(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 822 125**

(51) Int. CI.:

C12Q 1/6869 (2008.01)
C12Q 1/6886 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **24.12.2014 E 19163403 (9)**

(97) Fecha y número de publicación de la concesión europea: **15.07.2020 EP 3524694**

(54) Título: **Métodos y sistemas para detectar variantes genéticas**

(30) Prioridad:

28.12.2013 US 201361921456 P
05.03.2014 US 201461948509 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
29.04.2021

(73) Titular/es:

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

(72) Inventor/es:

ELTOUKHY, HELMY y
TALASAZ, AMIRALI

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos y sistemas para detectar variantes genéticas

ANTECEDENTES

La detección y cuantificación de polinucleótidos es importante para aplicaciones de biología molecular y médicas, como diagnóstico. La prueba genética es particularmente útil para una variedad de métodos diagnósticos. Por ejemplo, trastornos que están provocados por alteraciones genéticas raras (por ejemplo, variantes de secuencias) o cambios en etiquetaes epigenéticos, como cáncer y aneuploidía parcial o total, pueden detectarse o caracterizarse con más precisión con información de secuencia de ADN.

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

En particular, se han desarrollado muchos métodos para la estimación precisa de variación del número de copias, especialmente para muestras genómicas heterogéneas, como ADNg derivado de tumores o para ADNcf para muchas aplicaciones (por ejemplo, diagnóstico prenatal, de trasplante, inmune, metagenómico o cáncer). La mayoría de estos métodos incluyen la preparación de muestras mediante la cual los ácidos nucleicos originales se convierten en una biblioteca secuenciable, seguida de una secuenciación masivamente paralela y finalmente bioinformática para estimar la variación del número de copias en uno o más loci.
La WO2013142389 y Schmitt et al. (PNAS 21: 14508-14513,2012) divulgan un método para detectar mutaciones ultrararas mediante secuenciación de próxima generación. La WO2013181170 describe un método para generar una secuencia de consenso de cadena doble de error corregida.

SUMARIO

Aunque muchos de estos métodos son capaces de reducir o combatir los errores introducidos por la preparación de muestras y los procesos de secuenciación para todas las moléculas que se han convertido y secuenciado, estos métodos no son capaces de inferir en los recuentos de moléculas que se convirtieron pero no se secuenciaron. Como este recuento de moléculas convertidas pero no secuenciadas puede ser altamente variable de región a región genómica, estos recuentos pueden afectar de manera dramática y adversa a la sensibilidad que se puede lograr.

Para abordar este problema, puede convertirse el ácido desoxirribonucleico (ADN) de cadena doble de entrada mediante un proceso que etiqueta ambas mitades de la molécula de cadena doble individual, en algunos casos de forma diferente. Esto se puede realizar usando una variedad de técnicas, incluida ligadura de horquillas, burbujas o adaptadores bifurcados u otros adaptadores que tienen segmentos de cadenas doble y cadena sencilla (la porción no hibridada de un adaptador de burbuja, horquilla o bifurcación se considera en la presente de cadena sencilla) . Si se etiquetan correctamente, cada lado original de Watson y Crick (es decir, cadena) de la molécula de ADN de cadena doble de entrada puede etiquetarse e identificarse de manera diferente mediante el secuenciador y bioinformática posterior. Para todas las moléculas en una región particular, se pueden registrar los recuentos de moléculas donde se recuperaron ambos lados de Watson y Crick ("Pares") frente a aquellos en los que solo se recuperó la mitad ("Singletes"). El número de moléculas invisibles puede estimarse en base al número de Pares y Singletes detectados.

Por lo tanto, la invención proporciona un método para determinar una medida cuantitativa indicativa de una serie de moléculas de ácido desoxirribonucleico (ADN) de cadena doble individuales en una muestra que comprende:

(a) determinar una medida cuantitativa de moléculas de ADN individuales para las que se detectan ambas cadenas;
(b) determinar una medida cuantitativa de moléculas de ADN individuales para las que solo se detecta una de las cadenas de ADN;
(c) inferir de (a) y (b) anteriores una medida cuantitativa de moléculas de ADN individuales para las cuales no se detectó ninguna cadena; y
(d) usar (a)-(c) para determinar la medida cuantitativa indicativa de una serie de moléculas de ADN de cadena doble individuales en la muestra;

en donde determinar una medida cuantitativa de moléculas de ADN individuales comprende: (i) etiquetar dichas moléculas de ADN con un conjunto de etiquetas dúplex que etiquetan de manera diferente cadenas complementarias de una molécula de ADN de cadena doble en dicha muestra para proporcionar cadenas etiquetadas; y (ii) secuenciar por lo menos algunas de dichas cadenas etiquetadas para producir un conjunto de lecturas de secuencia.

