona un método para inferir la frecuencia de las llamadas de secuencia en una muestra de polinucleótidos que comprende: a. proporcionar por lo menos un conjunto de primeros polinucleótidos, en donde cada conjunto mapea para una secuencia de referencia diferente en uno o más genomas y, para cada conjunto de primeros polinucleótidos; i. amplificar los primeros polinucleótidos para producir un conjunto de polinucleótidos amplificados; ii. secuenciar un subconjunto del conjunto de polinucleótidos amplificados, para producir un conjunto de lecturas de secuenciación; iii. agrupar las lecturas de secuencia en familias, cada familia comprendiendo lecturas de secuencia de polinucleótidos amplificados amplificados a partir del mismo primer polinucleótido; b. inferir, para cada conjunto de primeros polinucleótidos, una frecuencia de llamada para una o más bases en el conjunto de primeros polinucleótidos, en donde inferir comprende: i. asignar, para cada familia, una puntuación de confianza para cada una de una pluralidad de llamadas, la puntuación de confianza teniendo en cuenta una frecuencia de la llamada entre los miembros de la familia; y ii. estimar una frecuencia de una o más llamadas teniendo en cuenta las puntuaciones de confianza de una o más llamadas asignadas a cada familia. Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los métodos mencionados anteriormente.

Esta divulgación también proporciona un método para comunicar información de secuencia sobre por lo menos una molécula de polinucleótido individual que comprende: a. proporcionar por lo menos una molécula de polinucleótido individual; b. codificar información de secuencia en la por lo menos una molécula de polinucleótido individual para producir una señal; c. pasar por lo menos parte de la señal a través de un canal para producir una señal recibida que comprende información de secuencia de nucleótidos sobre la por lo menos una molécula de polinucleótido individual, en donde la señal recibida comprende ruido y/o distorsión; d. decodificar la señal recibida para producir un mensaje que comprende información de secuencia sobre por lo menos una molécula de polinucleótido individual, en donde la decodificación reduce el ruido y/o la distorsión en el mensaje; y e. proporcionar el mensaje a un destinatario. En una realización, el ruido comprende llamadas de nucleótidos incorrectas. En otra realización, la distorsión comprende una amplificación desigual de la molécula de polinucleótido individual en comparación con otras moléculas de polinucleótidos individuales. En otra realización, la distorsión es resultado del sesgo de la amplificación o de la secuenciación. En otra realización, la por lo menos una molécula de polinucleótido individual es una pluralidad de moléculas de polinucleótidos individuales, y la decodificación produce un mensaje sobre cada molécula en la pluralidad. En otra realización, la codificación comprende amplificar la por lo menos una molécula de polinucleótido individual que se ha marcado opcionalmente, en donde la señal comprende una colección de moléculas amplificadas. En otra realización, el canal comprende un secuenciador de polinucleótidos y la señal recibida comprende lecturas de secuencia de una pluralidad de polinucleótidos amplificados a partir de la por lo menos una molécula de polinucleótido individual. En otra realización, la decodificación comprende agrupar lecturas de secuencia moléculas amplificadas, amplificadas a partir de cada una de las por lo menos una molécula de polinucleótido individual. En otra realización, la decodificación consiste de un método probabilístico o estadístico para filtrar la señal de secuencia generada. Esta divulgación también proporciona un sistema que comprende un medio legible por ordenador para realizar los métodos mencionados anteriormente.

En otra realización, los polinucleótidos se derivan de ADN o ARN genómico tumoral. En otra realización, los polinucleótidos se derivan de polinucleótidos libres de células, polinucleótidos exosomales, polinucleótidos bacterianos o polinucleótidos virales. Otra realización comprende además la detección y/o asociación de vías moleculares afectadas. En otra realización que comprende además la monitorización en serie del estado de salud o de enfermedad de un individuo. En otra realización, se infiere la filogenia de un genoma asociado con una enfermedad dentro de un individuo. Otra realización comprende además el diagnóstico, la monitorización o el tratamiento de una enfermedad. En otra realización, el régimen de tratamiento se selecciona o modifica en base a formas polimórficas detectadas o CNV o vías asociadas. En otra realización el tratamiento comprende una terapia de combinación.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los pasos siguientes: seleccionar regiones predefinidas en un genoma; acceder a lecturas de secuencia y enumerar el número de lecturas de secuencia en las regiones predefinidas; normalizar el número de lecturas de secuencia en las regiones predefinidas; y determinar el porcentaje de variación del número de copias en las regiones predefinidas.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuencia; b. filtrar lecturas que no alcanzan un umbral establecido; c. mapear lecturas de secuencia derivadas de la secuenciación en una secuencia de referencia; d. identificar un subconjunto de lecturas de secuencia mapeadas que se alinean con una variante de la secuencia de referencia en cada posición de base mapeable; e. para cada posición base mapeable, calcular una proporción de (a) una cantidad de lecturas de secuencia mapeadas que incluyen una variante en comparación con la secuencia de referencia, con (b) una cantidad de lecturas de secuencia totales para cada posición de base mapeable; f. normalizar las proporciones o la frecuencia de varianza para cada posición de base mapeable y determinar las potenciales variantes raras u otras alteraciones genéticas; y g. comparar el número resultante para cada una de las regiones con

potenciales variantes raras o mutaciones con números derivados de manera similar de un muestra de referencia.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; b. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiente a un polinucleótido único entre el conjunto de polinucleótidos originales marcados.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los pasos siguientes: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; b. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias consenso, cada secuencia de consenso correspondiendo a un polinucleótido único entre el conjunto de polinucleótidos originales marcados; c. filtrar de entre las secuencias de consenso aquellas que no logran alcanzar un umbral de calidad.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; y i. colapsar las lecturas de secuencia: 1. agrupar las lecturas de secuencias secuenciadas a partir de polinucleótidos de progenie amplificados en familias, cada familia amplificada a partir del mismo polinucleótido original marcado y, opcionalmente, 2. determinar una medida cuantitativa de las lecturas de secuencia en cada familia. En ciertas realizaciones, el código ejecutable realiza además los pasos de: b. determinar una medida cuantitativa de familias únicas; c. en base a (1) la medida cuantitativa de familias únicas y (2) la medida cuantitativa de lecturas de secuencia en cada grupo, infiriendo una medida de polinucleótidos originales marcados únicos en el conjunto. En ciertas realizaciones, el código ejecutable realiza además los pasos de: d. determinar una medida cuantitativa de formas polimórficas entre las familias; y e. en base a la medida cuantitativa determinada de formas polimórficas, deducir una medida cuantitativa de formas polimórficas en el número de polinucleótidos originales marcados únicos inferidos.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados de por lo menos un conjunto de polinucleótidos originales marcados que agrupan las lecturas de secuencias secuenciadas a partir de polinucleótidos amplificados en familias, cada familia amplificada del mismo primer polinucleótido en el conjunto; b. inferir una medida cuantitativa de familias en el conjunto; c. determinar la variación del número de copias comparando la medida cuantitativa de familias en cada conjunto.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los pasos siguientes: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados que agrupan las lecturas de secuencia en familias, cada familia comprendiendo lecturas de secuencia de polinucleótidos amplificados, amplificados a partir del mismo primer polinucleótido; b. inferir, para cada conjunto de primeros polinucleótidos, una frecuencia de llamada para una o más bases en el conjunto de primeros polinucleótidos, en donde inferir comprende: c. asignar, para cada familia, una puntuación de confianza para cada una de una pluralidad de llamadas, la puntuación de confianza teniendo en cuenta la frecuencia de la llamada entre los miembros de la familia; y d. estimar una frecuencia de una o más llamadas teniendo en cuenta las puntuaciones de confianza de una o más llamadas asignadas a cada familia.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a datos accediendo a un archivo de datos que comprende una señal recibida que comprende información de secuencia codificada de por lo menos una molécula de polinucleótido individual en donde la señal recibida comprende ruido y/o distorsión; b. decodificar la señal recibida para producir un mensaje que comprende información de secuencia sobre por lo menos una molécula de polinucleótido individual, en donde la decodificación reduce el ruido y/o la distorsión sobre cada polinucleótido individual en el mensaje; y c. escribir el mensaje que comprende información de secuencia sobre la por lo menos una molécula de polinucleótido individual en un archivo de ordenador.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que

comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; b. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiendo a un polinucleótido único entre el conjunto de polinucleótidos originales marcados; c. filtrar de entre las secuencias de consenso aquellas que no logran alcanzar un umbral de calidad.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los pasos siguientes: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; y b. colapsar las lecturas de secuencia: i. agrupando las lecturas de secuencias secuenciadas a partir de los polinucleótidos de progenie amplificados en familias, cada familia amplificada del mismo polinucleótido original marcado; y ii. opcionalmente, determinando una medida cuantitativa de lecturas de secuencia en cada familia. En ciertas realizaciones, el código ejecutable realiza además los pasos de: c. determinar una medida cuantitativa de familias únicas; d. en base a (1) la medida cuantitativa de familias únicas y (2) la medida cuantitativa de lecturas de secuencia en cada grupo, inferir una medida de polinucleótidos originales marcados únicos en el conjunto. En ciertas realizaciones, el código ejecutable realiza además los pasos de: e. determinar una medida cuantitativa de formas polimórficas entre las familias; y f. en base a la medida cuantitativa determinada de formas polimórficas, inferir una medida cuantitativa de formas polimórficas en el número de polinucleótidos originales marcados únicos inferidos. En ciertas realizaciones, el código ejecutable realiza además los pasos de: e. inferir la variación en el número de copias para la pluralidad de conjuntos en base a una comparación del número inferido de polinucleótidos originales marcados en cada conjunto mapeando para cada una de una pluralidad de secuencias de referencia.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los pasos siguientes: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados; b. agrupar las lecturas de secuencias secuenciadas a partir de polinucleótidos amplificados en familias, cada familia amplificada a partir del mismo primer polinucleótido del conjunto; c. inferir una medida cuantitativa de familias en el conjunto; d. determinar la variación en el número de copias comparando la medida cuantitativa de familias en cada conjunto.

Esta divulgación también proporciona un medio legible por ordenador en forma tangible no transitoria que comprende código ejecutable configurado para realizar los siguientes pasos: a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de un conjunto de polinucleótidos de progenie amplificados a partir de por lo menos un conjunto de polinucleótidos originales marcados que agrupan las lecturas de secuencia en familias, cada familia comprendiendo lecturas de secuencia de polinucleótidos amplificados, amplificados a partir de mismo primer polinucleótido; y b. inferir, para cada conjunto de primeros polinucleótidos, una frecuencia de llamada para una o más bases en el conjunto de primeros polinucleótidos, en donde inferir comprende: i. asignar, para cada familia, una puntuación de confianza para cada una de una pluralidad de llamadas, la puntuación de confianza teniendo en cuenta una frecuencia de la llamada entre los miembros de la familia; y ii. estimar una frecuencia de una o más llamadas teniendo en cuenta las puntuaciones de confianza de la una o más llamadas asignadas a cada familia.

Esta divulgación también proporciona un método que comprende: a. proporcionar una muestra que comprende entre 100 y 100.000 equivalentes haploides del genoma humano de polinucleótidos de ADN libre de células (ADNcf; y b. marcar los polinucleótidos con entre 2 y 1.000.000 de identificadores únicos. En ciertas realizaciones, el número de identificadores únicos es de por lo menos 3, por lo menos 5, por lo menos 10, por lo menos 15 o por lo menos 25 y como máximo 100, como máximo 1000 o como máximo 10.000. En ciertas realizaciones, el número de identificadores únicos es como máximo 100, como máximo 1000, como máximo 10.000, como máximo 100.000.

Esta divulgación también proporciona un método que comprende: a. proporcionar una muestra que comprende una pluralidad de equivalentes de genoma haploide humano de polinucleótidos fragmentados; b. determinar z, en donde z es una medida de la tendencia central (por ejemplo, media, mediana o moda) del número esperado de polinucleótidos duplicados que comienzan en cualquier posición en el genoma, en donde los polinucleótidos duplicados tienen las mismas posiciones de inicio y parada; y c. marcar polinucleótidos en la muestra con n identificadores únicos, en donde n está entre 2 y 100.000*z, 2 y 10.000*z, 2 y 1.000*z o 2 y 100*z.

Esta divulgación también proporciona un método que comprende: a. proporcionar por lo menos un conjunto de polinucleótidos originales marcados, y para cada conjunto de polinucleótidos originales marcados; b. producir una pluralidad de lecturas de secuencia para cada polinucleótido original marcado en el conjunto para producir un

conjunto de lecturas de secuenciación; y c. colapsar el conjunto de lecturas de secuenciación para generar un conjunto de secuencias de consenso, cada secuencia de consenso correspondiente a un polinucleótido único entre el conjunto de polinucleótidos originales marcados. En algunas realizaciones, las variantes en el número de copias identificadas son fraccionales (es decir, niveles no enteros) debido a la heterogeneidad en la muestra. En algunas realizaciones, se realiza el enriquecimiento de regiones seleccionadas. En algunas realizaciones, la información de variación en el número de copias se extrae simultáneamente en base a los métodos descritos en la presente. En algunas realizaciones, los métodos comprenden un paso inicial de hacer un cuello de botella para limitar el número de copias de partida iniciales o diversidad de polinucleótidos en la muestra.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 10% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 20% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 30% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 40% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 50% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 60% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 70% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 80% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética o cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 90% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 10% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 20% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de

alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 30% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 40% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 50% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 60% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 70% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 80% del genoma del individuo.

La divulgación también proporciona un método que comprende detectar la presencia o ausencia de alteración genética y la cantidad de variación genética en un individuo, en donde la detección se realiza con la ayuda de secuenciación de ácido nucleico libre de células, en donde se secuencia por lo menos el 90% del genoma del individuo.

En algunas realizaciones, la alteración genética es la variación en el número de copias o una o más mutaciones raras. En algunas realizaciones, la variación genética comprende una o más variantes causales y uno o más polimorfismos. En algunas realizaciones, la alteración genética y/o la cantidad de variación genética en el individuo pueden compararse con una alteración genética y/o cantidad de variación genética en uno o más individuos con una enfermedad conocida. En algunas realizaciones, la alteración genética y/o la cantidad de variación genética en el individuo pueden compararse con una alteración genética y/o cantidad de variación genética en uno o más individuos, sin una enfermedad. En algunas realizaciones, el ácido nucleico libre de células es ADN. En algunas realizaciones, el ácido nucleico libre de células es ARN. En algunas realizaciones, el ácido nucleico libre de células es ADN y ARN. En algunas realizaciones, La enfermedad es cáncer o precáncer. En algunas realizaciones, el método comprende además el diagnóstico o tratamiento de una enfermedad.

La divulgación también proporciona una composición que comprende entre 100 y 100.000 equivalentes del genoma haploide humano de polinucleótidos de ADNcf, en donde los polinucleótidos están marcados con entre 2 y 1.000.000 de identificadores únicos.

En algunas realizaciones, la composición comprende entre 1000 y 50.000 equivalentes del genoma humano haploide de polinucleótidos de ADNcf, en donde los polinucleótidos están marcados con entre 2 y 1.000 identificadores únicos. En algunas realizaciones, los identificadores únicos comprenden códigos de barras de nucleótidos. La divulgación también proporciona un método que comprende: a) proporcionar una muestra que comprende entre 100 y 100.000 equivalentes de genoma humano haploide de polinucleótidos de ADNcf; y b) marcar los polinucleótidos con entre 2 y 1.000.000 de identificadores únicos.

La divulgación también proporciona un sistema que comprende un medio legible por ordenador que comprende código ejecutable por máquina como se describe en la presente. La divulgación también proporciona un sistema que comprende un medio legible por ordenador que comprende código ejecutable por máquina que, tras la ejecución por un procesador informático, implementa un método como se describe en la presente.

Aspectos y ventajas adicionales de la presente divulgación se harán fácilmente evidentes para los expertos en esta técnica a partir de la siguiente descripción detallada, en la que solo se muestran y describen realizaciones ilustrativas de la presente divulgación. Como se entenderá, la presente divulgación es capaz de otras realizaciones diferentes, y sus varios detalles son capaces de modificaciones en varios aspectos obvios, todos sin apartarse de la

divulgación. Por consiguiente, los dibujos y la descripción deben considerarse de naturaleza ilustrativa y no restrictiva.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Las características novedosas de un sistema y los métodos de esta divulgación se exponen con particularidad en las reivindicaciones adjuntas. Se obtendrá una mejor comprensión de las características y ventajas de esta divulgación haciendo referencia a la siguiente descripción detallada que describe realizaciones ilustrativas, en las que se utilizan los principios de un sistema y métodos de esta divulgación, y los dibujos acompañantes de los cuales:

La **FIG. 1** es una representación en diagrama de flujo de un método de detección de la variación del número de copias usando una única muestra.
La **FIG. 2** es una representación en diagrama de flujo de un método de detección de la variación del número de copias usando muestras emparejadas.
La **FIG. 3** es una representación en diagrama de flujo de un método de detección de mutaciones raras (por ejemplo, variantes de un único nucleótido).
La **FIG. 4A** es un informe de detección de variación de número de copias gráfico generado a partir de un sujeto normal, no canceroso.
La **FIG. 4B** es un informe de detección de variación de número de copias gráfico generado a partir de un sujeto con cáncer de próstata.
La **FIG. 4C** es una representación esquemática del acceso habilitado para Internet de informes generados a partir del análisis de la variación del número de copias de un sujeto con cáncer de próstata.
La **FIG. 5A** es un informe de detección de variación de número de copias gráfico generado a partir de un sujeto con remisión de cáncer de próstata.
La **FIG. 5B** es un informe de detección de variación de número de copias gráfico generado a partir de un sujeto con recurrencia de cáncer de próstata.
La **FIG. 6A** es un informe de detección gráfico (por ejemplo, para variantes de un único nucleótido) generado a partir de varios experimentos de mezcla que usan muestras de ADN que contienen copias tanto de tipo salvaje como mutantes de MET y TP53.
La **FIG. 6B** es una representación gráfica logarítmica de los resultados de detección (por ejemplo, variante de un único nucleótido). Las mediciones del porcentaje de cáncer observado frente al esperado se muestran para varios experimentos de mezcla que usan muestras de ADN que contienen copias tanto de tipo salvaje como mutantes de MET, HRAS y TP53.
La **FIG. 7A** es un informe gráfico del porcentaje de dos (por ejemplo, variantes de un único nucleótido) en dos genes, PIK3CA y TP53, en un sujeto con cáncer de próstata en comparación con una referencia (control).
La **FIG. 7B** es una representación esquemática del acceso habilitado para Internet de los informes generados a partir del análisis (por ejemplo, variante de un único nucleótido) de un sujeto con cáncer de próstata.
La **FIG. 8** es una representación en diagrama de flujo de un método de análisis de material genético.
La **FIG. 9** es una representación en diagrama de flujo de un método para decodificar información en un conjunto de lecturas de secuencia para producir, con ruido y/o distorsión reducidos, una representación de información en un conjunto de polinucleótidos parentales marcados.
La **FIG. 10** es una representación en diagrama de flujo de un método para reducir la distorsión en la determinación de CNV a partir de un conjunto de lecturas de secuencia.
La **FIG. 11** es una representación en diagrama de flujo de un método para estimar la frecuencia de una base o secuencia de bases en un locus en una población de polinucleótidos parentales marcados de un conjunto de lecturas de secuencia.
La **FIG. 12** muestra un método para comunicar la información de secuencia.
La **FIG. 13** muestra las frecuencias alélicas menores detectadas en un panel completo de 70 kb en una titulación de ADNcf de LNCaP al 0,3% usando secuenciación estándar y flujos de trabajo de secuenciación digital. La secuenciación "analógica" estándar (Fig. 13A) enmascara todas las variantes raras verdaderas positivas en un ruido tremendo debido a los errores de PCR y secuenciación a pesar del filtrado Q30. La secuenciación digital (Fig. 13B) elimina toda el ruido de PCR y la secuenciación, revelando mutaciones verdaderas sin falsos positivos: los círculos verdes son puntos SNP en el ADNcf normal y los círculos rojos son mutaciones de LNCaP detectadas.
La **FIG. 14 :** Muestra la titulación de ADNcf de LNCaP.
La **FIG. 15** muestra un sistema informático que está programado o configurado de otra manera para implementar varios métodos de la presente divulgación.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

I. Descripción General

La presente divulgación proporciona un sistema y un método para la detección de mutaciones raras (por ejemplo, variaciones de nucleótidos individuales o múltiples) y variaciones en el número de copias en polinucleótidos

libres de células. En general, los sistemas y métodos comprenden la preparación de muestras, o la extracción y aislamiento de secuencias de polinucleótidos libres de células de un fluido corporal; la secuenciación posterior de los polinucleótidos libres de células mediante técnicas conocidas en la técnica; y la aplicación de herramientas bioinformáticas para detectar mutaciones raras y variaciones en el número de copias en comparación con una referencia. Los sistemas y métodos también pueden contener una base de datos o una colección de diferentes mutaciones raras o perfiles de variación del número de copias de diferentes enfermedades, que se usan como referencias adicionales para ayudar a la detección de mutaciones raras (por ejemplo, realización de perfiles de variación de nucleótidos individuales), realización de perfiles de la variación del número de copias o la realización de perfiles genéticos generales de una enfermedad.

Los sistemas y métodos pueden ser particularmente útiles en el análisis de ADN libre de células. En algunos casos, el ADN libre de células se extrae y se aísla de un fluido corporal fácilmente accesible como la sangre. Por ejemplo, el ADN libre de células puede extraerse usando una variedad de métodos conocidos en la técnica, que incluyen, pero no están limitados a, la precipitación con isopropanol y/o la purificación basada en sílice. El ADN libre de células puede extraerse de cualquier número de sujetos, como sujetos sin cáncer, sujetos con riesgo de cáncer, o sujetos que se sabe que tienen cáncer (por ejemplo, mediante de otros medios).

Después del paso de aislamiento/extracción, se puede realizar cualquiera de una serie de operaciones de secuenciación diferentes en la muestra de polinucleótido libre de células. Las muestras pueden procesarse antes de la secuenciación con uno o más reactivos (por ejemplo, enzimas, identificadores únicos (por ejemplo, códigos de barras), sondas, etc.). En algunos casos, si la muestra se procesa con un identificador único como un código de barras, las muestras o fragmentos de muestras pueden marcarse individualmente o en subgrupos con el identificador único. La muestra marcada puede usarse luego en una aplicación posterior tal como una reacción de secuenciación por la cual las moléculas individuales pueden rastrearse hasta las moléculas parentales.

Después de recopilar los datos de secuenciación de las secuencias de polinucleótidos libres de células, se pueden aplicar uno o más procesos bioinformáticos a los datos de secuencia para detectar características genéticas o aberraciones como la variación del número de copias, mutaciones raras (por ejemplo, variaciones de nucleótidos individuales o múltiples) o cambios en marcadores epigenéticos, incluyendo pero no limitados a, los perfiles de metilación. En algunos casos, en los que se desea un análisis de la variación del número de copias, los datos de secuencia pueden: 1) alinearse con un genoma de referencia; 2) filtrarse y mapearse; 3) repartirlos en ventanas o recipientes de secuencia; 4) contarse las lecturas de cobertura para cada ventana; 5) las lecturas de cobertura se pueden normalizar luego usando un algoritmo de modelado estocástico o estadístico; 6) y se puede generar un archivo de salida que refleje estados de número de copias discretos en varias posiciones en el genoma. En otros casos, en los que se desea un análisis de mutaciones raras, los datos de secuencia pueden 1) alinearse con un genoma de referencia; 2) filtrarse y mapearse; 3) calcularse la frecuencia de las bases de variantes en base a las lecturas de cobertura para esa base específica; 4) normalizarse la frecuencia de bases de variantes usando un algoritmo de modelado estocástico, estadístico o probabilístico; 5) y se puede generar un archivo de salida que refleje los estados de mutaciones en varias posiciones en el genoma.

Pueden tener lugar una variedad de reacciones y/o operaciones diferentes dentro de los sistemas y métodos divulgados en la presente, que incluyen, pero no están limitados a: secuenciación de ácidos nucleicos, cuantificación de ácidos nucleicos, optimización de la secuenciación, detección de la expresión génica, cuantificación de la expresión génica, realizar perfiles genómicos, realizar perfiles de cáncer o análisis de marcadores expresados. Además, los sistemas y métodos tienen numerosas aplicaciones médicas. Por ejemplo, puede usarse para la identificación, detección, diagnóstico, tratamiento, clasificación del estadio, o predicción de riesgo de varias enfermedades y trastornos genéticos y no genéticos, incluyendo el cáncer. Puede usarse para evaluar la respuesta de los sujetos a diferentes tratamientos de dichas enfermedades genéticas y no genéticas, o proporcionar información referente a la progresión de la enfermedad y el pronóstico.

La secuenciación de polinucleótidos puede compararse con un problema en la teoría de la comunicación. Un polinucleótido individual inicial o conjunto de polinucleótidos se considera como un mensaje original. Puede considerarse que el marcado y/o la amplificación codifican el mensaje original en una señal. La secuenciación puede considerarse como un canal de comunicación. La salida de un secuenciador, por ejemplo, lecturas de secuencia, puede considerarse como una señal recibida. El procesamiento bioinformático puede considerarse como un receptor que decodifica la señal recibida para producir un mensaje transmitido, por ejemplo, una secuencia o secuencias de nucleótidos. La señal recibida puede incluir artefactos, como ruido ya distorsión. El ruido puede considerarse como una adición aleatoria no deseada a una señal. La distorsión puede considerarse como una alteración en la amplitud de una señal o parte de una señal.

El ruido puede introducirse a través de errores al copiar y/o leer un polinucleótido. Por ejemplo, en un proceso de secuenciación, puede primero someterse un único polinucleótido a amplificación. La amplificación puede introducir errores, por lo que un subconjunto de los polinucleótidos amplificados puede contener, en un locus particular, una base que no es la misma que la base original en ese locus. Además, en el proceso de lectura una base en cualquier locus en particular puede leerse incorrectamente. Como consecuencia, la colección de lecturas de

secuencia puede incluir un cierto porcentaje de llamadas de base en un locus que no son las mismas que las de la base original. En las tecnologías de secuenciación típicas, esta tasa de error puede estar en dígitos únicos, por ejemplo, 2% -3%. Cuando se secuencia una colección de moléculas que se presume que tienen la misma secuencia, este ruido es lo suficientemente pequeño para que se pueda identificar la base original con una fiabilidad alta.

