(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 980 689**

(51) Int. Cl.:

C12Q 1/6869 (2008.01)
G16B 30/10 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **15.03.2014 E 21157571 (7)**
(97) Fecha y número de publicación de la concesión europea: **08.05.2024 EP 3882362**

(54) Título: **Métodos para la secuenciación de polinucleótidos libres de células**

(30) Prioridad:

15.03.2013 US 201361793997 P
13.07.2013 US 201361845987 P
16.08.2013 US 201313969260
04.09.2013 WO PCT/US2013/058061
05.03.2014 US 201461948530 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
02.10.2024

(73) Titular/es:

GUARDANT HEALTH, INC. (100.0%)
3100 Hanover Street
Palo Alto, CA 94304, US

(72) Inventor/es:

TALASAZ, AMIRALI;
ELTOUKHY, HELMY y
MORTIMER, STEFANIE ANN WARD

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

Observaciones:

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

DESCRIPCIÓN

Métodos para la secuenciación de polinucleótidos libres de células

ANTECEDENTES

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

La detección temprana y el seguimiento de las enfermedades genéticas, como el cáncer a menudo es útil y necesario en el éxito tratamiento o manejo de la enfermedad. Un enfoque puede incluir el seguimiento de una muestra derivada de ácidos nucleicos libres de células, una población de polinucleótidos que se pueden encontrar en diferentes tipos de fluidos corporales. En algunos casos, la enfermedad puede caracterizarse o detectarse basándose en 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 ácido nucleico, o el desarrollo de ciertas alteraciones genéticas raras. El ADN libre de células ("cfADN") se conoce en la técnica durante décadas y puede contener aberraciones genéticas asociadas con una enfermedad particular. Con mejoras en la secuenciación y técnicas para manipular ácidos nucleicos, existe una necesidad en la técnica de métodos y sistemas mejorados para usar ADN libre de células para detectar y controlar enfermedades.

Schmitt et al., PNAS (2012), 109(36):14508-13, analiza la "Detección de mutaciones ultrarraras mediante secuenciación de próxima generación" (título). US8209130B1 (Kennedy & Porreca; 2012) "proporciona un método para identificar una mutación en un ácido nucleico que implica la secuenciación del ácido nucleico para generar una pluralidad de lecturas de secuencia" (resumen). La WO2008/154317A1 (Pacific Biosciences of California, Inc.; 2008) analiza "Métodos y procesos para llamar bases en secuencia mediante métodos de incorporación" (título). La US2009/325239A1 (Si Lok, 2009) analiza "métodos para el mapeo de ácidos nucleicos e identificación de variaciones estructurales finas en ácidos nucleicos" (título).

RESUMEN

La invención se expone en el conjunto de reivindicaciones adjunto.

En algunas realizaciones, los polinucleótidos libres de células se aíslan de una muestra corporal que puede seleccionarse de un grupo consistente en sangre, plasma, suero, orina, saliva, excreciones de mucosa, esputo, heces y lágrimas.

En algunas realizaciones, los métodos de la divulgación también comprenden un paso para 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, los métodos de la divulgación pueden comprender enriquecer selectivamente regiones del genoma o transcriptoma del sujeto antes de la secuenciación. En otras realizaciones, los métodos de la divulgación comprenden enriquecer selectivamente regiones del genoma o transcriptoma del sujeto antes de la secuenciación. En otras realizaciones, los métodos de la divulgación comprenden enriquecer no selectivamente regiones del genoma o transcriptoma del sujeto antes de la secuenciación.

El código de barras es un polinucleótido, que comprende además un conjunto de oligonucleótidos que en combinación con la diversidad de moléculas secuenciadas de una región seleccionada permite la identificación de moléculas únicas y puede tener una longitud de por lo menos 3, 5, 10, 15, 20 25, 30, 35, 40, 45 o 50mer pares de bases.

En algunas realizaciones, la amplificación comprende la amplificación global o la amplificación del genoma completo.

En algunas realizaciones, las moléculas de secuencia de identidad única se detectan basándose en la información de la secuencia en las regiones inicial (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 la amplificación selectiva, la amplificación no selectiva, la amplificación por supresión o el enriquecimiento sustractivo.

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

En algunas realizaciones, el método puede comprender filtrar las lecturas con una precisión o puntuación de calidad menor que un umbral, por ejemplo, 90%, 99%, 99,9% o 99,99% y/o una 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 tienen un tamaño uniforme o sustancialmente uniforme, de aproximadamente 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb o 100kb. 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, una variante genética, mutación rara o variación en el número de copias se produce en una región del genoma seleccionada del grupo que consiste en fusiones de genes, duplicaciones de genes, deleciones de genes, translocaciones de genes, regiones de microsatélites, fragmentos de genes o combinación de los mismos. En otras realizaciones, una variante genética, mutación rara o variación del número de copias se produce en una región del genoma seleccionada del grupo que consiste en genes, oncogenes, genes supresores de tumores, promotores, elementos de secuencia reguladora o una combinación de los mismos. En algunas realizaciones, la variante es una variante de nucleótido, una sustitución de una sola base, o un indel, transversión, translocación, inversión, deleción, truncamiento o truncamiento génico pequeños 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 las propiedades únicas de las lecturas individuales.

En algunas realizaciones, se analizan muestras en intervalos de tiempo sucesivos del mismo sujeto y se comparan con los resultados de muestras anteriores. El método de la divulgación puede comprender además determinar la frecuencia de variación parcial en el número de copias, la pérdida de heterocigosidad, el análisis de la expresión génica, el análisis epigenético y el análisis de hipermetilación después de amplificar los polinucleótidos libres de células unidos a código de barras.

En algunas realizaciones, la variación del número de copias y el análisis de mutaciones raras se determinan en una muestra libre de células o sustancialmente libre de células obtenida de un sujeto usando secuenciación múltiplex, que comprende la realización de más de 10.000 reacciones de secuenciación; la secuenciación simultánea de por lo menos 10.000 lecturas diferentes; o la realización de análisis de datos en por lo menos 10.000 lecturas diferentes en todo el genoma. El método puede comprender la secuenciación multiplex que comprende la realización de 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 una o más de las metodologías de markov oculto, programación dinámica, máquina de vectores de soporte, red bayesiana, decodificación trellis, decodificación Viterbi, maximización de expectativas, filtrado Kalman o red neuronal.

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 sobre la base de las variantes descubiertas.

En algunas realizaciones, se modifica una terapia sobre la base del análisis de muestra más reciente. Además, los métodos de la divulgación pueden comprender 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, infección u otra anomalía tisular. En algunas realizaciones se analiza y monitoriza el sistema inmunitario del sujeto en instancias únicas o a lo largo del tiempo.

En algunas realizaciones, los métodos de la divulgación comprenden la identificación de una variante de la que se realiza un seguimiento mediante una prueba de imagenología (por ejemplo, CT, PET-TC, MRI, rayos X, ultrasonidos) para la localización de la anomalía tisular que se sospecha que provoca la variante identificada.

En algunas realizaciones, los métodos de la divulgación comprenden la realización de no-llamadas basadas en la población y la identificación de regiones de baja confianza. En algunas realizaciones, la obtención de los datos de medición para la cobertura de la secuencia comprende medir la profundidad de la cobertura de la secuencia en cada posición del genoma. En algunas realizaciones, la corrección de los datos de medición del sesgo de cobertura de la secuencia comprende el cálculo de la cobertura promediada por ventanas. En algunas realizaciones, la corrección de los datos de medición para el sesgo de cobertura de la secuencia comprende la realización de ajustes para tener en cuenta

el sesgo de GC en el proceso de construcción de bibliotecas y secuenciación. En algunas realizaciones, la corrección de los datos de medición para el sesgo de la cobertura de secuencia comprende realizar ajustes basados en el factor de ponderación adicional asociado a los mapeos individuales para compensar el sesgo.

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

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.

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

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

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

En algunas realizaciones, el método comprende el embotellamiento del material genético inicial de partida 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, una pluralidad de las secuencias de referencia proceden del mismo genoma.

En algunas realizaciones, el método comprende secuenciar un subconjunto del conjunto de polinucleótidos de la progenie amplificados suficiente para producir lecturas de secuencias 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 progenitores 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 progenitores amplificados secuenciados tiene el tamaño suficiente para que cualquier secuencia de nucleótidos representada en el conjunto de polinucleótidos progenitores marcados en un porcentaje que sea igual a la tasa de error de secuenciación por base porcentual de la plataforma de secuenciación usada, 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 posibilidades de estar representada entre el conjunto de secuencias consenso.

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

En algunas realizaciones, analizar comprende normalizar una medida (por ejemplo, un número) tomada de un conjunto de secuencias consenso frente a una medida tomada de un conjunto de secuencias consenso de una muestra de control.

En algunas realizaciones, analizar comprende detectar mutaciones, mutaciones raras, variantes de nucleótidos individuales, indeles, variaciones en el número de copias, transversiones, translocaciones, inversiones, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones génicas, fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de los ácidos nucleicos infección o cáncer.

En algunas realizaciones, los polinucleótidos comprenden ADN, ARN, una combinación de ambos o ADN más

ADNc derivado de ARN.

En algunas realizaciones se selecciona o se enriquece un cierto subconjunto de polinucleótidos sobre la base de la longitud del polinucleótido en pares de bases a partir 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 en un individuo, como una infección y/o un cáncer.

En algunas realizaciones, el método se lleva a cabo en combinación con la realización de perfiles del repertorio inmunitario.

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

En algunas realizaciones, el colapso comprende la detección y/o la corrección de errores, mellas o lesiones presentes en la cadena de sentido o antisentido de los polinucleótidos parentales marcados o de los polinucleótidos de la progenie amplificados.

En algunas realizaciones, se sospecha que el sujeto tiene una condición anormal. En algunas realizaciones, la condición anormal se selecciona del grupo que consiste en 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 génicas, fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de los ácidos nucleicos infección y cáncer.

En algunas realizaciones, el sujeto es una mujer embarazada. En algunas realizaciones, la variación en el número de copias o la mutación o variante genética rara es indicativa de una anomalía fetal. En algunas realizaciones, la anomalía fetal se selecciona del grupo que consiste en 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 génicas, fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de los ácidos nucleicos infección y cáncer.

En algunas realizaciones, los métodos comprenden además enriquecer selectivamente regiones del genoma o transcriptoma del sujeto antes de la secuenciación. En algunas realizaciones, los métodos comprenden además el enriquecimiento no selectivo de regiones del genoma o transcriptoma del sujeto antes de la secuenciación.

En algunas realizaciones, se analiza la totalidad del genoma o por lo menos el 85% del genoma. En algunas realizaciones, las variantes en el número de copias identificadas son fraccionarias (es decir, niveles no enteros) debido a la heterogeneidad de la muestra. En algunas realizaciones, se realiza un enriquecimiento de las regiones seleccionadas. En algunas realizaciones, la información sobre la variación del número de copias se extrae simultáneamente basándose en los métodos descritos en la presente. En algunas realizaciones, los métodos comprenden un paso inicial de embotellamiento de polinucleótidos para limitar el número de copias iniciales de partida o la diversidad de polinucleótidos en la muestra.

En algunas realizaciones, el material genético inicial de partida se proporciona en una cantidad de menos de 100 ng de ácido nucleico, la variación genética es una variación del número de copias/heterocigosidad y la detección se realiza con una resolución subcromosómica; por ejemplo, una resolución de por lo menos 100 megabases, una resolución de por lo menos 10 megabases, una resolución de por lo menos 1 megabase, una resolución de por lo menos 100 kilobases, una resolución de por lo menos 10 kilobases o una resolución de por lo menos 1 kilobase. En algunas realizaciones, el método comprende proporcionar una pluralidad de conjuntos de polinucleótidos parentales marcados, en donde cada conjunto es mapeable a una posición mapeable diferente en una secuencia de referencia. En algunas realizaciones, la posición mapeable en 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 algunas realizaciones, el marcador tumoral está presente en el conjunto de secuencias consenso a una frecuencia menor que la tasa de error introducida en el paso de amplificación. En algunas realizaciones, el por lo menos un conjunto es una pluralidad de conjuntos, y la posición mapeable de la secuencia de referencia comprende una pluralidad de posiciones mapeables en la secuencia de referencia, cada una de cuyas posiciones mapeables es el locus de un marcador tumoral. En algunas realizaciones, analizar comprende detectar la variación en el número de copias de las secuencias consenso entre por lo menos dos conjuntos de polinucleótidos parentales. En algunas realizaciones, analizar comprende detectar la presencia de variaciones de secuencia en comparación con las secuencias de referencia.

En algunas realizaciones, analizar 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 las secuencias consenso entre por lo menos dos conjuntos de polinucleótidos parentales. En algunas realizaciones, colapsar comprende: (i) agrupar las lecturas de secuencias secuenciadas a partir de polinucleótidos de la progenie amplificados en familias, cada familia amplificada a partir del mismo polinucleótido parental marcado; y (ii) determinar una secuencia consenso sobre la base de las lecturas de secuencias de una familia.

En algunas realizaciones, la alteración genética es una 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 una 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 la 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. Los métodos reivindicados pueden usarse en el diagnóstico o tratamiento de una enfermedad.

En algunas realizaciones de cualquiera de los métodos de la presente, el método comprende además la detección y/o asociación de vías moleculares afectadas. En algunas realizaciones de cualquiera de los métodos de la presente, el método comprende además la monitorización en serie del estado de salud o de enfermedad de un individuo. En algunas realizaciones se infiere la filogenia de un genoma asociado a una enfermedad dentro de un individuo. Los métodos descritos en la presente pueden ser útiles en el diagnóstico, monitorización o tratamiento de una enfermedad. Por ejemplo, el régimen de tratamiento puede seleccionarse o modificarse sobre la base de las formas polimórficas o CNV detectadas o de las vías asociadas.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Se obtendrá una mejor comprensión de las características y ventajas de esta divulgación haciendo referencia a la siguiente descripción detallada que establece realizaciones ilustrativas, en las que se utilizan los principios de los sistemas y métodos de esta divulgación, y los dibujos adjuntos de los cuales:

FIG. 1 es una representación de diagrama de flujo de un método de detección de la variación del número de copias utilizando una sola muestra.
FIG. 2 es una representación de diagrama de flujo de un método de detección de la variación del número de copias utilizando muestras pareadas.
FIG. 3 es una representación de diagrama de flujo de un método de detección de mutaciones raras (por ejemplo, variantes de un solo nucleótido).
FIG. 4A es un informe gráfico de detección de variación del número de copias generado a partir de un sujeto normal no canceroso.
FIG. 4B es un informe gráfico de detección de variación del número de copias generado a partir de un sujeto con cáncer de próstata.
FIG. 4C es una representación esquemática del acceso habilitado por Internet de informes generados a partir del análisis de variación del número de copias de un sujeto con cáncer de próstata.
FIG. 5A es un informe gráfico de detección de variación del número de copias generado a partir de un sujeto con remisión de cáncer de próstata.
FIG. 5B es un informe gráfico de detección de variación del número de copias generado a partir de un sujeto con cáncer de próstata recurrente.
FIG. 6A es un informe de detección gráfico (por ejemplo, para variantes de un solo nucleótido) generado a partir de varios experimentos de mezcla utilizando muestras de ADN que contienen copias tanto de tipo salvaje como mutantes de MET y TP53.
FIG. 6B es una representación gráfica logarítmica de los resultados de detección (por ejemplo, variante de un solo nucleótido). Se muestran las mediciones del porcentaje de cáncer observado frente al esperado para varios experimentos de mezcla utilizando muestras de ADN que contienen copias mutantes y de tipo salvaje de MET, HRAS y TP53.
FIG. 7A es un informe gráfico del porcentaje de dos (por ejemplo, variantes de un solo nucleótido) en dos genes, PIK3CA y TP53, en un sujeto con cáncer de próstata en comparación con una referencia (control).
FIG. 7B es una representación esquemática del acceso habilitado por Internet de informes generados a partir del análisis (por ejemplo, variante de un solo nucleótido) de un sujeto con cáncer de próstata.
FIG. 8 es una representación de diagrama de flujo de un método de análisis de material genético.
FIG. 9 es una representación de diagrama de flujo de un método de decodificación de 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 progenitores etiquetados.
FIG. 10 es una representación de diagrama de flujo de un método para reducir la distorsión en la determinación

de CNV a partir de un conjunto de lecturas de secuencia.

FIG. 11 es una representación de 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 parental marcada a partir de un conjunto de lecturas de secuencia.

FIG. 12 muestra un método de comunicar información de secuencia.

FIG. 13 muestra las frecuencias de los alelos menores detectados en un panel completo de 70 kb en una titulación de ADNc 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 la PCR y los errores de secuenciación a pesar del filtrado Q30. La secuenciación digital (Fig. 13B) elimina todo el ruido de secuenciación y PCR, revelando mutaciones verdaderas sin falsos positivos: los círculos verdes son puntos SNP en el cfADN normal y los círculos rojos son mutaciones detectadas en LNCaP.

FIG. 14: muestra la titulación de LNCap cfADN.

FIG. 15 muestra un sistema informático que está programado o configurado de otro modo para implementar varios métodos de la presente divulgación.

DESCRIPCIÓN DETALLADA

I. Visión general

De manera general, los métodos de la invención comprenden la preparación de muestras o la extracción y aislamiento de secuencias de polinucleótidos libres de células de un fluido corporal; secuenciación posterior de polinucleótidos libres de células mediante técnicas conocidas en la técnica; y 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 colección de diferentes mutaciones raras o perfiles de variación del número de copias de diferentes enfermedades, que se utilizarán como referencias adicionales para ayudar a la detección de mutaciones raras (p. ej., perfil de variación de un solo nucleótido), perfil de variación del número de copias o perfil genético general de una enfermedad.

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

Después de la etapa de aislamiento/extracción, cualquiera de un número de diferentes operaciones de secuenciación se puede realizar en la muestra de polinucleótido libre de células. Las muestras se pueden procesar antes de secuenciar 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 etiquetarse individualmente o en subgrupos con el identificador único. La muestra etiquetada se puede usar luego en una aplicación posterior, tal como una reacción de secuenciación mediante la cual las moléculas individuales se pueden rastrear hasta las moléculas parentales.

Después de los datos de secuenciación de secuencias de polinucleótidos libres de células se recogieron, uno o más procesos de bioinformática se puede aplicar a los datos de secuencia para detectar características genéticas o aberraciones tales como la variación del número de copias, raras mutaciones (por ejemplo, variaciones individuales o múltiples de nucleótidos) o cambios en los marcadores epigenéticos, incluidos, entre otros, los perfiles de metilación. En algunos casos, en los que se desea un análisis de variación del número de copias, los datos de la secuencia pueden: 1) alinearse con un genoma de referencia; 2) filtrado y mapeado; 3) dividido en ventanas o intervalos de secuencia; 4) lecturas de cobertura contadas para cada ventana; 5) las lecturas de cobertura pueden luego normalizarse utilizando un algoritmo de modelado estocástico o estadístico; 6) y se puede generar un archivo de salida que refleje los estados del número de copias discretas en varias posiciones del genoma. En otros casos, en los que se desea un análisis de mutaciones poco frecuentes, los datos de la secuencia pueden 1) alinearse con un genoma de referencia; 2) filtrado y mapeado; 3) frecuencia de las bases variantes calculadas en función de las lecturas de cobertura para esa base específica; 4) frecuencia base variante normalizada utilizando un algoritmo de modelado estocástico, estadístico o probabilístico; 5) y se puede generar un archivo de salida que refleje los estados de mutación en varias posiciones del genoma.

Puede ocurrir una variedad de reacciones y/operaciones diferentes dentro de los sistemas y métodos descritos en este documento, que incluyen, entre otros: secuenciación de ácido nucleico, cuantificación de ácido nucleico, optimización de secuenciación, detección de expresión génica, cuantificación de expresión génica, perfil genómico, elaboración de 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, estadificación o predicción del riesgo de diversas enfermedades y trastornos genéticos y no genéticos, incluido el cáncer.

Puede usarse para evaluar la respuesta del sujeto a diferentes tratamientos de dichas enfermedades genéticas y no genéticas, o proporcionar información sobre la progresión y el pronóstico de la enfermedad.

La secuenciación de polinucleótidos se puede comparar con un problema en teoría de la comunicación. Un polinucleótido individual inicial o un conjunto de polinucleótidos se considera un mensaje original. Se puede pensar en etiquetar y/o amplificar como codificar el mensaje original en una señal. La secuenciación se puede considerar como un canal de comunicación. La salida de un secuenciador, por ejemplo, lecturas de secuencia, se puede considerar como una señal recibida. El procesamiento bioinformático puede ser pensado como un receptor que decodifica la señal recibida para producir un mensaje transmitido, por ejemplo, una secuencia o secuencias de nucleótido. La señal recibida puede incluir artefactos, como ruido y distorsión. Se puede pensar en el ruido como una adición aleatoria no deseada a una señal. La distorsión se puede considerar como una alteración en la amplitud de una señal o parte de una señal.