Puede usarse un conjunto de adaptadores de bibliotecas para etiquetar las moléculas de interés (por ejemplo, por ligación, hibridación, etc.). El conjunto de adaptadores de bibliotecas puede comprender una pluralidad de moléculas de polinucleótidos con códigos de barras moleculares, en donde la pluralidad de moléculas de polinucleótidos son menores que o iguales a 80 bases de nucleótidos de longitud, en donde los códigos de barras moleculares son por lo menos de 4 bases de nucleótidos de longitud, y en donde (a) los códigos de barras moleculares son diferentes entre sí y tienen una distancia de edición de por lo menos 1 entre uno y otro; (b) los códigos de barras moleculares están localizados por lo menos una base de nucleótidos lejos de un extremo terminal de sus moléculas de polinucleótidos respectivas; (c) opcionalmente, por lo menos una base terminal es idéntica en todas las moléculas de polinucleótidos; y (d) ninguna de las moléculas de polinucleótidos contiene un motivo secuenciador completo.

Los adaptadores de bibliotecas pueden ser idéntica entre sí pero distintos para los códigos de barra moleculares. Cada uno de la pluralidad de adaptadores de bibliotecas puede comprender por lo menos una parte de cadena doble y por lo menos una parte de cadena sencilla (por ejemplo, una parte no complementaria o un saliente). La parte de cadena doble puede tener un código de barras molecular seleccionado de una colección de diferentes códigos de barra moleculares diferentes. El código de barras molecular dado puede tener un aleatorizador. Cada uno de los adaptadores de bibliotecas puede comprender además un código de barras de identificación de cadena en por lo menos una parte de cadena sencilla. El código de barras de identificación de cadenas puede incluir por lo menos 4 bases de nucleótidos. La parte de cadena sencilla puede tener un motivo secuenciador parcial. Los adaptadores de bibliotecas pueden no incluir un motivo secuenciador completo.

En algunos casos, ninguno de los adaptadores de bibliotecas contiene una secuencia para hibridar a una célula de flujo o formar una horquilla para secuenciación.

En algunos casos, todos los adaptadores de bibliotecas tienen un extremo termina con nucleótido(s) que son el mismo. El nucleótido(s) terminal idéntico puede ser de dos o más bases de nucleótidos de longitud.

Cada uno de los adaptadores de bibliotecas puede tener forma de Y, forma de burbuja o forma de horquilla. En algunos casos ninguno, de los adaptadores de bibliotecas contiene un motivo de identificación de muestras. Cada uno de los adaptadores de bibliotecas puede comprender una secuencia que puede hibridar selectivamente con un cebador universal. Cada uno de los adaptadores de bibliotecas puede comprender un código de barras molecular que es de por lo menos 5, 6, 7, 8, 9 y 10 bases de nucleótidos de longitud. Cada uno de los adaptadores de bibliotecas puede ser de 10 bases de nucleótidos a 80 de longitud, o de 30 a 70 bases de nucleótidos de longitud o de 40 a 60 bases de nucleótidos de longitud. En algunos casos, por lo menos 1, 2, 3, ó 4 bases terminales son idénticas en todos los adaptadores de bibliotecas. En algunos casos, por lo menos 4 bases terminales con idénticas en todos los adaptadores de bibliotecas.

La distancia de edición de los códigos de barra moleculares de los adaptadores de bibliotecas pueden ser una distancia de Hamming. La distancia de edición puede ser por lo menos 1, 2, 3, 4, ó 5. En algunos casos, la distancia de edición es con respecto a bases individuales de la pluralidad de moléculas de polinucleótidos. Los códigos de barras moleculares pueden estar localizados por lo menos 10 bases de nucleótidos alejados de un extremo terminal de un adaptador. La pluralidad de adaptadores de bibliotecas puede incluir por lo menos 2, 4, 6, 8, 10, 20, 30, 40 ó 50 códigos de barras moleculares diferentes, o de 2-100, 4-80, 6-60, ó 8-40 códigos de barras moleculares diferentes. En cualquiera de los casos de la presente, hay más polinucleótidos (por ejemplo, fragmentos de ADNcf) a ser etiquetados que códigos de barras moleculares diferentes de tal forma que el etiquetado no es único.

En algunos casos, el extremo terminal de un adaptador está configurado para la ligación (por ejemplo, con una molécula de ácido nucleico objetivo). En algunos casos, el extremo terminal de un adaptador es un extremo romo.

En algunos casos, los adaptadores se purifican y aíslan. La biblioteca puede comprender una o más bases de origen no natural.

Las moléculas de polinucleótidos pueden comprender una secuencia de cebadores posicionada 5' con respecto a los códigos de barras moleculares.

El conjunto de adaptadores de bibliotecas puede consistir esencialmente de la pluralidad de moléculas de

polinucleótidos.