Sin embargo, si una colección de polinucleótidos parentales incluye un subconjunto de polinucleótidos que tienen variantes de secuencia en un locus particular, el ruido puede ser un problema importante. Este puede ser el caso, por ejemplo, cuando el ADN libre de células incluye no solo el ADN de la línea germinal, sino también el ADN de otra fuente, como ADN fetal o ADN de una célula cancerosa. En este caso, si la frecuencia de moléculas con variantes de secuencia está en el mismo rango que la frecuencia de errores introducida por el proceso de secuenciación, entonces las verdaderas variantes de secuencia pueden no ser distinguibles del ruido. Esto podría interferir, por ejemplo, con la detección de variantes de secuencia en una muestra.

La distorsión puede manifestarse en el proceso de secuenciación como una diferencia en la intensidad de la señal, por ejemplo, el número total de lecturas de secuencia, producidas por moléculas en una población parental a la misma frecuencia. La distorsión puede introducirse, por ejemplo, a través del sesgo de amplificación, sesgo de GC o sesgo de secuenciación. Esto podría interferir con la detección de la variación del número de copias en una muestra. El sesgo de GC da como resultado la representación desigual de áreas ricas o pobres en contenido de GC en la lectura de secuencia.

Esta invención proporciona métodos para reducir los artefactos de secuenciación, como el ruido y/o la distorsión, en un proceso de secuenciación de polinucleótidos. Las lecturas de secuencia de agrupación en familias derivadas de moléculas individuales originales pueden reducir el ruido y/o la distorsión de una molécula individual única o de un conjunto de moléculas. Con respecto a una molécula individual, agrupar las lecturas en una familia reduce la distorsión, por ejemplo, indicando que muchas lecturas de secuencias representan en realidad una sola molécula en lugar de muchas moléculas diferentes. La lectura de secuencias colapsadas en una secuencia de consenso es una manera de reducir el ruido en el mensaje recibido de una molécula. El uso de funciones probabilísticas que convierten las frecuencias recibidas es otra manera. Con respecto a un conjunto de moléculas, agrupar las lecturas en familias y determinar una medida cuantitativa de las familias reduce la distorsión, por ejemplo, en la cantidad de moléculas en cada una de una pluralidad de loci diferentes. De nuevo, las lecturas de secuencias colapsadas de diferentes familias en secuencias de consenso eliminan los errores introducidos por error de amplificación y/o secuenciación. Además, la determinación de las frecuencias de las llamadas de base en base a las probabilidades derivadas de la información de la familia también reduce el ruido en el mensaje recibido de un conjunto de moléculas.

Se conocen métodos para reducir el ruido y/o la distorsión de un proceso de secuenciación. Estos incluyen, por ejemplo, secuencias de filtrado, por ejemplo, requerir que cumplan con un umbral de calidad, o reducir el sesgo de GC. Tales métodos se realizan típicamente en la colección de lecturas de secuencia que son la salida de un secuenciador, y se pueden realizar lecturas de lectura por secuencia, sin tener en cuenta la estructura de la familia (sub-colecciones de secuencias derivadas de una única molécula parental original). Ciertos métodos de esta invención reducen el ruido y la distorsión reduciendo el ruido y/o la distorsión dentro de familias de lecturas de secuencia, es decir, funcionando en lecturas de secuencia agrupadas en familias derivadas de una única molécula de polinucleótido parental. La reducción del artefacto de la señal a nivel de familia puede producir significativamente menos ruido y distorsión en el último mensaje que se proporciona que la reducción de artefactos realizada a un nivel de lectura de lectura por secuencia o en una salida de secuenciador como un todo.

La presente divulgación proporciona además métodos y sistemas para detectar con alta sensibilidad variación genética en una muestra de material genético inicial. Los métodos implican el uso de una o ambas de las siguientes herramientas: Primero, la conversión eficaz de polinucleótidos individuales en una muestra de material genético inicial en polinucleótidos parentales marcados listos para secuencia, para aumentar la probabilidad de que los polinucleótidos individuales en una muestra de material genético inicial sean representados en una muestra lista para secuencia. Esto puede producir información de secuencia sobre más polinucleótidos en la muestra inicial. En segundo lugar, la generación con alto rendimiento de secuencias de consenso para polinucleótidos parentales marcados mediante muestreo de alta velocidad de polinucleótidos de la progenie amplificada a partir de los polinucleótidos parentales marcados, y el colapso de las lecturas de secuencias generadas en secuencias de consenso que representan secuencias de polinucleótidos marcados parentales. Esto puede reducir el ruido introducido por el sesgo de amplificación y/o los errores de secuenciación, y puede aumentar la sensibilidad de la detección. El colapso se realiza en una pluralidad de lecturas de secuencia, generadas o a partir de lecturas de moléculas amplificadas, o de lecturas múltiples de una única molécula.

Los métodos de secuenciación implican típicamente la preparación de muestras, la secuenciación de polinucleótidos en la muestra preparada para producir lecturas de secuencia y la manipulación bioinformática de las lecturas de secuencia para producir información genética cuantitativa y/o cualitativa sobre la muestra. La preparación de muestras implica generalmente convertir polinucleótidos en una muestra en una forma compatible con la

plataforma de secuenciación usada. Esta conversión puede implicar el marcado de polinucleótidos. En ciertas realizaciones de esta invención, los marcadores comprenden marcadores de secuencia de polinucleótidos. Las metodologías de conversión usadas en la secuenciación pueden no ser 100% eficientes. Por ejemplo, no es infrecuente convertir polinucleótidos en una muestra con una eficiencia de conversión de aproximadamente el 1-5%, es decir, aproximadamente el 1-5% de los polinucleótidos en una muestra se convierten en polinucleótidos marcados. Los polinucleótidos que no se convierten en moléculas marcadas no se representan en una biblioteca marcada para la secuenciación. Por consiguiente, los polinucleótidos que tienen variantes genéticas representadas a baja frecuencia en el material genético inicial pueden no estar representados en la biblioteca marcada y, por lo tanto pueden no secuenciarse o detectarse. Aumentando la eficiencia de conversión, se aumenta la probabilidad de que un polinucleótido raro en el material genético inicial se represente en la biblioteca marcada y, por consiguiente, se detecte mediante secuenciación. Además, en lugar de abordar directamente el problema de la baja eficiencia de conversión de la preparación de la biblioteca, la mayoría de los protocolos hasta la fecha requieren más de 1 microgramo de ADN como material de entrada. Sin embargo, cuando el material de la muestra de entrada es limitado o se desea la detección de polinucleótidos con baja representación, la alta eficiencia de conversión puede secuenciar eficientemente la muestra y/o detectar adecuadamente tales polinucleótidos.

Esta divulgación proporciona métodos para convertir polinucleótidos iniciales en polinucleótidos marcados con una eficiencia de conversión de por lo menos el 10%, por lo menos el 20%, por lo menos el 30%, por lo menos el 40%, por lo menos el 50%, por lo menos el 60%, por lo menos el 80%, o por lo menos el 90%. Los métodos implican, por ejemplo, usar cualquiera de ligación de extremo romo, ligación del extremo pegajoso, sondas de inversión molecular, PCR, PCR basada en la ligación, PCR multiplex, ligación de cadena sencilla y circularización de cadena sencilla. Los métodos también pueden implicar limitar la cantidad de material genético inicial. Por ejemplo, la cantidad de material genético inicial puede ser inferior a 1 ug, inferior a 100 ng o inferior a 10 ng. Estos métodos se describen con más detalle en la presente.

La obtención de información cuantitativa y cualitativa precisa sobre los polinucleótidos en una biblioteca marcada puede dar como resultado una caracterización más sensible del material genético inicial. Típicamente, los polinucleótidos en una biblioteca marcada se amplifican y las moléculas amplificadas resultantes se secuencian. Dependiendo del rendimiento de la plataforma de secuenciación usada, solo un subconjunto de las moléculas en la biblioteca amplificada produce lecturas de secuencia. Entonces, por ejemplo, el número de moléculas amplificadas muestreadas para la secuenciación puede ser aproximadamente el 50% de los polinucleótidos únicos en la biblioteca marcada. Además, la amplificación puede sesgarse a favor o en contra de ciertas secuencias o ciertos miembros de la biblioteca marcada. Esto puede distorsionar la medición cuantitativa de las secuencias en la biblioteca marcada. Además, las plataformas de secuenciación pueden introducir errores en la secuenciación. Por ejemplo, las secuencias pueden tener una tasa de error por base del 0,5-1%. El sesgo de amplificación y los errores de secuenciación introducen ruido en el producto de la secuenciación final. Este ruido puede disminuir la sensibilidad de detección. Por ejemplo, las variantes de secuencia cuya frecuencia en la población marcada sea menor que la tasa de error de secuenciación pueden confundirse con ruido. Además, proporcionando lecturas de secuencias en cantidades mayores o menores que su número real en una población, el sesgo de amplificación puede distorsionar las mediciones de la variación del número de copias. Alternativamente, pueden producirse una pluralidad de lecturas de secuencia de un único polinucleótido sin amplificación. Esto puede hacerse, por ejemplo, con métodos de nanopore.

Esta divulgación proporciona métodos para detectar y leer con precisión polinucleótidos únicos en un grupo marcado. En ciertas realizaciones, esta divulgación proporciona polinucleótidos marcados en secuencia que, cuando se amplifican y secuencian, o cuando se secuencian una pluralidad de veces para producir una pluralidad de lecturas de secuencia, proporcionan información que permite el rastreo, o el colapso, de polinucleótidos de la progenie a la molécula de polinucleótido parental de marcador único. Las familias colapsadas de polinucleótidos de progenie amplificada reducen el sesgo de amplificación proporcionando información sobre las moléculas parentales únicas originales. El colapso también reduce los errores de secuenciación al eliminar de los datos de secuenciación secuencias mutantes de las moléculas de la progenie.

Detectar y leer polinucleótidos únicos en la biblioteca marcada puede implicar dos estrategias. En una estrategia, un subconjunto lo suficientemente grande del grupo de polinucleótidos de la progenie amplificada se secuencia de tal manera que, para un gran porcentaje de polinucleótidos parentales únicos marcados en el conjunto de polinucleótidos parentales marcados, hay una lectura de secuencia que se produce para al menos un polinucleótido de la progenie amplificada en una familia producida a partir de un polinucleótido parental marcado único. En una segunda estrategia, el conjunto de polinucleótidos de la progenie amplificado se muestrea para secuenciación a un nivel para producir lecturas de secuencia de múltiples miembros de la progenie de una familia derivada de un polinucleótido parental único. La generación de lecturas de secuencia de múltiples miembros de la progenie de una familia permite el colapso de las secuencias en secuencias parentales de consenso.

Así, por ejemplo, muestrear una serie de polinucleótidos de la progenie amplificada del conjunto de polinucleótidos de la progenie amplificada que es igual al número de polinucleótidos parentales marcados únicos en el conjunto de polinucleótidos parentales marcados (particularmente cuando el número es por lo menos 10.000)

producirá, estadísticamente, una lectura de secuencia para por lo menos una de la progenie de aproximadamente el 68% de los polinucleótidos parentales marcados en el conjunto, y aproximadamente el 40% de los polinucleótidos parentales marcados únicos en el conjunto original estará representada por al menos dos lecturas de secuencia de la progenie. En ciertas realizaciones, el conjunto de polinucleótidos de la progenie amplificada se muestrea lo suficiente para producir una media de cinco a diez lecturas de secuencia para cada familia. El muestreo del conjunto de la progenie amplificado de 10 veces la cantidad de moléculas del número de polinucleótidos parentales marcados únicos, producirá estadísticamente, la información de secuencia sobre el 99,995% de las familias, de las cuales el 99,95% del total de familias se cubrirá con una pluralidad de lecturas de secuencias. Puede construirse una secuencia de consenso a partir de los polinucleótidos de la progenie en cada familia para reducir drásticamente la tasa de error de la tasa de error de secuenciación por base nominal a una tasa posiblemente de muchos órdenes de magnitud más baja. Por ejemplo, si el secuenciador tiene una tasa de error por base aleatoria del 1% y la familia elegida tiene 10 lecturas, una secuencia de consenso construida a partir de estas 10 lecturas tendría una tasa de error de menos del 0,0001%. Por consiguiente, el tamaño de muestreo de la progenie amplificada que se va a secuenciar puede elegirse para garantizar que una secuencia que tenga una frecuencia en la muestra que no sea mayor que la tasa de error de secuenciación por base nominal a la tasa de la plataforma de secuenciación usada, tiene por lo menos un 99% de probabilidad de estar representada por al menos una lectura.

En otra realización, el conjunto de polinucleótidos de la progenie amplificada se muestrea a un nivel para producir una alta probabilidad, por ejemplo, por lo menos el 90%, de que una secuencia representada en el conjunto de polinucleótidos parentales marcados a una frecuencia que es aproximadamente la misma que la tasa de error de secuenciación por base de la plataforma de secuenciación usada está cubierta por al menos una lectura de secuencia y preferiblemente una pluralidad de lecturas de secuencia. Así que, por ejemplo, si la plataforma de secuenciación tiene una tasa de error por base del 0,2% en una secuencia o conjunto de secuencias, se representa en el conjunto de polinucleótidos parentales marcados a una frecuencia de aproximadamente el 0,2%, entonces el número de polinucleótidos en el grupo de la progenie amplificada que se está secuenciado puede ser aproximadamente X veces el número de moléculas únicas en el conjunto de polinucleótidos parentales marcados.

Estos métodos pueden combinarse con cualquiera de los métodos de reducción de ruido descritos. Incluyendo, por ejemplo, las lecturas de secuencias de calificación para su inclusión en el grupo de secuencias usadas para generar secuencias de consenso.

Esta información puede usarse ahora para análisis tanto cualitativos como cuantitativos. Por ejemplo, para el análisis cuantitativo, se determina una medida, por ejemplo, un recuento, de la cantidad de moléculas parentales marcadas que mapean en una secuencia de referencia. Esta medida puede compararse con una medida del mapeo de moléculas parentales marcadas a una región genómica diferente. Es decir, la cantidad de moléculas parentales marcadas que se mapean en una primera localización o posición mapeable a una secuencia de referencia, como el genoma humano, puede compararse con una medida de las moléculas parentales marcadas que mapean en una segunda localización o posición mapeable a una secuencia de referencia . Esta comparación puede revelar, por ejemplo, las cantidades relativas de las moléculas parentales que mapean en cada región. Esto, a su vez, proporciona una indicación de la variación del número de copias para moléculas que mapean en una región particular. Por ejemplo, Si la medida de los polinucleótidos que mapean en una primera secuencia de referencia es mayor que la medida de los polinucleótidos que mapean en una segunda secuencia de referencia, esto puede indicar que la población parental, y por extensión la muestra original, incluía polinucleótidos de células que mostraban aneuploidía. Las medidas se pueden normalizar frente a una muestra de control para eliminar varios sesgos. Las medidas cuantitativas pueden incluir, por ejemplo, número, recuento, frecuencia (ya sea relativa, inferida o absoluta).

Un genoma de referencia puede incluir el genoma de cualquier especie de interés. Las secuencias del genoma humano útiles como referencias pueden incluir el conjunto hg19 o cualquier conjunto hg anterior o disponible. Tales secuencias pueden consultarse usando el navegador del genoma disponible en genome.ucsc.edu/index.html. Otras especies de genomas incluyen, por ejemplo, PanTro2 (chimpancé) y mm9 (ratón).

Para el análisis cualitativo, las secuencias de un conjunto de polinucleótidos marcados que mapean en una secuencia de referencia pueden analizarse para detectar secuencias variantes y puede medirse su frecuencia en la población de polinucleótidos parentales marcados.

II. Preparación de la Muestra

A. Aislamiento y Extracción de Polinucleótidos

Los sistemas y métodos de esta divulgación pueden tener una amplia variedad de usos en la manipulación, preparación, identificación y/o cuantificación de polinucleótidos libres de células. Los ejemplos de polinucleótidos incluyen, pero no están limitados a: ADN, ARN, amplicones, ADNc, ADNds, ADNss, ADN plásmido, ADN cósmido, ADN de alto peso molecular (MW), ADN cromosómico, ADN genómico, ADN viral, ADN bacteriano, ADNmt (ADN

mitocondrial), ARNm, ARNr, ARNt, ARNn, ARNsi, ARNsn, ARNsno, ARNsca, ARNmicro, ARNds, ribozima, riboswitch y ARN viral (por ejemplo, ARN retroviral).

Los polinucleótidos libres de células pueden derivarse de una variedad de fuentes que incluyen fuentes humanas, de mamíferos, de mamíferos no humanos, de simios, de monos, de chimpancés, de reptiles, de anfibios o de aves. Además, las muestras pueden extraerse de una variedad de fluidos animales que contienen secuencias libres de células, que incluyen, pero no están limitadas a, sangre, suero, plasma, vítreo, esputo, orina, lágrimas, transpiración, saliva, semen, excreciones de mucosas, moco, fluido espinal, fluido amniótico, fluido linfático y similares. Los polinucleótidos libres de células pueden ser de origen fetal (a través de un fluido tomado de un sujeto embarazado), o pueden derivar de tejido del propio sujeto.

El aislamiento y la extracción de polinucleótidos libres de células pueden realizarse mediante la recolección de fluidos corporales usando una variedad de técnicas. En algunos casos, la recolección puede comprender la aspiración de un fluido corporal de un sujeto usando una jeringuilla. En otros casos, la recolección puede comprender pipetear o recolectar directamente el fluido en un recipiente de recolección.

Después de la recolección de líquido corporal, los polinucleótidos libres de células pueden aislarse y extraerse usando una variedad de técnicas conocidas en la técnica. En algunos casos, el ADN libre de células puede aislarse, extraerse y prepararse usando kits disponibles comercialmente, como el protocolo del kit de ácido nucleico circulante Qiagen Qiamp®. En otros ejemplos, pueden usarse el protocolo del kit de ensayo de ADNds HS Qiagen Qubit™, el kit Agilent™ DNA 1000 o la preparación de la biblioteca de secuenciación TruSeq™ protocolo de bajo rendimiento (LT).

Generalmente, los polinucleótidos libres de células se extraen y aíslan de los fluidos corporales a través de un paso de división en el que los ADN libres de células, como se encuentran en la solución, se separan de las células y otros componentes no solubles del fluido corporal. La partición división incluir, pero no está limitada a, técnicas como la centrifugación o la filtración. En otros casos, las células no se dividen a partir de ADN libre de células primero, sino que se lisan. En este ejemplo, el ADN genómico de las células intactas se divide mediante precipitación selectiva. Los polinucleótidos libres de células, incluido el ADN, pueden permanecer solubles y pueden separarse del ADN genómico insoluble y extraerse. Generalmente, después de la adición de tampones y otros pasos de lavado específicos para diferentes kits, el ADN puede precipitarse usando precipitación con isopropanol. Se pueden usar pasos de limpieza adicionales, como columnas a base de sílice para eliminar contaminantes o sales. Los pasos generales se pueden optimizar para aplicaciones específicas. Pueden añadirse polinucleótidos portadores a granel no específicos, por ejemplo, a lo largo de la reacción para optimizar ciertos aspectos del procedimiento, como el rendimiento.

El aislamiento y la purificación del ADN libre de células puede lograrse mediante cualquier medio, incluyendo, pero no limitado a, el uso de kits y protocolos comerciales proporcionados por compañías como Sigma Aldrich, Life Technologies, Promega, Affymetrix, IBI o similares. Los kits y protocolos también pueden no estar disponibles comercialmente.

Después del aislamiento, en algunos casos, los polinucleótidos libres de células se mezclan previamente con uno o más materiales adicionales, como uno o más reactivos (por ejemplo, ligasa, proteasa, polimerasa) antes de la secuenciación.

Un método para aumentar la eficiencia de conversión implica el uso de una ligasa diseñada para una reactividad óptima en el ADN de cadena sencilla, como un derivado de la ligasa de ADNss de ThermoPhage. Dichas ligasas omiten los pasos tradicionales en la preparación de bibliotecas de la reparación final y la formación de cola A que pueden tener eficiencias pobres y/o pérdidas acumuladas debido a los pasos de limpieza intermedios, y permite el doble de probabilidades de que el polinucleótido de inicio de sentido o antisentido se convierta en un polinucleótido apropiadamente marcado. También convierte polinucleótidos de cadena doble que pueden poseer salientes que pueden no ser lo suficientemente romos por la reacción de reparación final típica. Las condiciones de reacción óptimas para esta reacción de ADNss son: 1 x tampón de reacción (50 mM MOPS (pH 7,5), DTT 1 mM, MgCl2 5 mM, KCl 10 mM). Con ATP 50 mM, 25 mg/ml de BSA, MnCl2 2,5 mM, 200 pmol 85 nt de oligómero de ADNss y 5 U de ligasa de ADNss incubados a 65º C durante 1 hora. La amplificación posterior usando PCR puede convertir aún más la biblioteca de cadena sencilla marcada en una biblioteca de cadena doble y producir una eficiencia de conversión global muy por encima del 20%. Otros métodos para aumentar la tasa de conversión, por ejemplo, por encima del 10%, incluyen, por ejemplo, cualquiera de los siguientes, solos o en combinación: sondas de inversión molecular con apareamiento optimizado, ligación de extremo romo con un intervalo de tamaño de polinucleótido bien controlado, ligación de extremo pegajoso o un paso de amplificación multiplex frontal con o sin el uso de cebadores de fusión.

B. Codificación de Barras Molecular de Polinucleótidos Libres de Células

Los sistemas y métodos de esta divulgación también pueden permitir que los polinucleótidos libres de

células se marquen o rastreen para permitir la posterior identificación y origen del polinucleótido particular. Esta característica contrasta con otros métodos que usan reacciones agrupadas o multiplexadas y que solo proporcionan mediciones o análisis como una media de múltiples muestras. Aquí, la asignación de un identificador a polinucleótidos individuales o subgrupos de polinucleótidos puede permitir que se asigne una identidad única a las secuencias individuales o fragmentos de secuencias. Esto puede permitir la adquisición de datos de muestras individuales y no se limita a medias de muestras.

En algunos ejemplos, los ácidos nucleicos u otras moléculas derivadas de una cadena sencilla pueden compartir un marcador o identificador común y, por lo tanto, pueden identificarse posteriormente como derivadas de esa cadena. De manera similar, todos los fragmentos de una cadena sencilla de ácido nucleico pueden marcarse con el mismo identificador o marcador, permitiendo de este modo una identificación posterior de los fragmentos de la cadena parental. En otros casos, los productos de la expresión génica (por ejemplo, ARNm) pueden marcarse para cuantificar la expresión, por lo que se puede contar el código de barras o el código de barras en combinación con la secuencia a la que está unido. En otros casos más, los sistemas y métodos pueden usarse como control de amplificación por PCR. En tales casos, los productos de amplificación múltiple de una reacción de PCR se pueden marcar con el mismo marcador o identificador. Si los productos se secuencian posteriormente y demuestras diferencias de secuencia, las diferencias entre productos con el mismo identificador pueden atribuirse luego a un error de PCR.

Adicionalmente, pueden identificarse las secuencias individuales en base a las características de los datos de secuencia para las mismas lecturas. Por ejemplo, puede usarse la detección de datos de secuencia únicos en las partes de principio (inicio) y final (parada) de las lecturas de secuencia individuales, sola o en combinación, con la longitud o el número de pares de bases de cada lectura de secuencia para asignar identidades únicas a moléculas individuales. Los fragmentos de una cadena sencilla de ácido nucleico, a los que se les ha asignado una identidad única, pueden por lo tanto permitir la identificación posterior de fragmentos de la cadena parental. Esto se puede usar junto con hacer cuello de botella del material genético de partida inicial para limitar la diversidad.

Además, el uso de datos de secuencia únicos en las partes de principio (inicio) y final (parada) de las lecturas de secuenciación individuales y la longitud de lectura de secuenciación pueden usarse, solos o en combinación, con el uso de códigos de barras. En algunos casos, los códigos de barras pueden ser únicos como se describe en la presente. En otros casos, los códigos de barras en sí pueden no ser únicos. En este caso, el uso de códigos de barras no únicos, en combinación con los datos de secuencia en las partes de principio (inicio) y al final (parada) de las lecturas de secuencia individuales y la longitud de lectura de secuenciación puede permitir la asignación de una identidad única a secuencias individuales. De manera similar, a los fragmentos de una cadena sencilla de ácido nucleico a los que se les ha asignado una identidad única, pueden de este modo permitir la identificación posterior de los fragmentos de la cadena parental.

Generalmente, los métodos y sistemas proporcionados en la presente son útiles para la preparación de secuencias de polinucleótidos libres de células para una reacción de secuenciación de aplicación en sentido descendente. A menudo, un método de secuenciación es la secuenciación clásica de Sanger. Los métodos de secuenciación pueden incluir, pero no están limitados a: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de molécula individuales, secuenciación de nanoporos, secuenciación por semiconductores, secuenciación por ligación, secuenciación por hibridación, ARN-Seq (Illumina), Expresión génica digital (Helicos), secuenciación de próxima generación, secuenciación de moléculas individuales por síntesis (SMSS) (Helicos), secuenciación masivamente paralela, Matriz de moléculas individuales clonal (Solexa), secuenciación aleatoria, secuenciación de Maxim-Gilbert, caminata de cebadores, y cualquier otro método de secuenciación conocido en la técnica.

C. Asignación de Códigos de Barras a Secuencias de Polinucleótidos Libres de Células

Los sistemas y métodos divulgados en la presente pueden usarse en aplicaciones que implican la asignación de identificadores únicos o no únicos, o códigos de barras moleculares, a polinucleótidos libres de células. A menudo, el identificador es un oligonucleótido de código de barras que se utiliza para marcar el polinucleótido; pero, en algunos casos, se usan identificadores únicos diferentes. Por ejemplo, en algunos casos, el identificador único es una sonda de hibridación. En otros casos, el identificador único es un colorante, en cuyo caso la unión puede comprender la intercalación del colorante en la molécula de analito (como la intercalación en ADN o ARN) o la unión a una sonda marcada con el colorante. En otros casos más, el identificador único puede ser un oligonucleótido de ácido nucleico, en cuyo caso, la unión a las secuencias de polinucleótidos puede comprender una reacción de ligación entre el oligonucleótido y las secuencias o la incorporación a través de la PCR. En otros casos, la reacción puede comprender la adición de un isótopo metálico, ya sea directamente al analito o mediante una sonda marcada con el isótopo. Generalmente, la asignación de identificadores únicos o no únicos, o códigos de barras moleculares en las reacciones de esta divulgación puede seguir los métodos y sistemas descritos, por ejemplo, por las Solicitudes de Patente de Estados Unidos. 20010053519 , 20030152490 , 20110160078 y la Patente de Estados Unidos US 6.582.908.