El ruido puede ser introducido a través de los errores en la copia y/o la lectura de un polinucleótido. Por ejemplo, en un proceso de secuenciación, un único polinucleótido puede primero someterse a amplificación. La amplificación puede introducir errores, de modo que un subconjunto de 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 lugar particular puede leerse incorrectamente. Como consecuencia, la colección de lecturas de secuencia puede incluir un cierto porcentaje de llamadas de bases en un locus que no son iguales a la base original. En las tecnologías de secuenciación típicas, esta tasa de error puede ser de un solo dígito, por ejemplo, 2%-3%. Cuando se secuencia una colección de moléculas que se supone que tienen la misma secuencia, este ruido es lo suficientemente pequeño como para que se pueda identificar la base original con alta confiabilidad.

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 significativo. 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 el ADN fetal o el 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 introducidos por el proceso de secuenciación, entonces las variantes de secuencia verdaderas pueden no distinguirse del ruido. Esto podría interferir, por ejemplo, con la detección de variantes de secuencia en una muestra.

La distorsión se puede manifestar 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 se puede introducir, por ejemplo, a través de un sesgo de amplificación, un sesgo de GC o un 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 la secuencia.

Esta invención proporciona métodos para reducir los artefactos de secuenciación, tales como ruido y/o distorsión, en un proceso de secuenciación de polinucleótidos. La agrupación de lecturas de secuencias en familias derivadas de moléculas individuales originales puede reducir el ruido y/o la distorsión de una sola molécula individual o de un conjunto de moléculas. Con respecto a una sola molécula, la agrupación de lecturas en una familia reduce la distorsión, por ejemplo, indicando que muchas secuencias de lecturas representan en realidad una sola molécula en lugar de muchas moléculas diferentes. El colapso de las lecturas de secuencia en una secuencia de consenso es una forma de reducir el ruido en el mensaje recibido de una molécula. Usar funciones probabilísticas que conviertan las frecuencias recibidas es otra forma. Con respecto a un conjunto de moléculas, la agrupación de lecturas en familias y la determinación de 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, el colapso de las lecturas de secuencias de diferentes familias en secuencias de consenso elimina los errores introducidos por la amplificación y/o el error de secuenciación. Además, la determinación de las frecuencias de las llamadas de bases basadas en probabilidades derivadas de la información familiar también reduce el ruido en el mensaje recibido de un conjunto de moléculas.

Se conocen métodos de reducción de ruido y/o distorsión de un proceso de secuenciación. Estos incluyen, por ejemplo, filtrar secuencias, por ejemplo, exigir que cumplan un umbral de calidad o reducir el sesgo de GC. Estos métodos normalmente se realizan en la colección de lecturas de secuencia que son la salida de un secuenciador, y se pueden realizar lecturas de secuencia lectura por secuencia, sin tener en cuenta la estructura familiar (subcolecciones de secuencias derivadas de una única molécula parental original). Ciertos métodos de esta invención reducen el ruido y la distorsión al reducir el ruido y/o la distorsión dentro de las familias de lecturas de secuencia, es decir, que operan en lecturas de secuencia agrupadas en familias derivadas de una molécula de polinucleótido monoparental. La reducción de artefactos de señal a nivel familiar puede producir significativamente menos ruido y distorsión en el mensaje final que se proporciona que la reducción de artefactos realizada en un nivel de lectura de secuencia de lectura por secuencia o en la salida del secuenciador como un todo.

Los métodos de la invención también pueden ser útiles para la detección con variación genética de alta sensibilidad en una muestra de material genético inicial. Los métodos implican el uso de una o ambas de las siguientes herramientas: Primero, la conversión eficiente de polinucleótidos individuales en una muestra de material genético inicial

en polinucleótidos parentales etiquetados listos para la secuencia, a fin de aumentar la probabilidad de que los polinucleótidos individuales en una muestra de material genético inicial se representará en una muestra lista para la secuencia. Esto puede producir información de secuencia sobre más polinucleótidos en la muestra inicial. En segundo lugar, generación de alto rendimiento de secuencias consenso para polinucleótidos parentales marcados mediante muestreo de alta velocidad de polinucleótidos de la progenie amplificados a partir de los polinucleótidos parentales marcados, y colapso de la secuencia generada se lee en secuencias consenso que representan secuencias de polinucleótidos marcados parentales. Esto puede reducir el ruido introducido por el sesgo de amplificación y/o errores de secuencia, y puede aumentar la sensibilidad de detección. El colapso se realiza en una pluralidad de lecturas de secuencia, generadas a partir de lecturas de moléculas amplificadas o múltiples lecturas de una sola molécula.

Los métodos de secuenciación implican típicamente la preparación de la muestra, 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 la muestra normalmente implica convertir polinucleótidos en una muestra en una forma compatible con la plataforma de secuenciación utilizada. Esta conversión puede implicar marcar polinucleótidos. En determinadas realizaciones de esta invención, las etiquetas comprenden etiquetas de secuencia de polinucleótidos. Las metodologías de conversión utilizadas en la secuenciación pueden no ser 100% eficientes. Por ejemplo, no es raro convertir polinucleótidos en una muestra con una eficiencia de conversión de aproximadamente 1-5%, es decir, aproximadamente 1-5% de los polinucleótidos en una muestra se convierten en polinucleótidos marcados. Los polinucleótidos que no se convierten en moléculas etiquetadas no se representan en una biblioteca etiquetada para secuenciación. Por consiguiente, los polinucleótidos que tienen variantes genéticas representadas con baja frecuencia en el material genético inicial pueden no estar representados en la biblioteca etiquetada y, por lo tanto, no se puede secuenciar ni detectar. Al aumentar la eficiencia de conversión, aumenta la probabilidad de que un polinucleótido raro en el material genético inicial esté representado en la biblioteca etiquetada y, en consecuencia, se detecte mediante secuenciación. Además, en lugar de abordar directamente el problema de la baja eficiencia de conversión de la preparación de bibliotecas, 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 muestra de entrada es limitado o se desea la detección de polinucleótidos con baja representación, una alta eficiencia de conversión puede secuenciar eficientemente la muestra y/o detectar adecuadamente tales polinucleótidos.

Los métodos de la invención también implican la conversión de polinucleótidos iniciales en polinucleótidos etiquetados con una eficiencia de conversión de al menos 10%, al menos 20%, al menos 30%, al menos 40%, al menos 50%, al menos 60%, al menos 80% o al menos 90%. Los métodos implican, por ejemplo, el uso de cualquier ligación de extremos romos, ligación de extremos pegajosos, sondas de inversión molecular, PCR, PCR basada en ligación, PCR multiplex, ligación monocatenaria y circularización monocatenaria. Los métodos también pueden implicar la limitación de la cantidad de material genético inicial. Por ejemplo, la cantidad de material genético inicial puede ser menor de 1 ug, menor de 100 ng o menor de 10 ng. Estos métodos se describen con más detalle en este documento.

La obtención de información cuantitativa y cualitativa precisa sobre polinucleótidos en una biblioteca etiquetada puede resultar en una caracterización más sensible del material genético inicial. Normalmente, los polinucleótidos en una biblioteca etiquetada se amplifican y las moléculas amplificadas resultantes se secuencian. Dependiendo del rendimiento de la plataforma de secuenciación utilizada, solo un subconjunto de las moléculas en la biblioteca amplificada produce lecturas de secuencia. Así, por ejemplo, el número de moléculas amplificadas muestreadas para secuenciar puede ser aproximadamente solo el 50% de los polinucleótidos únicos en la biblioteca etiquetada. Además, la amplificación puede estar sesgada a favor o en contra de ciertas secuencias o ciertos miembros de la biblioteca etiquetada. Esto puede distorsionar la medición cuantitativa de secuencias en la biblioteca etiquetada. Además, las plataformas de secuenciación pueden introducir errores en la secuenciación. Por ejemplo, las secuencias pueden tener una tasa de error por base de 0,5-1%. El sesgo de amplificación y los errores de secuenciación introducen ruido en el producto de secuenciación final. Este ruido puede disminuir la sensibilidad de detección. Por ejemplo, las variantes de secuencia cuya frecuencia en la población etiquetada es menor que la tasa de error de secuenciación puede confundirse con ruido. Además, al proporcionar 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, se puede producir una pluralidad de lecturas de secuencia de un único polinucleótido sin amplificación. Esto se puede hacer, por ejemplo, con métodos de nanoporos.

Los métodos de la invención también implican detectar y leer polinucleótidos únicos en un grupo etiquetado con precisión. En ciertas realizaciones, esta divulgación proporciona polinucleótidos etiquetados con 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 permitió el rastreo o colapso de los polinucleótidos de la progenie en la molécula de polinucleótido parental de etiqueta única. El colapso de las familias de polinucleótidos de progenie amplificados reduce el sesgo de amplificación al proporcionar información sobre las moléculas parentales únicas originales. El colapso también reduce los errores de secuenciación al eliminar de los datos de secuenciación las secuencias mutantes de las moléculas de la progenie.

La detección y lectura de polinucleótidos únicos en la biblioteca etiquetada puede implicar dos estrategias. En una estrategia, un subconjunto suficientemente grande del conjunto de polinucleótidos de progenie amplificados se

secuencia de tal manera que, para un gran porcentaje de polinucleótidos parentales marcados únicos en el conjunto de polinucleótidos parentales marcados, hay una lectura de secuencia que se produce para al menos un polinucleótido de progenie amplificada en una familia producido a partir de un polinucleótido original marcado único. En una segunda estrategia, el conjunto de polinucleótidos de la progenie amplificado se muestrea para secuenciar a un nivel para producir lecturas de secuencia de múltiples miembros de la progenie de una familia derivada de un polinucleótido original único. La generación de lecturas de secuencia de múltiples miembros de la progenie de una familia permite el colapso de secuencias en secuencias parentales consenso.

Así, por ejemplo, el muestreo de un número de polinucleótidos de progenie amplificados a partir del conjunto del polinucleótidos de progenie amplificados que es igual al número de polinucleótidos parentales etiquetados únicos en el conjunto de polinucleótidos parentales etiquetados (particularmente cuando el número es de al menos 10.000) producirá, estadísticamente, una secuencia leída para al menos uno de la progenie de aproximadamente el 68% de los polinucleótidos progenitores etiquetados en el conjunto, y aproximadamente el 40% de los polinucleótidos progenitores etiquetados únicos en el grupo original estarán representados por al menos dos lecturas de secuencias de progenie. En determinadas realizaciones, el conjunto de polinucleótidos de progenie amplificados se muestrea suficientemente para producir un promedio de cinco a diez lecturas de secuencia para cada familia. El muestreo del conjunto de la progenie amplificada de 10 veces más moléculas que el número de polinucleótidos parentales marcados únicos producirá, estadísticamente, información de secuencia sobre el 99,995% de las familias, de las cuales el 99,95% del total de familias estará cubierto por una pluralidad de lecturas de secuencia. Se puede construir una secuencia de consenso a partir de los polinucleótidos de la progenie en cada familia para reducir drásticamente la tasa de error desde la tasa de error de secuenciación por base nominal a una tasa posiblemente muchos órdenes de magnitud más baja. Por ejemplo, si el secuenciador tiene una tasa de error aleatoria por base 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 inferior al 0,0001%. En consecuencia, el tamaño de muestreo de la progenie amplificada que se va a secuenciar se puede elegir de modo que se asegure una secuencia que tenga una frecuencia en la muestra que no sea mayor que la tasa de error de secuenciación nominal por base a una tasa de la plataforma de secuenciación utilizada al menos el 99% de probabilidad de estar representado por al menos una lectura.

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

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

Esta información se puede utilizar ahora tanto para el análisis cualitativo como cuantitativo. Por ejemplo, para el análisis cuantitativo, se determina una medida, por ejemplo, un recuento, de la cantidad de moléculas parentales marcadas que se mapean en una secuencia de referencia. Esta medida se puede comparar con una medida de moléculas parentales marcadas que se mapean en una región genómica diferente. Es decir, la cantidad de moléculas parentales etiquetadas que se mapean en una primera ubicación o posición mapeable en una secuencia de referencia, como el genoma humano, se puede comparar con una medida de moléculas parentales etiquetadas que se mapean en una segunda ubicación o posición mapeable en una secuencia de referencia. Esta comparación puede revelar, por ejemplo, las cantidades relativas de moléculas parentales que se mapean en cada región. Esto, a su vez, proporciona una indicación de la variación del número de copias para el mapeo de moléculas en una región particular. Por ejemplo, si la medida de mapeo de polinucleótidos a una primera secuencia de referencia es mayor que la medida de mapeo de polinucleótidos a una segunda secuencia de referencia, esto puede indicar que la población parental, y por extensión la muestra original, incluía polinucleótidos de células que presentan 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 ensamblaje de hg19 o cualquier ensamblaje de hg anterior o disponible. Estas secuencias se pueden interrogar utilizando el navegador del genoma disponible en genome.ucsc.edu/index.html. Otros genomas de especies incluyen, por ejemplo, PanTro2 (chimpancé) y mm9 (ratón).

Para el análisis cualitativo, las secuencias de un conjunto de polinucleótidos marcados de mapeo para una secuencia de referencia pueden ser analizados para secuencias variantes y su frecuencia en la población de polinucleótidos parentales marcados se pueden medir.

II. Preparación de muestras

A. Aislamiento y extracción de polinucleótidos

Los 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 se limitan a: ADN, ARN, amplicones, ADNc, ADNdc, ADNss, ADN plasmídico, 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, microARN, ARNdc, ribozima, riboswitch y ARN viral (por ejemplo, ARN retroviral).

Polinucleótidos libres de células se pueden derivar de una variedad de fuentes incluyendo fuentes de ser humano, mamífero, mamífero no humano, mono, chimpancé, reptil, anfibio, o aviar. Además, las muestras se pueden extraer de una variedad de fluidos animales que contienen secuencias libres de células, que incluyen, entre otros, sangre, suero, plasma, vítreo, esputo, orina, lágrimas, transpiración, saliva, semen, excreciones mucosas, moco, líquido cefalorraquídeo, amniótico líquido, líquido linfático y similares. Los polinucleótidos libres de células pueden ser de origen fetal (a través de un fluido extraído de una mujer embarazada) o pueden derivarse del tejido del propio sujeto.

El aislamiento y la extracción de polinucleótidos libres de células se pueden realizar mediante la recogida 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 jeringa. En otros casos, la recogida puede comprender el pipeteo o la recogida directa de fluido en un recipiente colector.

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

Generalmente, los polinucleótidos libres de células se extraen y se aíslan a partir de fluidos corporales a través de una etapa de separación en donde los ADN libres de células, tal como se encuentran en solución, se separan de las células y otros componentes no solubles del cuerpo líquido. El particionamiento puede incluir pero no se limita a técnicas tales como centrifugación o filtración. En otros casos, las células no se separan primero del ADN libre de células, sino que se lisan. En este ejemplo, el ADN genómico de 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 utilizar 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 de granelero no específicos, por ejemplo, a lo largo de la reacción para optimizar ciertos aspectos del procedimiento, como el rendimiento.

El aislamiento y purificación de ADN libre de células se puede lograr usando cualquier medio, incluido, entre otros, 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 estar disponibles no comercialmente.

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

Un método de aumentar la eficiencia de conversión implica el uso de una ligasa de ingeniería para la reactividad óptima en ADN monocatenario, tal como un ThermoPhage ADNss ligasa derivada. Dichas ligasas omiten los pasos tradicionales en la preparación de bibliotecas de reparación de extremos y colas A que pueden tener bajas eficiencias y/o pérdidas acumuladas debido a pasos de limpieza intermedios, y permiten el doble de probabilidad de que el polinucleótido inicial con sentido o antisentido sea convertido en un polinucleótido etiquetado apropiadamente. También convierte polinucleótidos bicatenarios que pueden poseer salientes que pueden no tener los extremos suficientemente romos por la reacción típica de reparación de extremos. Las condiciones de reacción óptimas para esta reacción de ADNss son: 1 x tampón de reacción (MOPS 50 mM (pH 7,5), DTT 1 mM, $MgCl_2$ 5 mM, KCl 10 mM). Con ATP 50 mM, 25 mg/ml de BSA, $MnCl_2$ 2,5 mM, oligómero de ADNss de 85 nt 200 pmol y ligasa de ADNss de 5 U incubados a 65ºC durante 1 hora. La posterior amplificación mediante PCR puede convertir aún más la biblioteca monocatenaria etiquetada en una biblioteca bicatenaria 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, solo o en combinación: sondas de inversión molecular optimizadas por hibridación, ligación de extremos romos con un intervalo de tamaño de polinucleótido bien controlado, ligación de extremos o un paso de amplificación multiplex inicial con o sin el uso de cebadores de fusión.

B. Codificación de barras molecular de polinucleótidos libres de células

Los métodos de esta descripción también pueden permitir que los polinucleótidos libres de células para ser etiquetados o seguidos con el fin de permitir la identificación y el origen del polinucleótido en particular subsiguiente. Esta característica contrasta con otros métodos que utilizan reacciones agrupadas o multiplex y que solo proporcionan mediciones o análisis como un promedio de múltiples muestras. Aquí, la asignación de un identificador a un individuo o subgrupos de polinucleótidos puede permitir que se asigne una identidad única a secuencias individuales o fragmentos de secuencias. Esto puede permitir la adquisición de datos de muestras individuales y no se limita a promedios de muestras.

En algunos ejemplos, los ácidos nucleicos u otras moléculas derivadas de una única hebra pueden compartir una etiqueta o identificador común y por lo tanto puede ser identificado más tarde como derivado de esa hebra. De manera similar, todos los fragmentos de una sola hebra de ácido nucleico pueden etiquetarse con el mismo identificador o etiqueta, permitiendo así la identificación posterior de los fragmentos de la hebra parental. En otros casos, los productos de expresión génica (p. ej., ARNm) pueden etiquetarse para cuantificar la expresión, mediante la cual 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, se pueden etiquetar múltiples productos de amplificación de una reacción de PCR con la misma etiqueta o identificador. Si los productos se secuencian posteriormente y muestran diferencias de secuencia, las diferencias entre productos con el mismo identificador pueden atribuirse a un error de PCR.

En general, los métodos proporcionados en este documento son útiles para la preparación de secuencias de polinucleótidos libres de células a una reacción de secuenciación de aplicación corriente abajo. A menudo, un método de secuenciación es la secuenciación clásica de Sanger. Los métodos de secuenciación pueden incluir entre otros: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de molécula única, secuenciación de nanoporos, secuenciación de semiconductores, secuenciación por ligación, secuenciación por hibridación, ARN-Seq (Illumina), Expresión génica digital (Helicos), secuenciación de próxima generación, secuenciación de molécula única por síntesis (SMSS) (Helicos), secuenciación masivamente paralela, matriz de molécula única clonal (Solexa), secuenciación de escopeta, secuenciación de Maxim-Gilbert, caminata de cebadores y cualquier otro método de secuenciación conocido en la técnica.

C. Asignación de códigos de barras para secuencias de polinucleótidos libres de células

Los sistemas y métodos descritos en este documento pueden utilizarse en aplicaciones que implican la asignación de únicos identificadores 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 usa para marcar el polinucleótido; pero, en algunos casos, se utilizan diferentes identificadores únicos. Por ejemplo, en algunos casos, el identificador único es una sonda de hibridación. En otros casos, el identificador único es un tinte, en cuyo caso la unión puede comprender la intercalación del tinte en la molécula del analito (como la intercalación en ADN o ARN) o la unión a una sonda marcada con el tinte. En otros casos más, el identificador único puede ser un oligonucleótido de ácido nucleico, en cuyo caso la unión a las secuencias polinucleotídicas puede comprender una reacción de ligación entre el oligonucleótido y las secuencias o incorporación mediante PCR. En otros casos, la reacción puede comprender la adición de un isótopo metálico, ya sea directamente al analito o mediante una sonda marcada con el isótopo. Generalmente, la asignación de identificadores únicos o no únicos, o códigos de barras moleculares en reacciones de esta divulgación puede seguir métodos y sistemas descritos por, por ejemplo, las solicitudes de patente de EE. UU. 20010053519, 20030152490, 20110160078 y la patente de EE. UU. 6,582,908.