También se proporciona un método que comprende (a) proporcionar una muestra que comprende un conjunto de moléculas de polinucleótidos de doble cadena, cada molécula de polinucleótidos de doble cadena incluyendo una primera y una segunda cadenas complementarias; (b) etiquetar dichas moléculas de polinucleótidos de cadena doble con un conjunto de etiquetas dúplex, en el que cada etiqueta dúplex etiqueta de manera diferente la primera y la segunda cadenas complementarias de una molécula de polinucleótidos de doble cadena en el conjunto; (c) secuenciar por lo menos algunas de las cadenas etiquetadas para producir un conjunto de lecturas de secuencia; (d) reducir y/o seguir la redundancia en el conjunto de lecturas de secuencia; (e) clasificar las lecturas de secuencia en lecturas emparejadas y lecturas no emparejadas, en donde (i) cada lectura emparejada corresponde a lecturas de secuencia generadas a partir de una primera cadena etiquetada y una segunda cadena complementaria etiquetada de manera diferente derivada de una molécula de polinucleótidos de cadena doble en dicho conjunto, y (ii) cada lectura no emparejada representa una primera cadena etiquetada que no tiene una segunda cadena complementaria etiquetada de manera diferente derivada de una molécula de polinucleótidos de doble cadena representada entre las lecturas de secuencia en el conjunto de lecturas de secuencia; (f) determinar medidas cuantitativas de (i) las lecturas emparejadas y (ii) las lecturas no emparejadas que mapean para cada uno de los uno o más loci genéticos; y (g) estimar con un procesador informático programado una medida cuantitativa de moléculas de polinucleótidos de doble cadena totales en el conjunto que mapean para cada uno de dichos uno o más loci genéticos en base a la medida cuantitativa de lecturas emparejadas y lecturas no emparejadas que mapean para cada locus.

El método puede comprender además (h) detectar la variación en el número de copias en la muestra determinando una medida cuantitativa total normalizada determinada en el paso (g) en cada uno del uno o más loci genéticos y determinar la variación en el número de copias en base a la medida normalizada. En algunas realizaciones, la muestra comprende moléculas de polinucleótidos de doble cadena procedentes sustancialmente de ácidos nucleicos libres de células. Las etiquetas dúplex no son adaptadores de secuenciación.

Reducir la redundancia en el conjunto de lecturas de secuencias puede comprender colapsar las lecturas de secuencias producidas a partir de productos amplificados de una molécula de polinucleótidos original en la muestra de vuelta a la molécula de polinucleótidos original. El método puede comprender además determinar una secuencia de consenso para la molécula de polinucleótidos original. El método puede comprender además identificar moléculas de polinucleótidos en uno o más loci genéticos que comprende una variante de secuencia. El método puede comprender además determinar una medida cuantitativa de lecturas emparejadas que mapean un locus, en donde ambas cadenas de la pareja comprenden una variante de secuencia. El método puede comprender además determinar una medida cuantitativa de moléculas emparejadas en las que sólo un miembro de la pareja lleva una variante de secuencia y/o determinar una medida cuantitativa de moléculas desemparejadas que llevan una variante de secuencia. La variante de secuencia puede seleccionarse del grupo que consiste de una variante de nucleótidos individual, una indel, una transversión, una translocación, una inversión, una deleción, una alteración estructural cromosómica, una fusión génica, una fusión de cromosomas, un truncamiento de genes, una amplificación de genes, una duplicación de genes y una lesión cromosómica.

La invención proporciona un método para determinar una medida cuantitativa indicativa de un número de fragmentos de ADN de cadena doble individuales en una muestra que comprende (a) determinar una medida cuantitativa de moléculas de ADN individuales para las que se detectan ambas cadenas; (b) determina una medida cuantitativa de moléculas de ADN individuales para las que sólo se detecta una de las cadenas de ADN; (c) inferir de (a) y (b) anteriores una medida cuantitativa de moléculas de ADN individuales para los que no se detectó ninguna cadena; y (d) usar (a)-(c) para determinar la medida cuantitativa indicativa de un número de fragmentos de ADN de cadena doble en la muestra, en donde determinar una medida cuantitativa de moléculas de ADN de cadena doble individuales comprende: (i) etiquetar dichas moléculas de ADN con un conjunto de etiquetas dúplex que etiquetan de manera diferente cadenas complementarias de una molécula de ADN de cadena doble en dicha muestra para proporcionar cadenas etiquetadas; y (ii) secuenciar por lo menos algunas de dichas cadenas etiquetadas para producir un conjunto de lecturas de secuencia.

En algunas realizaciones, el método comprende además detectar la variación del número de copias en la muestra determinando una medida cuantitativa normalizada determinada en el paso (d) en cada uno de los uno o más loci genéticos y determinar la variación del número de copias en base a la medida normalizada. La muestra puede comprender moléculas de polinucleótidos de cadena doble originarias sustancialmente de ácidos nucleicos libres de células.