A menudo, el método comprende unir los códigos de barras de oligonucleótidos a los analitos de ácidos nucleicos mediante una reacción enzimática que incluye, pero no está limitada a, una reacción de ligación. Por ejemplo, la enzima ligasa puede unir covalentemente un código de barras de ADN a ADN fragmentado (por ejemplo, ADN de alto peso molecular). Tras la unión de los códigos de barras, las moléculas pueden someterse a una reacción de secuenciación.

Sin embargo, también pueden usarse otras reacciones. Por ejemplo, pueden usarse cebadores de oligonucleótidos que contienen secuencias de códigos de barras en reacciones de amplificación (por ejemplo, PCR, qPCR, PCR con transcriptasa inversa, PCR digital, etc.) de los analitos de la plantilla de ADN, produciendo de este modo analitos marcados. Después de la asignación de códigos de barras a secuencias de polinucleótidos libres de células individuales, puede secuenciarse el grupo de moléculas.

En algunos casos, la PCR puede usarse para la amplificación global de secuencias de polinucleótidos libres de células. Esto puede comprender el uso de secuencias adaptadoras que pueden ligarse primero a diferentes moléculas seguido por amplificación por PCR usando cebadores universales. La PCR para la secuenciación puede realizarse por cualquier medio, incluyendo, pero no limitado a, el uso de kits comerciales proporcionados por Nugen (WGA kit), Life Technologies, Affymetrix, Promega, Qiagen y similares. En otros casos, pueden amplificarse solo ciertas moléculas objetivo dentro de una población de moléculas de polinucleótidos libres de células. Pueden usarse cebadores específicos, junto con ligación del adaptador, para amplificar selectivamente ciertos objetivos para la secuenciación en sentido descendente.

Los identificadores únicos (por ejemplo, códigos de barras de oligonucleótidos, anticuerpos, sondas, etc.) pueden introducirse en secuencias de polinucleótidos libres de células de forma aleatoria o no aleatoria. En algunos casos, se introducen a una proporción esperada de identificadores únicos en micropocillos. Por ejemplo, los identificadores únicos pueden cargarse de tal manera que se cargan más de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000,00 identificadores únicos por muestra de genoma. En algunos casos, los identificadores únicos pueden cargarse de tal manera que se cargan menos de aproximadamente 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 de identificadores únicos por muestra de genoma. En algunos casos, el número medio de identificadores únicos cargados por genoma de muestra es menor que, o mayor que, aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 de identificadores únicos por muestra de genoma

En algunos casos, los identificadores únicos pueden ser de una variedad de longitudes de tal manera que cada código de barras es por lo menos aproximadamente de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000 pares de bases En otros casos, los códigos de barras pueden comprender menos de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000 pares de bases.

En algunos casos, los identificadores únicos pueden ser oligonucleótidos de secuencia predeterminados o aleatorios o semi-aleatorios. En otros casos, puede usarse una pluralidad de códigos de barras de tal manera que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. En este ejemplo, los códigos de barras pueden ligarse a moléculas individuales, de tal manera que la combinación del código de barras y la secuencia a la que se puede ligar crea una secuencia única que puede rastrearse individualmente. Como se describe en la presente, la detección de códigos de barras no únicos en combinación con datos de secuencia de las partes de comienzo (inicio) y final (parada) de las lecturas de secuencia puede permitir la asignación de una identidad única a una molécula particular. La longitud, o número de pares de bases, de una lectura de secuencia individual también puede usarse para asignar una identidad única a dicha molécula. Como se describe en la presente, los fragmentos de una cadena sencilla de ácido nucleico a los que se les ha asignado una identidad única, pueden, por tanto, permitir la identificación posterior de los fragmentos de la cadena parental. De esta manera los polinucleótidos en la muestra pueden ser marcados de manera única o sustancialmente única.

Los identificadores únicos pueden usarse para marcar una amplia variedad de analitos incluyendo, pro no limitados a, moléculas de ARN o ADN. Por ejemplo, los identificadores únicos (por ejemplo, oligonucleótidos de código de barras) pueden unirse a cadenas completas de ácidos nucleicos o a fragmentos de ácidos nucleicos (por ejemplo, ADN genómico fragmentado, ARN fragmentado). Los identificadores únicos (por ejemplo, oligonucleótidos) también pueden unirse a productos de expresión génica, ADN genómico, ADN mitocondrial, ARN, ARNm y similares.

En muchas aplicaciones, puede ser importante determinar si las secuencias de polinucleótidos libres de células individuales reciben cada una un identificador único diferente (por ejemplo, código de barras de oligonucleótidos). Si la población de identificadores únicos introducidos en los sistemas y métodos no es significativamente diversa, es posible que diferentes analitos se marcen con identificadores idénticos. Los sistemas y métodos divulgados en la presente pueden permitir la detección de secuencias de polinucleótidos libres de células marcadas con el mismo identificador. En algunos casos, pueden incluirse secuencias de referencia con la población de secuencias de polinucleótidos libres de células a ser analizadas. La secuencia de referencia puede ser, por

ejemplo, un ácido nucleico con una secuencia conocida y una cantidad conocida. Si los identificadores únicos son códigos de barras de oligonucleótidos y los analitos son ácidos nucleicos, los analitos marcados pueden secuenciarse y cuantificarse posteriormente. Estos métodos pueden indicar si a uno o más fragmentos y/o analitos se les puede haber asignado un código de barras idéntico.

Un método divulgado en la presente puede comprender usar reactivos necesarios para la asignación de códigos de barras a los analitos. En el caso de reacciones de ligación, pueden cargarse reactivos que incluyen, pero no están limitados a, enzima ligasa, tampón, oligonucleótidos adaptadores, una pluralidad de códigos de barras de ADN identificadores únicos y similares en los sistemas y métodos. En el caso de enriquecimiento, pueden usarse reactivos que incluyen pero no están limitados a una pluralidad de cebadores de PCR, oligonucleótidos que contienen una secuencia de identificación única, o secuencia de código de barras, ADN polimerasa, DNTP y tampón y similares en la preparación para la secuenciación.

En general, el método y el sistema de esta divulgación pueden utilizar los métodos de la Patente de Estados Unidos 7.537.897 en el uso de códigos de barras moleculares para contar moléculas o analitos.

En una muestra que comprende ADN libre de células (ADNcf) de una pluralidad de genomas, hay cierta probabilidad de que más de un polinucleótido de diferentes genomas tenga las mismas posiciones de inicio y parada ("duplicados" o "cognados"). El número probable de duplicados que comienzan en cualquier posición es una función del número de equivalentes del genoma haploide en una muestra y la distribución de los tamaños de los fragmentos. Por ejemplo, el ADNcf tiene un pico de fragmentos de aproximadamente 160 nucleótidos, y la mayoría de los fragmentos en este pico varían de aproximadamente 140 nucleótidos a 180 nucleótidos. Por consiguiente, el ADNcf de un genoma de aproximadamente 3 billones de bases (por ejemplo, el genoma humano) puede estar compuesto de casi 20 millones ($2x10^7$) fragmentos de polinucleótidos. Una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 equivalentes de genoma humano haploide. (De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes del genoma humano haploide). Una muestra que contiene aproximadamente 10.000 (10^4) equivalentes del genoma haploide de dicho ADN puede tener aproximadamente 200 billones ($2x10^{11}$) de moléculas de polinucleótidos individuales. Se ha determinado empíricamente que en una muestra de aproximadamente 10.000 equivalentes del genoma haploide del ADN humano, hay aproximadamente 3 polinucleótidos duplicados comenzando en cualquier posición dada. Por tanto, tal colección puede contener una diversidad de aproximadamente $6x10^{10}$-$8x10^{10}$ (aproximadamente 60 billones-80 billones, por ejemplo, aproximadamente 70 billones ($7x10^{10}$)) moléculas de polinucleótidos secuenciadas diferencialmente.

La probabilidad de identificar correctamente las moléculas depende del número inicial de equivalentes del genoma, la distribución de la longitud de las moléculas secuenciadas, la uniformidad de la secuencia y el número de marcadores. Cuando el recuento de marcadores es igual a uno, es decir, equivalente a no tener marcadores únicos o no marcar. La tabla siguiente enumera la probabilidad de identificar correctamente una molécula como única asumiendo una distribución de tamaño libre de células típica como la anterior.

Recuento de marcadores	% de marcadores correctamente identificados de forma única
1000 equivalentes del genoma haploide humano	
1	96.9643
4	99.2290
9	99.6539
16	99.8064
25	99.8741

100	99.9685
3000 equivalentes del genoma haploide humano	
1	91.7233
4	97.8178
9	99.0198
16	99.4424
25	99.6412
100	99.9107

En este caso, tras secuenciar el ADN genómico, puede no ser posible determinar qué lecturas de secuencia se derivan de qué moléculas parentales. Este problema puede disminuirse marcando las moléculas parentales con un número suficiente de identificadores únicos (por ejemplo, el recuento de marcadores), de tal manera que existe la posibilidad de que dos moléculas duplicadas, es decir, moléculas que tienen las mismas posiciones de inicio y parada, lleven diferentes identificadores únicos por lo que esas lecturas de secuencia pueden rastrearse hasta las moléculas parentales particulares. Un enfoque para este problema es marcar de manera única cada una, o casi cada una, de las diferentes moléculas parentales en la muestra. Sin embargo, dependiendo del número de equivalentes de genes haploides y la distribución de los tamaños de los fragmentos en la muestra, esto puede requerir billones de identificadores únicos diferentes.

Este método puede ser engorroso y costoso. Una población de polinucleótidos en una muestra de ADN genómico fragmentado se puede marcar con n identificadores únicos diferentes, en donde n es por lo menos 2 y no más de 100.000*z, en donde z es una medida de la tendencia central (por ejemplo, media, mediana, moda) de un número esperado de moléculas duplicadas que tienen las mismas posiciones de inicio y parada. En ciertas realizaciones, n es por lo menos cualquiera de 2*z, 3*z, 4*z, 5*z, 6*z, 7*z, 8*z, 9*z, 10*z, 11*z, 12*z, 13*z, 14*z, 15*z, 16*z, 17*z, 18*z, 19*z, o 20*z (por ejemplo, límite inferior). En otras realizaciones, n no es mayor que 100.000*z, 10.000*z, 1000*z o 100*z (por ejemplo, límite superior). Por tanto, n puede variar entre cualquier combinación de estos límites inferior y superior. En ciertas realizaciones, n está entre 5*z y 15*z, entre 8*z y 12*z, o aproximadamente 10*z. Por ejemplo, un equivalente de genoma humano haploide tiene aproximadamente 3 picogramos de ADN. Una muestra de aproximadamente 1 microgramo de ADN contiene aproximadamente 300.000 equivalentes de genoma humano haploide. El número n puede estar entre 15 y 45, entre 24 y 36 o aproximadamente 30. Pueden lograrse mejoras en la secuenciación siempre que por lo menos algunos de los polinucleótidos duplicados o cognados lleven identificadores únicos, es decir, lleven marcadores diferentes. Sin embargo, en ciertas realizaciones, el número de marcadores usados se selecciona de tal manera que haya por lo menos un 95% de probabilidad de que todas las moléculas duplicadas que comienzan en cualquier posición lleven identificadores únicos. Por ejemplo, una muestra que comprende aproximadamente 10.000 equivalentes del genoma humano haploide del ADNcf puede etiquetarse con aproximadamente 36 identificadores únicos. Los identificadores únicos pueden comprender seis códigos de barras de ADN únicos. Unidos a ambos extremos de un polinucleótido, se producen 36 identificadores únicos posibles. Las muestras marcadas de este modo pueden ser aquellas con un intervalo de aproximadamente 10 ng a cualquiera de aproximadamente 100 ng, aproximadamente 1 µg, aproximadamente 10 µg de polinucleótidos fragmentados, por ejemplo, ADN genómico, por ejemplo, ADNcf.

Por consiguiente, esta divulgación también proporciona composiciones de polinucleótidos de ADNcf marcados. Un conjunto de polinucleótidos en la composición que mapea en una posición base mapeable a un genoma puede estar marcado de manera no única, es decir, el número de identificadores diferentes puede ser por lo menos 2 y menos que el número de polinucleótidos que mapean en la posición base mapeable. Una composición de entre aproximadamente 10 ng y aproximadamente 10 µg (por ejemplo, cualquiera de aproximadamente 10 ng-1 µg, aproximadamente 10 ng-100 ng, aproximadamente 100 ng-10 µg, aproximadamente 100 ng-1 µg, aproximadamente 1 µg-10 µg) puede soportar entre 2, 5, 10, 50 o 100 a cualquiera de 100, 1000, 10.000 o 100.000 identificadores diferentes. Por ejemplo, pueden usarse entre 5 y 100 identificadores diferentes para marcar los polinucleótidos en dicha composición.

III. Plataformas de Secuenciación de Ácidos Nucleicos

Después de la extracción y el aislamiento de los polinucleótidos libres de células de fluidos corporales, se pueden secuenciar las secuencias libres de células. A menudo, un método de secuenciación es la secuenciación clásica de Sanger. Los métodos de secuenciación pueden incluir, pero no están limitados a: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de moléculas individuales, secuenciación por nanoporos, secuenciación por semiconductores, secuenciación por ligación, secuenciación por hibridación, ARN-

Seq (Illumina), expresión génica digital (Helicos), secuenciación de próxima generación, secuenciación de moléculas individuales por síntesis (SMSS) (Helicos), secuenciación masivamente paralela, matriz de moléculas individuales clonal (Solexa), secuenciación aleatoria, secuenciación de Maxim-Gilbert, caminata de cebadores, secuenciación usando las plataformas PacBio, SOLiD, Ion Torrent o Nanopore y cualquier otro método de secuenciación conocido en la técnica.

En algunos casos, las reacciones de secuenciación de varios tipos, como se describe en la presente, pueden comprender una variedad de unidades de procesamiento de muestras. Las unidades de procesamiento de muestras pueden incluir, pero no están limitadas a múltiples carriles, múltiples canales, múltiples pocillos u otro medio para procesar múltiples conjuntos de muestras de manera sustancialmente simultánea. Además, la unidad de procesamiento de muestras puede incluir múltiples cámaras de muestras para permitir el procesamiento de múltiples ejecuciones simultáneamente.

En algunos ejemplos, pueden realizarse reacciones de secuenciación simultáneas usando secuenciación multiplex. En algunos casos, los polinucleótidos libres de células pueden secuenciarse con por lo menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. En otros casos, los polinucleótidos libres de células pueden secuenciarse con menos de 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. Las reacciones de secuenciación pueden realizarse secuencialmente o simultáneamente. El análisis de datos posterior puede realizarse en todas o parte de las reacciones de secuenciación. En algunos casos, el análisis de datos puede realizarse en por lo menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. En otros casos, el análisis de datos puede realizarse en menos de 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación

En otros ejemplos, el número de reacciones de secuencia puede proporcionar cobertura para diferentes cantidades del genoma. En algunos casos, la cobertura de la secuencia del genoma puede ser por lo menos el 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95 %, 99%, 99,9% o 100%. En otros casos, la cobertura de secuencia del genoma puede ser menor del 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95 %, 99%, 99,9% o 100%.

En algunos ejemplos, la secuenciación puede realizarse en polinucleótidos libres de células que pueden comprender una variedad de tipos diferentes de ácidos nucleicos. Los ácidos nucleicos pueden ser polinucleótidos u oligonucleótidos. Los ácidos nucleicos incluyen, pero no están limitados a ADN o ARN, de cadena sencilla o cadena doble o un par ARN/ADNc.

IV. Estrategia de Análisis de Polinucleótidos

La Fig. 8. es un diagrama, **800**, que muestra una estrategia para analizar polinucleótidos en una muestra de material genético inicial. En el paso **802**, se proporciona una muestra que contiene material genético inicial. La muestra puede incluir ácido nucleico objetivo en baja abundancia. Por ejemplo, el ácido nucleico de un genoma normal o de tipo salvaje (por ejemplo, un genoma de la línea germinal) puede predominar en una muestra que también incluye no más del 20%, no más del 10%, no más del 5%, no más de 1 %, no más del 0,5% o no más del 0,1% de ácido nucleico de por lo menos otro genoma que contiene variación genética, por ejemplo, un genoma de cáncer o un genoma fetal, o un genoma de otra especie. La muestra puede incluir, por ejemplo, ácido nucleico libre de células o células que comprenden ácido nucleico. El material genético inicial puede constituir no más de 100 ng de ácido nucleico. Esto puede contribuir a un sobre-muestreo apropiado de los polinucleótidos originales mediante la secuenciación o el proceso de análisis genético. Alternativamente, la muestra se puede tapar o hacer cuello de botella artificialmente para reducir la cantidad de ácido nucleico a no más de 100 ng o puede enriquecerse selectivamente para analizar solo las secuencias de interés. La muestra puede modificarse para producir selectivamente lecturas de secuencia de moléculas que mapean en cada una o más localizaciones seleccionadas en una secuencia de referencia. Una muestra de 100 ng de ácido nucleico puede contener aproximadamente 30.000 equivalentes del genoma haploide humano, es decir, moléculas que juntas proporcionan una cobertura de un genoma humano de 30.000 veces.

En el paso **804**, el material genético inicial se convierte en un conjunto de polinucleótidos parentales marcados. El marcado puede incluir unir marcadores secuenciados a moléculas en el material genético inicial. Los marcadores secuenciados pueden seleccionarse de tal manera que todos los polinucleótidos únicos que mapean en la misma localización en una secuencia de referencia tengan un marcador de identificación único. La conversión puede realizarse a alta eficiencia, por ejemplo, por lo menos al 50%.

En el paso **806**, el conjunto de polinucleótidos parentales marcados se amplifica para producir un conjunto de polinucleótidos de la progenie amplificado. La amplificación puede ser, por ejemplo, de 1.000 veces.

En el paso **808**, se muestrea el conjunto de polinucleótidos de la progenie amplificado para la secuenciación. La tasa de muestreo se elige de tal manera que las lecturas de la secuencia producidas tanto (1)

cubran un número objetivo de moléculas únicas en el conjunto de polinucleótidos parentales marcados como (2) cubran moléculas únicas en el conjunto de polinucleótidos parentales marcados a unas veces de cobertura objetivo (por ejemplo, cobertura de 5 a 10 veces de los polinucleótidos parentales.

En el paso **810**, el conjunto de lecturas de secuencia se colapsa para producir un conjunto de secuencias de consenso correspondientes a los polinucleótidos parentales marcados únicos. Las lecturas de secuencia pueden calificarse para su inclusión en el análisis. Por ejemplo, las lecturas de secuencia que no cumplen con las puntuaciones de control de calidad pueden eliminarse del grupo. Las lecturas de secuencia pueden clasificarse en familias que representan lecturas de las moléculas de progenie derivadas de una molécula parental única particular. Por ejemplo, una familia de polinucleótidos de la progenie amplificada puede constituir aquellas moléculas amplificadas derivadas de un único polinucleótido parental. Al comparar las secuencias de la progenie en una familia, puede deducirse una secuencia de consenso del polinucleótido parental original. Esto produce un conjunto de secuencias de consenso que representan polinucleótidos parentales únicos en el conjunto marcado.

En el paso **812**, el conjunto de secuencias de consenso se analiza usando cualquiera de los métodos analíticos descritos en la presente. Por ejemplo, las secuencias de consenso que mapean en una localización de secuencia de referencia particular pueden analizarse para detectar casos de variación genética. Las secuencias de consenso que mapean secuencias de referencia particulares puede medirse y normalizarse frente a muestras de control. Las medidas del mapeo de moléculas en secuencias de referencia pueden compararse a través de un genoma para identificar áreas en el genoma en las que varía el número de copias, o se pierde heterocigosidad.

La Fig. 9 es un diagrama que presenta un método más genérico para extraer información de una señal representada por una colección de lecturas de secuencia. En este método, después de secuenciar los polinucleótidos de progenie amplificada, las lecturas de secuencia se agrupan en familias de moléculas amplificadas a partir de una molécula de identidad única (910). Esta agrupación puede ser un punto de partida para métodos de interpretar la información en la secuencia para determinar el contenido de los polinucleótidos parentales marcados con mayor exactitud, por ejemplo, menos ruido y/o distorsión.

El análisis de la colección de lecturas de secuencia permite hacer inferencias sobre la población de polinucleótidos parentales a partir de la cual se generaron las lecturas de secuencia. Tales inferencias pueden ser útiles ya que la secuenciación generalmente implica leer solo un subconjunto parcial de los polinucleótidos amplificados totales globales. Por lo tanto, no se puede estar seguro de que cada polinucleótido parental esté representado por al menos una secuencia leída en la colección de lecturas de secuencia.

Una de estas inferencias es el número de polinucleótidos parentales únicos en el grupo original. Dicha inferencia puede hacerse en base al número de familias únicas en las que se pueden agrupar las lecturas de secuencia y el número de lecturas de secuencia en cada familia. En este caso, una familia se refiere a una colección de lecturas de secuencias que se pueden rastrear hasta un polinucleótido parental original. La inferencia puede hacerse usando métodos estadísticos bien conocidos. Por ejemplo, si la agrupación produce muchas familias, cada una representada por una o unas pocas progenies, entonces se puede inferir que la población original incluía polinucleótidos parentales más únicos que no se secuenciaron. Por otro lado, si la agrupación produce solo unas pocas familias, cada familia representada por muchas progenies, se puede inferir que la mayoría de los polinucleótidos únicos en la población parental están representados por al menos un grupo de lecturas de secuencia en esa familia.

Otra de tales inferencias es la frecuencia de una base o secuencia de bases en un locus particular en un grupo original de polinucleótidos. Dicha inferencia puede hacerse en base al número de familias únicas en las que se pueden agrupar las lecturas de secuencia y el número de lecturas de secuencia en cada familia. Analizando las llamadas de base en un locus en una familia de lecturas de secuencia, se asigna una puntuación de confianza a cada llamada o secuencia de base particular. Luego, teniendo en consideración la puntuación de confianza para cada llamada de base en una pluralidad de familias, se determina la frecuencia de cada base o secuencia en el locus.

V. Detección de la Variación del Número de Copias

A. Detección de la Variación del Número de Copia Usando una Muestra Individual

La Fig. 1 es un diagrama, **100**, que muestra una estrategia para la detección de la variación del número de copias en un único sujeto. Como se muestra en la presente, los métodos de detección de la variación del número de copias pueden implementarse de la siguiente manera. Después de la extracción y el aislamiento de los polinucleótidos libres de células en el paso **102,** se puede secuenciar una única muestra individual mediante una plataforma de secuenciación de ácidos nucleicos conocida en la técnica en el paso **104**. Este paso genera una pluralidad de lecturas de secuencias de fragmentos genómicos. En algunos casos, estas lecturas de secuencias pueden contener información de códigos de barras. En otros ejemplos, no se utilizan códigos de barras. Después de la secuenciación, a las lecturas se les asigna una puntuación de calidad. Una puntuación de calidad puede ser una

representación de lecturas que indica si esas lecturas pueden ser útiles en un análisis posterior en base a un umbral. En algunos casos, algunas lecturas no tienen la calidad o la longitud suficientes para realizar el paso de mapeo siguiente. Las lecturas de secuenciación con una puntuación de calidad de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse de los datos. En otros casos, las lecturas de secuencia asignadas a una calidad con una puntuación menor del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos. En el paso **106**, las lecturas de fragmentos genómicos que cumplen con un umbral de puntuación de calidad especificado se mapean en un genoma de referencia, o una secuencia plantilla que se sabe que no contiene variaciones en el número de copias. Después de la alineación del mapeo, a las lecturas de secuencia se les asigna una puntuación de mapeo. Una puntuación de mapeo puede ser una representación o lecturas mapeadas de nuevo a la secuencia de referencia que indica si cada posición es o no mapeable únicamente. En casos, las lecturas pueden ser secuencias no relacionadas con el análisis de la variación del número de copias. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Las lecturas de la secuenciación con una puntuación de mapeo de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a una puntuación de mapeo menor del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos.

Después del filtrado de datos y el mapeo, la pluralidad de lecturas de secuencia genera una región cromosómica de cobertura. En el paso **108**, estas regiones cromosómicas pueden dividirse en ventanas o recipientes de longitud variable. Una ventana o recipiente puede tener por lo menos de 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, o 1000 kb. Una ventana o recipiente también puede tener bases de hasta 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, o 1000 kb. Una ventana o recipiente también puede tener aproximadamente 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, o 1000 kb.

Para la normalización de la cobertura en el paso **110**, cada ventana o recipiente se selecciona para contener aproximadamente el mismo número de bases mapeables. En algunos casos, cada ventana o recipiente en una región cromosómica puede contener el número exacto de bases mapeables. En otros casos, cada ventana o recipiente puede contener un número diferente de bases mapeables. Además, cada ventana o recipiente puede no superponerse con una ventana o recipiente adyacente. En otros casos, una ventana o recipiente puede superponerse con otra ventana o recipiente adyacente. En algunos casos, una ventana o recipiente puede superponerse por lo menos en 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, o 1000 bp En otros casos, una ventana o recipiente puede superponerse hasta 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, o 1000 bp. En algunos casos, una ventana o recipiente puede superponerse aproximadamente bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, o 1000 bp,

En algunos casos, cada una de las regiones ventana puede estar dimensionada de tal manera que contenga aproximadamente el mismo número de bases mapeables de manera única. La capacidad de mapeo de cada base que comprende una región ventana se determina y se usa para generar un archivo de capacidad de mapeo que contiene una representación de lecturas de las referencias que se mapean de nuevo en la referencia de cada archivo. El archivo de capacidad de mapeo contiene una fila por cada posición, que indica si cada posición es únicamente o no mapeable.