El método comprende la unión de códigos de barras de oligonucleótidos a analitos de ácido nucleico a través de una reacción enzimática que incluye pero no se limita a una reacción de ligación. Por ejemplo, la enzima de ligasa puede unir covalentemente un código de barras de ADN a ADN fragmentado (por ejemplo, ADN de alto peso molecular). Después de adjuntar los códigos de barras, las moléculas pueden someterse a una reacción de secuenciación.

En algunos casos, la PCR se puede usar para la amplificación global de las secuencias de polinucleótidos libres de células. Esto puede comprender el uso de secuencias adaptadoras que se pueden ligar primero a diferentes moléculas seguido de amplificación por PCR usando cebadores universales. La PCR para secuenciación se puede realizar usando cualquier medio, incluyendo pero no limitado al uso de kits comerciales proporcionados por Nugen (kit WGA), Life Technologies, Affymetrix, Promega, Qiagen y similares. En otros casos, solo se pueden amplificar determinadas moléculas diana dentro de una población de moléculas de polinucleótidos libres de células. Pueden usarse cebadores específicos, junto con la ligación del adaptador, para amplificar selectivamente ciertas dianas para la secuenciación corriente abajo.

Se usan una pluralidad de códigos de barras de manera que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. Los códigos de barras se pueden ligar a moléculas individuales de modo que la

combinación del código de barras y la secuencia a la que se puede ligar crea una secuencia única que se puede rastrear individualmente. Como se describe en el presente documento, la detección de códigos de barras no únicos en combinación con datos de secuencia de porciones de inicio (comienzo) y final (parada) de lecturas de secuencia puede permitir la asignación de una identidad única a una molécula particular. La longitud, o el número de pares de bases, de una secuencia individual leída también puede usarse para asignar una identidad única a dicha molécula. Como se describe en el presente documento, a los fragmentos de una sola hebra de ácido nucleico a los que se les ha asignado una identidad única, pueden permitir así la identificación posterior de los fragmentos de la hebra parental. De esta manera, los polinucleótidos de la muestra pueden etiquetarse de forma única o sustancialmente única.

En general, el método y el sistema de esta divulgación pueden utilizar los métodos de la patente de EE. UU. US 7,537,897 en el uso de códigos de barras moleculares para contar moléculas o analitos, que se incorpora en su totalidad aquí como referencia.

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

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

Recuento de etiquetas	Etiqueta% Identificada correctamente 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 que no sea posible determinar qué lecturas de secuencia se derivan de qué moléculas parentales. Este problema se puede disminuir marcando las moléculas parentales con un número suficiente de identificadores únicos (por ejemplo, el recuento de etiquetas) de modo que exista la probabilidad de que dos moléculas duplicadas, es decir, moléculas que tienen las mismas posiciones de inicio y parada, tengan identificadores únicos diferentes, por lo que esas lecturas de secuencia son rastreables hasta moléculas parentales particulares. Un enfoque para este problema es etiquetar de forma única todas o casi todas las moléculas parentales diferentes de 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 miles de millones de identificadores únicos diferentes.

El método anterior puede ser engorroso y caro. Los inventores de la presente invención se han dado cuenta inesperadamente de que los fragmentos polinucleotídicos individuales en una muestra de ácido nucleico genómico (p. ej., muestra de ADN genómico) pueden identificarse de forma única marcando con identificadores no únicos, p. ej., marcando de forma no exclusiva los fragmentos polinucleotídicos individuales. Como se usa en este documento, una colección de moléculas puede considerarse "etiquetada de forma única" si cada una de al menos el 95% de las moléculas de la colección lleva una etiqueta de identificación ("identificador") que no es compartida por ninguna otra molécula de la colección ("etiqueta única" o "identificador único"). Se puede considerar que una colección de moléculas está "etiquetada de forma no única" si cada una de al menos el 1%, al menos el 5%, al menos el 10%, al menos el 15%, al menos el 20%, al menos el 25%, al menos el 30%, al menos el 35%, al menos el 40%, al menos el 45%, o al menos o aproximadamente el 50% de las moléculas de la colección lleva una etiqueta de identificación que es compartida por al menos otra molécula de la colección ("etiqueta no única" o "identificador no exclusivo"). En algunas realizaciones, para una población no etiquetada de forma única, no más del 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% o 50% de las moléculas están marcadas de forma única. En algunas realizaciones, para el etiquetado único, se utilizan al menos dos veces más etiquetas diferentes que el número estimado de moléculas en la muestra. El número de etiquetas de identificación diferentes utilizadas para etiquetar moléculas en una colección puede variar, por ejemplo, entre 2, 4, 8, 16 o 32 en el extremo inferior del rango y cualquiera de 50, 100, 500, 1000, 5000 y 10.000 en el extremo superior del rango. Entonces, por ejemplo, una colección de entre 100 mil millones y 1 mil millones de moléculas puede etiquetarse con entre 4 y 100 etiquetas de identificación diferentes.

La presente descripción proporciona métodos y composiciones en las que una población de polinucleótidos en una muestra de ADN genómico fragmentado se etiqueta con n diferentes identificadores únicos. En algunas realizaciones, n es al menos 2 y no más de 100.000*z, donde z es una medida de tendencia central (por ejemplo, media, mediana, moda) de un número esperado de moléculas duplicadas que tienen las mismas posiciones de inicio y finalización. En algunas realizaciones, z es 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más de 10. En algunas realizaciones, z es menos de 10, menos de 9, menos de 8, menos que 7, menos de 6, menos de 5, menos de 4, menos de 3. En ciertas realizaciones, n es al 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 oscilar entre cualquier combinación de estos límites superior e inferior. En ciertas realizaciones, n está entre 5*z y 15*z, entre 8*z y 12*z, o aproximadamente 10*z. Por ejemplo, un genoma humano haploide equivalente tiene aproximadamente 3 picogramos de ADN. Una muestra de aproximadamente 1 microgramo de ADN contiene aproximadamente 300.000 equivalentes del genoma humano haploide. En algunas realizaciones, el número n puede estar entre 5 y 95, 6 y 80, 8 y 75, 10 y 70, 15 y 45, entre 24 y 36 o aproximadamente 30. En algunas realizaciones, el número n es menor que 96. Por ejemplo, el número n puede ser mayor o igual que 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3435, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94 o 95. En algunas situaciones, el número n puede ser mayor que cero pero menor que 100, 99, 98, 97, 96, 95, 94, 93, 92, 91 o 90. En algunos ejemplos, el número n es 64. El número n puede ser menos de 75, menos de 50, menos de 40, menos de 30, menos de 20, menos de 10 o menos de 5. Se pueden lograr mejoras en la secuenciación siempre que al menos algunos de los polinucleótidos duplicados o afines llevan identificadores únicos, es decir, llevan etiquetas diferentes. Sin embargo, en determinadas realizaciones, el número de etiquetas utilizadas se selecciona de modo que exista al menos un 95% de posibilidades de que todas las moléculas duplicadas que comprenden las mismas secuencias de inicio y final lleven identificadores únicos.

Algunas realizaciones proporcionan métodos para realizar una reacción de ligación en donde los polinucleótidos parentales en una muestra se mezclan con una mezcla de reacción que comprende y diferentes oligonucleótidos de código de barras, en donde y = a raíz cuadrada de n. La ligación puede resultar en la unión aleatoria de oligonucleótidos de código de barras a polinucleótidos parentales en la muestra. A continuación, la mezcla de reacción se puede incubar en condiciones de ligación suficientes para efectuar la ligación de los oligonucleótidos de código de barras a los polinucleótidos parentales de la muestra. En algunas realizaciones, los códigos de barras aleatorios seleccionados de los diferentes oligonucleótidos de códigos de barras se ligan a ambos extremos de los polinucleótidos parentales. La ligación aleatoria de los códigos de barras y a uno o ambos extremos de los polinucleótidos parentales puede resultar en la producción de identificadores únicos y2. Por ejemplo, una muestra que comprende aproximadamente 10.000 equivalentes del genoma humano haploide de cfADN puede etiquetarse con aproximadamente 36 identificadores únicos. Los identificadores únicos pueden comprender seis códigos de barras de ADN únicos. La unión de 6 códigos de barras únicos a ambos extremos de un polinucleótido puede resultar en la producción de 36 identificadores únicos posibles.

En algunas realizaciones, una muestra que comprende de aproximadamente 10.000 equivalentes de genoma humano haploide de ADN se marca con 64 identificadores únicos, en donde los 64 identificadores únicos se producen por la ligación de 8 códigos de barras únicos a ambos extremos de polinucleótidos parentales. La eficacia de ligación de la reacción puede ser superior al 10%, superior al 20%, superior al 30%, superior al 40%, superior al 50%, superior al 60%, superior al 70%, superior al 80% o superior al 90%. Las condiciones de ligación pueden comprender el uso de adaptadores bidireccionales que pueden unirse a cualquier extremo del fragmento y aún ser amplificables. Las condiciones de ligación pueden comprender ligación de extremos romos, en contraposición a colas con adaptadores bifurcados. Las condiciones

de ligación pueden comprender una valoración cuidadosa de una cantidad de oligonucleótidos adaptadores y/o de códigos de barras. Las condiciones de ligación pueden comprender el uso de más de 2X, más de 5X, más de 10X, más de 20X, más de 40X, más de 60X, más de 80X, (p. ej., ~100X) exceso molar de oligonucleótidos de adaptador y/o código de barras en comparación con una cantidad de fragmentos de polinucleótidos parentales en la mezcla de reacción. Las condiciones de ligación pueden comprender el uso de una ADN ligasa T4 (p. ej., módulo de ultra ligación NEBNExt). En un ejemplo, se utilizan 18 microlitros de mezcla maestra de ligasa con ligación de 90 microlitros (18 partes de las 90) y potenciador de ligación. Por consiguiente, marcar polinucleótidos parentales con n identificadores únicos puede comprender el uso de un número y códigos de barras diferentes, en los que y = una raíz cuadrada de n. Las muestras marcadas de esta manera 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, cfADN. El número y de códigos de barras utilizados para identificar los polinucleótidos originales en una muestra puede depender de la cantidad de ácido nucleico en la muestra.

III. Plataformas de secuenciación de ácido nucleico

Después de la extracción y el aislamiento de polinucleótidos libres de células de los fluidos corporales, las secuencias libres de células se pueden secuenciar. A menudo, un método de secuenciación es la secuenciación clásica de Sanger. Los métodos de secuenciación pueden incluir entre otros: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de una sola molécula, secuenciación de nanoporos, secuenciación de semiconductores, secuenciación por ligación, secuenciación por hibridación, ARN-Seq (Illumina), Digital Gene Expression (Helicos), Secuenciación de próxima generación, Secuenciación de una sola molécula por síntesis (SMSS) (Helicos), Secuenciación masivamente paralela, Matriz de moléculas simples clonales (Solexa), Secuenciación de escopeta, Secuenciación de Maxim-Gilbert con cebador, caminata, secuenciación usando 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 el presente documento, pueden comprender una variedad de muestras unidades de procesamiento. Las unidades de procesamiento de muestras pueden incluir pero no se limitan a múltiples carriles, múltiples canales, múltiples pozos u otro medio de procesar múltiples conjuntos de muestras sustancialmente simultáneamente. Además, la unidad de procesamiento de muestras puede incluir múltiples cámaras de muestras para permitir el procesamiento de múltiples ejecuciones simultáneamente.

En algunos ejemplos, las reacciones de secuenciación simultáneas se pueden realizar usando secuenciación multiplex. En algunos casos, los polinucleótidos libres de células pueden secuenciarse con al menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10.000, 50.000, 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, 10.000, 50.000, 100.000 reacciones de secuenciación. Las reacciones de secuenciación se pueden realizar de forma secuencial o simultánea. Se pueden realizar análisis de datos posteriores en todas o en parte de las reacciones de secuenciación. En algunos casos, el análisis de datos se puede realizar en al menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10.000, 50.000, 100.000 reacciones de secuenciación. En otros casos, el análisis de datos se puede realizar en menos de 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10.000, 50.000, 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 de al menos 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99,9% o 100%. En otros casos, la cobertura de la secuencia del genoma puede ser inferior al 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99,9% o 100%.

En algunos ejemplos, la secuenciación se puede realizar en polinucleótidos libres de células que pueden comprender una variedad de diferentes tipos de ácidos nucleicos. Los ácidos nucleicos pueden ser polinucleótidos u oligonucleótidos. Los ácidos nucleicos incluidos, pero no se limitan a ADN o ARN, monocatenario o bicatenario o un par ARN/ADNc.

IV. Estrategia de análisis de polinucleótidos

Fig 8. es un diagrama, **800**, que muestra una estrategia para el análisis de 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 diana en baja abundancia. Por ejemplo, el ácido nucleico de un genoma normal o de tipo silvestre (p. ej., 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 al 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 no puede constituir más de 100 ng de ácido nucleico. Esto puede contribuir a un sobremuestreo adecuado de los polinucleótidos originales mediante el proceso de secuenciación o análisis genético. Alternativamente, la muestra se puede tapar o embotellar artificialmente para reducir la cantidad de ácido nucleico a no más de 100 ng o

enriquecerse selectivamente para analizar solo las secuencias de interés. La muestra se puede modificar para producir de forma selectiva lecturas de secuencia de moléculas que se mapean en cada una de una o más ubicaciones seleccionadas en una secuencia de referencia. Una muestra de 100 ng de ácido nucleico puede contener aproximadamente 30.000 equivalentes de genoma haploide humano, es decir, moléculas que, juntas, proporcionan una cobertura de 30.000 veces más de un genoma humano.

En el paso **804** el material genético inicial se convierte en un conjunto de polinucleótidos parentales etiquetados. El etiquetado puede incluir unir etiquetas secuenciadas a moléculas en el material genético inicial. Las etiquetas secuenciadas pueden seleccionarse de modo que todos los polinucleótidos únicos que se mapean en la misma ubicación en una secuencia de referencia tengan una etiqueta de identificación única. La conversión se puede realizar con alta eficiencia, por ejemplo al menos 50%.

En el paso **806,** el conjunto de polinucleótidos parentales marcados se amplifica para producir un conjunto de polinucleótidos de progenie amplificados. La amplificación puede ser, por ejemplo, 1000 veces.

En el paso **808,** el conjunto de polinucleótidos de progenie amplificados se muestrea para la secuenciación. La frecuencia de muestreo se elige de modo que las lecturas de secuencia producidas (1) cubran un número objetivo de moléculas únicas en el conjunto de polinucleótidos parentales marcados y (2) cubran moléculas únicas en el conjunto de polinucleótidos parentales marcados en un pliegue de cobertura diana (p. ej., de 5 a 10 veces la cobertura de polinucleótidos parentales.

En el paso **810,** el conjunto de secuencia de lecturas se colapsa para producir un conjunto de secuencias de consenso correspondientes a polinucleótidos parentales etiquetados únicos. Las lecturas de secuencia pueden ser cualificadas para su inclusión en el análisis. Por ejemplo, las lecturas de secuencia que no cumplan con una puntuación de control de calidad se pueden eliminar del conjunto. Las lecturas de secuencia se pueden clasificar en familias que representan lecturas de moléculas de progenie derivadas de una molécula parental única en particular. Por ejemplo, una familia de polinucleótidos de progenie amplificados puede constituir aquellas moléculas amplificadas derivadas de un polinucleótido de un solo padre. Al comparar las secuencias de la progenie en una familia, una secuencia de consenso del par original t puede deducirse el polinucleótido. Esto produce un conjunto de secuencias consenso que representan polinucleótidos parentales únicos en el grupo marcado.

En el paso **812,** el conjunto de secuencias de consenso se analiza usando cualquiera de los métodos analíticos descritos en el presente documento. Por ejemplo, el mapeo de secuencias de consenso en una ubicación de secuencia de referencia particular puede analizarse para detectar instancias de variación genética. El mapeo de secuencias de consenso con secuencias de referencia particulares se puede medir y normalizar frente a muestras de control. Las medidas de las moléculas que mapean las secuencias de referencia se pueden comparar en un genoma para identificar áreas en el genoma en las que el número de copias varía o se pierde la heterocigosidad.

La figura 9 es un diagrama que presenta un método más genérico de extraer información de una señal representada por una colección de lecturas de secuencia. En este método, después de secuenciar polinucleótidos de progenie amplificados, 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 los métodos de interpretación de la información en la secuencia para determinar el contenido de los polinucleótidos parentales marcados con mayor fidelidad, por ejemplo, menos ruido y/o distorsión.

El análisis de la colección de lecturas de secuencia permite hacer inferencias sobre la población de polinucleótidos parental a partir de la cual se generaron las lecturas de secuencia. Tales inferencias pueden ser útiles porque la secuenciación implica típicamente leer sólo un subconjunto parcial de los polinucleótidos amplificados totales globales. Por lo tanto, no se puede estar seguro de que todos los polinucleótidos parentales estarán representados por al menos una secuencia leída en la colección de lecturas de secuencia.

Una tal inferencia es el número de polinucleótidos parentales únicas en el grupo original. Esta inferencia se puede hacer basándose en el 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 secuencia rastreables hasta un polinucleótido parental original. La inferencia se puede hacer utilizando métodos estadísticos bien conocidos. Por ejemplo, si el agrupamiento produce muchas familias, cada una representada por una o unas pocas progenies, entonces se puede inferir que la población original incluía más polinucleótidos parentales únicos que no fueron secuenciados. Por otro lado, si el agrupamiento produce solo unas pocas familias, cada familia representada por muchos descendientes, entonces se puede inferir que la mayoría de los polinucleótidos únicos en la población original están representados por al menos un grupo de lectura de secuencia en esa familia.

Otra inferencia de este tipo es la frecuencia de una base o secuencia de bases en un locus particular en un grupo original de polinucleótidos. Esta inferencia se puede hacer basándose en el número de familias únicas en las que se pueden agrupar las lecturas de secuencia y el número de lecturas de secuencia en cada familia. Al analizar 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 en particular. Luego, teniendo en cuenta 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 copias usando una sola muestra

La figura 1 es un diagrama, **100**, que muestra una estrategia para la detección de la variación del número de copias en un solo sujeto. Como se muestra en el presente documento, los métodos de detección de variación del número de copias se pueden implementar como sigue. Después de la extracción y el aislamiento de polinucleótidos libres de células en el paso **102,** se puede secuenciar una única muestra única 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 secuencias de lectura 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 análisis posteriores basados en un umbral. En algunos casos, algunas lecturas no tienen la calidad o la longitud suficiente para realizar el siguiente paso de mapeo. Las lecturas de secuenciación con una puntuación de calidad de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera de los datos. En otros casos, las lecturas de secuenciación asignadas a una puntuación de calidad inferior al 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse del conjunto de datos. En el paso **106,** las lecturas del fragmento genómico que cumplen un umbral de puntuación de calidad especificado se mapean en un genoma de referencia, o una secuencia de plantilla que se sabe que no contiene variaciones en el número de copias. Después de mapear la alineación, 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 de forma única. En casos, las lecturas pueden ser secuencias no relacionadas con el análisis de variación del número de copias. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Las lecturas de secuenciación con una puntuación de mapeo de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% se pueden filtrar del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a un mapeo puntuados por debajo del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos.

Después de filtrado y de asignación de datos, la pluralidad de lecturas de secuencia genera una región cromosómica de la cobertura. En el paso **108,** estas regiones cromosómicas se pueden dividir en ventanas o intervalos de longitud variable. Una ventana o intervalo puede tener al 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 intervalo 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 intervalo 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 cobertura en el paso **110,** se selecciona cada ventana o bin para contener aproximadamente el mismo número de bases asignables. En algunos casos, cada ventana o intervalo de una región cromosómica puede contener el número exacto de bases cartografiables. En otros casos, cada ventana o intervalo puede contener un número diferente de bases asignables. Además, cada ventana o intervalo puede no superponerse con una ventana o intervalo adyacente. En otros casos, una ventana o intervalo puede superponerse con otra ventana o intervalo adyacente. En algunos casos, una ventana o intervalo puede superponerse en al menos 1 pb, 2, pb, 3 pb, 4 pb, 5, pb, 10 pb, 20 pb, 25 pb, 50 pb, 100 pb, 200 pb, 250 pb, 500 pb o 1000 pb. En otros casos, una ventana o intervalo puede superponerse hasta en 1 pb, 2, pb, 3 pb, 4 pb, 5, pb, 10 pb, 20 pb, 25 pb, 50 pb, 100 pb, 200 pb, 250 pb, 500 pb o 1000 pb. En algunos casos, una ventana o intervalo puede superponerse en aproximadamente 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 de ventana puede dimensionar de manera que contienen aproximadamente el mismo número de bases asignables de forma única. La capacidad de asignación de cada base que comprende una región de ventana se determina y se utiliza para generar un archivo de capacidad de asignación que contiene una representación de las lecturas de las referencias que se asignan de nuevo a la referencia para cada archivo. El archivo de mapeo contiene una fila por cada posición, lo que indica si cada posición es o no mapeable de forma única.