En algunas realizaciones, el método comprende además clasificar las lecturas de secuencias en lecturas emparejadas y lecturas desemparejadas, en donde (i) cada lectura emparejada corresponde a las lecturas de secuencias generadas de una primera cadena etiquetada y una cadena complementaria etiquetada de manera diferente de una molécula de polinucleótidos de cadena doble en el conjunto, y (ii) cada lectura desemparejada representa una primera cadena etiquetada que no tiene cadena complementaria etiquetada de manera diferente derivada de una molécula de polinucleótidos de cadena doble representada entre las lecturas de secuencias en el

conjunto de lecturas de secuencias. En algunas realizaciones el método comprende además determinar medidas cuantitativas de (i) las lecturas emparejadas y (ii) las lecturas desemparejadas que mapean para cada uno de los uno o más loci genéticos para determinar una medida cuantitativa de moléculas de ADN de cadena doble totales en la muestra que mapean para cada uno de los uno o más loci genéticos en base a la medida cuantitativa de lecturas emparejadas que mapean cada locus.

Aspectos y ventajas adicionales de la presente divulgación serán fácilmente aparentes para los expertos en la técnica a partir de la siguiente descripción detallada.

BREVE DECRIPCION DE LOS DIBUJOS

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

La **FIG. 1** es una representación de diagrama de flujo de un método de la presente divulgación para determinar la variación del número de copias (CNV);
La **FIG. 2** representa el mapeo de parejas y singletes a Locus A y Locus B en un genoma;
La **FIG. 3** muestra una secuencia de referencia que codifica un Locus A genético;
Las **FIGs. 4A-C** muestran amplificación, secuenciación, reducción de la redundancia y emparejamiento de moléculas complementarias;
La **FIG. 5** muestra la confianza aumentada en la detección de variantes de secuencias emparejando lecturas de cadenas de Watson y Crick;
La **FIG. 6** muestra un sistema informático que está programado o configurado de otra manera para implementar varios métodos de la presente divulgación;
La **FIG. 7** es una representación esquemática de un sistema para analizar una muestra que comprende ácidos nucleicos de un usuario, incluyendo un secuenciador; software bioinformático y conexión a internet para informar del análisis mediante, por ejemplo, un dispositivo portátil o un ordenador de sobremesa;
La **FIG. 8** es una representación de diagrama de flujo de un método para determinar CNV usando pruebas agrupadas y grupos de control; y
Las **FIGs. 9A-9C** ilustran esquemáticamente un método para etiquetar una molécula de polinucleótidos con un adaptador de bibliotecas y posteriormente un adaptador de secuenciación.

DESCRIPCION DETALLADA

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

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

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

El término "sujeto", como se usa en la presente, se refiere de manera general a un animal, como una

especie mamífera (por ejemplo humano) o especie aviar (por ejemplo pájaro), u otro organismo, como una planta. Más específicamente, el sujeto puede ser un vertebrado, un mamífero, un ratón, un primate, un simio o un humano. Los animales incluyen, pero no están limitados a, animales de granja, animales de deportes, y mascotas. Un sujeto puede ser un individuo sano, un individuo que tiene o se sospecha que tiene una enfermedad o predisposición a la enfermedad, o un individuo que necesita terapia o se sospecha que necesita terapia. Un sujeto puede ser un paciente.

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

Los términos "adaptador(es), y "etiqueta(s)" se usan como sinónimos a lo largo de esta especificación. Un adaptador o etiqueta puede acoplarse a una secuencia de polinucleótidos a ser "etiquetada" por cualquier enfoque incluyendo ligación, hibridación, u otros enfoques.

El término "adaptador de bibliotecas" como se usa en la presente, se refiere generalmente a una molécula (por ejemplo, polinucleótido) cuya identidad (por ejemplo, secuencia) puede usarse para diferenciar polinucleótidos en una muestra biológica (también "muestra" en la presente).

El término "adaptador de secuenciación", como se usa en la presente, se refiere generalmente a una molécula (por ejemplo, polinucleótido) que está adaptada para permitir que un instrumento de secuenciación secuencie un polinucleótido objetivo, como interactuando con el polinucleótido objetivo para permitir la secuenciación. El adaptador de secuenciación permite que el polinucleótido objetivo se secuencie por el instrumento de secuenciación. En un ejemplo, el adaptador de secuenciación comprende una secuencia de nucleótidos que hibrida o enlaza con un polinucleótido de captura unido a un soporte sólido de un sistema de secuenciación, como una célula de flujo. En otro ejemplo, el adaptador de secuenciación comprende una secuencia de nucleótidos que hibrida o enlaza con un polinucleótido para generar un giro de horquilla, que permite que el polinucleótido objetivo sea secuenciado por un sistema de secuenciación. El adaptador de secuenciación puede incluir un motivo secuenciador, que puede ser una secuencia de nucleótidos que es complementaria a una secuencia de célula de flujo de otra molécula (por ejemplo, polinucleótido) y utilizable por el sistema de secuenciación para secuenciar el polinucleótido objetivo. El motivo secuenciador puede incluir también una secuencia cebador para su uso en la secuenciación, como secuenciación por síntesis. El motivo secuenciador puede incluir la secuencia(s) necesaria para acoplar un adaptador de bibliotecas a un sistema de secuenciación y secuenciar el polinucleótido objetivo.