Además, las ventanas predefinidas, conocidas a lo largo del genoma por ser difíciles de secuenciar, o que contienen un sesgo de GC sustancialmente alto, pueden ignorarse del conjunto de datos. Por ejemplo, las regiones que se sabe que se encuentran cerca del centrómero de los cromosomas (es decir, el ADN centromérico) se sabe contienen secuencias altamente repetitivas que pueden producir resultados falsos positivos. Estas regiones pueden ignorarse. Otras regiones del genoma, como las regiones que contienen una concentración inusualmente alta de otras secuencias altamente repetitivas, como el ADN microsatélite, pueden filtrarse del conjunto de datos.

El número de ventanas analizadas también puede variar. En algunos casos, se analizan por lo menos 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5.000, 10.000, 20.000, 50.000 o 100.000 ventanas. En otros casos, el número de ventanas analizadas es de hasta 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5.000, 10.000, 20.000, 50.000 o 100.000 ventanas.

Para un genoma ejemplar derivado de secuencias de polinucleótidos libres de células, el siguiente paso comprende determinar la cobertura de lectura para cada región de la ventana. Esto puede realizarse usando lecturas o con códigos de barras o sin códigos de barras. En los casos sin códigos de barras, los pasos de mapeo anteriores proporcionarán cobertura de diferentes posiciones base. Se pueden contar las lecturas de secuencia que tienen suficientes puntuaciones de mapeo y calidad y se encuentran dentro de las ventanas de cromosomas que no están filtradas. Al número de lecturas de cobertura se les puede asignar una puntuación por cada posición mapeable. En los casos que implican códigos de barras, todas las secuencias con el mismo código de barras, propiedades físicas o combinación de las dos pueden colapsarse en una lectura, ya que todas se derivan de la molécula parental de la muestra. Este paso reduce los sesgos que pueden haberse introducido durante cualquiera de los pasos anteriores, tales como pasos que implican la amplificación. Por ejemplo, si una molécula se amplifica 10 veces pero otra se

amplifica 1000 veces, cada molécula solo se representa una vez después del colapso, negando de este modo el efecto de una amplificación desigual. Solo las lecturas con códigos de barras únicos se pueden contar para cada posición mapeable e influir en la puntuación asignada.

Las secuencias de consenso pueden generarse a partir de familias de lecturas de secuencia por cualquier método conocido en la técnica. Tales métodos incluyen, por ejemplo, métodos lineales o no lineales de construcción de secuencias de consenso (tales como votación, promedio, estadística, detección de probabilidad máxima a posteriori o máxima, programación dinámica, bayesiano, Markov oculto o métodos de máquina de vectores de soporte, etc.) derivados de la teoría de la comunicación digital, teoría de la información o la bioinformática.

Después de que se ha determinado la cobertura de lectura de la secuencia, se aplica un algoritmo de modelado estocástico para convertir la cobertura de lectura de secuencia de ácido nucleico normalizada para cada región de ventana a los estados de número de copias discretas. En algunos casos, este algoritmo puede comprender uno o más de los siguientes: modelo de Markov oculto, programación dinámica, máquina de vectores de soporte, red bayesiana, decodificación de entramados, decodificación de Viterbi, maximización de expectativas, metodologías de filtrado de Kalman y redes neurales.

En el paso **112**, los estados de número de copias discretas de cada región de ventana pueden utilizarse para identificar la variación del número de copias en las regiones cromosómicas. En algunos casos, todas las regiones de ventana adyacentes con el mismo número de copias se pueden fusionar en un segmento para informar de la presencia o ausencia del estado de variación del número de copias. En algunos casos, varias ventanas pueden filtrarse antes de fusionarse con otros segmentos.

En el paso **114**, la variación del número de copias se puede informar como un gráfico, que indica varias posiciones en el genoma y un aumento o disminución correspondiente o el mantenimiento de la variación del número de copias en cada posición respectiva. Adicionalmente, la variación del número de copias puede usarse para informar de una puntuación porcentual que indica la cantidad de material de enfermedad (o ácidos nucleicos que tienen una variación del número de copias) que existe en la muestra de polinucleótidos libres de células.

En la Fig. 10 se muestra un método para determinar la variación del número de copias. En ese método, después de agrupar las lecturas de la secuencia en familias generadas a partir de un único polinucleótido parental (1010), las familias se cuantifican, por ejemplo, determinando el número de familias que mapean en cada una de una pluralidad de localizaciones de secuencias de referencia diferentes. Las CNV pueden determinarse directamente comparando una medida cuantitativa de familias en cada uno de una pluralidad de loci diferentes (1016b). Alternativamente, se puede inferir una medida cuantitativa de familias en la población de polinucleótidos parentales marcados usando tanto una medida cuantitativa de familias como una medida cuantitativa de miembros de la familia en cada familia, por ejemplo, como se ha tratado anteriormente. Entonces, la CNV puede determinarse comparando la medida inferida de cantidad en la pluralidad de loci. En otras realizaciones, se puede tomar un enfoque híbrido por medio del cual puede hacerse una inferencia similar de la cantidad original después de la normalización para el sesgo representativo durante el proceso de secuenciación, como el sesgo de GC, etc.

B. Detección de la Variación del Número de Copias Usando Muestras Emparejadas

La detección de la variación del número de copias con muestras emparejadas comparte muchos de los pasos y parámetros del enfoque de muestra única descrito en la presente. Sin embargo, como se muestra en **200** de la Fig. 2, la detección de la variación del número de copias usando muestras emparejadas requiere la comparación de la cobertura de secuencia con una muestra de control en lugar de compararla con la capacidad de mapeo prevista del genoma. Este enfoque puede ayudar en la normalización a través de las ventanas.

La Fig. 2 es un diagrama, **200** que muestra una estrategia para la detección de la variación del número de copias en un sujeto emparejado. Como se muestra en la presente, los métodos de detección de la variación del número de copias pueden implementarse de la siguiente manera. En el paso **204**, puede secuenciarse una única muestra individual mediante una plataforma de secuenciación de ácidos nucleicos conocida en la técnica después de la extracción y el aislamiento de la muestra en el paso **202**. Este paso genera una pluralidad de lecturas de secuencias de fragmentos genómicos. Adicionalmente, se toma una muestra de muestra o control de otro sujeto. En algunos casos, el sujeto de control puede ser un sujeto que no se sabe que tiene una enfermedad, mientras que el otro sujeto puede tener o estar en riesgo de contraer una enfermedad particular. En algunos casos, estas lecturas de secuencia pueden contener información de códigos de barras. En otros ejemplos, no se utilizan códigos de barras. Después de la secuenciación, se asigna una puntuación de calidad a las lecturas. En algunos casos, algunas lecturas no tienen la suficiente calidad o longitud para realizar el paso de mapeo posterior. Las lecturas de secuencia con una puntuación de calidad de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos. En otros casos, las lecturas de secuencia asignadas a una calidad con una puntuación menor que el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos. En el paso **206,** las lecturas de fragmentos genómicos que cumplen con un umbral de puntuación de calidad especificado se mapean en un genoma de referencia, o una secuencia plantilla que se sabe que no contiene variaciones en el número de copias.

Después de la alineación de mapeo, a las lecturas de secuencia se les asigna una puntuación de mapeo. En casos, las lecturas pueden ser secuencias no relacionadas con el análisis de la variación del número de copias. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Las lecturas de secuencia con una puntuación de mapeo de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos. En otros casos, las lecturas de secuencia asignadas a mapeo con una puntuación menor del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos.

Después del filtrado de datos y el mapeo la pluralidad de lecturas de secuencia genera una región cromosómica de cobertura para cada uno de los sujetos de prueba y control. En el paso **208**, estas regiones cromosómicas pueden dividirse en ventanas o recipientes longitud variable. Una ventana o recipiente puede tener por lo menos 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, o 1000 kb. Una ventana o recipiente también puede ser menor de 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, o 1000 kb.

Para la normalización de la cobertura en el paso **210**, cada ventana o recipiente se selecciona para contener aproximadamente el mismo número de bases mapeables para cada uno de los sujetos de prueba y control. En algunos casos, cada ventana o recipiente en una región cromosómica puede contener el número exacto de bases mapeables. En otros casos, cada ventana o recipiente puede contener un número diferente de bases mapeables. Además, cada ventana o recipiente puede no superponerse con una ventana o recipiente adyacente. En otros casos, una ventana o recipiente puede superponerse con otra ventana o recipiente adyacente. En algunos casos, una ventana o recipiente puede superponerse por lo menos en 1 bp, 2 bp, 3 bp, 4 bp, 5 bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, o 1000 bp En otros casos, una ventana o recipiente puede superponerse en menos de 1 bp, 2 bp, 3 bp, 4 bp, 5 bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, o 1000 bp.

En algunos casos, cada una de las regiones ventana se dimensiona de tal manera que contenga aproximadamente el mismo número de bases únicamente mapeables para cada uno de los sujetos de prueba y control. La capacidad de mapeo de cada base que comprende una región ventana se determina y se usa para generar un archivo de capacidad de mapeo que contiene una representación de lecturas de las referencias que se mapean de nuevo a la referencia de cada archivo. El archivo de capacidad de mapeo contiene una fila por cada posición, que indica si cada posición es únicamente o no mapeable.

Adicionalmente, las ventanas predefinidas, conocidas en todo el genoma por ser difíciles de secuenciar, o que contienen un sesgo de GC sustancialmente alto, se filtran del conjunto de datos. Por ejemplo, se sabe que las regiones que se sabe que se encuentran cerca del centrómero de los cromosomas (es decir, el ADN centromérico) contienen secuencias altamente repetitivas que pueden producir resultados falsos positivos. Estas regiones pueden filtrarse. Otras regiones del genoma, como las regiones que contienen una concentración inusualmente alta de otras secuencias altamente repetitivas, como el ADN microsatélite, pueden filtrarse del conjunto de datos.

El número de ventanas analizadas también puede variar. En algunos casos, se analizan por lo menos 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5.000, 10.000, 20.000, 50.000 o 100.000 ventanas. En otros casos, se analizan menos de 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5.000, 10.000, 20.000, 50.000 o 100.000 ventanas.

Para un genoma ejemplar derivado de secuencias de polinucleótidos libres de células, el siguiente paso comprende determinar la cobertura de lectura para cada región de ventana para cada uno de los sujetos de prueba y control. Esto se puede realizar usando lecturas con códigos de barras o sin códigos de barras. En los casos sin códigos de barras, los pasos de mapeo anteriores proporcionarán cobertura de diferentes posiciones de base. Se pueden contar las lecturas de secuencia que tienen suficientes puntuaciones de mapeo y calidad y se encuentran dentro de las ventanas de cromosomas que no están filtradas. Al número de lecturas de cobertura puede asignársele una puntuación por cada posición mapeable. En los casos que implican códigos de barras, todas las secuencias con el mismo código de barras pueden colapsarse en una lectura, ya que todas se derivan de la molécula parental de la muestra. Este paso reduce los sesgos que pueden haberse introducido durante cualquiera de los pasos anteriores, como en pasos que implican la amplificación. Solo las lecturas con códigos de barras únicos pueden contarse para cada posición mapeable e influir en la puntuación asignada. Por esta razón, es importante que el paso de ligación del código de barras se realice de una manera optimizada para producir la cantidad más baja de sesgo.

Al determinar la cobertura de lectura de ácidos nucleicos para cada ventana, la cobertura de cada ventana puede normalizarse mediante la cobertura media de esa muestra. Usando dicho enfoque, puede ser deseable secuenciar tanto el sujeto de prueba como el control bajo condiciones similares. La cobertura de lectura para cada ventana puede expresarse luego como una relación sobre ventanas similares

Las relaciones de cobertura de lecturas de ácidos nucleicos para cada ventana del sujeto de prueba pueden determinarse dividiendo la cobertura de lectura de cada región ventana de la muestra de prueba con la cobertura de lectura de una región ventana correspondiente del control.

Una vez que se han determinado las relaciones de cobertura de las lecturas de secuencia, se aplica un

algoritmo de modelado estocástico para convertir las relaciones normalizadas para cada región ventana en estados de números de copias discretos. En algunos casos, este algoritmo puede comprender un modelo oculto de Markov. En otros casos, el modelo estocástico puede comprender programación dinámica, máquina de vectores de soporte, modelado bayesiano, modelado probabilístico, decodificación de entramados, decodificación de Viterbi, maximización de esperanza, metodologías de filtrado de Kalman o redes neurales.

En el paso **212,** los estados de números de copias discretas de cada región ventana se pueden utilizar para identificar la variación del número de copias en las regiones cromosómicas. En algunos casos, todas las regiones ventana adyacentes con el mismo número de copias se pueden fusionar en un segmento para informar de la presencia o ausencia del estado de variación del número de copias. En algunos casos, pueden filtrarse varias ventanas antes de fusionarse con otros segmentos.

En el paso **214,** se puede informar de la variación del número de copias como un gráfico, que indica varias posiciones en el genoma y un aumento o disminución correspondiente o el mantenimiento de la variación del número de copias en cada posición respectiva. Adicionalmente, la variación en el número de copias puede usarse para informar de una puntuación porcentual que indica la cantidad de material de la enfermedad existente en la muestra de polinucleótidos libres de células.

VI. Detección de Mutaciones Raras

La detección de mutaciones raras comparte características similares a ambos enfoques de variación del número de copias Sin embargo, como se muestra en la Fig. 3, **300**, la detección de mutaciones raras usa la comparación de la cobertura de secuencia con una muestra de control o secuencia de referencia en lugar de compararla con la capacidad de mapeo relativa del genoma. Este enfoque puede ayudar en la normalización a través de ventanas.

Generalmente, la detección de mutaciones raras puede realizarse en regiones enriquecidas selectivamente del genoma o transcriptoma purificadas y aisladas en el paso **302.** Como se describe en la presente, regiones específicas, que pueden incluir, pero no está limitado a, genes, oncogenes, genes supresores de tumores, promotores, elementos de secuencia reguladores, regiones no codificantes, ARNmi, ARNsn y similares pueden amplificarse selectivamente de una población total de polinucleótidos libres de células. Esto se puede realizar como se describe en la presente. En un ejemplo, puede usarse secuenciación multiplex, con o sin marcadores de código de barras para secuencias de polinucleótidos individuales. En otros ejemplos, la secuenciación puede realizarse usando cualquier plataforma de secuenciación de ácidos nucleicos conocida en la técnica. Este paso genera una pluralidad de lecturas de secuencias de fragmentos genómicos como en el paso **304**. Además, se obtiene una secuencia de referencia de una muestra de control, tomada de otro sujeto. En algunos casos, el sujeto de control puede ser un sujeto que se sabe que no tiene aberraciones o enfermedades genéticas conocidas. En algunos casos, estas lecturas de secuencia pueden contener información de códigos de barras. En otros ejemplos, no se utilizan códigos de barras. Después de la secuenciación, a las lecturas se les asigna un puntuación de calidad . Una puntuación de calidad puede ser una representación de lecturas que indica si esas lecturas pueden ser útiles en un análisis posterior en base a un umbral. En algunos casos, algunas lecturas no tienen la calidad o la longitud suficientes para realizar el paso de mapeo posterior. Las lecturas de secuencia con una puntuación de calidad de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ser ignoradas del conjunto de datos. En otros casos, las lecturas de secuencia asignadas con una puntuación de calidad de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% puede ser ignoradas del conjunto de datos. En el paso **306**, las lecturas de fragmentos genómicos que cumplen con un umbral de puntuación de calidad especificado se mapean a un genoma de referencia, o una secuencia de referencia que se sabe que no contiene mutaciones raras. Después de la alineación del mapeo, a las lecturas de secuencia se les asigna una puntuación de mapeo. Una puntuación de mapeo puede ser una representación o lecturas mapeadas de nuevo a la secuencia de referencia indicando si cada posición es o no mapeable de manera única. En casos, las lecturas pueden ser secuencias no relacionadas con análisis de mutaciones raras. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Las lecturas de secuencia con una puntuación de mapeo de por lo menos el 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ser ignoradas del conjunto de datos. En otros casos, las lecturas de secuencia asignadas a u mapeo puntuado con menos del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden ignorarse del conjunto de datos.

Para cada base mapeable, las bases que no cumplen con el umbral mínimo para la capacidad de mapeo, o bases de baja calidad, pueden reemplazarse por las bases correspondientes que se encuentran en la secuencia de referencia.

Después del filtrado de datos y el mapeo, se analizan bases de variantes encontradas entre las lecturas de secuencia obtenidas del sujeto y la secuencia de referencia.

Para un genoma ejemplar derivado de secuencias de polinucleótidos libres de células, el siguiente paso comprende determinar la cobertura de lectura para cada posición base mapeable. Esto se puede realizar usando o

lecturas con códigos de barras o sin códigos de barras. En los casos sin códigos de barras, los pasos de mapeo anteriores proporcionarán cobertura de diferentes posiciones de bases. Se pueden contar las lecturas de secuencia que tengan puntuaciones suficientes de mapeo y calidad. Al número de lecturas de cobertura se le puede asignar una puntuación por cada posición mapeable. En los casos que implican códigos de barras, todas las secuencias con el mismo código de barras se pueden colapsar en una lectura de consenso, ya que todas se derivan de la molécula parenteral de muestra. La secuencia para cada base se alinea como la lectura de nucleótido más dominante para esa localización específica. Además, el número de moléculas únicas puede contarse en cada posición para derivar una cuantificación simultánea en cada posición. Este paso reduce los sesgos que pueden haberse introducido durante cualquiera de los pasos anteriores, como los pasos que implican amplificación. Solo las lecturas con códigos de barras únicos pueden contarse para cada posición mapeable e influir en la puntuación asignada.

Una vez que se puede determinar la cobertura de lectura y se identifican las bases de variante en relación con la secuencia de control en cada lectura, la frecuencia de las bases de variante puede calcularse como el número de lecturas que contienen la variante dividido por el número total de lecturas. Esto puede expresarse como una relación para cada posición mapeable a el genoma.

Para cada posición de base, las frecuencias de los cuatro nucleótidos, citosina, guanina, timina, adenina se analizan en comparación con la secuencia de referencia. Se aplica un algoritmo de modelado estocástico o estadístico para convertir las relaciones normalizadas para cada posición mapeable para reflejar los estados de frecuencia para cada variante de base. En algunos casos, este algoritmo puede comprender uno o más de los siguientes: modelo de Markov oculto, programación dinámica, máquina de vectores de soporte, modelado bayesiano o probabilístico, decodificación de entramado, decodificación de Viterbi, maximización de esperanza, metodologías de filtrado de Kalman y redes neurales.

En el paso **312**, los estados de mutaciones raras discretos de cada posición de base se pueden utilizar para identificar una variante de base con una alta frecuencia de varianza en comparación con el valor de referencia de la secuencia de referencia. En algunos casos, el valor de referencia puede representar una frecuencia de por lo menos el 0,0001%, 0,001%, 0,01%, 0,1%, 1,0%, 2,0%, 3,0%, 4,0% 5,0%, 10% o 25%. En otros casos, el valor de referencia puede representar una frecuencia de por lo menos el 0,0001%, 0,001%, 0,01%, 0,1%, 1,0%, 2,0%, 3,0%, 40% 5,0%. 10%, o 25%. En algunos casos, todas las posiciones de bases adyacentes con la variante o la mutación de base pueden fusionarse en un segmento para informar de la presencia o ausencia de una mutación rara. En algunos casos, varias posiciones pueden filtrarse antes de fusionarse con otros segmentos.

Después del cálculo de las frecuencias de varianza para cada posición de base, la variante con la mayor desviación para una posición específica en la secuencia derivada del sujeto en comparación con la secuencia de referencia se identifica como una mutación rara. En algunos casos, una mutación rara puede ser una mutación de cáncer. En otros casos, una mutación rara podría estar relacionada con un estado de enfermedad.

Una mutación o variante rara puede comprender una aberración genética que incluye, pero no está limitados a, una sustitución de base única, o indeles pequeños, transversiones, translocaciones, inversión, deleciones, truncamientos o truncamientos de genes. En algunos casos, una mutación rara puede tener como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 o 20 nucleótidos de longitud. En otros casos, una mutación rara puede tener por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 o 20 nucleótidos de longitud.

En el paso **314**, la presencia o ausencia de una mutación puede reflejarse en forma gráfica, indicando varias posiciones en el genoma y un aumento o disminución correspondiente o el mantenimiento de una frecuencia de mutación en cada posición respectiva. Adicionalmente, pueden usarse mutaciones raras para informar de una puntuación porcentual que indica la cantidad de material de la enfermedad que existe en la muestra de polinucleótidos libres de células. Una puntuación de confianza puede acompañar a cada mutación detectada, dadas las estadísticas conocidas de las varianzas típicas en las posiciones informadas en las secuencias de referencia de no enfermedad. Las mutaciones también pueden clasificarse por orden de abundancia en el sujeto o clasificarse por importancia clínicamente prevenible.

La Fig. 11 muestra un método para inferir la frecuencia de una base o secuencia de bases en un locus particular en una población de polinucleótidos. Las lecturas de secuencia se agrupan en familias generadas a partir de un polinucleótido marcado original (1110). Para cada familia, se asigna una puntuación de confianza a una o más bases en el locus. La puntuación de confianza puede asignarse mediante cualquiera de una serie de métodos estadísticos conocidos asignados y puede basarse, por lo menos en parte, en la frecuencia con la que aparece una base entre las lecturas de secuencia que pertenecen a la familia (1112). Por ejemplo, la puntuación de confianza puede ser la frecuencia con la que aparece la base entre las lecturas de secuencia. Como otro ejemplo, para cada familia, se puede construir un modelo oculto de Markov, de tal manera que se pueda tomar una decisión de máxima probabilidad o máxima a posteriori en base a la frecuencia de ocurrencia de una base particular en una familia individual. Como parte de este modelo, también pueden generarse la probabilidad de error y la puntuación de confianza resultante para una decisión particular. Una frecuencia de la base en la población original puede asignarse luego en base a las puntuaciones de confianza entre las familias (1114).

VII. Aplicaciones

A. Detección Temprana de Cáncer

Pueden detectarse numerosos cánceres usando los métodos y sistemas descritos en la presente. Las células cancerosas, como la mayoría de las células, pueden caracterizarse por una tasa de rotación, en la que las células viejas mueren y se reemplazan por células más nuevas. Generalmente las células muertas, en contacto con la vasculatura en un sujeto dado, pueden liberar ADN o fragmentos de ADN en el torrente sanguíneo. Esto también es cierto para las células cancerosas durante varias etapas de la enfermedad. Las células cancerosas también pueden caracterizarse, dependiendo de la etapa de la enfermedad, por diversas aberraciones genéticas, como la variación del número de copias y mutaciones raras. Este fenómeno puede usarse para detectar la presencia o ausencia de cánceres en individuos usando los métodos y sistemas descritos en la presente.

Por ejemplo, puede extraerse sangre de sujetos con riesgo de cáncer y prepararse como se describe en la presente para generar una población de polinucleótidos libres de células. En un ejemplo, esto podría ser ADN libre de células. Los sistemas y métodos de la divulgación pueden emplearse para detectar mutaciones raras o variaciones en el número de copias que pueden existir en ciertos cánceres presentes. El método puede ayudar a detectar la presencia de células cancerosas en el cuerpo, a pesar de la ausencia de síntomas u otras características distintivas de la enfermedad.

Los tipos y la cantidad de cánceres que pueden detectarse pueden incluir, pero no están limitados a, cánceres de la sangre, cánceres del cerebro, cánceres de la piel, cánceres de la nariz, cánceres de la garganta, cánceres del hígado, cánceres de los huesos, linfomas, cánceres de páncreas, cánceres de la piel, cánceres del intestino, cánceres rectales, cánceres de tiroides, cánceres de vejiga, cánceres de riñón, cánceres de boca, cánceres de estómago, tumores en estado sólido, tumores heterogéneos, tumores homogéneos y similares.

En la detección temprana de cánceres, puede utilizarse para detectar cánceres cualquiera de los sistemas o métodos descritos en la presente, incluyendo la detección de mutaciones raras o la detección de la variación del número de copias. Estos sistemas y métodos pueden usarse para detectar cualquier número de aberraciones genéticas que puedan provocar o ser el resultado de cánceres. Estas pueden incluir, pero no están limitadas a, mutaciones, mutaciones raras, indeles, variaciones en el número de copias, transversiones, translocaciones, inversión, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones genéticas, fusiones cromosómicas, truncamientos genéticos, amplificación de genes, duplicaciones de genes, lesiones cromosómicas, lesiones de ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la infección por metilación de los ácidos nucleicos y el cáncer.

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

B. Monitorización y Pronóstico del Cáncer

Los sistemas y métodos proporcionados en la presente pueden usarse para monitorizar cánceres ya conocidos u otras enfermedades en un sujeto particular. Esto puede permitir que un sujeto o profesional adapte las opciones de tratamiento de acuerdo con el progreso de la enfermedad. En este ejemplo, los sistemas y métodos descritos en la presente pueden usarse para construir perfiles genéticos de un sujeto particular del curso de la enfermedad. En algunos casos, los cánceres pueden progresar, volviéndose más agresivos y genéticamente inestables. En otros ejemplos, los cánceres pueden permanecer benignos, inactivos, latentes o en remisión. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión, la remisión o la recurrencia de la enfermedad.

Además, los sistemas y métodos descritos en la presente pueden ser útiles para determinar la eficacia de una opción de tratamiento particular. En un ejemplo, las opciones de tratamiento exitosas pueden realmente aumentar la cantidad de variación del número de copias o mutaciones raras detectadas en la sangre del sujeto si el tratamiento tiene éxito, ya que pueden morir más cánceres y desprenderse del ADN. En otros ejemplos, esto puede no ocurrir. En otro ejemplo, quizás ciertas opciones de tratamiento pueden estar correlacionadas con los perfiles genéticos de los cánceres a lo largo del tiempo. Esta correlación puede ser útil para seleccionar una terapia. Además, si se observa que un cáncer está en remisión después del tratamiento, los sistemas y métodos descritos en la presente pueden ser útiles para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

Por ejemplo, las mutaciones que tienen lugar dentro de un intervalo de frecuencia que comienza en el nivel de umbral pueden determinarse a partir del ADN en una muestra de un sujeto, por ejemplo, un paciente. Las mutaciones pueden ser, por ejemplo, mutaciones relacionadas con el cáncer. La frecuencia puede variar de, por ejemplo, por lo menos del 0,1%, por lo menos del 1%, o por lo menos del 5% al 100%. La muestra puede ser, por ejemplo, ADN libre de células o una muestra de tumor. Se puede prescribir un curso de tratamiento en base a cualquiera o todas las mutaciones que tengan lugar dentro del intervalo de frecuencia, incluyendo, por ejemplo, sus frecuencias. Se puede tomar una muestra del sujeto en cualquier momento posterior. Se pueden determinar las mutaciones que tienen lugar dentro del intervalo original de frecuencia o un intervalo diferente de frecuencia. El curso del tratamiento puede ajustarse en base a las mediciones posteriores.