Además, ventanas predefinidas, conocidas por todo el genoma a ser difícil de secuenciar, o contiene un sesgo sustancialmente alto GC, pueden ser filtradas a partir 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, ADN centromérico) contienen secuencias muy 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 al menos 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5000, 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, 5000, 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 la determinación de cobertura de lectura para cada región de la ventana. Esto se puede realizar utilizando 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 tengan suficientes puntuaciones de mapeo y calidad y que se encuentren dentro de las ventanas cromosómicas que no están filtradas. Al número de lecturas de cobertura se le puede asignar una puntuación por cada posición mapeable. En los casos que involucran códigos de barras, todas las secuencias con el mismo código de barras, propiedades físicas o combinación de los dos pueden colapsarse en una sola lectura, ya que todas se derivan de la muestra de la molécula madre. Este paso reduce los sesgos que pueden haberse introducido durante cualquiera de los pasos anteriores, como los 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, anulando así el efecto de amplificación desigual.

Las secuencias de consenso se pueden generar a partir de familias de secuencia de lecturas por cualquier método conocido en la técnica. Dichos métodos incluyen, por ejemplo, métodos lineales o no lineales para construir secuencias de consenso (como votación, promediado, estadístico, detección máxima a posteriori o de máxima verosimilitud, programación dinámica, Bayesiano, Markov oculto o métodos de máquina de vectores de soporte, etc.) derivado de la teoría de la comunicación digital, la teoría de la información o la bioinformática.

Después de determinarse la cobertura de lectura de secuencia, se aplica un algoritmo de modelado estocástico para convertir la cobertura de lecturas 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 trellis, decodificación de Viterbi, maximización de expectativas, metodologías de filtrado de Kalman y redes neuronales.

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

En el paso **114,** la variación del número de copias puede ser reportada como un gráfico, indicando diferentes posiciones en el genoma y un correspondiente aumento o disminución o mantenimiento de la variación del número de copias en cada posición respectiva. Además, la variación del número de copias puede usarse para informar una puntuación porcentual que indique cuánto material patológico (o ácidos nucleicos que tienen una variación del número de copias) existe en la muestra de polinucleótido libre de células.

Un método de determinar la variación del número de copias se muestra en la Fig. 10. En ese método, después de agrupar lecturas de secuencia en familias generadas a partir de un polinucleótido parental único (1010), las familias se cuantifican, por ejemplo, por determinación del número de familias que se mapean en cada una de una pluralidad de ubicaciones de secuencia de referencia diferentes. Las CNV se pueden determinar directamente comparando una medida cuantitativa de familias en cada una 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 discutió anteriormente. Entonces, la CNV se puede determinar comparando la medida de cantidad inferida en la pluralidad de loci. En otras formas de realización, un enfoque híbrido puede ser tomado por el que una inferencia similar de cantidad original puede hacerse siguiendo la normalización de sesgo de representación durante el proceso de secuenciación, tales como el sesgo de GC, etc.

B. Detección de variación del número de copias mediante el uso de una muestra pareada

La detección de variación del número de copias de muestra pareada comparte muchos de los pasos y parámetros como el enfoque de muestra única descrito en este documento. Sin embargo, como se muestra en **200** de la figura 2 de la detección de variación del número de copias usando muestras pareadas, se requiere la comparación de la cobertura de secuencia con una muestra de control en lugar de compararla con la capacidad de mapeo predicha del genoma. Este enfoque puede ayudar a la normalización en todas las ventanas. La figura 2 es un diagrama **200** que muestra una estrategia para la detección de la variación del número de copias en sujetos emparejados. Como se muestra en el presente documento, los métodos de detección de variación del número de copias se pueden implementar como sigue. En el paso **204,** se puede secuenciar una única muestra única mediante una plataforma de secuenciación de ácidos nucleicos conocida en la técnica después de la extracción y aislamiento de la muestra en el paso **202.** Este paso genera una pluralidad de lecturas de secuencias de fragmentos genómicos. Además, se toma una muestra o una muestra de 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 tener una enfermedad en 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, a las lecturas se les asigna una puntuación de calidad. En algunos casos, algunas lecturas no tienen la calidad o la longitud suficiente para realizar el siguiente paso de mapeo. Las lecturas de secuenciación con una puntuación de calidad de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a una calidad puntuada menos del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos. En el paso **206,** las lecturas del fragmento genómico que cumplen un umbral de puntuación de calidad especificado se mapean en un genoma de referencia, o una secuencia de plantilla que se sabe que no contiene variaciones en el número de copias. Después de mapear la alineación, 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 variación del número de copias. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Lecturas de secuenciación con una puntuación de mapeo de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a un mapeo se puntuaron por debajo del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos.

Después de filtrar y de asignación de datos, la pluralidad de lecturas de secuencia genera una región cromosómica de la cobertura para cada uno de los sujetos de prueba y de control. En el paso **208,** estas regiones cromosómicas se pueden dividir en ventanas o intervalos de longitud variable. Una ventana o intervalo puede tener al 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 intervalo también puede tener 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.

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

En algunos casos, cada una de las regiones de la ventana tiene un tamaño que contiene aproximadamente el mismo número de bases para cada uno de los sujetos de prueba y control. La capacidad de asignación de cada base que comprende una región de ventana se determina y se utiliza para generar un archivo de capacidad de asignación que contiene una representación de las lecturas de las referencias que se asignan de nuevo a la referencia para cada archivo. El archivo de mapeo contiene una fila por cada posición, lo que indica si cada posición es o no mapeable de forma única.

Además, ventanas predefinidas, conocidas por todo el genoma a ser difícil de secuenciar, o contienen un sesgo sustancialmente alto GC, se filtran a partir 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, ADN centromérico) contienen secuencias muy repetitivas que pueden producir resultados falsos positivos. Estas regiones se pueden filtrar. 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.

También puede variar el número de ventanas analizadas. En algunos casos, se analizan al menos 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5000, 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, 5000, 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 la determinación de la cobertura de lectura para cada región de ventana para cada uno de los sujetos de prueba y de control. Esto se puede realizar utilizando 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 tengan suficientes puntuaciones de mapeo y calidad y que se encuentren dentro de las ventanas cromosómicas que no están filtradas. Al número de lecturas de cobertura se le puede asignar una puntuación por cada posición mapeable. En los casos que involucran códigos de barras, todas las secuencias con el mismo código de barras pueden colapsarse en una sola lectura, ya que todas se derivan de la muestra de la molécula madre. Este paso reduce los sesgos que pueden haberse introducido durante cualquiera de los pasos anteriores, como los 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 menor cantidad de sesgo:

Al determinar la cobertura de lectura de ácido nucleico para cada ventana, la cobertura de cada ventana se puede normalizar mediante la cobertura media de esa muestra. Usando tal enfoque, puede ser deseable secuenciar tanto al sujeto de prueba como al control en condiciones similares. La cobertura de lectura para cada ventana se puede expresar luego como una relación entre ventanas similares. Las relaciones de cobertura de lectura de ácido nucleico para cada

ventana del sujeto de prueba se pueden determinar dividiendo la cobertura de lectura de cada región de la ventana de la muestra de prueba con la cobertura de lectura de una región de ventana correspondiente del amplio de control.

Después de la secuencia de lectura se han determinado ratios de cobertura, se aplica un algoritmo de modelado estocástico para convertir los coeficientes normalizados para cada región de ventana en estados de número de copias discretas. En algunos casos, este algoritmo puede comprender un modelo de Markov oculto. En otros casos, el modelo estocástico puede comprender programación dinámica, máquina de vectores de soporte, modelado bayesiano, modelado probabilístico, decodificación de Trellis, decodificación de Viterbi, maximización de expectativas, metodologías de filtrado de Kalman o redes neuronales.

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

En el paso **214,** la variación del número de copia puede ser reportado como gráfico, indicando diferentes posiciones en el genoma y un correspondiente aumento o disminución o mantenimiento de la variación del número de copias en cada posición respectiva. Además, la variación del número de copias puede usarse para informar una puntuación porcentual que indique cuánto material patológico existe en la muestra de polinucleótido libre de células.

VI. Detección de mutación rara

La detección de mutaciones raras comparte características similares a medida que se acerca la variación del número de copias. Sin embargo, como se muestra en la figura 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 a la normalización en todas las ventanas.

En general, la detección de mutaciones raras se puede realizar en regiones enriquecidas selectivamente del genoma o el transcriptoma purificado y aislado en el paso **302.** Como se describe en el presente documento, regiones específicas, que pueden incluir pero no se limitan a genes, oncogenes, genes supresores de tumores, los promotores, elementos de secuencia reguladora, regiones no codificantes, ARNmi, ARNsn y similares pueden amplificarse selectivamente a partir de una población total de polinucleótidos libres de células. Esto se puede realizar como se describe en este documento. En un ejemplo, se puede usar secuenciación multiplex, con o sin etiquetas de código de barras para secuencias de polinucleótidos individuales. En otros ejemplos, la secuenciación se puede realizar 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 una puntuación de calidad. Una puntuación de calidad puede ser una representación de lecturas que indica si esas lecturas pueden ser útiles en análisis posteriores basados en un umbral. En algunos casos, algunas lecturas no tienen la calidad o la longitud suficiente para realizar el siguiente paso de mapeo. Las lecturas de secuenciación con una puntuación de calidad de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a una calidad puntuada al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos. En el paso 306, las lecturas del fragmento genómico que cumplen un umbral de puntuación de calidad especificado se mapean en un genoma de referencia, o una secuencia de referencia que se sabe que no contiene mutaciones raras. Después de mapear la alineación, 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 de forma única. En casos, las lecturas pueden ser secuencias no relacionadas con el análisis de mutaciones raras. Por ejemplo, algunas lecturas de secuencia pueden originarse a partir de polinucleótidos contaminantes. Las lecturas de secuenciación con una puntuación de mapeo de al menos 90%, 95%, 99%, 99,9%, 99,99% o 99,999% se pueden filtrar del conjunto de datos. En otros casos, las lecturas de secuenciación asignadas a un mapeo puntuados por debajo del 90%, 95%, 99%, 99,9%, 99,99% o 99,999% pueden filtrarse fuera del conjunto de datos.

Para cada base mapeable, bases que no cumplan con el umbral mínimo para mapeabilidad, o bases de baja calidad, puede ser reemplazada por las bases correspondientes como se encuentra en la secuencia de referencia.

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

Para un genoma ejemplar derivado de secuencias de polinucleótidos libres de células, el siguiente paso comprende la determinación de la cobertura de lecturas para cada posición de la base mapeable. Esto se puede realizar utilizando 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 tengan un mapeo suficiente y puntajes de calidad. Al número de lecturas de cobertura se le puede asignar una puntuación por cada posición mapeable. En los casos que involucran códigos de barras, todas las secuencias con el mismo código de barras pueden colapsarse en una lectura de consenso, ya que todas se derivan de la molécula madre de muestra. La secuencia para cada base se alinea como el nucleótido más dominante leído para esa ubicación específica. Además, el número de moléculas únicas se puede contar en cada posición para obtener 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 la amplificación. Solo las lecturas con códigos de barras únicos pueden contarse para cada posición mapeable e influir en la puntuación asignada.

Una vez que la cobertura de lectura puede ser determinada y son bases variantes relativas a la secuencia de control en cada lectura identificada, la frecuencia de bases variantes puede ser calculada como el número de lecturas que contiene la variante dividida por el número total de lecturas. Esto puede expresarse como una proporción para cada posición mapeable en 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 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 trellis, decodificación de Viterbi, maximización de expectativas, metodologías de filtrado de Kalman y redes neuronales.

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

Después del cálculo de las frecuencias de varianza para cada posición de base, la variante con una desviación más grande 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 poco común puede ser una mutación cancerosa. En otros casos, una mutación rara puede estar correlacionada con un estado de enfermedad.

Una mutación rara o variante puede comprender una aberración genética que incluye, pero no se limita a una base de sustitución única, o pequeños indeles, 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 al 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 se pueden reflejar en forma gráfica, indicando diferentes posiciones en el genoma y un correspondiente aumento o disminución o mantenimiento de una frecuencia de mutación en cada posición respectiva. Además, se pueden usar mutaciones raras para informar una puntuación porcentual que indique cuánto material patológico existe en la muestra de polinucleótido libre de células. Una puntuación de confianza puede acompañar a cada mutación detectada, dadas las estadísticas conocidas de variaciones típicas en posiciones informadas en secuencias de referencia que no son enfermedades. Las mutaciones también pueden clasificarse en orden de abundancia en el sujeto o clasificarse por importancia clínicamente procesable.

La figura 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 etiquetado 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 varios métodos estadísticos conocidos y puede basarse, al menos en parte, en la frecuencia con la que aparece una base entre las lecturas de secuencia pertenecientes a la familia (1112). Por ejemplo, la puntuación de confianza puede ser la frecuencia con la que aparece la base entre las lecturas de la secuencia. Como otro ejemplo, para cada familia, se puede construir un modelo de Markov oculto, de modo que se pueda tomar una decisión de máxima probabilidad o máxima a posteriori basada en la frecuencia de ocurrencia de una base particular en una sola familia. Como parte de este modelo, también se pueden generar la probabilidad de error y la puntuación de confianza resultante para una decisión en particular. Luego, se puede asignar una frecuencia de la base en la población original con base en las puntuaciones de confianza entre las familias (1114).

VII. Aplicaciones

A. Detección temprana del cáncer

Numerosos tipos de cáncer pueden ser detectados usando los métodos y sistemas descritos en el presente documento. Las células cancerosas, como la mayoría de las células, se pueden caracterizar por una tasa de renovación, en donde las células viejas mueren y son reemplazadas por células más nuevas. Generalmente, las células muertas, en contacto con la vasculatura de un sujeto dado, pueden liberar ADN o fragmentos de ADN al torrente sanguíneo. Esto también se aplica a las células cancerosas durante varias etapas de la enfermedad. Las células cancerosas también pueden caracterizarse, dependiendo del estadio de la enfermedad, por diversas aberraciones genéticas, como la variación del número de copias, así como mutaciones raras. Este fenómeno puede usarse para detectar la presencia o ausencia de cánceres en individuos usando los métodos y sistemas descritos en este documento.

Por ejemplo, la sangre de los sujetos con riesgo de cáncer puede ser extraida y preparada como se describe en el presente documento 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 de la enfermedad.

Los tipos y número de cánceres que se pueden detectar puede incluir pero no se limitan a cánceres de la sangre, cánceres cerebrales, cánceres de pulmón, cáncer de piel, cáncer de la nariz, cánceres de garganta, cáncer de hígado, cánceres de hueso, linfomas, cánceres pancreáticos, cánceres de la piel, cánceres de intestino, cánceres de recto, cánceres de tiroides, cánceres de vejiga, cánceres de riñón, cánceres de boca, cánceres de estómago, tumores de estado sólido, tumores heterogéneos, tumores homogéneos y similares.

En la detección temprana de cánceres, cualquiera de los sistemas o métodos descritos en este documento, incluyendo la detección de mutaciones raras o el número de copias de detección de variación, puede ser utilizado para detectar tipos de cáncer. Estos sistemas y métodos pueden usarse para detectar cualquier número de aberraciones genéticas que pueden causar o resultar de cánceres. Estos pueden incluir entre otros, 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 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 del ácido nucleico y cáncer.

Además, los sistemas y métodos descritos en este documento 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 médicos ayudar a caracterizar mejor una forma específica de cáncer. A menudo, los cánceres son heterogéneos tanto en composición como en estadios. 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 pistas al sujeto o al médico con respecto al pronóstico de un tipo específico de cáncer.

B. Seguimiento y pronóstico de cáncer

Los sistemas y métodos proporcionados en este documento pueden usarse para controlar los cánceres ya conocidos, u otras enfermedades en un sujeto particular. Esto puede permitir que un sujeto o un médico adapten las opciones de tratamiento de acuerdo con el progreso de la enfermedad. En este ejemplo, los sistemas y métodos descritos en este documento 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 o en remisión. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión, remisión o recurrencia de la enfermedad.

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

Por ejemplo, las mutaciones que ocurren dentro de un rango de frecuencia que comienza en el nivel umbral se pueden determinar 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, al menos 0,1%, al menos 1% o al menos 5% a 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 ocurren dentro del rango de frecuencia

incluyendo, por ejemplo, sus frecuencias. Se puede tomar una muestra del sujeto en cualquier momento posterior. Se pueden determinar las mutaciones que ocurren dentro del rango de frecuencia original o en un rango de frecuencia diferente. El curso del tratamiento se puede ajustar en función de las mediciones posteriores.

C. Detección temprana y seguimiento de otras enfermedades o estados patológicos

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

Además, los sistemas y métodos de esta descripción también se pueden utilizar para controlar infecciones sistémicas sí mismos, como las que pueden ser causadas por un patógeno tal como una bacteria o virus. Se puede utilizar la variación del número de copias o incluso la detección de mutaciones raras 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 mutar en formas más virulentas durante el curso de la infección.

Sin embargo, otro ejemplo para el que pueden utilizarse el sistema y métodos de esta descripción es el seguimiento de los sujetos de trasplante. Generalmente, 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 perfilar las actividades de rechazo del cuerpo huésped, cuando las células inmunes intentan destruir el tejido trasplantado. Esto puede ser útil para controlar el estado del tejido trasplantado, así como para alterar el curso del tratamiento o la prevención del rechazo.

Además, los métodos de la descripción se pueden usar para caracterizar la heterogeneidad de una condición anormal en un sujeto, comprendiendo el método la generación de un perfil genético de polinucleótidos extracelulares en el sujeto, en donde el perfil genético comprende una pluralidad de datos que resultan de la variación del número de copias y análisis de mutaciones raras. En algunos casos, incluidos, entre otros, el 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. Nuevamente, 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 descripción pueden utilizarse para generar un perfil, huellas digitales o un conjunto de datos que son una suma de la información genética procedente de diferentes células en una enfermedad heterogénea. Este conjunto de datos puede comprender la variación de número de copias y análisis de mutaciones raras solos o en combinación.

D. Detección temprana y la vigilancia de otras enfermedades o estados de enfermedad de origen fetal

Además, los sistemas y métodos de la descripción se pueden usar para diagnosticar, pronosticar, monitorear u observar cánceres u otras enfermedades de origen fetal. Es decir, estas metodologías pueden emplearse en una mujer embarazada para diagnosticar, pronosticar, monitorear u observar cánceres u otras enfermedades en un sujeto no nacido cuyo ADN y otros polinucleótidos pueden co-circular con moléculas maternas.

VIII. Terminología

La terminología utilizada en el mismo es para el propósito de describir solamente realizaciones particulares y no pretende ser limitante de sistemas y métodos de esta descripción. Como se usa en el presente documento, las formas singulares "un", “una”, “el” y “ella” pretenden incluir las formas plurales también, a menos que el contexto indique claramente lo contrario. Además, en la medida en que los términos "que incluye", "incluye", "que tiene", "tiene", "con" o variantes de los mismos se utilizan en la descripción detallada y/o en las reivindicaciones, se pretende que dichos términos incluyan una manera similar al término "que comprende".

Se describen varios aspectos de sistemas y métodos de esta descripción 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 los sistemas y métodos pueden practicarse 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 ocurrir en diferentes órdenes y/o simultáneamente con otros actos o eventos. Además, no todos los actos o eventos ilustrados son necesarios para implementar una metodología de acuerdo con esta información.

Los rangos se pueden expresar en el presente documento desde "aproximadamente" un valor particular y/o hasta "aproximadamente" otro valor particular. Cuando se expresa tal intervalo, otra realización incluye desde un valor particular y/o hasta el 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 rangos son significativos tanto en relación con el otro punto final como independientemente del otro punto final. El término "aproximadamente" como se usa en este documento se refiere a un rango que es el 15% más o menos de un valor numérico establecido dentro del contexto del uso particular. Por ejemplo, alrededor de 10 incluirían un rango de 8,5 a 11,5.

Sistemas informáticos

Los métodos de la presente divulgación pueden ser implementados utilizando, o con la ayuda de los sistemas informáticos. La figura 15 muestra un sistema informático 1501 que está programado o configurado de otro modo para implementar los métodos de la presente divulgación. El sistema informático 1501 puede regular varios aspectos de la preparación, secuenciación y/o análisis de muestras. En algunos ejemplos, el sistema informático 1501 está configurado para realizar la preparación y el análisis de muestras, incluida la secuenciación de ácidos nucleicos.