Como se usan en la presente los términos "por lo menos", "como mucho" o "aproximadamente", cuando preceden a una serie, se refieren a cada miembro de la serie, a menos que se identifique lo contrario.

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

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

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

1. <u>Métodos para procesar y/o analizar una muestra de ácidos nucleicos</u>

Un aspecto de la presente divulgación proporciona métodos para determinar una alternancia genómica en una muestra de ácidos nucleicos de un sujeto. La **FIG. 1** muestra un método para determinar la variación del número de copias (CNV). El método puede implementarse para determinar otras alternancias, como SNVs.

A. <u>Aislamiento de Polinucleótidos</u>

Los métodos divulgados en la presente pueden comprender aislar uno o más polinucleótidos. Un polinucleótido puede comprender cualquier tipo de ácido nucleico, por ejemplo, una secuencia de ácido nucleico genómico, o una secuencia artificial (por ejemplo, una secuencia no encontrada en un ácido nucleico genómico). Por ejemplo, una secuencia artificial puede contener nucleótidos no naturales. También, un polinucleótido puede

comprender tanto ácido nucleico genómico como una secuencia artificial, en cualquier parte. Por ejemplo, un polinucleótido puede comprender del 1 al 99% de ácido nucleico genómico y del 99% al 1% de secuencia artificial, donde el total suma hasta el 100%. Por tanto, también se contemplan las fracciones de los porcentajes. Por ejemplo, se contempla una proporción del 99,1% al 0,9%.

Un polinucleótido puede comprender cualquier tipo de ácido nucleico, como ADN y/o ARN. Por ejemplo, si un polinucleótido es ADN, puede ser ADN genómico, ADN complementario (ADNc), o cualquier otro ácido desoxirribonucleico. Un polinucleótido puede ser también ADN libre de células (ADNcf). Por ejemplo, el polinucleótido puede ser ADN circulante. El ADN circulante puede comprender ADN tumoral circulante (ADNct). Un polinucleótido puede ser de cadena doble o cadena sencilla. Alternativamente, un polinucleótido puede comprender una combinación de una parte de cadena doble y una parte de cadena sencilla.

Los polinucleótidos no tienen que ser libres de células. En algunos casos, los polinucleótidos pueden aislarse a partir de una muestra. Por ejemplo, en el paso (102) (**FIG. 1**), se aíslan polinucleótidos de cadena doble a partir de una muestra. Una muestra puede ser cualquier muestra biológica aislada de un sujeto. Por ejemplo, una muestra puede comprender, sin limitación, fluidos corporales, sangre completa, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejidos, fluido sinovial, fluido linfático, fluido de ascitis, fluido intersticial o extracelular, el fluido en los espacios entre las células, incluyendo fluido crevicular gingival, médula ósea, fluido cefalorraquídeo, saliva, mucosas, esputo, semen, sudor, orina o cualquier otro fluido corporal. Un fluido corporal puede incluir saliva, sangre o suero. Por ejemplo, un polinucleótido puede ser ADN libre de células aislado de un fluido corporal, por ejemplo, sangre o suero. Una muestra también puede ser una muestra tumoral, que se puede obtener de un sujeto por varios enfoques, que incluyen, pero no están limitados a, venopunción, excreción, eyaculación, masaje, biopsia, aspiración con aguja, lavado, raspado, incisión quirúrgica o intervención u otros enfoques.

Una muestra puede comprender varias cantidades de ácido nucleico que contiene equivalentes genómicos. Por ejemplo, una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 (10^4) equivalentes de genoma humano haploides y, en el caso de ADNcf, aproximadamente 200 billones ($2x10^{11}$) moléculas de polinucleótidos individuales. De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes de genoma humano haploides y, en el caso de ADNcf, aproximadamente 600 billones de moléculas individuales.

Una muestra puede comprender ácidos nucleicos de diferentes fuentes. Por ejemplo, una muestra puede comprender ADN de la línea germinal o ADN somático. Una muestra puede comprender ácidos nucleicos que llevan mutaciones. Por ejemplo, una muestra puede comprender ADN que lleva mutaciones en la línea germinal y/o mutaciones somáticas. Una muestra también puede comprender ADN que lleva mutaciones asociadas al cáncer (por ejemplo, mutaciones somáticas asociadas al cáncer).