C. Detección Temprana y Monitorización de Otras Enfermedades o Estados de Enfermedades

Los métodos y sistemas descritos en la presente pueden no estar limitados a la detección de mutaciones raras y variaciones en el número de copias asociadas solo con los cánceres. Varias otras enfermedades e infecciones pueden resultar en otros tipos de afecciones que pueden ser adecuadas para la detección y la monitorización tempranos. Por ejemplo, en ciertos casos, los trastornos genéticos o las enfermedades infecciosas pueden provocar cierto mosaicismo genético dentro de un sujeto. Este mosaicismo genético puede provocar variación en el número de copias y mutaciones raras que podrían ser observadas. En otro ejemplo, el sistema y los métodos de la divulgación también pueden usarse para monitorizar los genomas de las células inmunes dentro del cuerpo. Las células inmunes, como las células B, pueden experimentar una expansión clonal rápida ante la presencia de ciertas enfermedades. Las expansiones clonales pueden monitorizarse usando la detección de variación del número de copias y pueden monitorizarse ciertos estados inmunes. En este ejemplo, el análisis de la variación del número de copias puede realizarse a lo largo del tiempo para producir un perfil de cómo puede estar progresando una enfermedad particular.

Además, los sistemas y métodos de esta divulgación también pueden usarse para monitorizar las mismas infecciones sistémicas, como las que pueden ser provocadas por un patógeno como una bacteria o virus. La variación en el número de copias o incluso la detección de mutaciones raras pueden usarse para determinar cómo está cambiando una población de patógenos durante el curso de la infección. Esto puede ser particularmente importante durante las infecciones crónicas, como las infecciones por VIH/SIDA o hepatitis, en las que los virus pueden cambiar el estado del ciclo de vida y/o mutarse a formas más virulentas durante el curso de la infección.

Otro ejemplo más para el cual pueden usarse el sistema y los métodos de esta divulgación es la monitorización de sujetos de trasplante. En general, el tejido trasplantado sufre un cierto grado de rechazo por parte del cuerpo tras el trasplante. Los métodos de esta divulgación pueden usarse para determinar o realizar perfiles de las actividades de rechazo del cuerpo del huésped, ya que las células inmunes intentan destruir el tejido trasplantado. Esto puede ser útil para monitorizar el estado del tejido trasplantado, así como para alterar el curso del tratamiento o la prevención del rechazo.

Además, los métodos de la divulgación pueden usarse para caracterizar la heterogeneidad de una condición anormal en un sujeto, el método comprendiendo generar un perfil genético de polinucleótidos extracelulares en el sujeto, en donde el perfil genético comprende una pluralidad de datos resultantes de los análisis de variación del número de copias y mutaciones raras. En algunos casos, incluyendo pero no limitados a cáncer, una enfermedad puede ser heterogénea. Las células de la enfermedad pueden no ser idénticas. En el ejemplo del cáncer, se sabe que algunos tumores comprenden diferentes tipos de células tumorales, algunas células en diferentes etapas del cáncer. En otros ejemplos, la heterogeneidad puede comprender múltiples focos de enfermedad. De nuevo, en el ejemplo del cáncer, puede haber múltiples focos tumorales, quizás donde uno o más focos son el resultado de metástasis que se han diseminado desde un sitio primario.

Los métodos de esta divulgación pueden usarse para generar o realizar perfiles, huellas o conjuntos de datos que es una suma de información genética derivada de diferentes células en una enfermedad heterogénea. Este conjunto de datos puede comprender los análisis de variación del número de copias y de mutaciones raras solo o en combinación.

D. Detección Temprana y Monitorización de otras Enfermedades o Estados de Enfermedades de Origen Fetal

Además, los sistemas y métodos de la divulgación pueden usarse para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades de origen fetal. Es decir, estas metodologías pueden emplearse en una paciente embarazada para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades en un sujeto no nato cuyo ADN y otros polinucleótidos pueden co-circular con moléculas maternas.

VIII. Terminología

La terminología usada en la presente tiene el propósito de describir realizaciones particulares solamente y

no se pretende que limite los sistemas y métodos de esta divulgación. Como se usa en la presente, las formas singulares "un", "una" y "el" se pretende que incluyan también las formas en plural, a menos que el contexto indique claramente lo contrario. Además, en la medida en que los términos "incluyendo", "incluye", "teniendo", "tiene", "con", o variantes de los mismos se usan en la descripción detallada y/o las reivindicaciones, se pretende que tales términos incluyan de manera similar al término " comprendiendo".

Varios aspectos de los sistemas y métodos de esta divulgación se han descrito anteriormente con referencia a aplicaciones ejemplares para ilustración. Debe entenderse que se establecen numerosos detalles, relaciones y métodos específicos para proporcionar una comprensión completa de los sistemas y métodos. Sin embargo, un experto en la técnica relevante reconocerá fácilmente que se pueden poner en práctica sistemas y métodos sin uno o más de los detalles específicos o con otros métodos. Esta divulgación no está limitada por el orden ilustrado de actos o eventos, ya que algunos actos pueden tener lugar en órdenes diferentes y/o concurrentemente con otros actos o eventos. Además, no se requieren todos los actos o eventos ilustrados para implementar una metodología de acuerdo con esta divulgación.

Los intervalos pueden expresarse en este documento a partir de "aproximadamente" un valor particular, y/o a "aproximadamente" otro valor particular. Cuando se expresa tal intervalo, otra realización incluye desde un valor particular y/o al otro valor particular. De manera similar, cuando los valores se expresan como aproximaciones, mediante el uso del antecedente "aproximadamente", se entenderá que el valor particular forma otra realización. Se entenderá además que los puntos finales de cada uno de los intervalos son significativos tanto en relación con el otro punto final, como independientemente del otro punto final. El término "aproximadamente", como se usa en la presente, se refiere a un intervalo que es el 15% más o menos de un valor numérico expresado en el contexto del uso particular. Por ejemplo, alrededor de 10 incluiría un rango de 8,5 a 11,5.

Sistemas Informáticos

Los métodos de la presente divulgación pueden implementarse usando, o con la ayuda de, sistemas informáticos. La FIG. 15 muestra un sistema informático 1501 que está programado o configurado de otra manera para implementar los métodos de la presente divulgación. El sistema informático 1501 puede regular varios aspectos de preparación, secuenciación y/o análisis de muestras. En algunos ejemplos, el sistema informático 1501 está configurado para realizar la preparación de las muestras y el análisis de las muestras, incluyendo la secuenciación de los ácidos nucleicos.

El sistema informático 1501 incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador informático" en la presente) 1505, que puede ser un procesador de un solo núcleo o de múltiples núcleos, o una pluralidad de procesadores para el procesamiento en paralelo. El sistema informático 1501 también incluye memoria o localización de memoria 1510 (por ejemplo, memoria de acceso aleatorio, memoria de solo lectura, memoria flash), unidad de almacenamiento electrónico 1515 (por ejemplo, disco duro), interfaz de comunicación 1520 (por ejemplo, adaptador de red) para comunicarse con uno o más de otros sistemas, y dispositivos periféricos 1525, como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónicos. La memoria 1510, la unidad de almacenamiento 1515, la interfaz 1520 y los dispositivos periféricos 1525 están en comunicación con la CPU 1505 a través de un bus de comunicación (líneas sólidas), como una placa base. La unidad de almacenamiento 1515 puede ser una unidad de almacenamiento de datos (o repositorio de datos) para almacenar datos. El sistema informático 1501 puede acoplarse operativamente a una red informática ("red") 1530 con la ayuda de la interfaz de comunicación 1520. La red 1530 puede ser Internet, una internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. La red 1530 en algunos casos es una red de telecomunicaciones y/o datos. La red 1530 puede incluir uno o más servidores informáticos, que pueden habilitar la computación distribuida, como la computación en la nube. La red 1530, en algunos casos con la ayuda del sistema informático 1501, puede implementar una red peer-to-peer, lo que puede permitir que los dispositivos acoplados al sistema informático 1501 se comporten como un cliente o un servidor.

La CPU 1505 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 1510. Los ejemplos de operaciones realizadas por la CPU 1505 pueden incluir búsqueda, decodificación, ejecución y reescritura.

La unidad de almacenamiento 1515 puede almacenar archivos, como controladores, bibliotecas y programas guardados. La unidad de almacenamiento 1515 puede almacenar programas generados por usuarios y sesiones grabadas, así como también salidas asociadas con los programas. La unidad de almacenamiento 1515 puede almacenar datos de usuario, por ejemplo, preferencias de usuario y programas de usuario. El sistema informático 1501 en algunos casos puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático 1501, como las que se encuentran en un servidor remoto que está en comunicación con el sistema informático 1501 a través de una intranet o Internet.

El sistema informático 1501 puede comunicarse con uno o más sistemas informáticos remotos a través de

la red 1530. Por ejemplo, el sistema informático 1501 puede comunicarse con un sistema informático remoto de un usuario (por ejemplo, un operador). Los ejemplos de sistemas informáticos remotos incluyen ordenadores personales (por ejemplo, ordenadores personales portátiles), pizarras o tabletas (por ejemplo, iPad de Apple®, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, iPhone de Apple®, dispositivo con Android, Blackberry®), o asistentes digitales personales. El usuario puede acceder al sistema informático 1501 a través de la red 1530.

Los métodos como se describen en la presente pueden implementarse a través del código ejecutable de la máquina (por ejemplo, procesador informático) almacenado en una localización de almacenamiento electrónico del sistema informático 1501, como por ejemplo, en la memoria 1510 o en la unidad de almacenamiento electrónico 1515. El código ejecutable por máquina o legible por máquina puede proporcionarse en forma de software. Durante el uso, el código puede ser ejecutado por el procesador 1505. En algunos casos, el código puede recuperarse de la unidad de almacenamiento 1515 y almacenarse en la memoria 1510 para que el procesador 1505 pueda acceder fácilmente. En algunas situaciones, la unidad de almacenamiento electrónico 1515 puede ser excluido, y las instrucciones ejecutables por máquina se almacenan en la memoria 1510.

El código puede pre-compilarse 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 suministrarse en un lenguaje de programación que se puede seleccionar para permitir que el código se ejecute de manera pre-compilada o compilada.

Los aspectos de los sistemas y métodos proporcionados en la presente, como el sistema informático 1501, pueden incorporarse en la programación. Varios aspectos de la tecnología pueden considerarse como "productos" o "artículos de fabricación", típicamente en forma de código ejecutable en máquina (o procesador) y/o datos asociados que se transmiten 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 cualquiera 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 puede comunicarse a través de Internet o 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 host a la plataforma informática de un servidor de aplicaciones. Por tanto, otro tipo de medios que pueden llevar los elementos del software incluyen ondas ópticas, eléctricas y electromagnéticas, como las que se usan en las interfaces físicas entre dispositivos locales, a través de redes fijas por cable y ópticas y a través de varios enlaces aéreos. Los elementos físicos que transportan tales ondas, como enlaces cableados o inalámbricos, enlaces ópticos o similares, también pueden considerarse como medios que llevan el software. Tal como se usa en la presente, a menos que esté restringido a medios de "almacenamiento" tangibles, no transitorios, los términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participe en la provisión de instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, como un código ejecutable por ordenador, puede tomar muchas formas, incluyendo, pero no limitadas a, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión física. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, como cualquiera de los dispositivos de almacenamiento en cualquier ordenador o similar, como los que se pueden usar para implementar bases de datos, etc. que se muestran en los dibujos. Los medios de almacenamiento volátiles incluyen 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 forman un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden tomar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o de luz, como las generadas durante las comunicaciones de datos por radio frecuencia (RF) e infrarrojos (IR). Las formas comunes de medios legibles por ordenador incluyen, por ejemplo, un disquete, un disco flexible, un disco duro, una cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, cinta de papel de tarjetas perforadas, cualquier otro medio de almacenamiento físico con patrones de orificios, una memoria RAM, una ROM, una PROM y una EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporte datos o instrucciones, cables o enlaces que transporten onda portadora, o cualquier otro medio desde el cual un ordenador pueda leer el código y/o los datos de programación. 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 1501 puede incluir o estar en comunicación con una pantalla electrónica que comprende una interfaz de usuario (UI) para proporcionar, por ejemplo, uno o más resultados de análisis de muestras. Los ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en web.

EJEMPLOS

Ejemplo 1 - Pronóstico y Tratamiento del Cáncer de Próstata

Se toma una muestra de sangre de un sujeto con cáncer de próstata. Anteriormente, un oncólogo determina que el sujeto tiene cáncer de próstata en estadio II y recomienda un tratamiento. El ADN libre de células se extrae, se aísla, se secuencia y se analiza cada 6 meses después del diagnóstico inicial.

Se extrae ADN libre de células y se aísla de la sangre mediante el protocolo del kit Qiagen Qubit. Se añade un ADN portador para aumentar los rendimientos. El ADN se amplifica mediante PCR y cebadores universales. Se secuencian 10 ng de ADN usando un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq. El 90% del genoma del sujeto se cubre mediante la secuenciación de ADN libre de células.

Los datos de secuencia se ensamblan y analizan para determinar la variación del número de copias. Las lecturas de secuencia se mapean y se comparan con un individuo sano (control). En base al número de lecturas de secuencia, las regiones cromosómicas se dividen en regiones no superpuestas de 50 kb. Las lecturas de secuencia se comparan entre sí y se determina una proporción para cada posición mapeable.

Se aplica un modelo oculto de Markov para convertir los números de copias en estados discretos para cada ventana.

Se generan informes, mapeando las posiciones del genoma y la variación del número de copias que se muestran la Fig. 4A (para un individuo sano) y en la Fig. 4B para el sujeto con cáncer.

Estos informes, en comparación con otros perfiles de sujetos con resultados conocidos, indican que este cáncer particular es agresivo y resistente al tratamiento. La carga tumoral libre de células es del 21%. El sujeto es monitorizado durante 18 meses. En el mes 18, el perfil de variación del número de copias comienza a aumentar dramáticamente, desde la carga tumoral libre de células del 21% al 30%. Se hace una comparación con los perfiles genéticos de otros sujetos de próstata. Se determina que este aumento en la variación del número de copias indica que el cáncer de próstata está avanzando de la etapa II a la etapa III. El régimen de tratamiento original según lo prescrito ya no trata el cáncer. Se prescribe un nuevo tratamiento.

Además, estos informes se envían y se accede a ellos de forma electrónica a través de Internet. El análisis de los datos de secuencia se realiza en un sitio diferente a la localización del sujeto. El informe se genera y se transmite a la localización del sujeto. A través de una computadora con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral (Fig. 4C).

Ejemplo 2 - Remisión y Recurrencia del Cáncer de Próstata.

Se toma una muestra de sangre de un superviviente al cáncer de próstata. El sujeto había sido sometido anteriormente a numerosas rondas de quimioterapia y radiación. El sujeto en el momento de la prueba no presentó síntomas o problemas de salud relacionados con el cáncer. Las exploraciones y los análisis estándar revelan que el sujeto no tiene cáncer.

Se extrae ADN libre de células y se aísla de la sangre usando el protocolo del kit Qiagen TruSeq. Se añade un ADN portador para aumentar los rendimientos. El ADN se amplifica usando PCR y cebadores universales. Se secuencian 10 ng de ADN usando un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq. Se añaden 12mer de códigos de barras a moléculas individuales usando un método de ligación.

Los datos de secuencias se ensamblan y analizan para determinar la variación del número de copias. Las lecturas de secuencia se mapean y se comparan con un individuo sano (control). En base al número de lecturas de secuencia, las regiones cromosómicas se dividen en regiones no superpuestas de 40 kb. Las lecturas de secuencia se comparan entre sí y se determina una proporción para cada posición mapeable.

Las secuencias con código de barras no únicos se colapsan en una única lectura para ayudar a normalizar el sesgo de la amplificación.

Se aplica un modelo oculto de Markov para convertir los números de copias en estados discretos para cada ventana.

Se generan informes, mapeando las posiciones del genoma y la variación del número de copias que se muestran en la Fig. 5A, para un sujeto con cáncer en remisión y en la Fig. 5B para un sujeto con cáncer en recurrencia.

Este informe en comparación con otros perfiles de sujetos con resultados conocidos indica que, en el mes 18, se detecta un análisis de mutación rara para la variación del número de copias con una carga tumoral libre de células del 5%. Un oncólogo prescribe tratamiento de nuevo.

Ejemplo 3 - Cáncer de Tiroides y Tratamiento

Se sabe que un sujeto tiene cáncer de tiroides en estadio IV y se somete a un tratamiento estándar, incluyendo radioterapia con 1-131. Las tomografías computarizadas no son concluyentes en cuanto a si la radioterapia está destruyendo las masas cancerosas. Se extrae sangre antes y después de la última sesión de radiación.

Se extrae ADN libre de células y se aísla de la sangre mediante el protocolo del kit Qiagen Qubit. Se añade una muestra de ADN a granel no específico a las reacciones de preparación de la muestra para aumentar los rendimientos.

Se sabe que el gen BRAF puede estar mutado en la posición del aminoácido 600 en este cáncer de tiroides. A partir de la población de ADN libre de células, el ADN de BRAF se amplifica selectivamente usando cebadores específicos para el gen. Se añaden códigos de barras de 20 mer se agregan a la molécula parental como control para contar las lecturas.

Se secuencian 10 ng de ADN usando un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq.

Los datos de secuencia se ensamblan y analizan para detectar la variación del número de copias. Las lecturas de secuencia se mapean y se comparan con un individuo sano (control). En base al número de lecturas de secuencia, según se determina contando las secuencias de códigos de barras, las regiones cromosómicas se dividen en regiones no superpuestas de 50 kb. Las lecturas de secuencia se comparan entre sí y se determina una proporción para cada posición mapeable.

Se aplica un modelo oculto de Markov para convertir los números de copias en estados discretos para cada ventana.

Se genera un informe, mapeando las posiciones del genoma y la variación del número de copias.

Se comparan los informes generados antes y después del tratamiento. El porcentaje de carga de células tumorales aumenta del 30% al 60% después de la sesión de radiación. Se determina que el aumento en la carga tumoral es un aumento en la necrosis del tejido canceroso frente al tejido normal como resultado del tratamiento. Los oncólogos recomiendan que el sujeto continúe el tratamiento prescrito.

Ejemplo 4 - Sensibilidad de Detección de Mutaciones Raras

Para determinar los intervalos de detección de mutaciones raras presentes en una población de ADN, se realizan experimentos de mezcla. Las secuencias de ADN, algunas que contienen copias de tipo salvaje de los genes TP53, HRAS y MET y otras que contienen copias con mutaciones raras en los mismos genes, se mezclan entre sí a distintas proporciones. Las mezclas de ADN se preparan de tal manera que las proporciones o porcentajes de ADN mutante a ADN de tipo salvaje varía del 100% al 0,01%.

Se secuencian 10 ng de ADN para cada experimento de mezcla usando un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq.

Los datos de secuencia se ensamblan y analizan para detectar mutaciones raras. Las lecturas de secuencia se mapean y se comparan con una secuencia de referencia (control). En base al número de lecturas de secuencia, se determina la frecuencia de variación para cada posición mapeable.

Se aplica un modelo oculto de Markov para convertir la frecuencia de varianza para cada posición mapeable a estados discretos para la posición de base.

Se genera un informe, mapeando las posiciones de base del genoma y el porcentaje de detección de la mutación rara sobre el valor de referencia, según se determina por la secuencia de referencia (Fig. 6A).

Los resultados de varios experimentos de mezcla que varían del 0,1% al 100% se representan en un gráfico a escala logarítmica, con un porcentaje medido de ADN con una mutación rara representada como una función del porcentaje real de ADN con una mutación rara (Fig. 6B). Los tres genes, TP53, HRAS y MET están representados. Se encuentra una fuerte correlación lineal entre las poblaciones de mutaciones raras medidas y esperadas. Adicionalmente, con estos experimentos se encuentra un umbral de sensibilidad más bajo de

aproximadamente el 0,1% del ADN con una mutación rara en una población de ADN no mutado (Fig. 6B).

Ejemplo 5 - Detección de Mutaciones Raras en un Sujeto con Cáncer de Próstata

Se cree que un sujeto tiene cáncer de próstata en etapa temprana. Otras pruebas clínicas proporcionan resultados no concluyentes. Se extrae sangre del sujeto y se extrae ADN libre de células, se aísla, se prepara y se secuencia.

Se seleccionó un panel de varios oncogenes y genes supresores de tumores para la amplificación selectiva usando un kit de PCR TaqMan © (Invitrogen) usando cebadores específicos de los genes. Las regiones de ADN amplificadas incluyen el ADN que contiene los genes PIK3CA y TP53.

Se secuencian 10 ng de ADN usando un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq.

Los datos de secuencia se ensamblan y analizan para detectar mutaciones raras. Las lecturas de secuencia se mapean y se comparan con una secuencia de referencia (control). En base al número de lecturas de secuencia, se determinó la frecuencia de varianza para cada posición mapeable.

Se aplica un modelo oculto de Markov para convertir la frecuencia de varianza para cada posición mapeable a estados discretos para cada posición base.

Se genera un informe que mapeando las posiciones de las bases genómicas y el porcentaje de detección de la mutación rara sobre el valor de referencia, según se determina por la secuencia de referencia (La figura 7A). Se encuentran mutaciones raras con una incidencia del 5% en dos genes, PIK3CA y TP53, respectivamente, indicando que el sujeto tiene un cáncer en etapa temprana. Se inicia el tratamiento.

Además, estos informes se envían y se accede a ellos electrónicamente a través de Internet. El análisis de los datos de secuencia se realiza en un sitio diferente a la localización del sujeto. El informe se genera y se transmite a la localización del sujeto. A través de un ordenador con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral (Fig. 7B).

Ejemplo 6 - Detección de Mutaciones Raras en Sujetos con Cáncer Colorrectal

Se cree que un sujeto tiene cáncer colorrectal en etapa intermedia. Otras pruebas clínicas proporcionan resultados no concluyentes. Se extrae sangre del sujeto y se extrae ADN libre de células.

Se usan 10 ng del material genético libre de células que se extrae de un solo tubo de plasma. El material genético inicial se convierte en un conjunto de polinucleótidos parentales marcados. El marcado incluía unir los marcadores requeridos para la secuenciación, así como identificadores no únicos para rastrear las moléculas de la progenie hasta los ácidos nucleicos parentales. La conversión se realiza a través de una reacción de ligación optimizada como se ha descrito anteriormente y el rendimiento de la conversión se confirma observando el perfil de tamaño de las moléculas después de la ligación. El rendimiento de la conversión se mide como el porcentaje de moléculas iniciales de partida que tienen ambos extremos ligados con marcadores. La conversión usando este enfoque se realiza con una eficiencia alta, por ejemplo, por lo menos el 50%.

La biblioteca marcada se amplifica por PCR y se enriquece para los genes más asociados con el cáncer colorrectal (por ejemplo, KRAS, APC, TP53, etc.) y el ADN resultante se secuencia mediante un enfoque de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq.

Los datos de secuencia se ensamblan y analizan para detectar mutaciones raras. Las lecturas de secuencia se colapsan en grupos familiares que pertenecen a una molécula parental (y se corrigen por error al colapsar) y se mapean usando una secuencia de referencia (control). En base al número de lecturas de secuencia, se determina la frecuencia de variaciones raras (sustituciones, inserciones, eliminaciones, etc.) y las variaciones en el número de copias y la heterocigosidad (cuando sea apropiado) para cada posición asignable.

Se genera un informe, mapeando las posiciones de bases genómicas y el porcentaje de detección de la mutación rara sobre el valor de referencia según se determina por la secuencia de referencia. Las mutaciones raras se encuentran con una incidencia del 0,3-0,4% en dos genes, KRAS y FBXW7, respectivamente, indicando que el sujeto tiene cáncer residual. Se inicia el tratamiento.

Además, estos informes se envían y se accede a ellos electrónicamente a través de Internet. El análisis de los datos de secuencia se realiza en un sitio diferente a la localización del sujeto. El informe se genera y se transmite a la localización del sujeto. A través de un ordenador con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral.

Ejemplo 7 - Tecnología de Secuenciación Digital

Las concentraciones de ácidos nucleicos desprendidos de tumores son típicamente tan bajas que las tecnologías de secuenciación de próxima generación actuales solo pueden detectar tales señales esporádicamente o en pacientes con una carga tumoral terminalmente alta. La razón principal es que tales tecnologías están plagadas de tasas de error y sesgos que pueden ser de órdenes de magnitud superiores a lo que se requiere para detectar de manera confiable las alteraciones genéticas de novo asociadas con el cáncer en el ADN circulante. Aquí se muestra una nueva metodología de secuenciación, la tecnología de secuenciación digital (DST), que aumenta la sensibilidad y especificidad de la detección y cuantificación de ácidos nucleicos derivados de tumores raros entre los fragmentos de la línea germinal en por lo menos 1-2 órdenes de magnitud.

La arquitectura DST está inspirada en los sistemas de comunicación digital de tecnología avanzada que combaten el ruido y la distorsión altos provocados por los canales de comunicación modernos y son capaces de transmitir información digital sin problemas a velocidades de datos extremadamente altas. De manera similar, los flujos de trabajo de próxima generación actuales están plagados de ruido y distorsión extremadamente altos (debido a la preparación de la muestra, la amplificación basada en PCR y la secuenciación). La secuenciación digital es capaz de eliminar el error y la distorsión creados por estos procesos y producir una representación casi perfecta de todas las variantes raras (incluyendo las CNV).