El sistema de ordenador 1501 incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador de ordenador" en este documento) 1505, que puede ser un solo núcleo o procesador multi núcleo, o una pluralidad de procesadores para el procesamiento paralelo. El sistema informático 1501 también incluye memoria o ubicación 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, tales 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 continuas), como una placa base. La unidad de almacenamiento 1515 puede ser una unidad de almacenamiento de datos (o depósito 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 permitir la informática distribuida, como la informática en la nube. La red 1530, en algunos casos con la ayuda del sistema informático 1501, puede implementar una red de igual a igual, 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 ser incorporados en un programa o software. Las instrucciones pueden almacenarse en una ubicación de memoria, tal como la memoria 1510. Los ejemplos de operaciones realizadas por la CPU 1505 pueden incluir buscar, decodificar, ejecutar y reescribir.

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 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 ubicadas 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, operador). Ejemplos de sistemas informáticos remotos incluyen computadoras personales (p. ej., PC portátil), tablet o tabletas (p. ej., IPad de Apple®, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (p. ej., IPhone de Apple®, dispositivo habilitado para Android, Blackberry®) o asistentes digitales personales. El usuario puede acceder al sistema de ordenador 1501 a través de la red 1530.

Los métodos como se describe aquí pueden ser implementados a modo de máquina (por ejemplo, procesador de ordenador) código almacenado ejecutable en una ubicación de almacenamiento electrónico del sistema de ordenador 1501, tales como, por ejemplo, en la memoria 1510 o la unidad de almacenamiento electrónica 1515. El código ejecutable 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 ser recuperado de la unidad de almacenamiento 1515 y almacenado en la memoria 1510 para que el procesador 1505 tenga fácil acceso. En algunas situaciones, la unidad de almacenamiento electrónica 1515 se puede excluir, y las instrucciones ejecutables por máquina se almacenan en la

memoria 1510.

El código puede ser pre-compilado y configurado para su uso con una máquina que tiene una procesadora adaptada para ejecutar el código, o se puede compilar en tiempo de ejecución. El código se puede suministrar en un lenguaje de programación que se puede seleccionar para permitir que el código se ejecute de una manera precompilada o como compilada.

Los aspectos de los sistemas y métodos proporcionados en este documento, tales como el sistema informático 1501, pueden incorporarse en programación. Varios aspectos de la tecnología se pueden considerar como "productos" o "artículos manufacturados" típicamente en forma de código ejecutable de máquina (o procesador) y/o datos asociados que se transportan o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina se puede almacenar en una unidad de almacenamiento electrónica, como una memoria (por ejemplo, memoria de solo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir parte o toda la memoria tangible de las computadoras, 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 el software o partes del mismo pueden comunicarse a través de Internet o de otras redes de telecomunicaciones. Tales comunicaciones, por ejemplo, pueden permitir la carga del software desde un ordenador o procesador a otro, por ejemplo, desde un servidor de gestión u ordenador principal a la plataforma informática de un servidor de aplicaciones. Por tanto, otro tipo de medio que puede llevar los elementos de software incluye ondas ópticas, eléctricas y electromagnéticas, como las que se utilizan a través de interfaces físicas entre dispositivos locales, a través de redes terrestres alámbricas y ópticas y a través de varios enlaces aéreos. Los elementos físicos que transportan dichas ondas, como enlaces por cable o inalámbricos, enlaces ópticos o similares, también pueden considerarse medios que llevan el software. Tal como se usa en el presente documento, a menos que se limite a medios de "almacenamiento" tangibles y no transitorios, términos tales como "medio legible" de computadora o máquina se refieren a cualquier medio que participe en proporcionar instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, tal como un código ejecutable por ordenador, puede adoptar muchas formas, incluyendo, pero no limitado a un medio de almacenamiento tangible, un medio de onda portadora o medio de transmisión físico. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, tales como cualquiera de los dispositivos de almacenamiento en cualquier computadora o similares, tales como los que se pueden usar para implementar las bases de datos, etc. mostrados en los dibujos. Los medios de almacenamiento volátiles incluyen la memoria dinámica, como la memoria principal de dicha plataforma informática. Los medios de transmisión tangibles incluyen cables coaxiales; cable de cobre y fibra óptica, incluidos los cables que componen un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden adoptar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o luminosas, como las generadas durante las comunicaciones de datos por radiofrecuencia (FR) e infrarrojos (IR). Por lo tanto, las formas comunes de medios legibles por computadora 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, papel para tarjetas perforadas cinta, cualquier otro medio de almacenamiento físico con patrones de orificios, una RAM, una ROM, una PROM y EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporta datos o instrucciones, cables o enlaces que transportan dicha onda portadora, o cualquier otro medio desde el cual una computadora pueda leer el código y/o datos de programación. Muchas de estas formas de medios legibles por ordenador pueden estar implicadas en el transporte de 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 un interfaz de usuario (IU) para proporcionar, por ejemplo, uno o más resultados de análisis de muestras. Los ejemplos de IU incluyen, sin limitación, una interfaz gráfica de usuario (IGU) y una interfaz de usuario basada en web.

EJEMPLOS

Ejemplo 1 - Pronóstico y tratamiento de cáncer de próstata

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

ADN libre de células se extrae y se aísla de la sangre usando 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 utilizando un método de secuenciación masivamente paralelo con un secuenciador personal Illumina MiSeq. El 90% del genoma del sujeto está cubierto mediante la secuenciación de ADN libre de células.

Datos de secuencia se ensamblan y se analizan para la variación del número de copia. Las lecturas de secuencia se mapean y comparan con un individuo sano (control). Según el número de lecturas de secuencia, las regiones cromosómicas se dividen en regiones de 50 kb que no se solapan. Las lecturas de secuencia se comparan entre sí y se

determina una relación para cada posición mapeable.

Un modelo Hidden Markov se aplica para convertir números de copias en estados discretos para cada ventana.

Los informes se generan, las posiciones del genoma de mapeo y el número de copias muestran una variación en la figura 4A (para un individuo sano) y la figura 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 monitoreado durante 18 meses. En el mes 18, el perfil de variación del número de copias comienza a aumentar drásticamente, desde una carga tumoral libre de células del 21% al 30%. Se realiza una comparación con los perfiles genéticos de otros sujetos prostáticos. Se determina que este aumento en la variación del número de copias indica que el cáncer de próstata está avanzando del estadio II al estadio III. El régimen de tratamiento original prescrito ya no trata el cáncer. Se prescribe un nuevo tratamiento.

Además, estos informes se presentan y se puede acceder por vía electrónica a través de Internet. El análisis de los datos de la secuencia se produce en un sitio diferente al del sujeto. El informe se genera y se transmite a la ubicación del sujeto. A través de una computadora habilitada para Internet, el sujeto accede a los informes que reflejan su carga tumoral (figura 4C).

Ejemplo 2 - Remisión y recurrencia del cáncer de próstata.

Una muestra de sangre se extrae de un sobreviviente de cáncer de próstata. El sujeto se había sometido previamente a numerosas rondas de quimioterapia y radiación. El sujeto al momento de la prueba no presentaba síntomas o problemas de salud relacionados con el cáncer. Las exploraciones y los análisis estándar revelan que el sujeto no tiene cáncer.

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

Datos de secuencia se ensamblan y se analizan para la variación del número de copia. Las lecturas de secuencia se mapean y comparan con un individuo sano (control). Según el número de lecturas de secuencia, las regiones cromosómicas se dividen en regiones de 40 kb que no se solapan. Las lecturas de secuencia se comparan entre sí y se determina una relación para cada posición mapeable.

Secuencias no únicas con código de barras se contraen en una lectura única para ayudar a normalizar el sesgo de amplificación.

Un modelo Hidden Markov se aplica para convertir números de copias en estados discretos para cada ventana.

Se generan informes, posiciones del genoma de mapeo y la variación del número de copias se muestra en la figura 5A, para un sujeto con cáncer en remisión y la figura 5B para un sujeto con cáncer en la recurrencia.

Este informe en comparación con otros perfiles de sujetos con resultados conocidos indica que al mes 18, se detecta el análisis de mutaciones raras para la variación del número de copias en la carga tumoral libre de célula de 5%. Un oncólogo prescribe el tratamiento nuevamente.

Ejemplo 3 - Cáncer de tiroides y tratamiento

Un sujeto es conocido por tener cáncer de tiroides de Etapa IV y se somete a tratamiento estándar, incluyendo la radiación terapia con 1-131. Las tomografías computarizadas no son concluyentes en cuanto a si la radioterapia está destruyendo masas cancerosas. Se extrae sangre antes y después de la última sesión de radiación.

ADN libre de células se extrae y se aísla de la sangre usando el protocolo del kit Qiagen Qubit. Se agrega una muestra de ADN a granel no específico a las reacciones de preparación de la muestra que aumentan los rendimientos.

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

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

Datos de secuencia se ensamblan y se analizan para la detección de la variación del número de copia. Las lecturas de secuencia se mapean y comparan con un individuo sano (control). Según el número de lecturas de secuencia, según se determina contando las secuencias de códigos de barras, las regiones cromosómicas se dividen en regiones de 50 kb que no se solapan. Las lecturas de secuencia se comparan entre sí y se determina una relación para cada posición mapeable.

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

Se genera un informe, posiciones del genoma de mapeo y variación del número de copia.

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 de la carga tumoral es un aumento de la necrosis del tejido canceroso frente al tejido normal como resultado del tratamiento. Los oncólogos recomiendan que el sujeto continúe con el tratamiento prescrito.

Ejemplo 4 - Sensibilidad de detección de mutaciones raras

Con el fin de determinar los intervalos de detección de mutaciones raras presentes en una población de ADN, se llevan a cabo los experimentos de mezcla. Las secuencias de ADN, algunas que contienen copias de tipo silvestre de los genes TP53, HRAS y MET y algunas que contienen copias con mutaciones raras en los mismos genes, se mezclan en distintas proporciones. Las mezclas de ADN se preparan de manera que las proporciones o porcentajes de ADN mutante a ADN de tipo silvestre oscilen entre el 100% y el 0,01%.

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

Datos de secuencia se ensamblan y se analizan para la detección de mutaciones raras. Las lecturas de secuencia se mapean y comparan con una secuencia de referencia (control). En función del número de lecturas de secuencia, se determina la frecuencia de variación para cada posición mapeable.

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

Un informe se genera, mapeando las posiciones de bases del genoma y detección de porcentaje de la mutación rara sobre la línea de base como se determina por la secuencia de referencia (figura 6A).

Los resultados de varios experimentos de mezcla que van desde 0,1% a 100% están representados en una escala logarítmica gráfica, con el porcentaje medido de ADN con una mutación rara graficada como una función del porcentaje real de ADN con una mutación rara (figura 6B). Están representados los tres genes, TP53, HRAS y MET. Se encuentra una fuerte correlación lineal entre las poblaciones de mutaciones raras medidas y esperadas. Además, con estos experimentos se encuentra un umbral de sensibilidad más bajo de aproximadamente el 0,1% de ADN con una mutación rara en una población de ADN no mutado (figura 6B).

Ejemplo 5 - Detección de mutaciones raras en sujetos con cáncer de próstata

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

Un panel de varios oncogenes y genes supresores de tumor se seleccionan para la amplificación selectiva utilizando un kit TaqMan © PCR (Invitrogen) utilizando cebadores específicos de genes. Las regiones de ADN amplificadas incluyen ADN que contiene genes PIK3CA y TP53.

10 ng de ADN se secuencia utilizando un enfoque de secuenciación masiva en paralelo con un secuenciador personal Illumina MiSeq.

Datos de secuencia se ensamblan y se analizan para la detección de mutaciones raras. Las lecturas de secuencia se mapean y comparan con una secuencia de referencia (control). En función del número de lecturas de secuencia, se determinó la frecuencia de variación para cada posición mapeable.

Un modelo Hidden Markov se aplica a la frecuencia de conversión de varianza para cada posición mapeable en estados discretos para cada posición de base.

Un informe se genera, mapeando las posiciones de bases genómicas y detección de porcentaje de la mutación rara sobre la línea de base como se determina por la secuencia de referencia (figura 7A). Se encuentran mutaciones raras

con una incidencia del 5% en dos genes, PIK3CA y TP53, respectivamente, lo que indica que el sujeto tiene un cáncer en etapa temprana. Se inicia el tratamiento.

Además, estos informes se presentan y se puede acceder por vía electrónica a través de Internet. El análisis de los datos de la secuencia se produce en un sitio diferente al del sujeto. El informe se genera y se transmite a la ubicación del sujeto. A través de una computadora con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral (figura 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 el ADN libre de células.

Se utilizan 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 etiquetado incluyó la unión de etiquetas necesarias para la secuenciación, así como identificadores no únicos para rastrear las moléculas de la progenie hasta los ácidos nucleicos parentales. La conversión se realiza mediante una reacción de ligación optimizada como se describe anteriormente y el rendimiento de conversión se confirma al observar el perfil de tamaño de las moléculas después de la ligación. El rendimiento de conversión se mide como el porcentaje de moléculas iniciales iniciales que tienen ambos extremos ligados con etiquetas. La conversión que usa este enfoque se realiza con alta eficiencia, por ejemplo, en al menos 50%.

La biblioteca etiquetada se amplificó por PCR y se enriqueció para los genes más asociados con el cáncer colorrectal, (por ejemplo, KRAS, APC, TP53, etc) y el ADN resultante se secuenció utilizando un enfoque de secuenciación masiva en paralelo con un secuenciador personal Illumina MiSeq.

Datos de secuencia se ensamblan y se analizan para la detección de mutaciones raras. Las lecturas de secuencia se colapsan en grupos familiares que pertenecen a una molécula madre (así como se corrigen los errores al colapsar) y se mapean usando una secuencia de referencia (control). Basándose en el número de lecturas de secuencia, se determina la frecuencia de variaciones raras (sustituciones, inserciones, deleciones, etc.) y variaciones en el número de copias y heterocigosidad (cuando sea apropiado) para cada posición mapeable.

Un informe se genera, mapeando las posiciones de bases genómicas y detección de porcentaje de mutaciones raras sobre la línea de base como se determina por la secuencia de referencia. Se encuentran mutaciones raras con una incidencia de 0,3-0,4% en dos genes, KRAS y FBXW7, respectivamente, lo que indica que el sujeto tiene cáncer residual. Se inicia el tratamiento.

Además, estos informes se presentan y se puede acceder por vía electrónica a través de Internet. El análisis de los datos de la secuencia se produce en un sitio diferente al del sujeto. El informe se genera y se transmite a la ubicación del sujeto. A través de una computadora 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 generados por tumor son normalmente tan bajas que tecnologías de secuenciación de nueva generación actuales sólo pueden detectar tales señales de forma esporádica o en pacientes con carga tumoral terminalmente alta. La razón principal es que estas tecnologías están plagadas de tasas de error y sesgos que pueden ser órdenes de magnitud más altas de 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, Tecnología de secuenciación digital (DST), que aumenta la sensibilidad y la especificidad de la detección y cuantificación de ácidos nucleicos derivados de tumores raros entre los fragmentos de la línea germinal en al menos 1-2 órdenes de magnitud.

La arquitectura DST está inspirada en los sistemas de comunicación digital de última generación que combaten el alto ruido y la distorsión causados 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 actuales de próxima generación están plagados de un ruido y una distorsión extremadamente altos (debido a la preparación de muestras, la amplificación y secuenciación basada en PCR). La secuenciación digital puede eliminar el error y la distorsión creados por estos procesos y producir una representación casi perfecta de todas las variantes raras (incluidas las CNV).

Preparación de biblioteca de alta diversidad

A diferencia de protocolos de preparación de bibliotecas de secuenciación convencionales, por los que la mayoría de fragmentos de ADN circulante extraídos se pierden debido a la conversión biblioteca ineficiente, nuestro flujo de trabajo de Tecnología de Secuenciación Digital permite que la gran mayoría de moléculas de partida sea convertida y se

secuenciada. Esto es de vital importancia para la detección de variantes raras, ya que puede haber solo un puñado de moléculas mutadas somáticamente en un tubo de sangre completo de 10 ml. El eficiente proceso de conversión de biología molecular desarrollado permite la mayor sensibilidad posible para la detección de variantes raras.

Panel oncogénico completo y accionable

El flujo de trabajo diseñado alrededor de la plataforma DST es flexible y altamente sintonizable ya que las regiones diana pueden ser tan pequeñas como exones individuales o tan anchas como exomas completos (o incluso genomas completos). Un panel estándar consta de todas las bases exónicas de 15 genes relacionados con el cáncer procesables y la cobertura de los exones "calientes" de 36 genes supresores de onco/tumor adicionales (p. ej., Exones que contienen al menos una o más mutaciones somáticas informadas en COSMIC).

Ejemplo 8: Estudios analíticos

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

Estudios preclínicos

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

Estudios clínicos piloto

Correlación de la biopsia del tumor vs. mutaciones somáticas de ADN circulante

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

Tabla 1

ID del paciente	Etapa	Genes mutantes en un tumor emparejado	Porcentaje de cfADN mutante
CRC Nº 1	II-B	TP53	0,2%
CRC Nº 2	II-C	KRAS SMAD4	0,6% 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	III-C	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 Nº 1	IV	BRAF	0,2%
Melanoma Nº 2	IV	APC	0,3%
		EGFR	0,9%
		MI C	10,5%
Melanoma Nº 3	IV	BRAF	3,3%
Melanoma Nº 4	IV	BRAF	0,7%

REIVINDICACIONES

1. Un método que comprende:

a) proporcionar material genético inicial de partida que comprenda polinucleótidos libres de células;
b) convertir los polinucleótidos del material genético inicial de partida en polinucleótidos parentales marcados ligando oligonucleótidos que comprenden códigos de barras no únicos a los polinucleótidos libres de células, de tal manera que la combinación del código de barras y la secuencia del polinucleótido libre de células cree una secuencia única que pueda rastrearse individualmente, en donde las condiciones de ligación comprenden el uso de más de un exceso molar de 80X de oligonucleótidos de código de barras en comparación con los polinucleótidos libres de células;
c) amplificar los polinucleótidos parentales marcados para producir polinucleótidos de progenie amplificados;
d) secuenciar un subconjunto de los polinucleótidos de progenie amplificados para producir lecturas de secuencia;
e) agrupar las lecturas de secuencia en familias, cada familia generada a partir de un único polinucleótido parental marcado del paso (b); y
f) producir una representación de la información en los polinucleótidos parentales marcados y/o el material genético inicial de partida con ruido y/o distorsión reducidos en comparación con las lecturas de secuencia dentro de una familia.

2. El método de la reivindicación 1, en donde el paso f) comprende inferir la frecuencia de una base o secuencia en un locus particular en el material genético inicial de partida basándose en el número de familias únicas en las que se agrupan las lecturas de secuencia y el número de lecturas de secuencia en cada familia.

3. El método de la reivindicación 2, en donde se asigna una puntuación de confianza a cada llamada de base o secuencia en una familia de lecturas de secuencia analizando las llamadas de base en el locus en una familia de lecturas de secuencia.

4. El método de la reivindicación 3, en donde la frecuencia de cada base o secuencia en el locus se determina teniendo en cuenta la puntuación de confianza para cada llamada de base en una pluralidad de las familias.

5. El método de cualquiera de las reivindicaciones anteriores, en donde los oligonucleótidos del código de barras se ligan aleatoriamente a ambos extremos de los polinucleótidos libres de células.

6. El método de cualquiera de las reivindicaciones precedentes, en donde la detección de los códigos de barras no únicos en combinación con los datos de secuencia de las partes inicial (inicio) y final (parada) de las lecturas de secuencia permite la asignación de una identidad única a una molécula concreta.

7. El método de cualquiera de las reivindicaciones anteriores, que comprende además inferir el número de polinucleótidos parentales únicos en el material genético inicial de partida basándose en el número de familias únicas en las que pueden agruparse las lecturas de secuencia y el número de lecturas de secuencia en cada familia.

8. El método de cualquiera de las reivindicaciones anteriores, en donde los códigos de barras comprenden oligonucleótidos de por lo menos 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 o 50 pares de bases de longitud.

9. El método de cualquiera de las reivindicaciones anteriores, en donde los polinucleótidos libres de células son ADN libre de células, opcionalmente en donde el ADN libre de células se extrae y aísla de la sangre.

10. El método de cualquiera de las reivindicaciones anteriores, que comprende además generar secuencias de consenso a partir de las familias de lecturas de secuencias.