B. <u>Etiquetado</u>

Los polinucleótidos divulgados en la presente pueden ser etiquetados. Por ejemplo, en el paso (104) (**FIG. 1**) los polinucleótidos de cadena doble se etiquetan con etiquetas dúplex, etiquetas que marcan de manera diferente las cadenas complementarias (es decir, las cadenas "Watson" y "Crick") de una molécula de cadena doble. En una realización las etiquetas dúplex son polinucleótidos que tienen partes complementarias o no complementarias.

Las etiquetas pueden ser cualquier tipo de moléculas unidas a un polinucleótido, incluyendo, pero no limitado a, ácidos nucleicos, compuestos químicos, sondas fluorescentes o sondas radiactivas. Las etiquetas también pueden ser oligonucleótidos (por ejemplo, ADN o ARN). Las etiquetas pueden comprender secuencias conocidas, secuencias desconocidas, o ambas. Una etiqueta puede comprender secuencias aleatorias, secuencias predeterminadas, o ambas. Una etiqueta puede ser de cadena doble o de cadena sencilla. Una etiqueta de cadena doble puede ser una etiqueta dúplex. Una etiqueta de cadena doble puede comprender dos cadenas complementarias. Alternativamente, una etiqueta de cadena doble puede comprender una parte hibridada y una parte no hibridada. La etiqueta de cadena doble puede tener forma de Y, por ejemplo, la parte hibridada está en un extremo de la etiqueta y la parte no hibridada está en el extremo opuesto de la etiqueta. Un ejemplo de ello son los "adaptadores Y" utilizado en la secuenciación Illumina. Otros ejemplos incluyen adaptadores con forma de horquilla o adaptadores con forma de burbuja. Los adaptadores con forma de burbuja tienen secuencias no complementarias flanqueadas en ambos lados por secuencias complementarias.

El etiquetado divulgado en la presente puede realizarse usando cualquier método. Un polinucleótido puede etiquetarse con un adaptador mediante hibridación. Por ejemplo, el adaptador puede tener una secuencia de nucleótidos que es complementaria a por lo menos una parte de una secuencia del polinucleótido. Como alternativa, un polinucleótido puede etiquetarse con un adaptador mediante ligación.

Por ejemplo, el etiquetado puede comprender el uso de uno o más enzimas. El enzima puede ser una ligasa. La ligasa puede ser una ADN ligasa. Por ejemplo, la ADN ligasa puede ser una ADN ligasa T4, una ADN

ligasa de E. coli y/o una ligasa de mamífero. La ligasa de mamífero puede ser ADN ligasa I, ADN ligasa III o ADN ligasa IV. La ligasa también puede ser una ligasa termoestable. Las etiquetas pueden ligarse a un extremo romo de un polinucleótido (ligación de extremos romos). Alternativamente, las etiquetas pueden ligarse a un extremo adhesivo de un polinucleótido (ligación de extremos adhesivo). La eficacia de la ligación puede aumentarse mediante la optimización de varias condiciones. La eficacia de la ligación puede aumentarse optimizando el tiempo de reacción de la ligación. Por ejemplo, el tiempo de reacción de la ligación puede ser inferior a 12 horas, por ejemplo, inferior a 1, inferior a 2, inferior a 3, inferior a 4, inferior a 5, inferior a 6, inferior a 7, inferior a 8, inferior a 9, inferior a 10, inferior a 11, inferior a 12, inferior a 13, inferior a 14, inferior a 15, inferior a 16, inferior a 17, inferior a 18, inferior a 19, o inferior a 20 horas. En un ejemplo particular, el tiempo de reacción de la ligación es inferior a 20 horas. La eficiencia de la ligación puede aumentarse optimizando la concentración de ligasa en la reacción. Por ejemplo, la concentración de ligasa puede ser de por lo menos 10, por lo menos 50, por lo menos 100, por lo menos 150, por lo menos 200, por lo menos 250, por lo menos 300, por lo menos 400, por lo menos 500, o por lo menos 600 unidades/microlitro. La eficiencia también puede optimizarse añadiendo o variando la concentración de un enzima adecuado para la ligación, cofactores enzimáticos u otros aditivos, y/o optimizando una temperatura de una solución que tiene el enzima. La eficiencia también puede optimizarse variando el orden de adición de varios componentes de la reacción. El final de la secuencia de etiquetado puede comprender dinucleótido para aumentar la eficiencia de ligación. Cuando la etiqueta comprende una parte no complementaria (por ejemplo, adaptador en forma de Y), la secuencia en la parte complementaria del adaptador de etiqueta puede comprender una o más secuencias seleccionadas que promueven la eficiencia de ligación. Preferiblemente, tales secuencias están localizadas en el extremo terminal de la etiqueta. Dichas secuencias pueden comprender 1, 2, 3, 4, 5, o 6 bases terminales. La solución de reacción con alta viscosidad (por ejemplo, un número de Reynolds bajo) también puede usarse para aumentar la eficiencia de la ligación. Por ejemplo, la solución puede tener un número de Reynolds menor de 3000, menor de 2000, menor de 1000, menor de 900, menor de 800, menor de 700, menor de 600, menor de 500, menor de 400, menor de 300, menor de 200, menor de 100, menor de 50, menor de 25, o menor de 10. También se contempla que se puede usar una distribución de fragmentos aproximadamente unificada (por ejemplo, una desviación estándar ajustada) para aumentar la eficiencia de ligación. Por ejemplo, la variación en el tamaño de los fragmentos puede variar en menos del 20%, menos del 15%, menos del 10%, menos del 5% o menos del 1%. El marcado también puede comprender extensión del cebador, por ejemplo, mediante reacción en cadena de polimerasa (PCR). El etiquetado también puede comprender cualquier PCR basado en ligación, PCR múltiplex, ligación de cadena sencilla, o circularización de cadena sencilla.