Preparación de Biblioteca de Alta Diversidad

A diferencia de los protocolos de preparación de bibliotecas de secuenciación convencionales, en los que la mayoría de los fragmentos de ADN circulantes extraídos se pierden debido a la conversión de bibliotecas ineficiente, nuestro flujo de trabajo de tecnología de secuenciación digital permite que la gran mayoría de las moléculas de partida se conviertan y secuencien. Esto es críticamente importante para la detección de variantes raras, ya que solo puede haber un puñado de moléculas mutadas somáticamente en un tubo de 10 ml completo de sangre. El eficiente proceso de conversión de biología molecular desarrollado permite la mayor sensibilidad posible para la detección de variantes raras.

Panel de Oncogén Procesable Integral

El flujo de trabajo diseñado alrededor de la plataforma DST es flexible y altamente ajustable, ya que las regiones objetivo pueden ser tan pequeñas como exones individuales o tan amplias como exomas completos (o incluso genomas completos). Un panel estándar consta de todas las bases exónicas de 15 genes procesables relacionados con el cáncer y la cobertura de los exones "calientes" de otros 36 oncogenes/genes supresores de tumores (por ejemplo, exones que contienen por lo menos una o más mutaciones somáticas informadas en el COSMIC) .

Ejemplo 8: Estudios analíticos

Para estudiar el rendimiento de nuestra tecnología, se evaluó su sensibilidad en muestras analíticas. Adicionamos cantidades variables de ADN de línea celular de cáncer de LNCaP en un fondo de ADNcf normal y fuimos capaces de detectar con éxito mutaciones somáticas hasta una sensibilidad del 0,1% (ver Figura 13).

Estudios Preclínicos

Se investigó la concordancia del ADN circulante con el ADNg de tumor en modelos de xenoinjerto humano en ratones. En siete ratones CTC negativos, cada uno con uno de dos tumores diferentes de cáncer de mama humano, todas las mutaciones somáticas detectadas en el ADNg del tumor también se detectaron en el ADNcf de la sangre del ratón usando DST validando aún más la utilidad del ADNcf para realización de perfiles genéticos tumorales no invasivos.

Estudios Clínicos Piloto

Correlación de Biopsia Tumoral frente a Mutaciones Somáticas de ADN Circulante

Se inició un estudio piloto en muestras humanas de diferentes tipos de cánceres. Se investigó la concordancia de los perfiles de mutación tumoral derivados del ADN libre de células circulante con los derivados de muestras de biopsias tumorales emparejadas. Se encontró una concordancia superior al 93% entre el tumor y los perfiles de mutación somática de ADNcf en cánceres colorrectales y de melanoma en 14 pacientes (Tabla 1).

Tabla 1

ID del paciente	Etapa	Genes mutantes en el tumor emparejado	Porcentaje de ADNcf mutante
CRC Nº 1	II-B	TP53	0.2%
CRC Nº 2	II-C	KRAS	0.6%
		SMAD4	1.5%
		GNAS	1.4%
		FBXW7	0.8%
CRC Nº 3	III-B	KRAS	1.1%
		TP53	1.4%
		PIK3CA	1.7%
		APC	0.7%
CRC Nº 4	III-B	KRAS	0.3%
		TP53	0.4%
CRC Nº 5	III-B	KRAS	0.04%
CRC Nº 6	3-100	KRAS	0.03%
CRC Nº 7	IV	PIK3CA	1.3%
		KRAS	0.6%
		TP53	0.8%
CRC Nº 8	IV	APC	0.3%
		SMO	0.6%
		TP53	0.4%
		KRAS	0.0%
CRC Nº 9	IV	APC	47.3%
		APC	40.2%
		KRAS	37.7%
		PTEN	0.0%
		TP53	12.9%
CRC Nº 10	IV	TP53	0.9%
Melanoma Nº1	IV	BRAF	0.2%
Melanoma Nº2	IV	APC	0.3%
		EGFR	0.9%
		MI C	10.5%
Melanoma Nº3	IV	BRAF	3.3%
Melanoma Nº4	IV	BRAF	0.7%

Debe entenderse a partir de lo anterior que, aunque se han ilustrado y descrito implementaciones particulares, pueden hacerse varias modificaciones a las mismas y se contemplan en la presente. Tampoco 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, las descripciones e ilustraciones de las realizaciones preferidas en la presente no deben interpretarse en un sentido limitativo. Además, debe entenderse que todos los aspectos de la invención no están limitados a las representaciones, configuraciones o proporciones relativas específicas expuestas en la presente que dependen de una variedad de condiciones y variables. Varias modificaciones en la forma y en el detalle de las realizaciones de la invención serán aparentes para los expertos en la técnica.

REIVINDICACIONES

1. Un método para determinar la variación en el número de copias en una muestra que incluye polinucleótidos libres de células, el método comprendiendo:

a. proporcionar por lo menos dos conjuntos de polinucleótidos libres de células, que mapean para diferentes posiciones mapeables en una secuencia de referencia en un genoma, y, para los conjuntos de polinucleótidos libres de células;

i. marcar de forma no única los polinucleótidos libres de células con un conjunto de códigos de barras moleculares;
ii. amplificar los polinucleótidos libres de células para producir polinucleótidos amplificados;
iii. secuenciar un subconjunto del conjunto de polinucleótidos amplificados, para producir un conjunto de lecturas de secuenciación;
iv. agrupar el conjunto de lecturas de secuenciación secuenciadas a partir de polinucleótidos amplificados en familias que corresponden a lecturas de secuenciación de polinucleótidos amplificados a partir del mismo polinucleótido libre de células;
v. inferir una medida cuantitativa de familias en los conjuntos; y

b. determinar la variación en el número de copias en base a la medida cuantitativa de las familias en los conjuntos.

2. El método de la reivindicación 1, en el que la muestra se extrae de un fluido seleccionado del grupo que consiste de sangre, plasma, suero, vítreo, esputo, orina, lágrimas, transpiración, saliva, semen, excreciones mucosales, moco, líquido cefalorraquídeo, líquido amniótico y líquido linfático.

3. El método de la reivindicación 1 o la reivindicación 2, en el que los polinucleótidos libres de células comprenden polinucleótidos derivados de ADN genómico tumoral.

4. El método de cualquiera de las reivindicaciones 1-3, en el que los códigos de barras moleculares se unen a los polinucleótidos libres de células a través de una reacción enzimática como una reacción de ligadura.

5. El método de la reivindicación 4, en el que hay entre 2 y 1.000.000 de códigos de barras moleculares diferentes en el conjunto de códigos de barras moleculares, y en el que la muestra comprende entre 100 y 100.000 equivalentes de genoma haploide de polinucleótidos de ADN libre de células (ADNcf).

6. El método de cualquiera de las reivindicaciones 1-5, que comprende además regiones selectivamente enriquecidas de un genoma o transcriptoma del sujeto antes de la secuenciación.

7. El método de cualquiera de las reivindicaciones 1-6, que comprende además filtrar las lecturas de secuenciación con una puntuación de precisión o calidad inferior a un umbral y/o puntuación de mapeo inferior a un umbral.

8. El método de cualquiera de las reivindicaciones 1-7, en el que las lecturas de secuenciación se agrupan en familias en base a la secuencia de código de barras no única en combinación con los datos de secuencia en las partes de principio (inicio) y final (parada) de las lecturas de secuenciación, opcionalmente combinando además la longitud de las lecturas de secuenciación.

9. El método de cualquiera de las reivindicaciones 1-8, en el que inferir una medida cuantitativa de familias en el conjunto comprende determinar el número de familias que mapean para diferentes loci de referencia.

10. El método de la reivindicación 1, que comprende además inferir una medida cuantitativa del número de lecturas de secuencia dentro de las familias.

11. El método de cualquiera de las reivindicaciones 1-9, en el que la medida cuantitativa se normaliza para el sesgo representacional durante el proceso de secuenciación.

12. El método de cualquiera de las reivindicaciones 1-11, en el que la medida cuantitativa es un recuento.

13. Un medio legible por ordenador que comprende código ejecutable por máquina no transitorio que, tras la ejecución por un procesador informático, implementa un método, el método comprendiendo:

a. acceder a un archivo de datos que comprende una pluralidad de lecturas de secuenciación, en donde las lecturas de secuencia se derivan de polinucleótidos de progenie amplificados a partir de polinucleótidos libres de células originales marcados de manera no única;
b. agrupar las lecturas de secuenciación secuenciadas a partir de los polinucleótidos de progenie en familias

que comprenden lecturas de secuenciación de polinucleótidos de progenie amplificados a partir del mismo polinucleótido libre de células original marcado;

c. inferir una medida cuantitativa de familias en los polinucleótidos libres de células originales marcados de manera no única; y

d. determinar la variación en el número de copias comparando la medida cuantitativa de familias en los polinucleótidos libres de células originales marcados de manera no única.

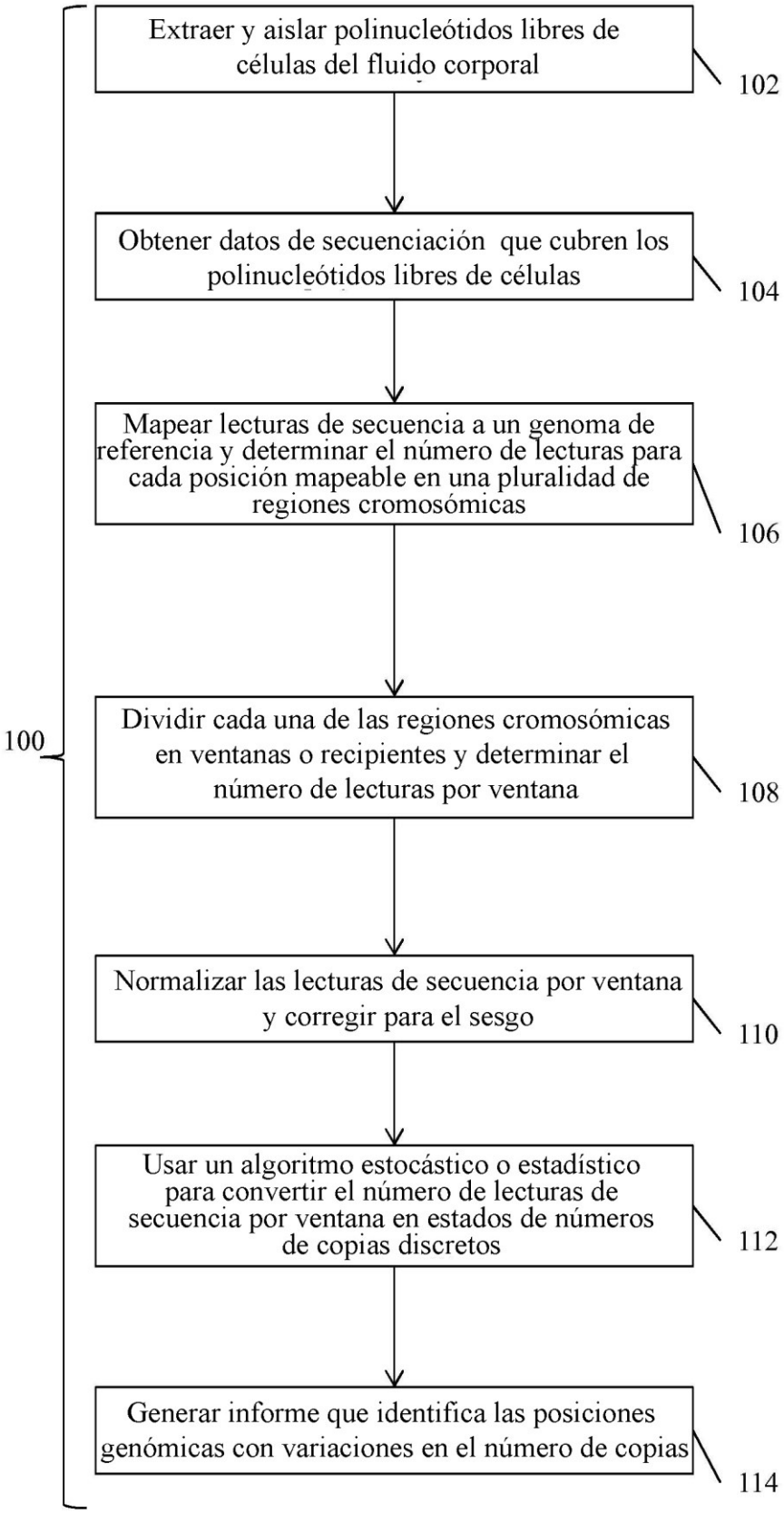

Extraer y aislar polinucleótidos libres de células del fluido corporal
102
Obtener datos de secuenciación que cubren los polinucleótidos libres de células
104
Mapear lecturas de secuencia a un genoma de referencia y determinar el número de lecturas para cada posición mapeable en una pluralidad de regiones cromosómicas
106
Dividir cada una de las regiones cromosómicas en ventanas o recipientes y determinar el número de lecturas por ventana
108
Normalizar las lecturas de secuencia por ventana y corregir para el sesgo
110
Usar un algoritmo estocástico o estadístico para convertir el número de lecturas de secuencia por ventana en estados de números de copias discretos
112
Generar informe que identifica las posiciones genómicas con variaciones en el número de copias
114
100

Fig. 1

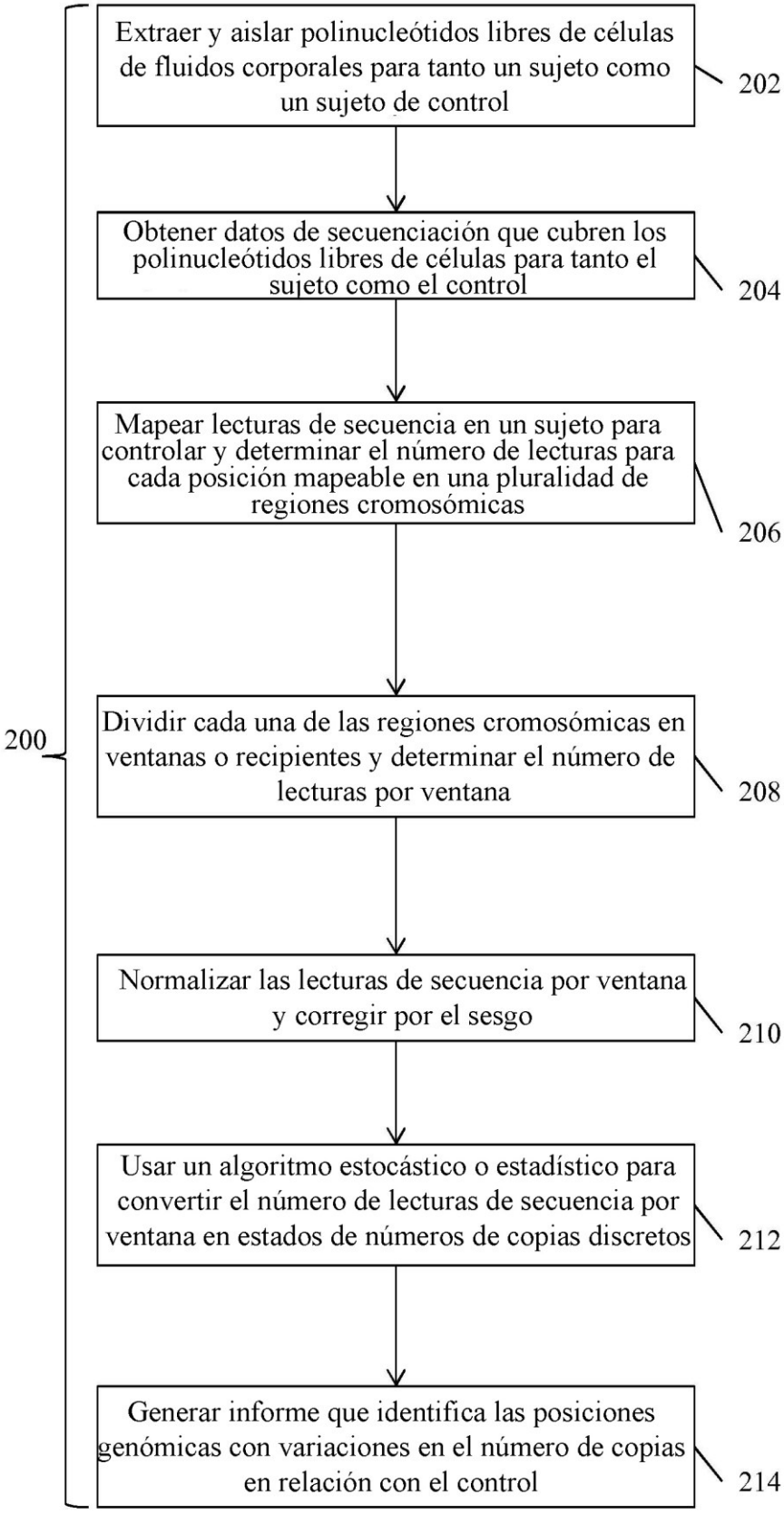

Extraer y aislar polinucleótidos libres de células de fluidos corporales para tanto un sujeto como un sujeto de control
202
Obtener datos de secuenciación que cubren los polinucleótidos libres de células para tanto el sujeto como el control
204
Mapear lecturas de secuencia en un sujeto para controlar y determinar el número de lecturas para cada posición mapeable en una pluralidad de regiones cromosómicas
206
Dividir cada una de las regiones cromosómicas en ventanas o recipientes y determinar el número de lecturas por ventana
208
Normalizar las lecturas de secuencia por ventana y corregir por el sesgo
210
Usar un algoritmo estocástico o estadístico para convertir el número de lecturas de secuencia por ventana en estados de números de copias discretos
212
Generar informe que identifica las posiciones genómicas con variaciones en el número de copias en relación con el control
214
200

Fig. 2

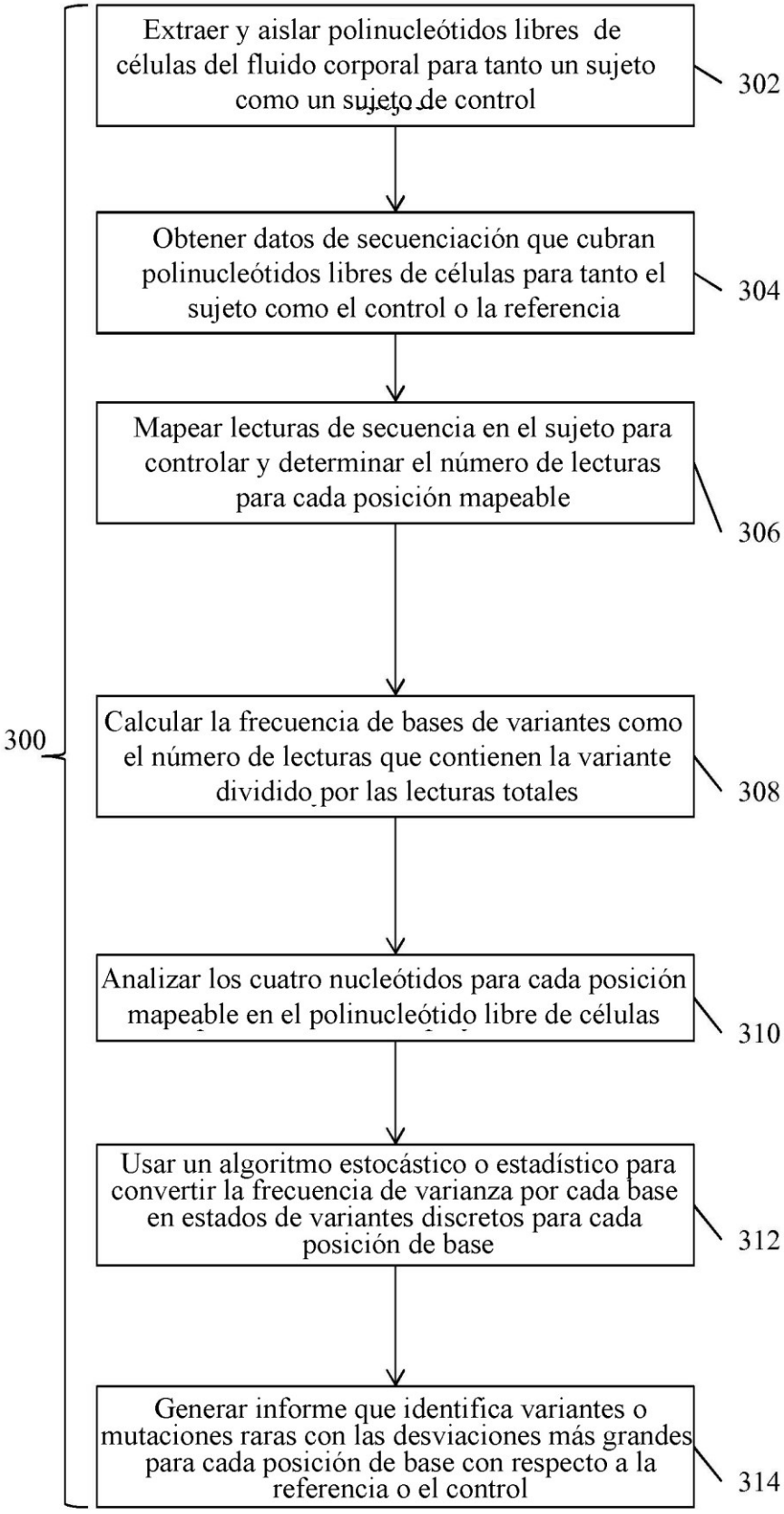
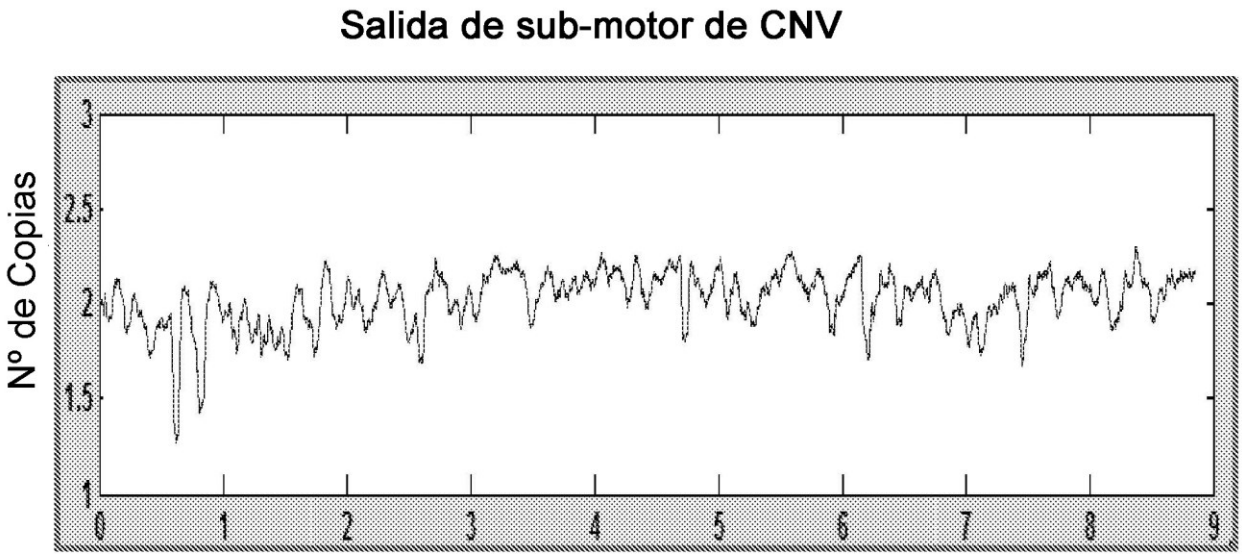
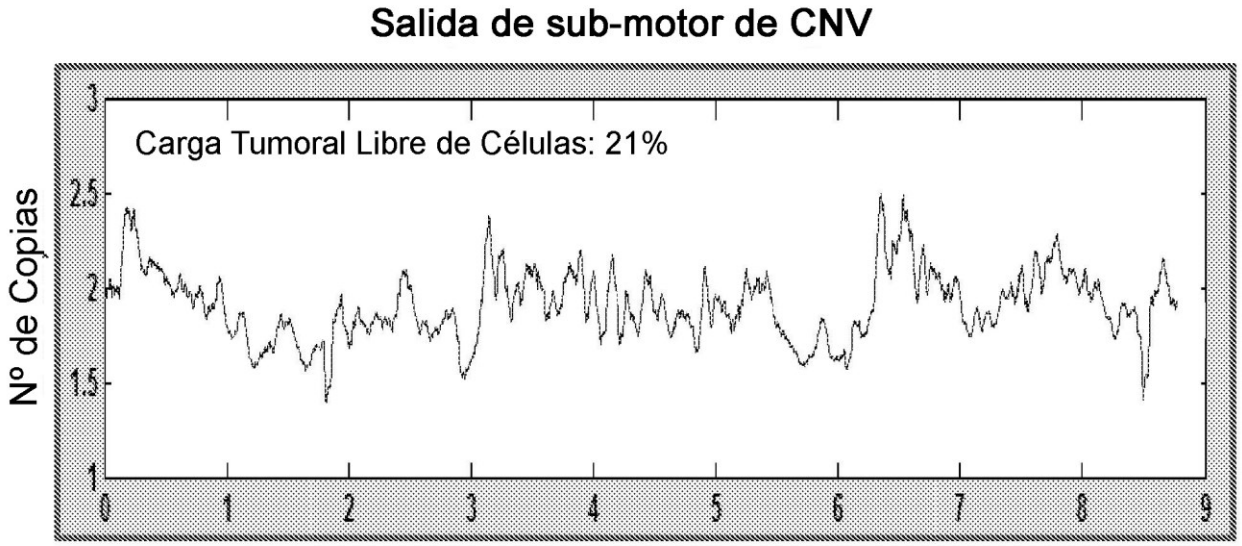

Extraer y aislar polinucleótidos libres de células del fluido corporal para tanto un sujeto como un sujeto de control
302
Obtener datos de secuenciación que cubran polinucleótidos libres de células para tanto el sujeto como el control o la referencia
304
Mapear lecturas de secuencia en el sujeto para controlar y determinar el número de lecturas para cada posición mapeable
306
300
Calcular la frecuencia de bases de variantes como el número de lecturas que contienen la variante dividido por las lecturas totales
308
Analizar los cuatro nucleótidos para cada posición mapeable en el polinucleótido libre de células
310
Usar un algoritmo estocástico o estadístico para convertir la frecuencia de varianza por cada base en estados de variantes discretos para cada posición de base
312
Generar informe que identifica variantes o mutaciones raras con las desviaciones más grandes para cada posición de base con respecto a la referencia o el control
314