11. El método de la reivindicación 10, en donde la generación de las secuencias de consenso comprende métodos lineales o no lineales de construcción de secuencias de consenso, como métodos de votación, promedio, estadísticos, de detección máxima a posteriori o de máxima verosimilitud, de programación dinámica, bayesianos, de Markov oculto o de máquina de vectores de soporte.

12. El método de cualquiera de las reivindicaciones anteriores, en donde el material genético inicial de partida comprende (i) no más de 100 ng de polinucleótidos; o (ii) de 10 ng a cualquiera de 100 ng, 1 µg o 10 µg de ADN libre de células.

13. El método de cualquiera de las reivindicaciones anteriores, que comprende además enriquecer selectivamente regiones del genoma o transcriptoma de un sujeto antes de la secuenciación.

14. El uso del método de cualquiera de las reivindicaciones anteriores para la identificación, detección, diagnóstico, estadificación o predicción del riesgo de cáncer; o para evaluar la respuesta del sujeto a diferentes tratamientos contra el cáncer; o proporcionar información relativa a la progresión y el pronóstico del cáncer.

15. El uso del método de cualquiera de las reivindicaciones 1 a 13:

a. para construir un perfil genético del sujeto, del que procede el fluido corporal, a lo largo de una enfermedad; o
b. para generar un perfil, huella dactilar o conjunto de datos que sea una suma de la información genética información derivada de diferentes células en una enfermedad heterogénea del sujeto del que deriva el fluido corporal.

16. El uso de acuerdo con la reivindicación 15, en donde el perfil permite al sujeto o a un profesional adaptar las opciones de tratamiento de acuerdo con el progreso de la enfermedad.

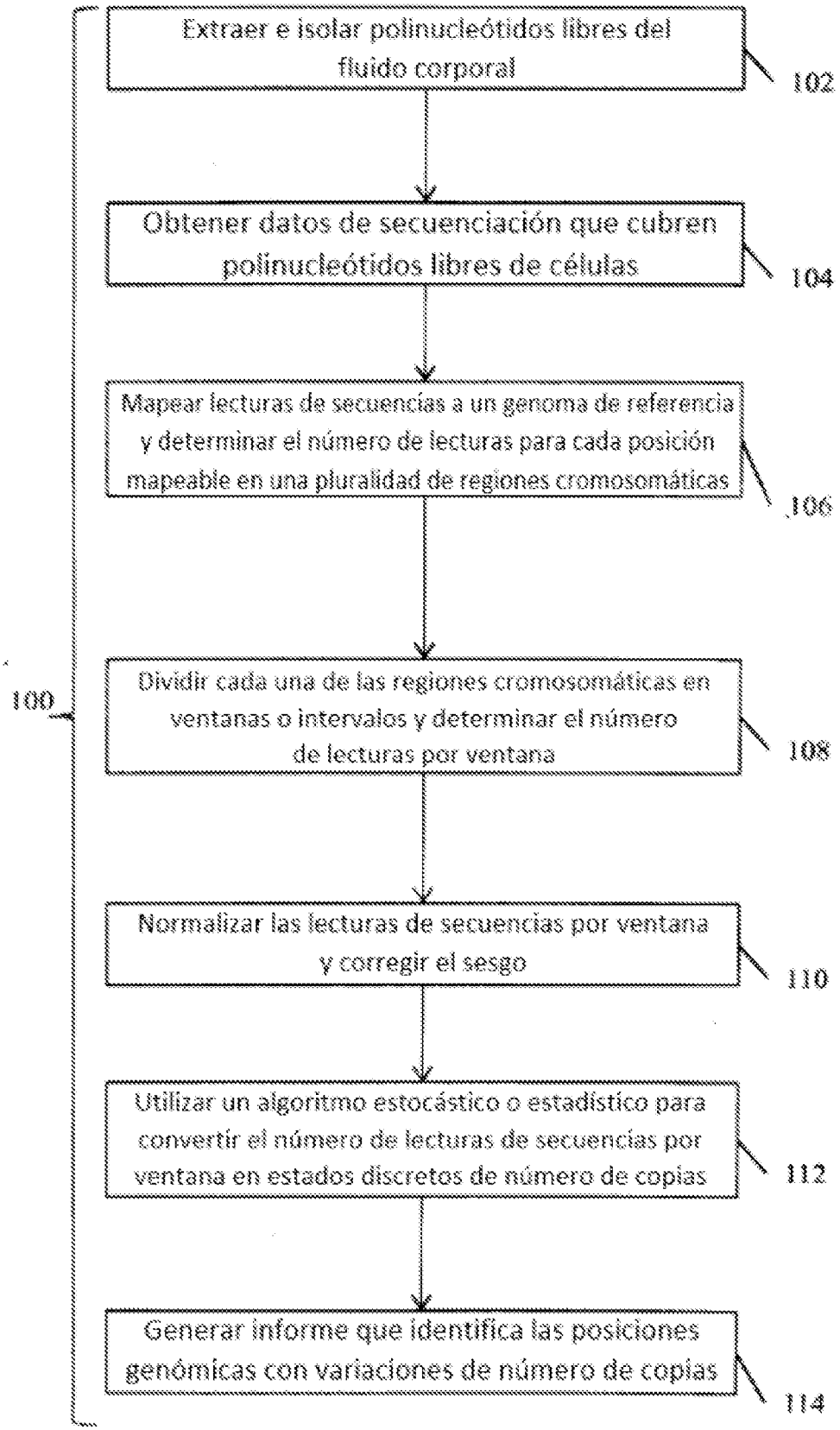

Extraer e isolar polinucleótidos libres del fluido corporal
102
Obtener datos de secuenciación que cubren polinucleótidos libres de células
104
Mapear lecturas de secuencias a un genoma de referencia y determinar el número de lecturas para cada posición mapeable en una pluralidad de regiones cromosomáticas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

Fig. 1

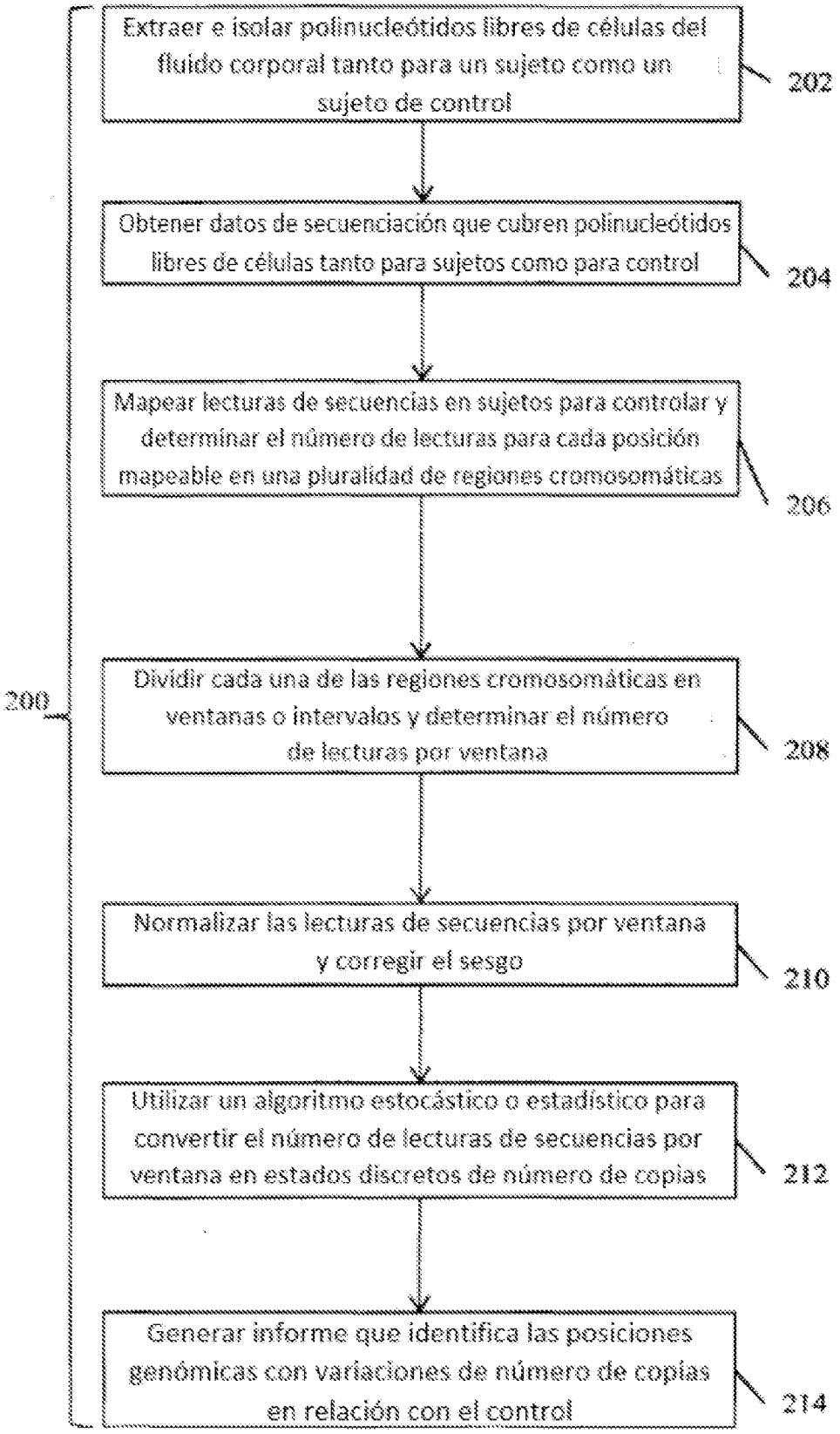

Extraer e isolar polinucleótidos libres de células del fluido corporal tanto para un sujeto como un sujeto de control
202
Obtener datos de secuenciación que cubren polinucleótidos libres de células tanto para sujetos como para control
204
Mapear lecturas de secuencias en sujetos para controlar y determinar el número de lecturas para cada posición mapeable en una pluralidad de regiones cromosomáticas
206
200
Dividir cada una de las regiones cromosomáticas en ventanas o intervalos y determinar el número de lecturas por ventana
208
Normalizar las lecturas de secuencias por ventana y corregir el sesgo
210
Utilizar un algoritmo estocástico o estadístico para convertir el número de lecturas de secuencias por ventana en estados discretos de número de copias
212
Generar informe que identifica las posiciones genómicas con variaciones de número de copias en relación con el control
214

Fig. 2

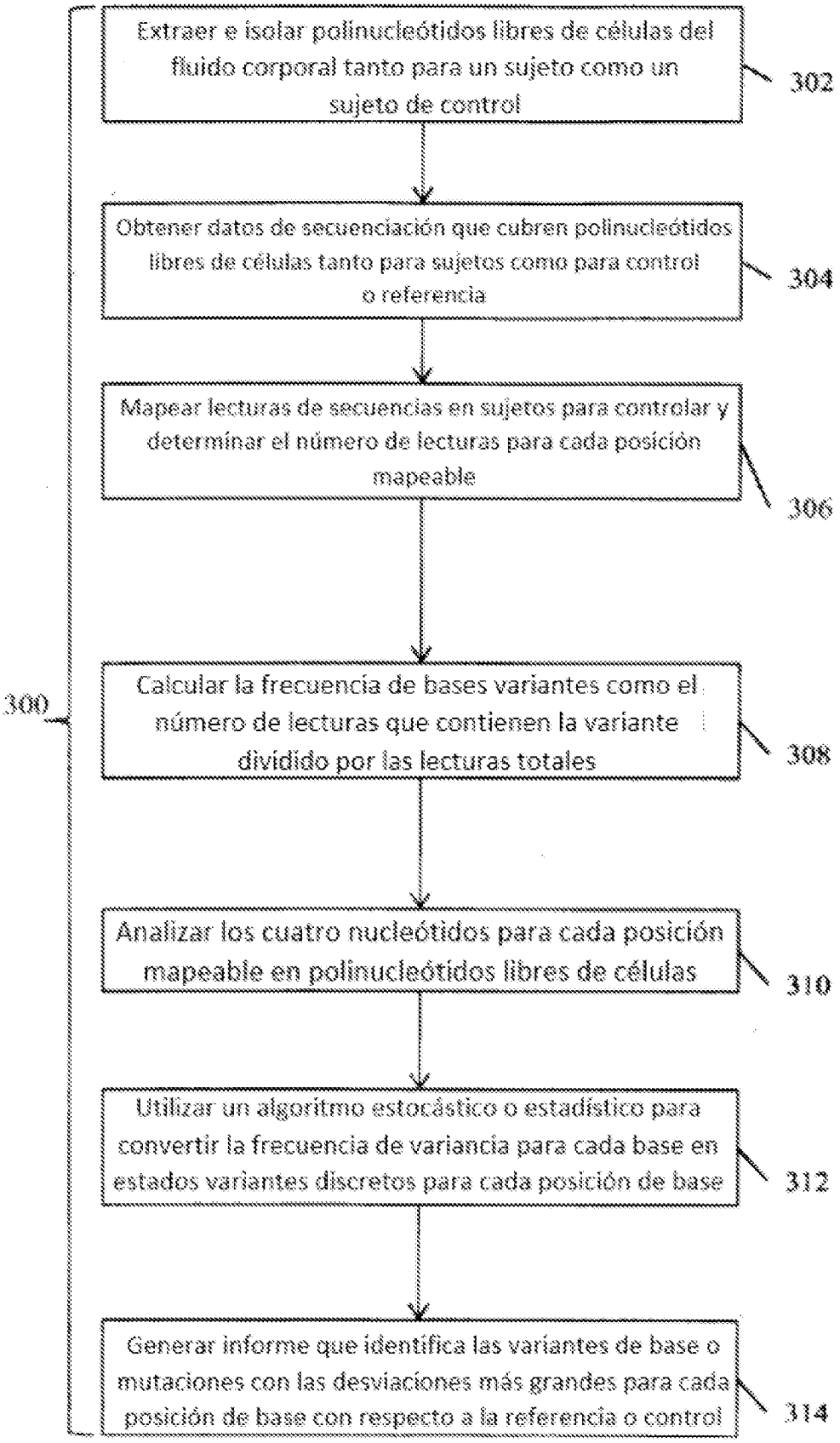
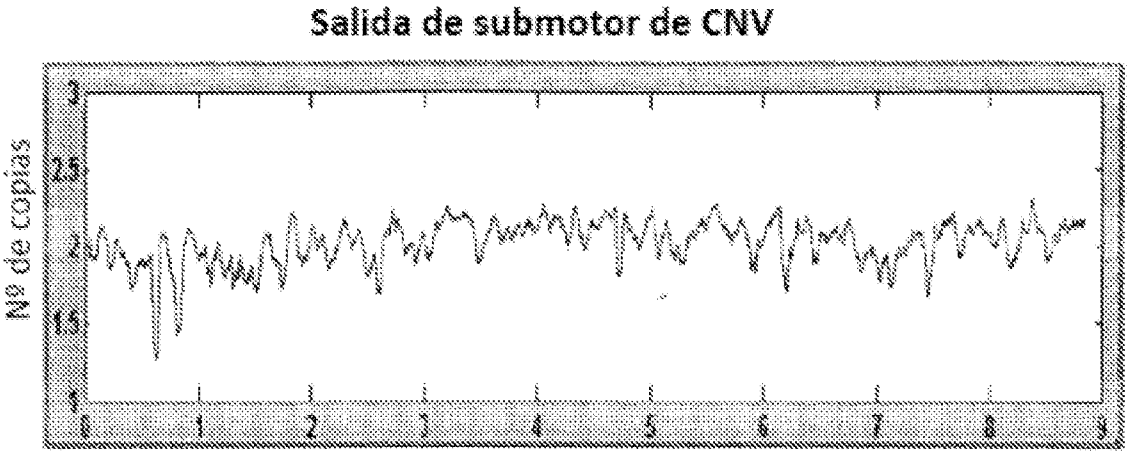
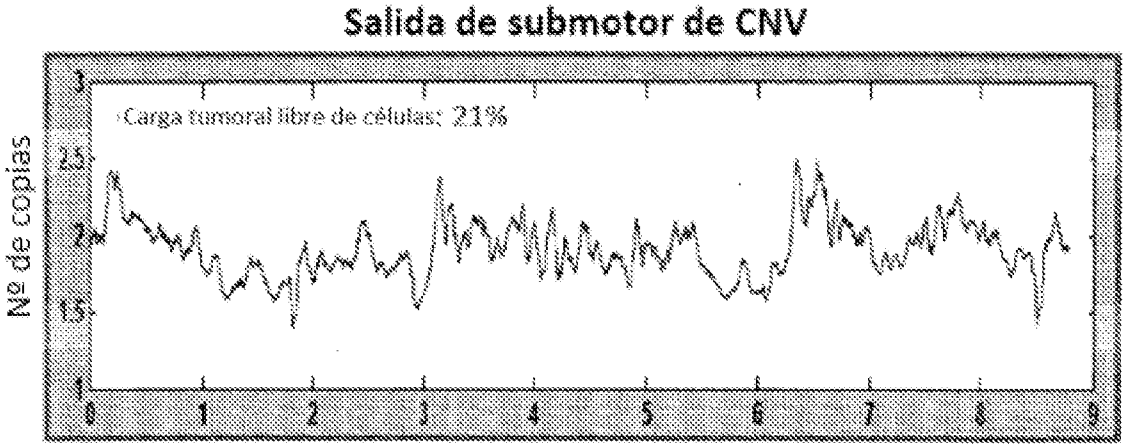

Extraer e isolar polinucleótidos libres de células del fluido corporal tanto para un sujeto como un sujeto de control
302
Obtener datos de secuenciación que cubren polinucleótidos libres de células tanto para sujetos como para control o referencia
304
Mapear lecturas de secuencias en sujetos para controlar y determinar el número de lecturas para cada posición mapeable
306
Calcular la frecuencia de bases 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 polinucleótidos libres de células
310
Utilizar un algoritmo estocástico o estadístico para convertir la frecuencia de variancia para cada base en estados variantes discretos para cada posición de base
312
Generar informe que identifica las variantes de base o mutaciones con las desviaciones más grandes para cada posición de base con respecto a la referencia o control
314
300

Fig. 3



Muestra Normal 1

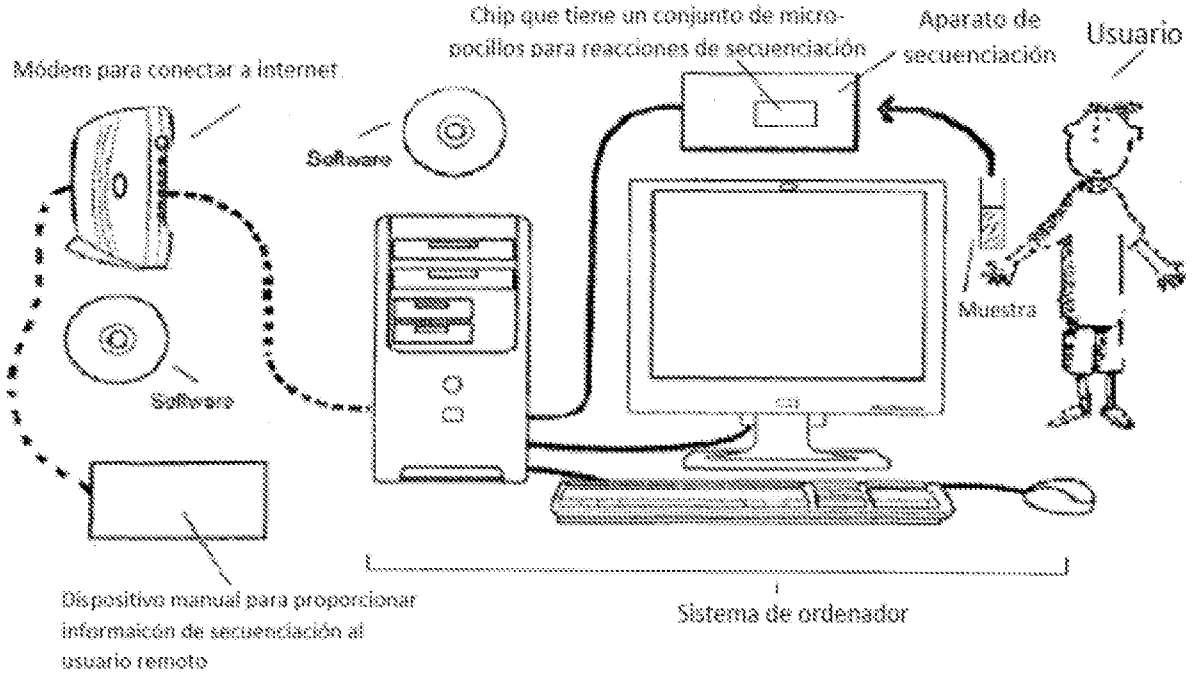
A



Paciente con cáncer de próstata 1

B

Fig. 4


Chip que tiene un conjunto de micro-pocillos para reacciones de secuenciación
Aparato de secuenciación
Usuario
Módem para conectar a Internet
Software
Muestra
Software
Dispositivo manual para proporcionar información de secuenciación al usuario remoto
Sistema de ordenador