En algunos casos, las etiquetas de la presente comprenden códigos de barras moleculares. Dichos códigos de barras moleculares pueden usarse para diferenciar polinucleótidos en una muestra. Preferiblemente los códigos de barras moleculares son diferentes entre sí. Por ejemplo, los códigos de barras moleculares pueden tener una diferencia entre ellos que se puede caracterizar por una distancia de edición predeterminada o una distancia de Hamming. En algunos casos, los códigos de barras moleculares de la presente tienen una distancia de edición mínima de 1, 2, 3, 4, 5, 6, 7, 8, 9, ó 10. Para mejorar más la eficiencia de la conversión (por ejemplo, etiquetado) de moléculas no etiquetadas a moléculas etiquetadas, una preferiblemente utiliza etiquetas cortas. Por ejemplo, en algunas realizaciones, una etiqueta de adaptador de biblioteca puede tener una longitud de hasta 65, 60, 55, 50, 45, 40, o 35 bases de nucleótidos. Una colección de dichos códigos de barras de bibliotecas cortos incluye preferiblemente un número de códigos de barras moleculares diferentes, por ejemplo, por lo menos 2, 4, 6, 8, 10, 12, 14, 16, 18, ó 20 códigos de barras diferentes con una distancia de edición mínima de 1, 2, 3 o más.

Por tanto, una colección de moléculas puede incluir una o más etiquetas. En algunos casos, algunas moléculas en una colección pueden incluir una etiqueta de identificación ("identificador") como un código de barras molecular que no es compartido por ninguna otra molécula en la colección. Por ejemplo, en algunos casos de una colección de moléculas, por lo menos el 50%, por lo menos el 51%, por lo menos el 52%, por lo menos el 53%, por lo menos el 54%, por lo menos el 55%, por lo menos el 56%, por lo menos el 57 %, por lo menos el 58%, por lo menos el 59%, por lo menos el 60%, por lo menos el 61%, por lo menos el 62%, por lo menos el 63%, por lo menos el 64%, por lo menos el 65%, por lo menos el 66%, por lo menos el 67 %, por lo menos el 68%, por lo menos el 69%, por lo menos el 70%, por lo menos el 71%, por lo menos el 72%, por lo menos el 73%, por lo menos el 74%, por lo menos el 75%, por lo menos el 76%, por lo menos el 77 %, por lo menos el 78%, por lo menos el 79%, por lo menos el 80%, por lo menos el 81%, por lo menos el 82%, por lo menos el 83%, por lo menos el 84%, por lo menos el 85%, por lo menos el 86%, por lo menos el 87 %, por lo menos el 88%, por lo menos el 89%, por lo menos el 90%, por lo menos el 91%, por lo menos el 92%, por lo menos el 93%, por lo menos el 94%, por lo menos el 95%, por lo menos el 96%, por lo menos el 97%, por lo menos el 98%, por lo menos el 99% o el 100% de las moléculas de la colección pueden incluir un identificador o código de barras molecular que no es compartido por ninguna otra molécula en la colección. Como se usa en la presente, se considera que una colección de moléculas está "etiquetada de manera única" si cada una de por lo menos el 95% de las moléculas en la colección lleva un identificador que no se comparte por ninguna otra molécula en la colección ("etiqueta única" o "identificador único"). Una colección de moléculas se considera "etiquetada de manera no única" si cada una de por lo menos el 1%, por lo menos el 5%, por lo menos el 10%, por lo menos el 15%, por lo menos el 20%, por lo menos el 25% , por lo menos el 30%, por lo menos el 35%, por lo menos el 40%, por lo menos el 45%, o por lo menos o aproximadamente el 50% de las moléculas en la colección lleva una etiqueta identificadora o un código de barras molecular que es compartido por lo menos por otra molécula en la colección ("etiqueta no única" o "identificador no único"). Por consiguiente, en una

población etiquetada de manera no única, no más del 1% de las moléculas están etiquetadas de manera única. Por ejemplo, en una población etiquetada de manera no única, no más del 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, o 50% de las moléculas pueden ser etiquetadas de manera única.