Fig. 3



Muestra Normal 1

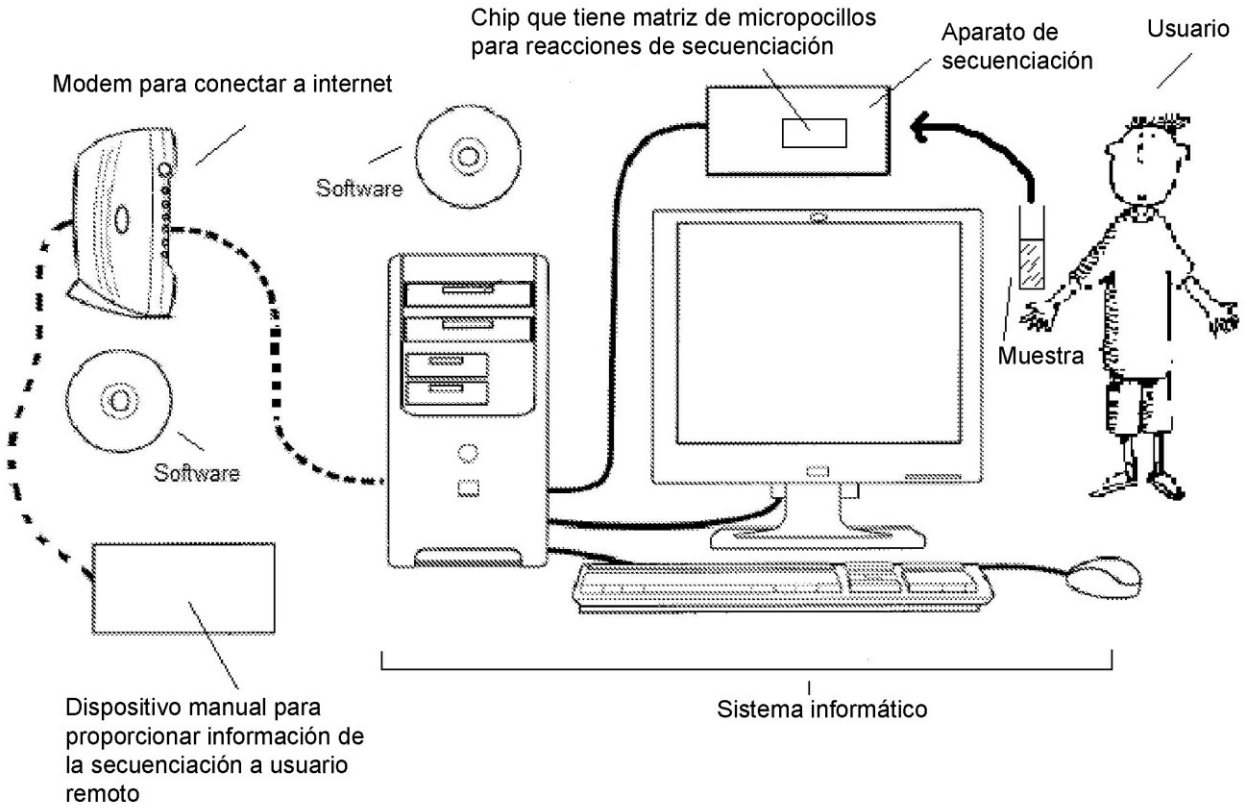
A



Paciente con Cáncer de Próstata 1

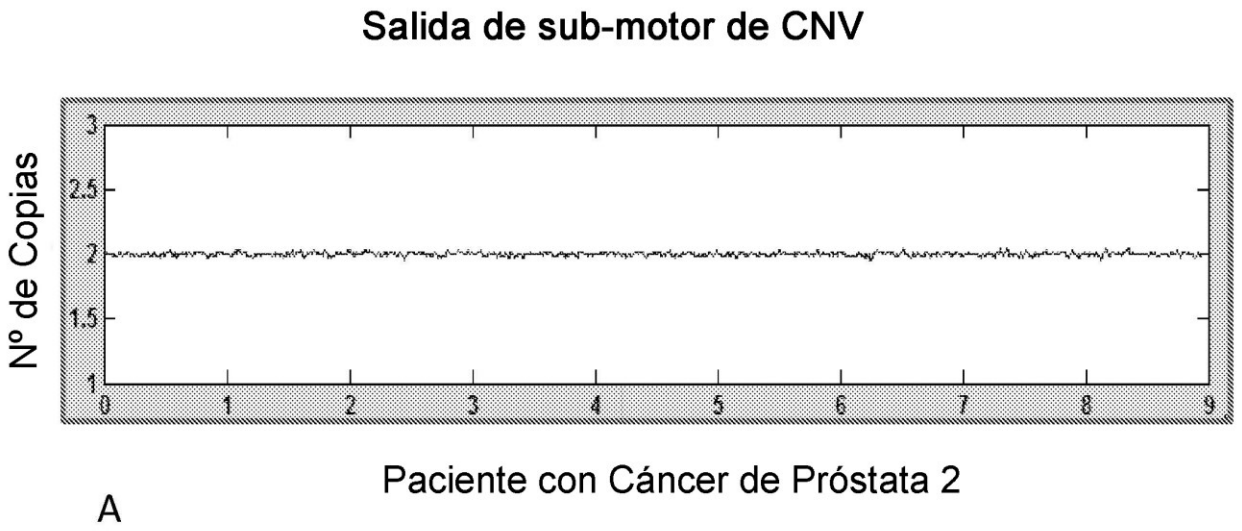
B

Fig. 4

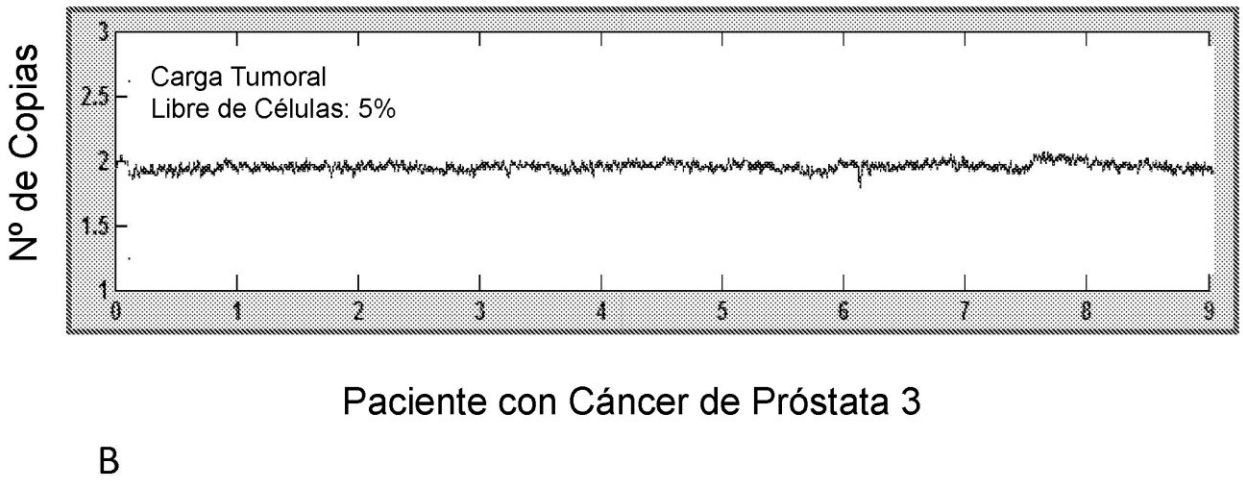

Chip que tiene matriz de micropocillos para reacciones de secuenciación
Aparato de secuenciación
Usuario
Modem para conectar a internet
Software
Software
Muestra
Dispositivo manual para proporcionar información de la secuenciación a usuario remoto
Sistema informático

C

Fig. 4

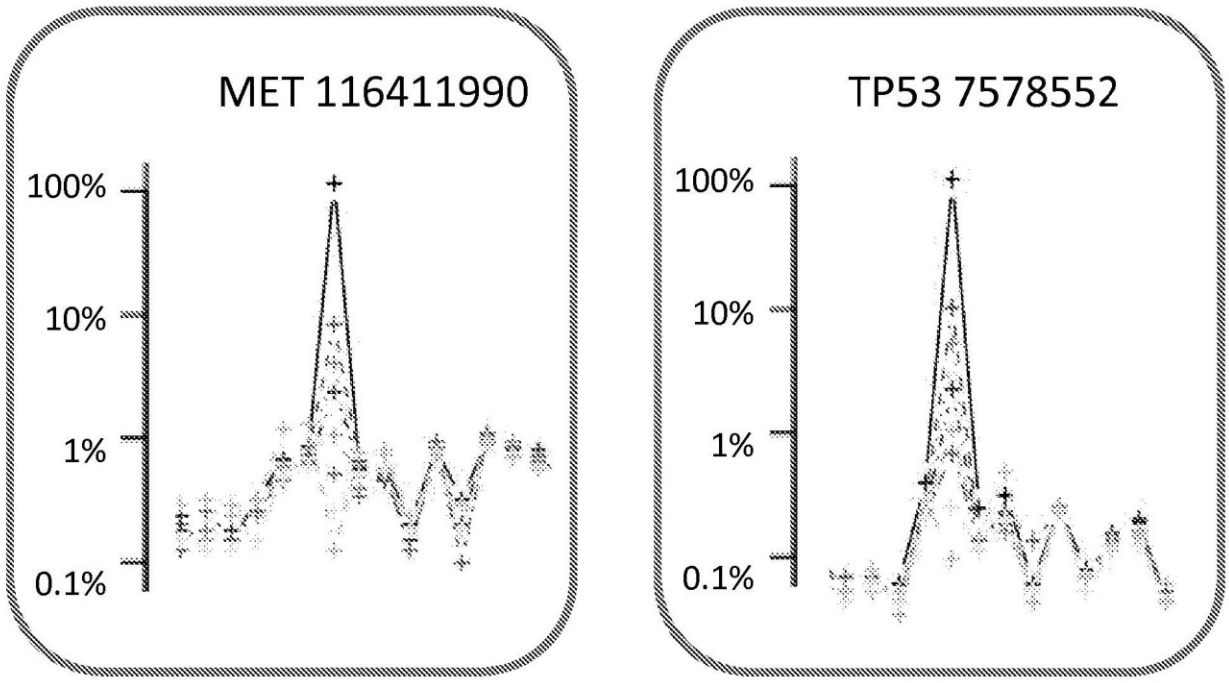

Salida de sub-motor de CNV
Nº de Copias
3
2.5
2
1.5
1
0
1
2
3
4
5
6
7
8
9
Paciente con Cáncer de Próstata 2

A

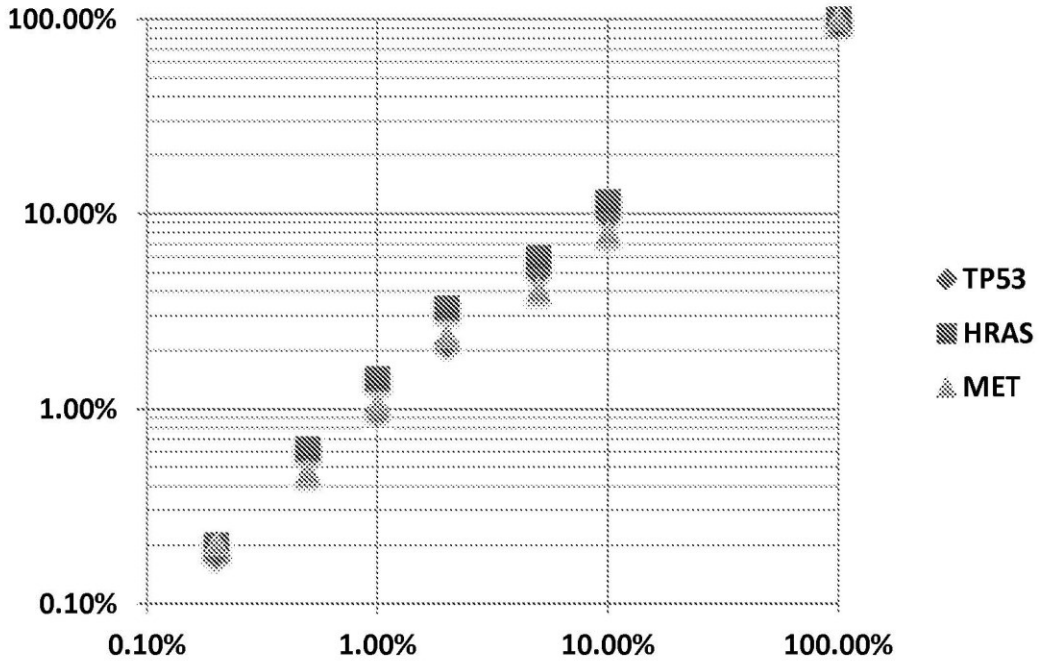

Nº de Copias
Carga Tumoral
Libre de Células: 5%
3
2.5
2
1.5
1
0
1
2
3
4
5
6
7
8
9
Paciente con Cáncer de Próstata 3

B

Fig. 5

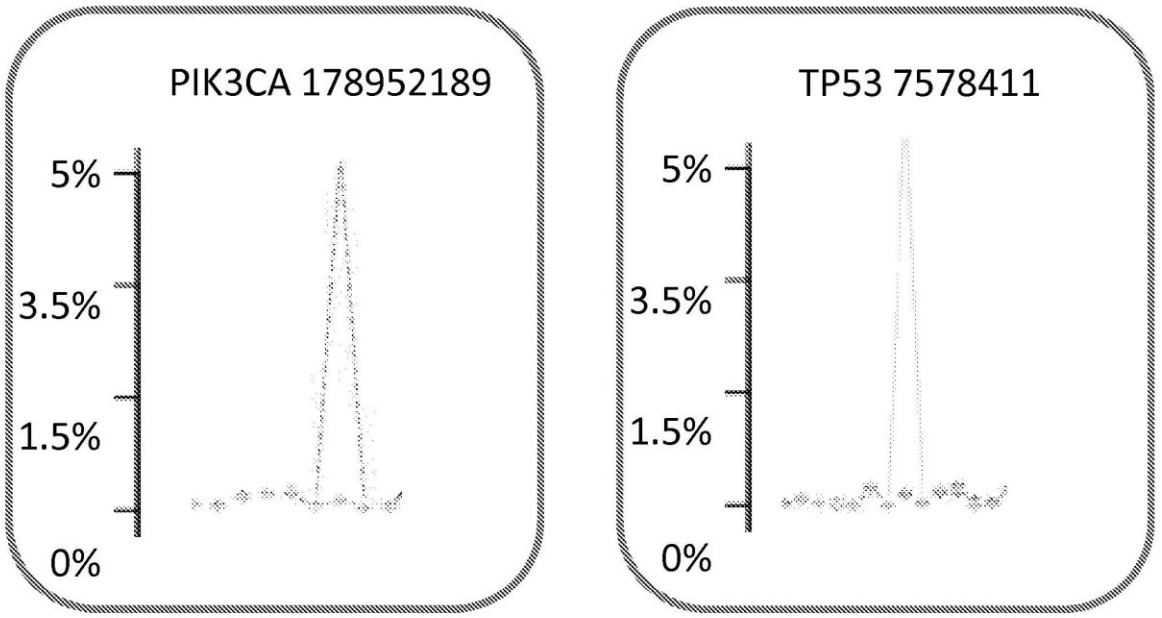

MET 116411990
100%
10%
1%
0.1%
TP53 7578552
100%
10%
1%
0.1%

A

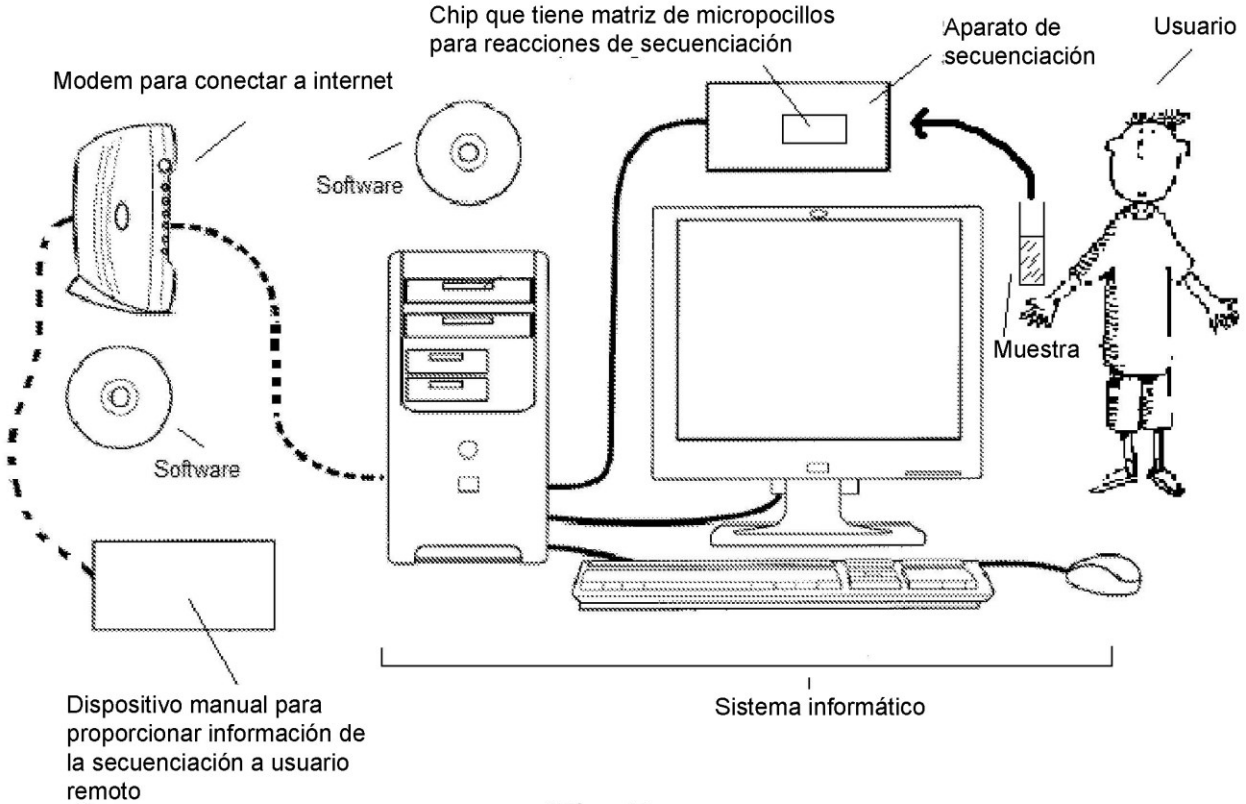

100.00%
10.00%
1.00%
0.10%
0.10%
1.00%
10.00%
100.00%
TP53
HRAS
MET

B

Fig. 6


PIK3CA 178952189
5%
3.5%
1.5%
0%
TP53 7578411
5%
3.5%
1.5%
0%

A


Chip que tiene matriz de micropocillos para reacciones de secuenciación
Aparato de secuenciación
Usuario
Modem para conectar a internet
Software
Muestra
Software
Dispositivo manual para proporcionar información de la secuenciación a usuario remoto
Sistema informático

B

Fig. 7

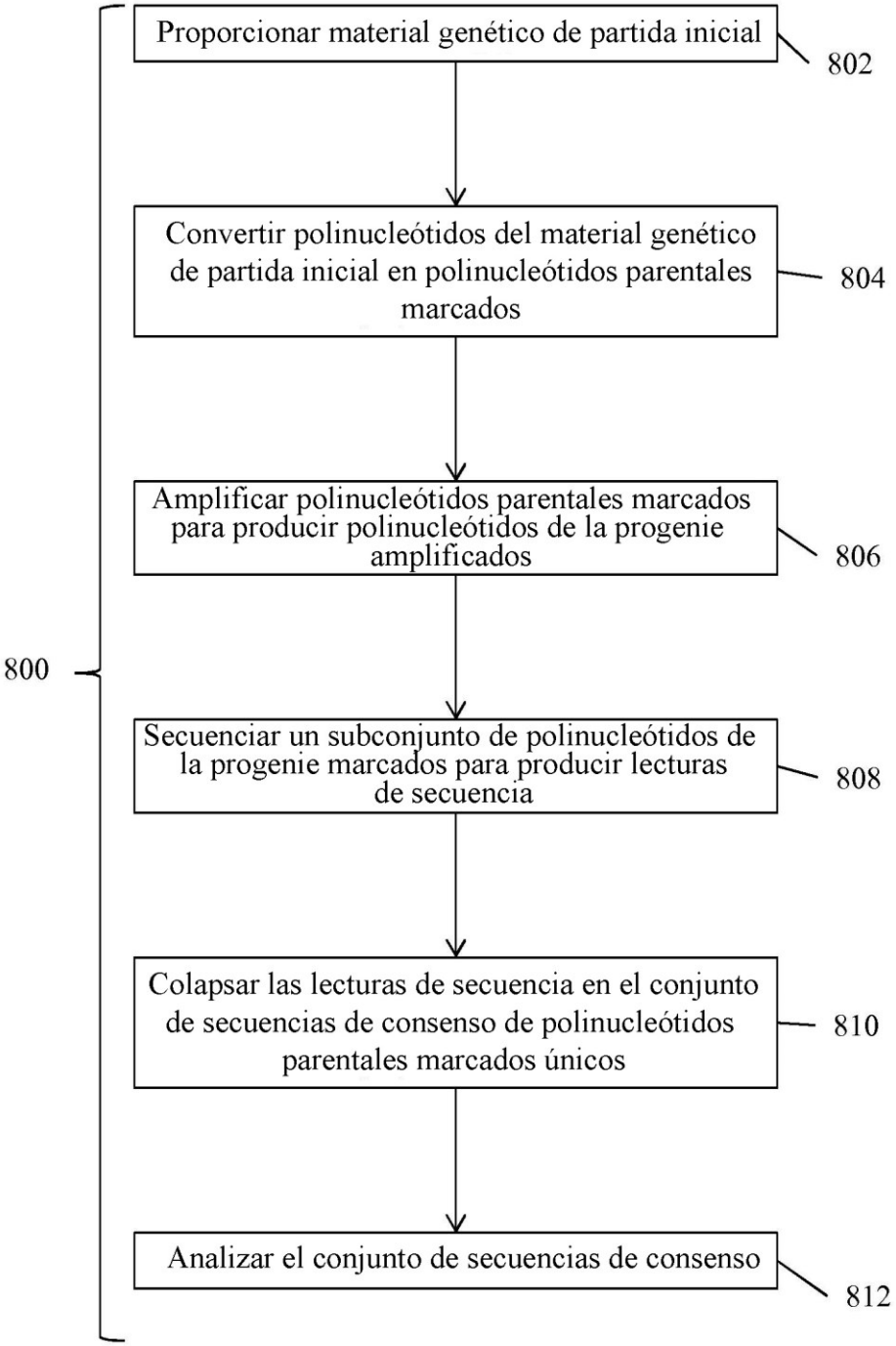

Proporcionar material genético de partida inicial
802
Convertir polinucleótidos del material genético de partida inicial en polinucleótidos parentales marcados
804
Amplificar polinucleótidos parentales marcados para producir polinucleótidos de la progenie amplificados
806
800
Secuenciar un subconjunto de polinucleótidos de la progenie marcados para producir lecturas de secuencia
808
Colapsar las lecturas de secuencia en el conjunto de secuencias de consenso de polinucleótidos parentales marcados únicos
810
Analizar el conjunto de secuencias de consenso
812

Fig. 8

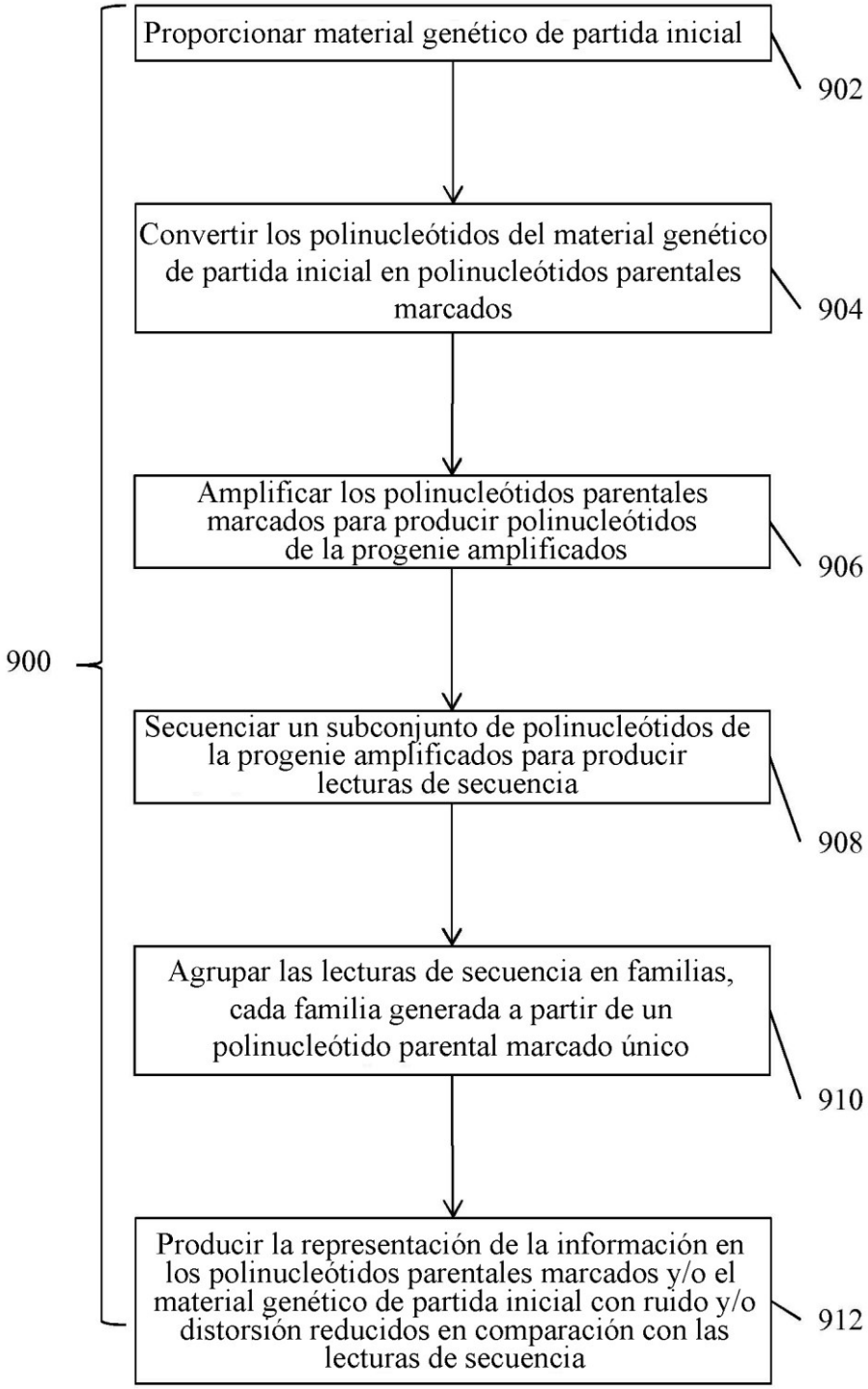

Proporcionar material genético de partida inicial
902
Convertir los polinucleótidos del material genético de partida inicial en polinucleótidos parentales marcados
904
Amplificar los polinucleótidos parentales marcados para producir polinucleótidos de la progenie amplificados
906
900
Secuenciar un subconjunto de polinucleótidos de la progenie amplificados para producir lecturas de secuencia
908
Agrupar las lecturas de secuencia en familias, cada familia generada a partir de un polinucleótido parental marcado único
910
Producir la representación de la información en los polinucleótidos parentales marcados y/o el material genético de partida inicial con ruido y/o distorsión reducidos en comparación con las lecturas de secuencia
912