C

Fig. 4

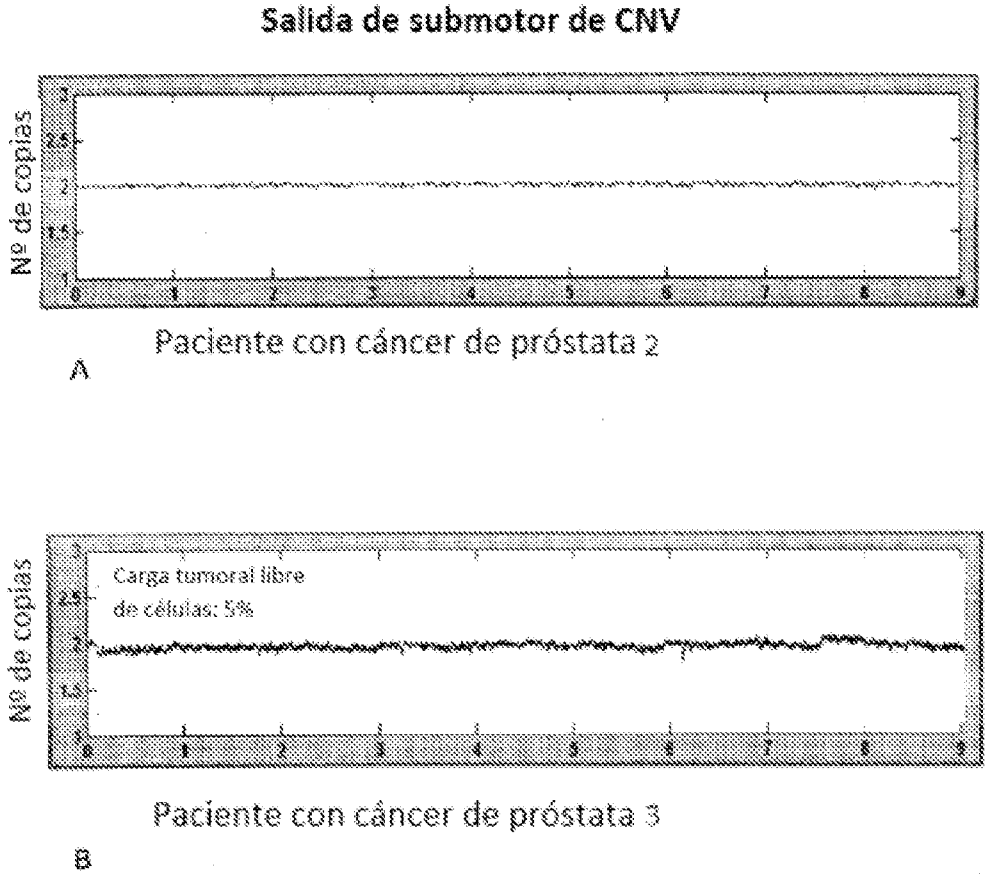
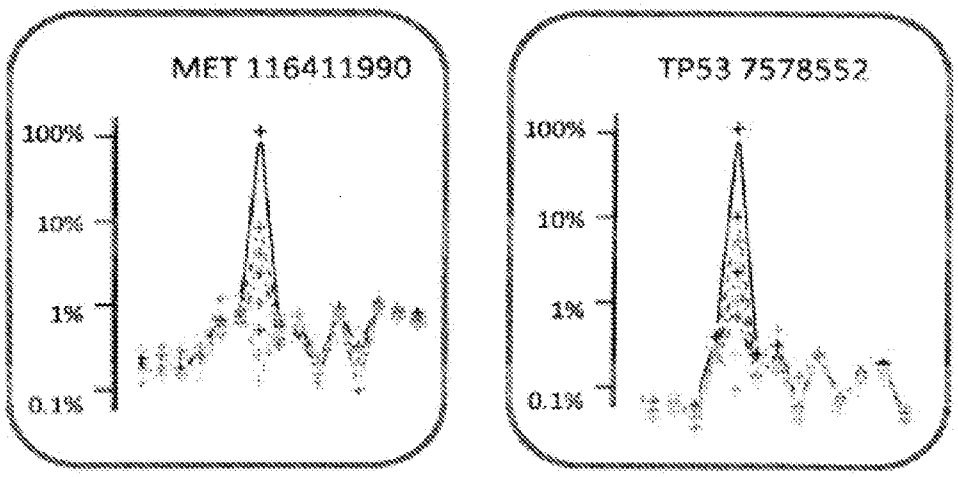
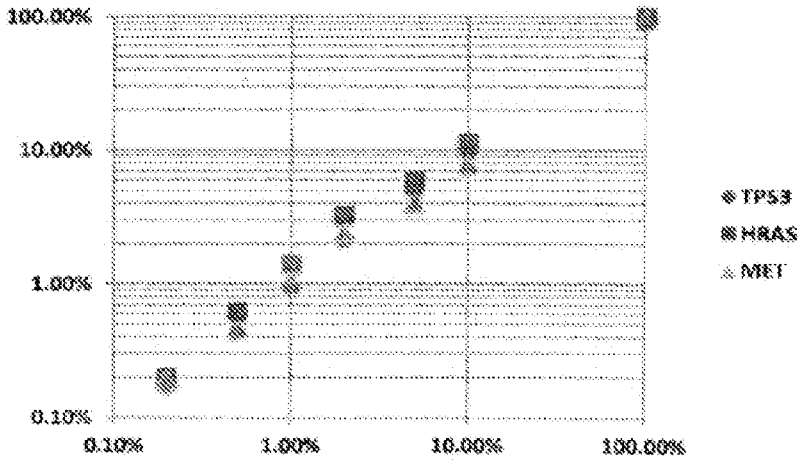

Salida de submotor de CNV
Nº de copias
Paciente con cáncer de próstata 2
A
Nº de copias
Carga tumoral libre
de células: 5%
Paciente con cáncer de próstata 3
B

Fig. 5

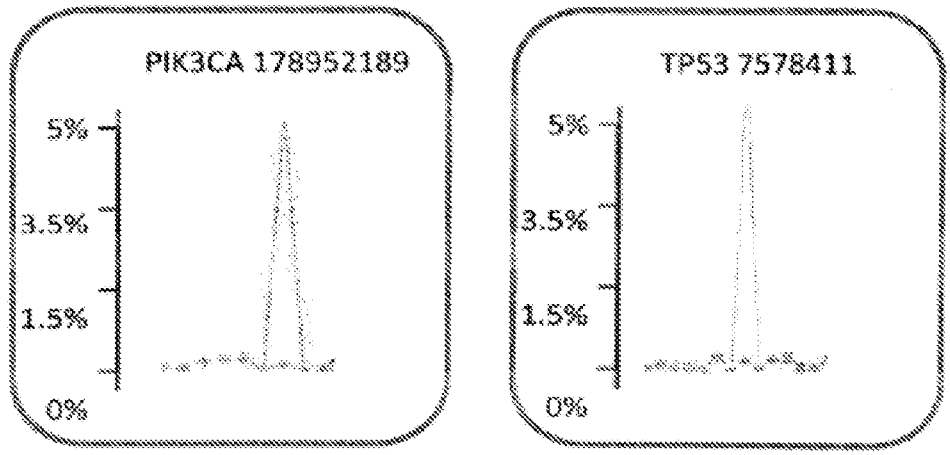

MET 116411990
100%
10%
1%
0.1%
TP53 7578552
100%
10%
1%
0.1%

A

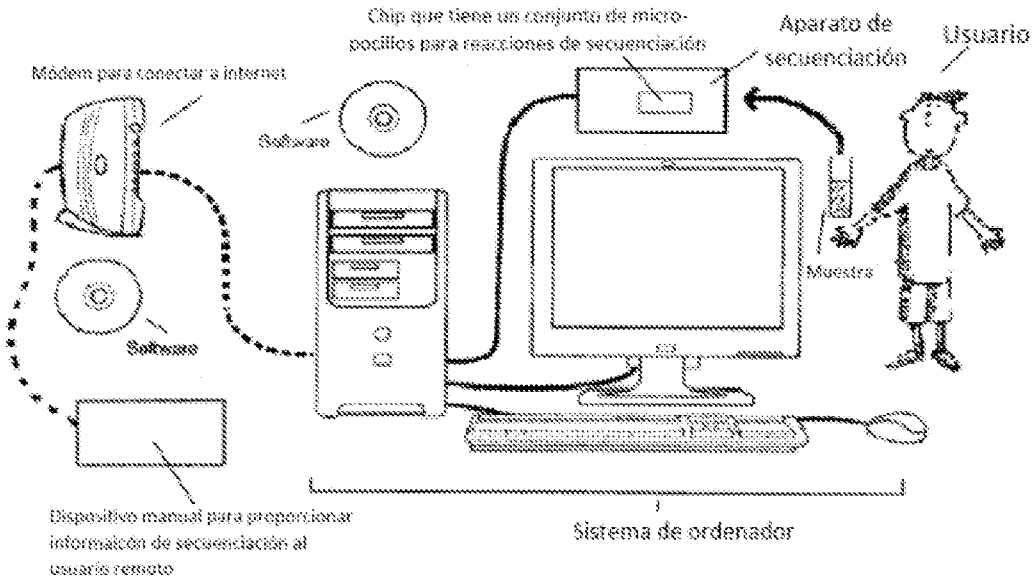

100.00%
10.00%
1.00%
0.10%
0.10%
1.00%
10.00%
100.00%
TP53
HRAS
MET

B

Fig. 6


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

A


Chip que tiene un conjunto de micro-pocillos para reacciones de secuenciación
Aparato de secuenciación
Usuario
Módem para conectar a Internet
Software
Software
Muestra
Dispositivo manual para proporcionar información de secuenciación al usuario remoto
Sistema de ordenador

B

Fig. 7

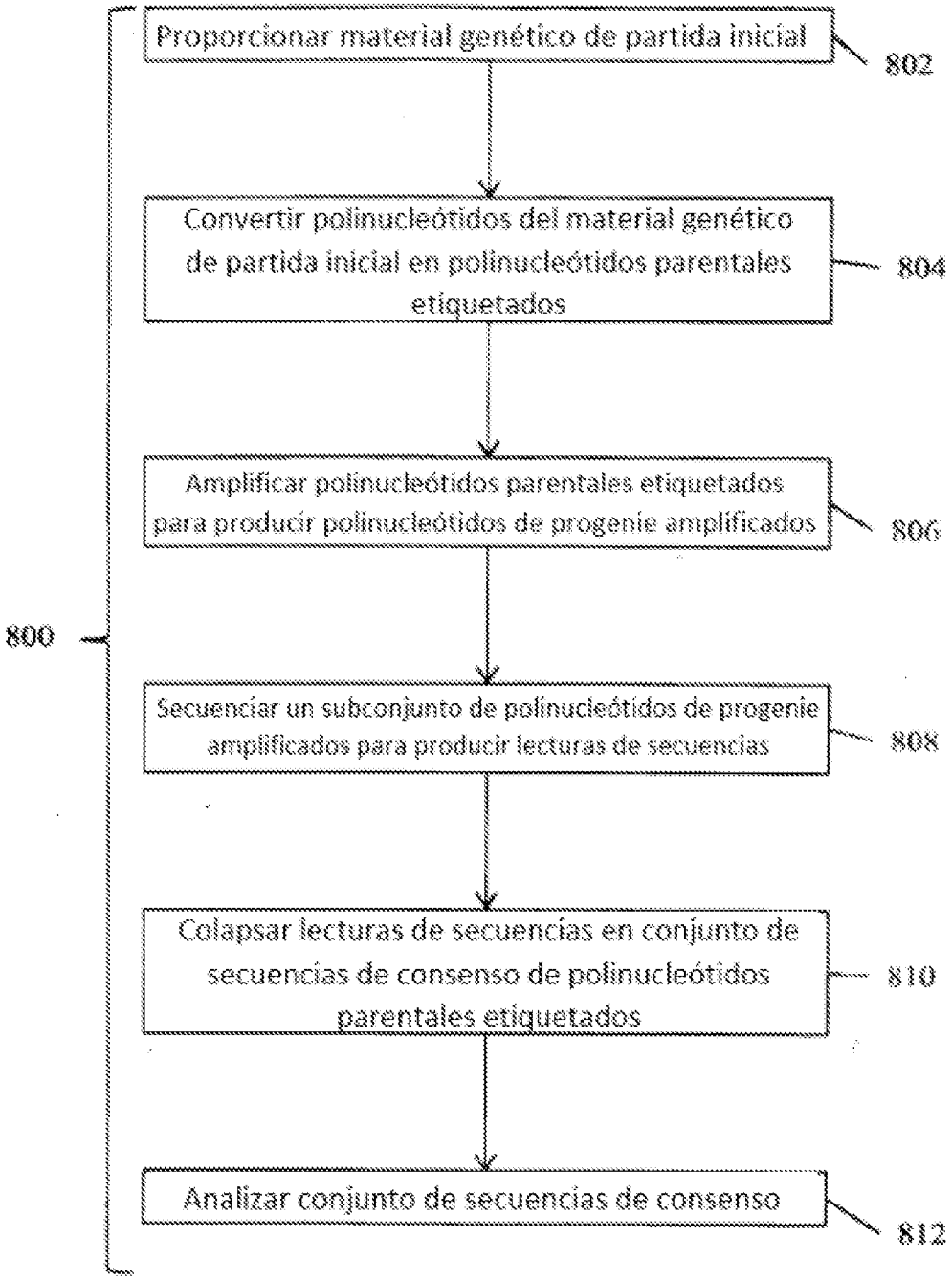

Proporcionar material genético de partida inicial
802
Convertir polinucleótidos del material genético de partida inicial en polinucleótidos parentales etiquetados
804
Amplificar polinucleótidos parentales etiquetados para producir polinucleótidos de progenie amplificados
806
800
Secuenciar un subconjunto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

Fig. 8

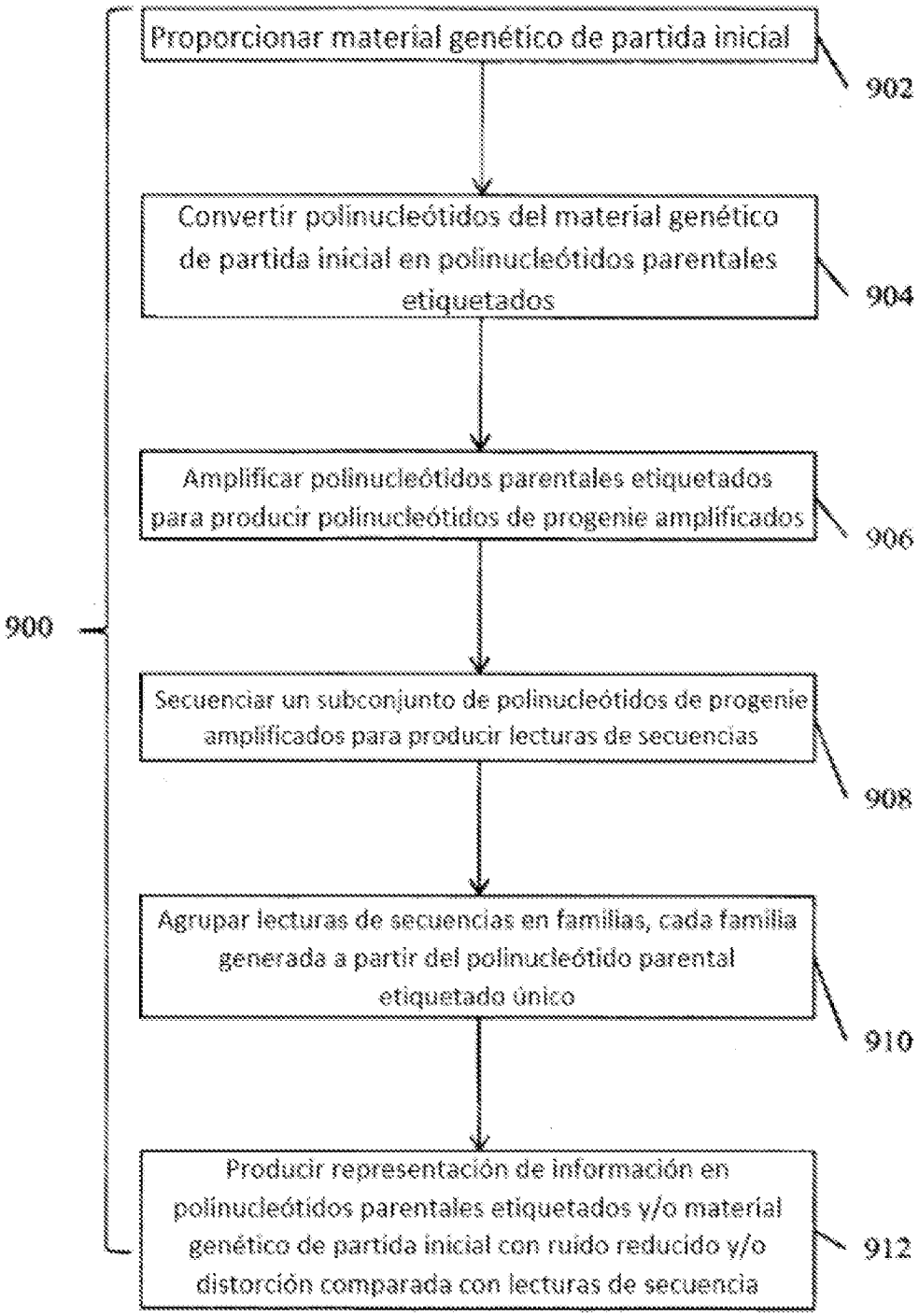

Proporcionar material genético de partida inicial
902
Convertir polinucleótidos del material genético de partida inicial en polinucleótidos parentales etiquetados
904
Amplificar polinucleótidos parentales etiquetados para producir polinucleótidos de progenie amplificados
906
900
Secuenciar un subconjunto de polinucleótidos de progenie amplificados para producir lecturas de secuencias
908
Agrupar lecturas de secuencias en familias, cada familia generada a partir del polinucleótido parental etiquetado único
910
Producir representación de información en polinucleótidos parentales etiquetados y/o material genético de partida inicial con ruido reducido y/o distorción comparada con lecturas de secuencia
912