Fig. 9

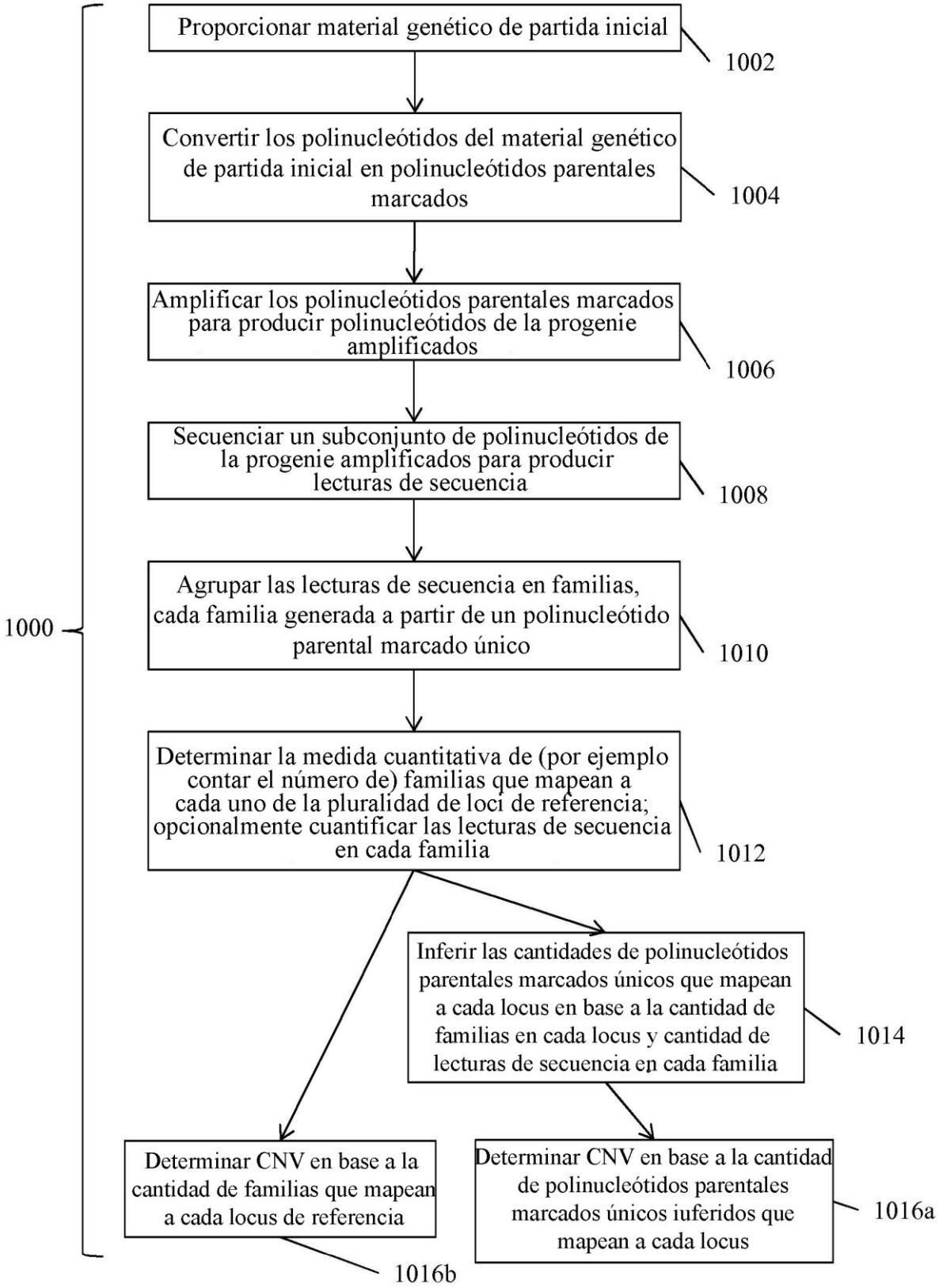

Proporcionar material genético de partida inicial
1002
Convertir los polinucleótidos del material genético de partida inicial en polinucleótidos parentales marcados
1004
Amplificar los polinucleótidos parentales marcados para producir polinucleótidos de la progenie amplificados
1006
Secuenciar un subconjunto de polinucleótidos de la progenie amplificados para producir lecturas de secuencia
1008
Agrupar las lecturas de secuencia en familias, cada familia generada a partir de un polinucleótido parental marcado único
1010
1000
Determinar la medida cuantitativa de (por ejemplo contar el número de) familias que mapean a cada uno de la pluralidad de loci de referencia; opcionalmente cuantificar las lecturas de secuencia en cada familia
1012
Inferir las cantidades de polinucleótidos parentales marcados únicos que mapean a cada locus en base a la cantidad de familias en cada locus y cantidad de lecturas de secuencia en cada familia
1014
Determinar CNV en base a la cantidad de familias que mapean a cada locus de referencia
1016b
Determinar CNV en base a la cantidad de polinucleótidos parentales marcados únicos iuferidos que mapean a cada locus
1016a

Fig. 10

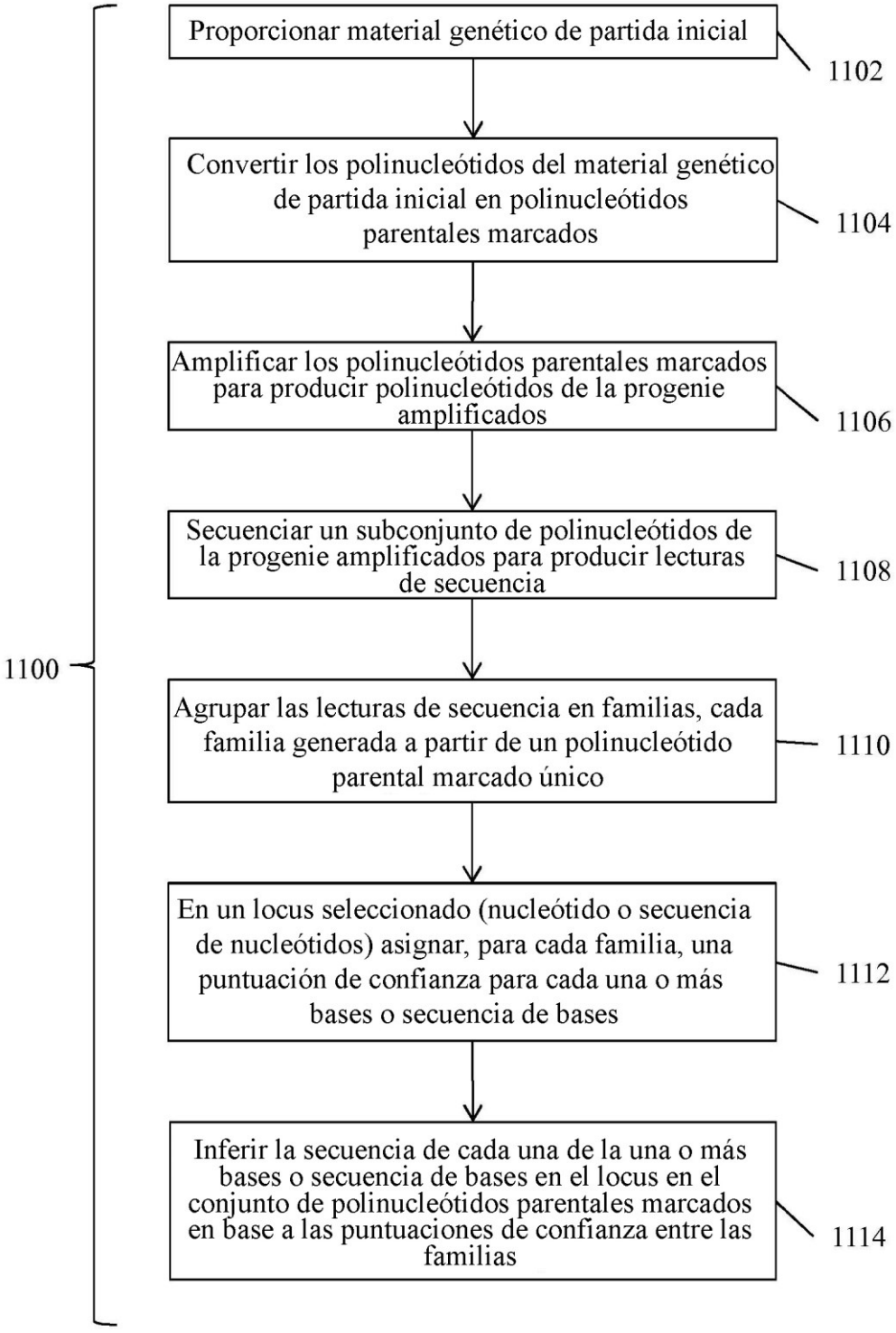

Proporcionar material genético de partida inicial
1102
Convertir los polinucleótidos del material genético de partida inicial en polinucleótidos parentales marcados
1104
Amplificar los polinucleótidos parentales marcados para producir polinucleótidos de la progenie amplificados
1106
Secuenciar un subconjunto de polinucleótidos de la progenie amplificados para producir lecturas de secuencia
1108
1100
Agrupar las lecturas de secuencia en familias, cada familia generada a partir de un polinucleótido parental marcado único
1110
En un locus seleccionado (nucleótido o secuencia de nucleótidos) asignar, para cada familia, una puntuación de confianza para cada una o más bases o secuencia de bases
1112
Inferir la secuencia de cada una de la una o más bases o secuencia de bases en el locus en el conjunto de polinucleótidos parentales marcados en base a las puntuaciones de confianza entre las familias
1114

Fig. 11

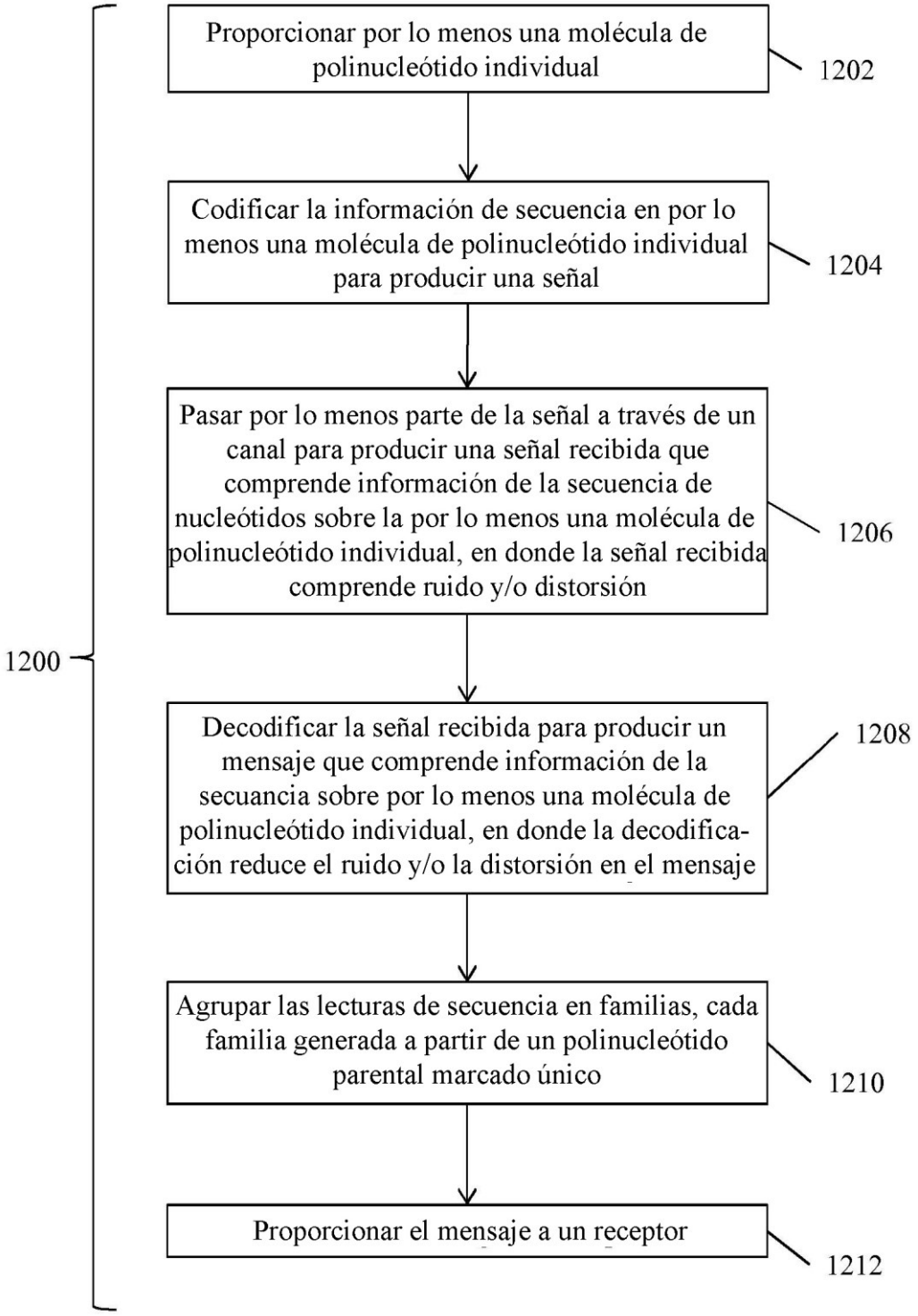

Proporcionar por lo menos una molécula de polinucleótido individual
1202
Codificar la información de secuencia en por lo menos una molécula de polinucleótido individual para producir una señal
1204
Pasar por lo menos parte de la señal a través de un canal para producir una señal recibida que comprende información de la secuencia de nucleótidos sobre la por lo menos una molécula de polinucleótido individual, en donde la señal recibida comprende ruido y/o distorsión
1206
1200
Decodificar la señal recibida para producir un mensaje que comprende información de la secuancia sobre por lo menos una molécula de polinucleótido individual, en donde la decodificación reduce el ruido y/o la distorsión en el mensaje
1208
Agrupar las lecturas de secuencia en familias, cada familia generada a partir de un polinucleótido parental marcado único
1210
Proporcionar el mensaje a un receptor
1212

Fig. 12

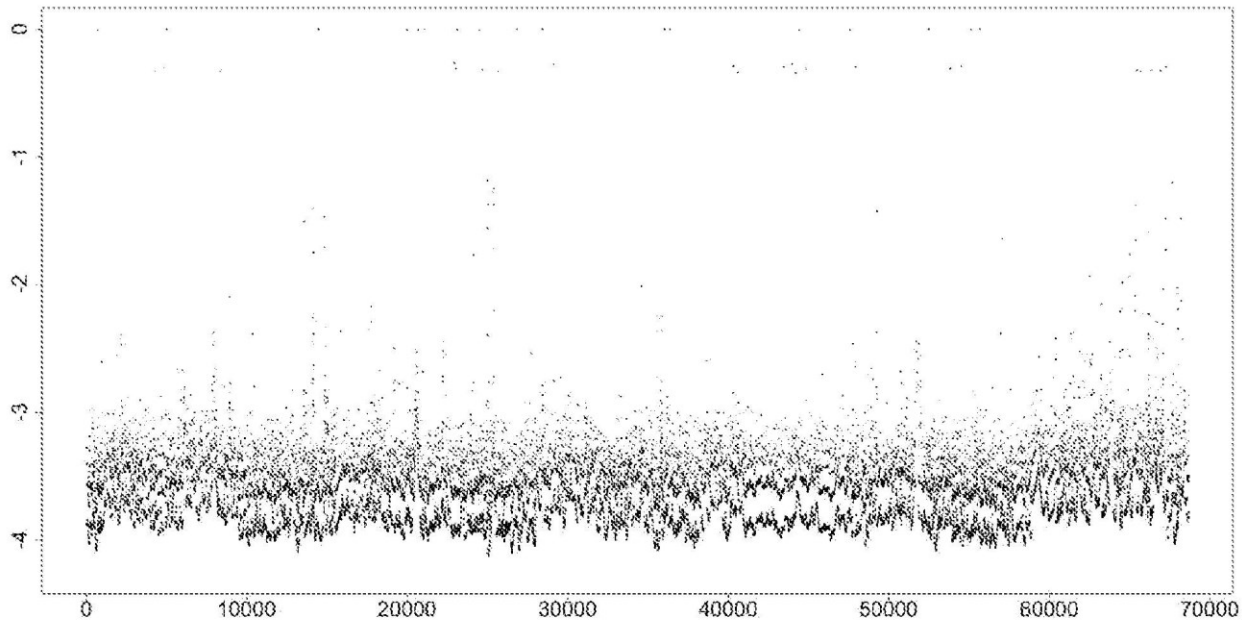


Fig. 13A

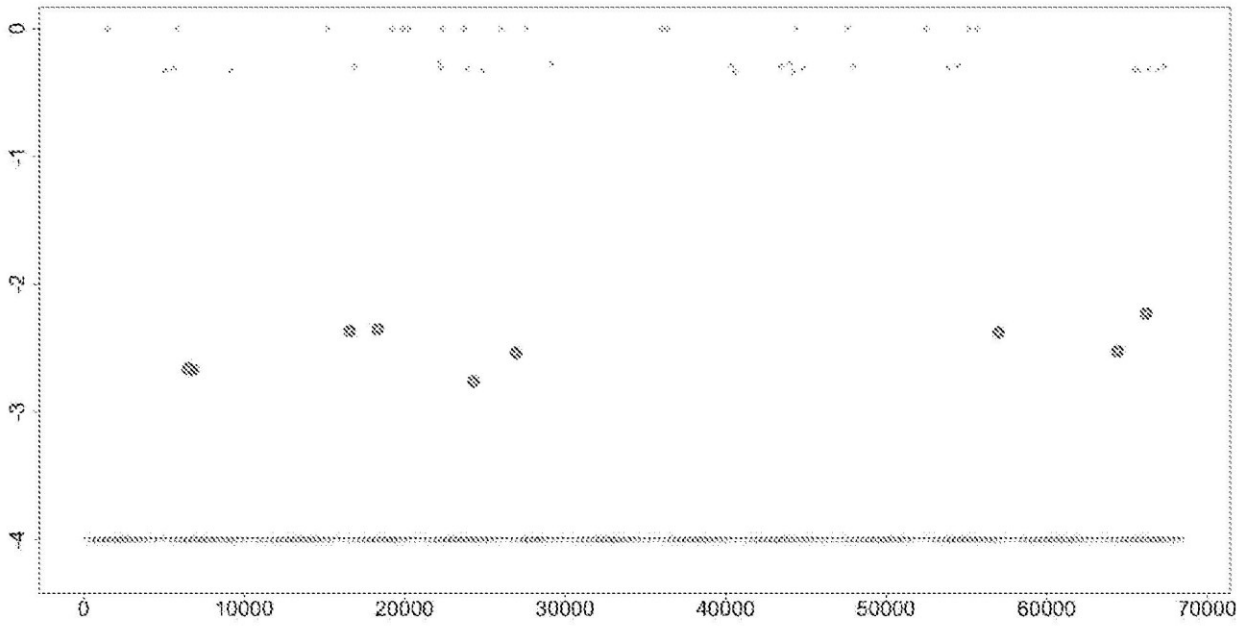


Fig. 13B

Fig. 13

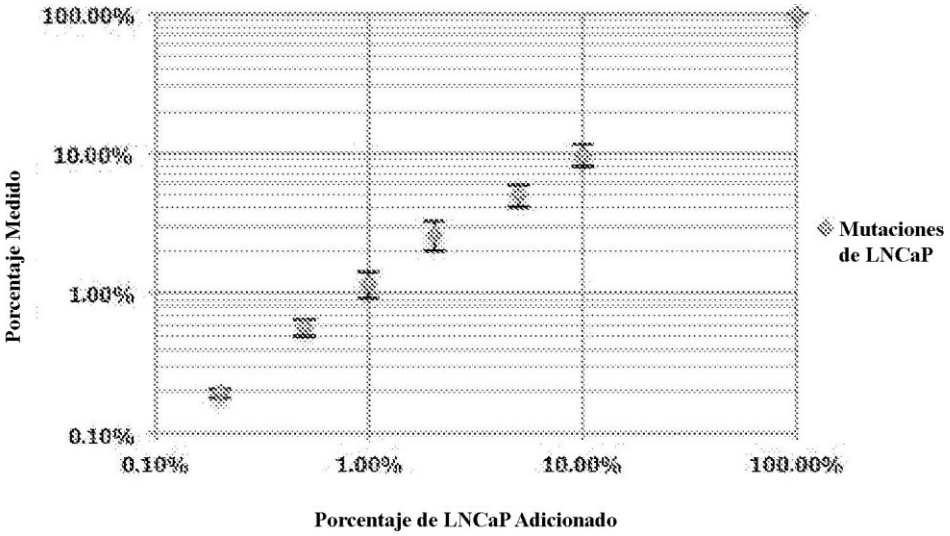

100.00%
10.00%
1.00%
0.10%
Porcentaje Medido
0.10%
1.00%
10.00%
100.00%
Porcentaje de LNCaP Adicionado
Mutaciones de LNCaP

Fig. 14

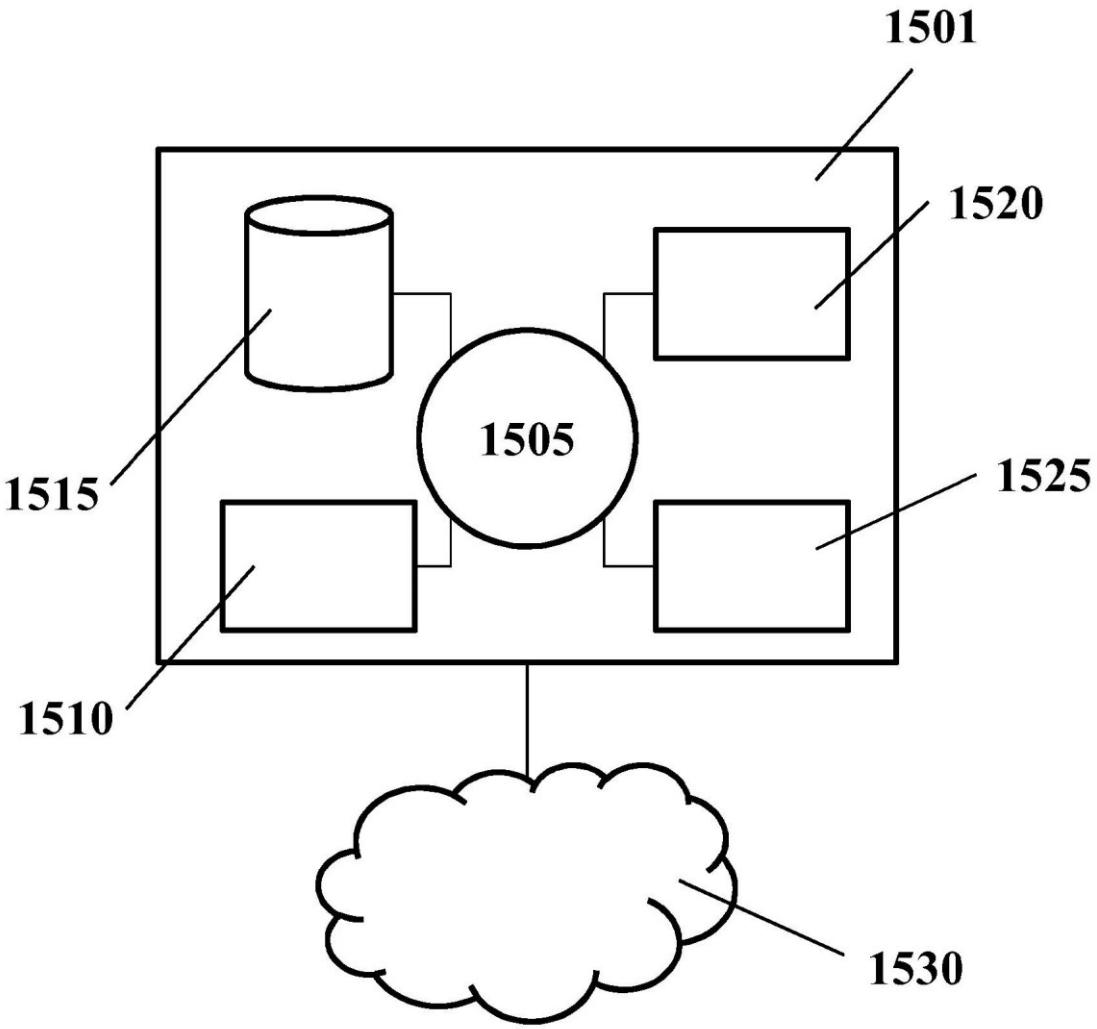


Fig. 15