Fig. 9

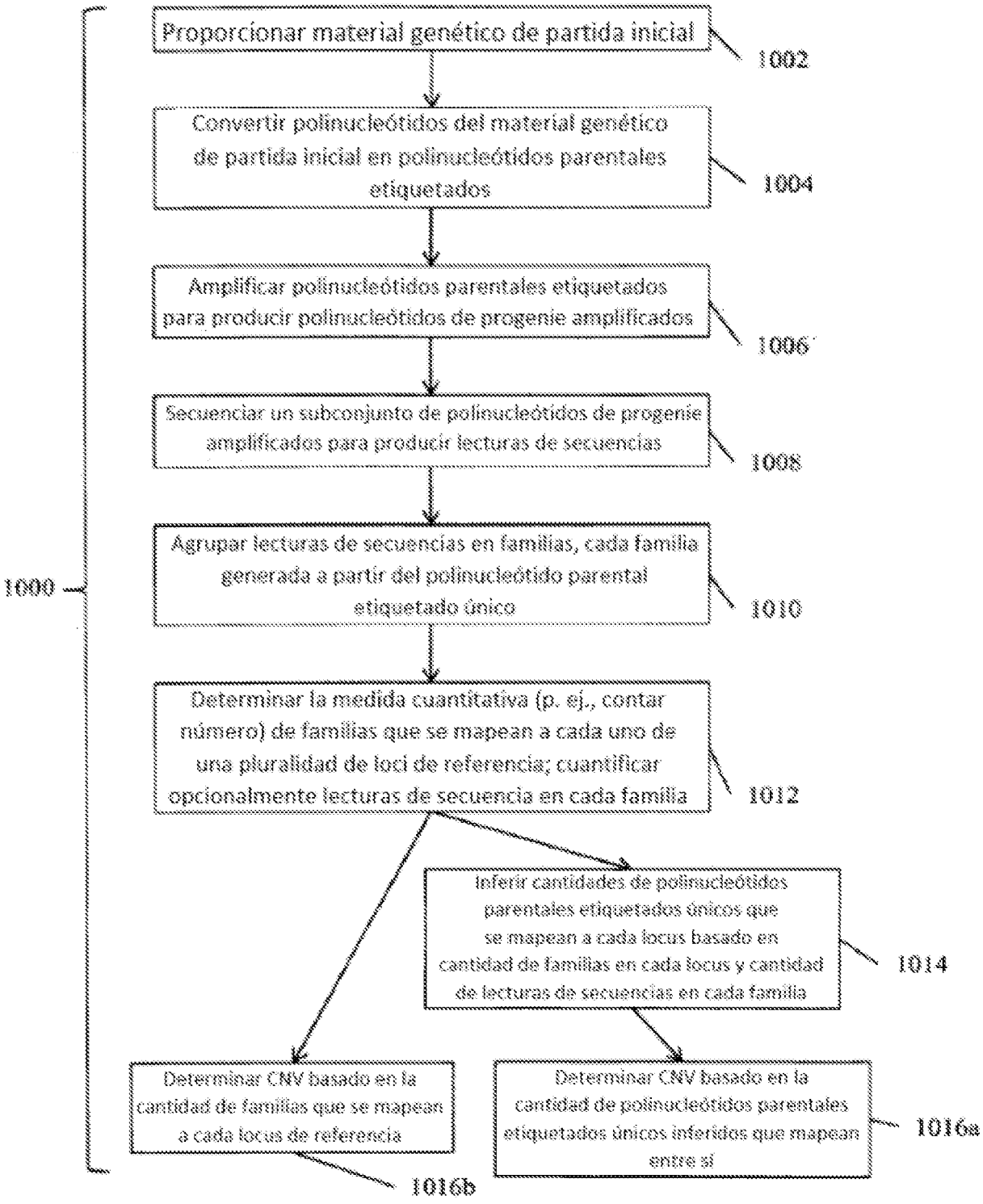

Proporcionar material genético de partida inicial
1002
Convertir polinucleótidos del material genético de partida inicial en polinucleótidos parentales etiquetados
1004
Amplificar polinucleótidos parentales etiquetados para producir polinucleótidos de progenie amplificados
1006
Secuenciar un subconjunto de polinucleótidos de progenie amplificados para producir lecturas de secuencias
1008
Agrupar lecturas de secuencias en familias, cada familia generada a partir del polinucleótido parental etiquetado único
1010
1000
Determinar la medida cuantitativa (p. ej., contar número) de familias que se mapean a cada uno de una pluralidad de loci de referencia; cuantificar opcionalmente lecturas de secuencia en cada familia
1012
Inferir cantidades de polinucleótidos parentales etiquetados únicos que se mapean a cada locus basado en cantidad de familias en cada locus y cantidad de lecturas de secuencias en cada familia
1014
Determinar CNV basado en la cantidad de familias que se mapean a cada locus de referencia
1016b
Determinar CNV basado en la cantidad de polinucleótidos parentales etiquetados únicos inferidos que mapean entre sí
1016a

Fig. 10

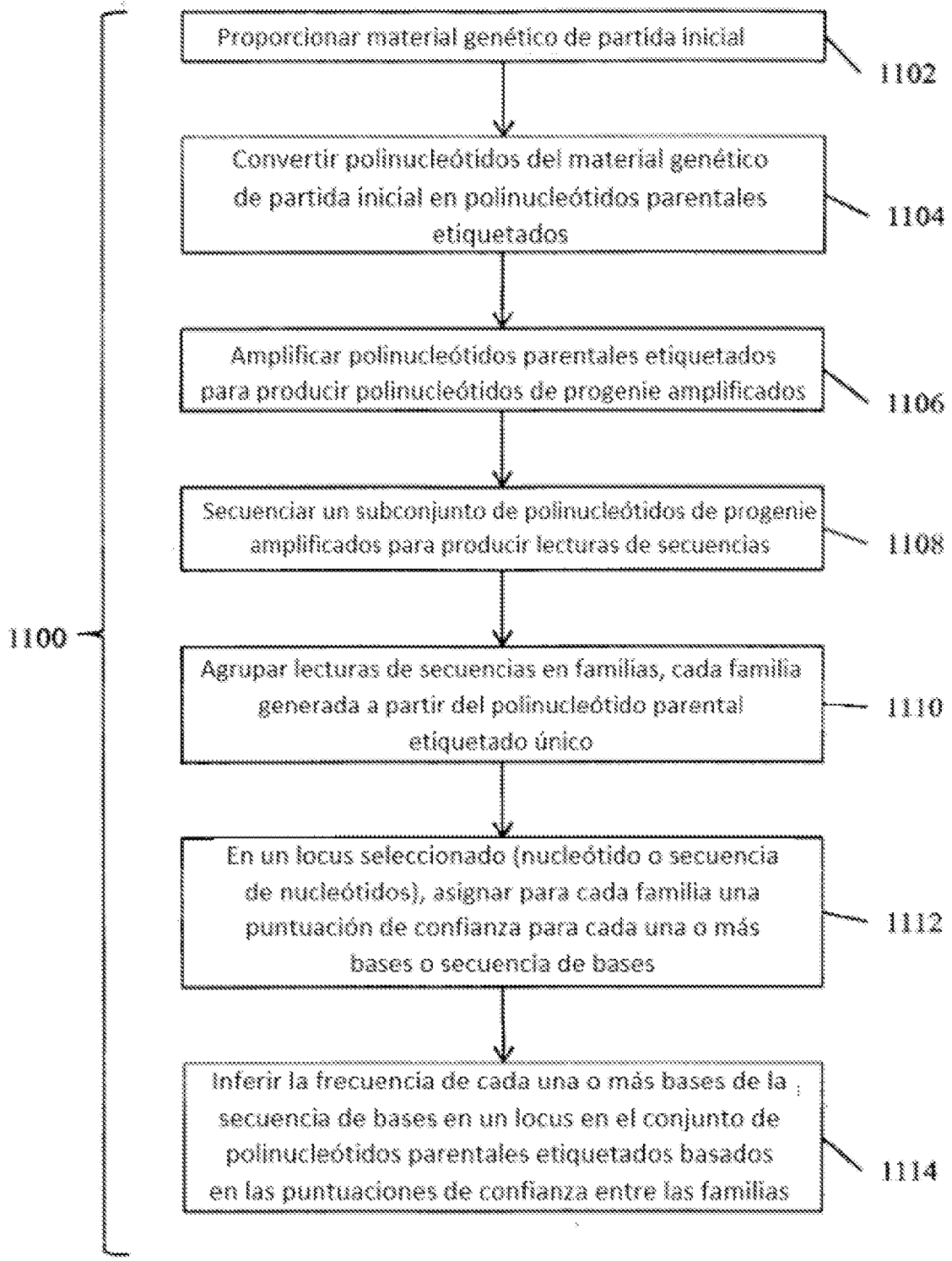

Proporcionar material genético de partida inicial
1102
Convertir polinucleótidos del material genético de partida inicial en polinucleótidos parentales etiquetados
1104
Amplificar polinucleótidos parentales etiquetados para producir polinucleótidos de progenie amplificados
1106
Secuenciar un subconjunto de polinucleótidos de progenie amplificados para producir lecturas de secuencias
1108
1100
Agrupar lecturas de secuencias en familias, cada familia generada a partir del polinucleótido parental etiquetado 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 frecuencia de cada una o más bases de la secuencia de bases en un locus en el conjunto de polinucleótidos parentales etiquetados basados en las puntuaciones de confianza entre las familias
1114

Fig. 11

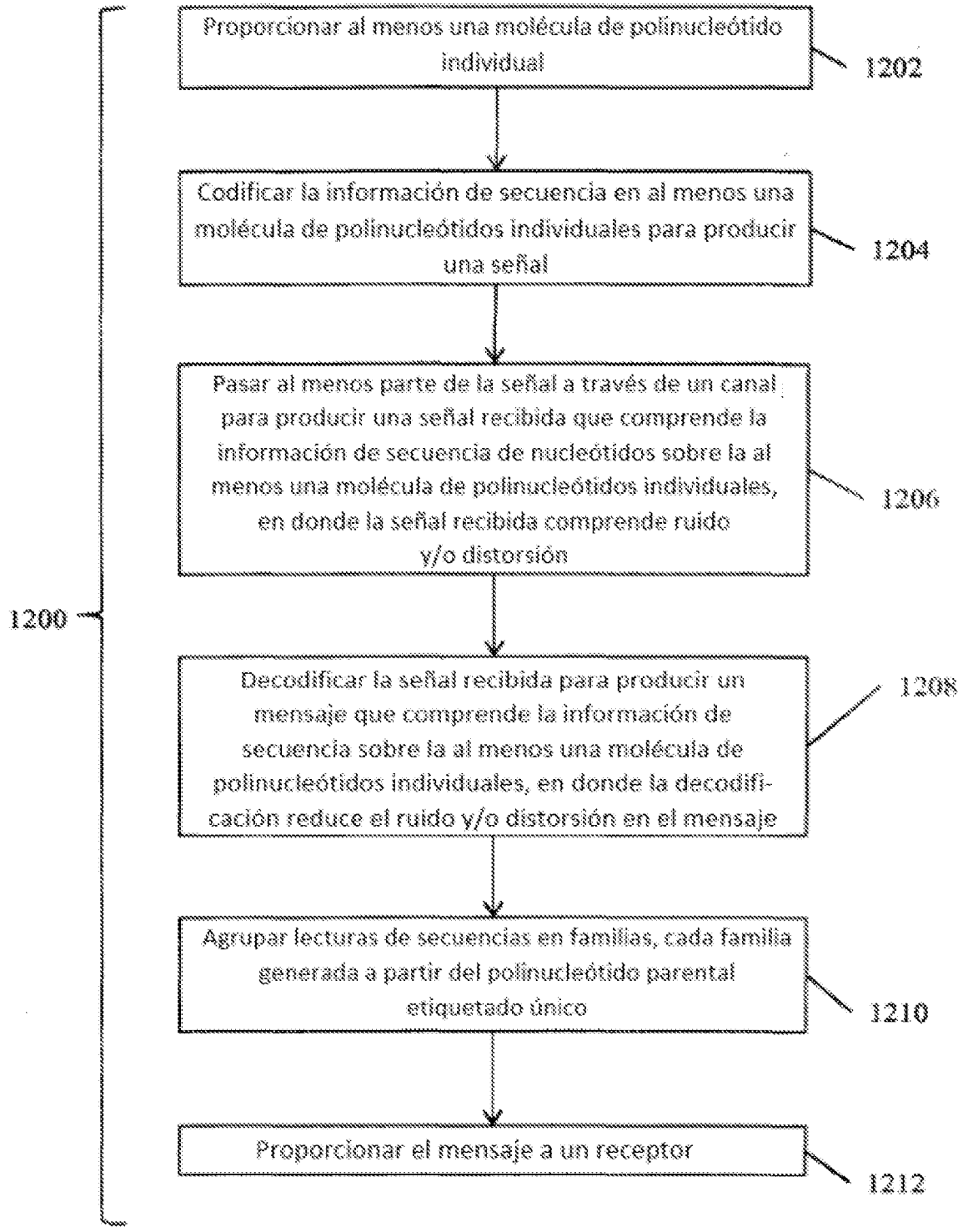

Proporcionar al menos una molécula de polinucleótido individual
1202
Codificar la información de secuencia en al menos una molécula de polinucleótidos individuales para producir una señal
1204
Pasar al menos parte de la señal a través de un canal para producir una señal recibida que comprende la información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

Fig. 12

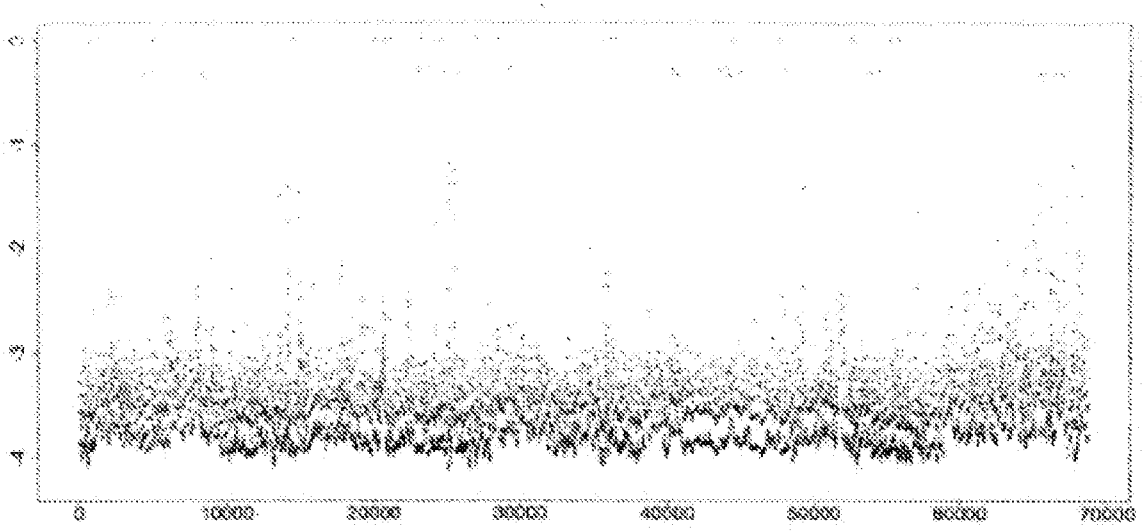


Fig. 13A

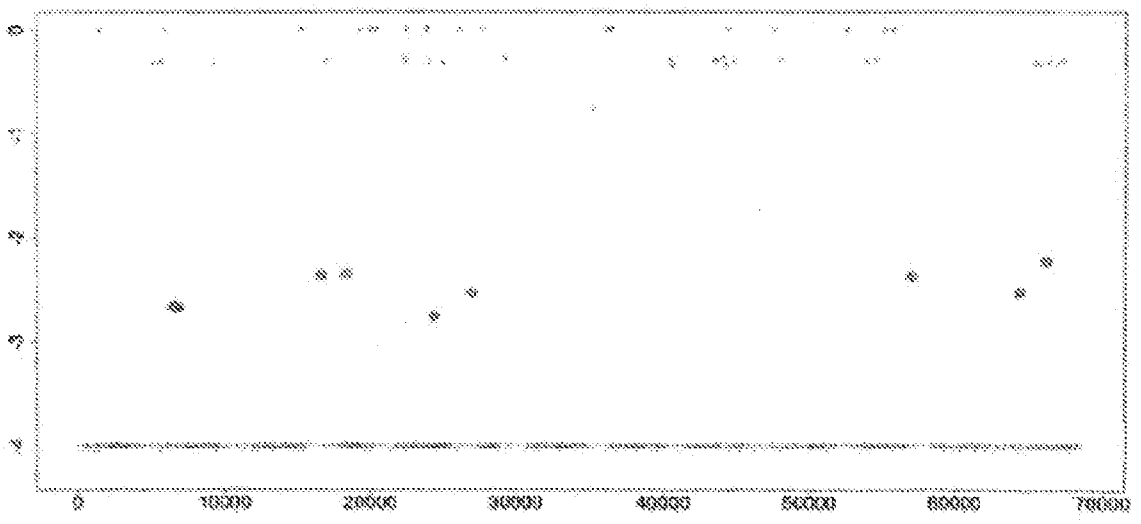


Fig. 13B

Fig. 13

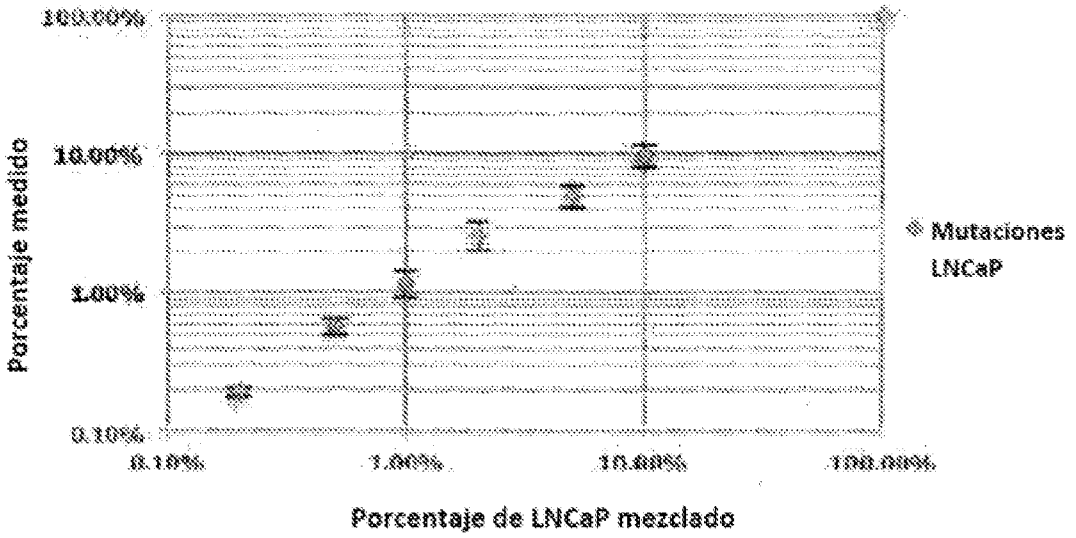

100.00%
10.00%
1.00%
0.10%
Porcentaje medido
0.10%
1.00%
10.00%
100.00%
Porcentaje de LNCaP mezclado
Mutaciones LNCaP

Fig. 14

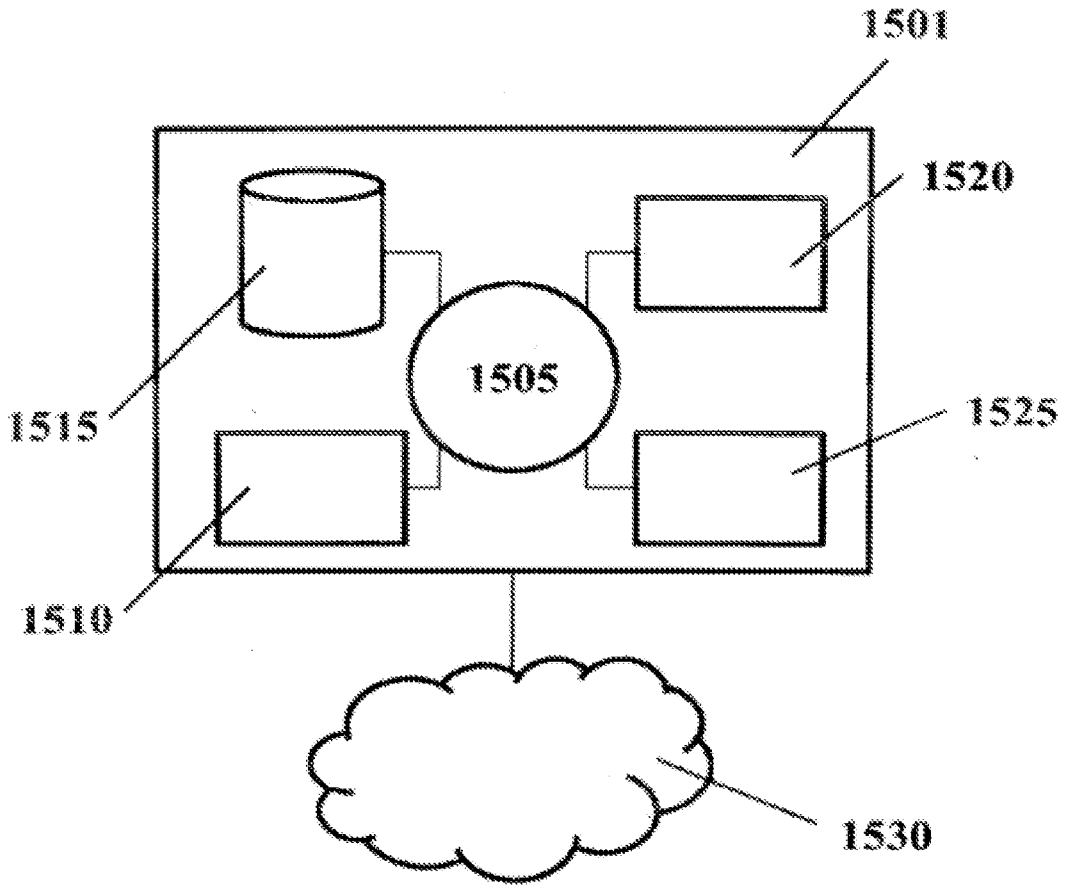


Fig. 15