Se pueden usar un número de etiquetas diferentes en base al número estimado de moléculas en una muestra. En algunos métodos de etiquetado, el número de etiquetas diferentes puede ser por lo menos el mismo que el número estimado de moléculas en la muestra. En otros métodos de etiquetado, el número de etiquetas diferentes puede ser de por lo menos dos, tres, cuatro, cinco, seis, siete, ocho, nueve, diez, cien o mil veces más que el número estimado de moléculas en la muestra. En etiquetado único, pueden usarse por lo menos dos veces (o más) más etiquetas diferentes que el número estimado de moléculas en la muestra.

Las moléculas en la muestra pueden estar etiquetadas de manera no única. En tales casos, se usa un número menor de etiquetas o códigos de barras moleculares que el número de moléculas en la muestra a ser etiquetada. Por ejemplo, no se utilizan más de 100, 50, 40, 30, 20 ó 10 etiquetas únicas o códigos de barras moleculares para etiquetar una muestra compleja como una muestra de ADN libre de células con muchos más fragmentos diferentes.

El polinucleótido a ser etiquetado puede fragmentarse, ya sea de manera natural o usando otros enfoques, como, por ejemplo, cizallamiento. Los polinucleótidos pueden fragmentarse mediante ciertos métodos, incluyendo pero no limitados a, cizallamiento mecánico, pasar la muestra a través de una jeringuilla, sonicación, tratamiento térmico (por ejemplo, durante 30 minutos a 90° C) y/o tratamiento con nucleasa (por ejemplo, usando DNasa, RNasa, endonucleasa, exonucleasa, y/o enzima de restricción).

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

Los polinucleótidos etiquetados pueden comprender secuencias asociadas con el cáncer. Las secuencias asociadas con el cáncer pueden comprender variación de nucleótido único (SNV), variación del número de copias (CNV), inserciones, deleciones y/o reordenamientos.

Los polinucleótidos pueden comprender secuencias asociadas con el cáncer, como leucemia linfoblástica aguda (ALL), leucemia mieloide aguda (AML), carcinoma adrenocortical, sarcoma de Kaposi, cáncer anal, carcinoma de células basales, cáncer de vías biliares, cáncer de vejiga, cáncer de hueso, osteosarcoma, histiocitoma fibroso maligno, glioma de tronco encefálico, cáncer de cerebro, craneofaringioma, ependimoblastoma, ependimoma, meduloblastoma, meduloepitelioma, tumor del parénquima pineal, cáncer de mama, tumor bronquial, linfoma de Burkitt, linfoma no de Hodgkin, tumor carcinoide, cáncer cervical, cordoma, leucemia linfocítica crónica (CLL), leucemia mielógena crónica (CML), cáncer de colon, cáncer colorrectal, linfoma de células T cutáneo, carcinoma ductal in situ, cáncer de endometrio, cáncer de esófago, sarcoma de Ewing, cáncer de ojo, melanoma intraocular, retinoblastoma, histiocitoma fibroso, cáncer de vesícula biliar, cáncer gástrico, glioma, leucemia de células pilosas, cáncer de cabeza y cuello, cáncer de corazón, cáncer hepatocelular (de hígado), linfoma de Hodgkin, cáncer de hipofaringe, cáncer de riñón, cáncer de laringe, cáncer de labios, cáncer de cavidad oral, cáncer de pulmón, carcinoma de células no pequeñas, carcinoma de células pequeñas, melanoma, cáncer de boca, síndromes mielodisplásicos, mieloma múltiple, meduloblastoma, cáncer de cavidad nasal, cáncer de seno paranasal, neuroblastoma, cáncer nasofaríngeo, cáncer oral, cáncer orofaríngeo, osteosarcoma, cáncer de ovario, cáncer de páncreas, papilomatosis, paraganglioma, cáncer paratiroideo, cáncer de pene, cáncer de faringe, tumor pituitario, neoplasia de células plasmáticas, cáncer de próstata, cáncer rectal, cáncer de células renales, rabdomiosarcoma , cáncer de glándula salival, síndrome de Sezary, cáncer de piel, no melanoma, cáncer de intestino delgado, sarcoma de tejidos blandos, carcinoma de células escamosas, cáncer testicular, cáncer de garganta, timoma, cáncer de tiroides, cáncer de uretra, cáncer uterino, sarcoma uterino, cáncer vaginal, cáncer de vulva, macroglobulinemia de Waldenstrom y/o tumor de Wilms.

Un equivalente de genoma humano haploide tiene aproximadamente 3 picogramos de ADN. Una muestra de aproximadamente 1 microgramo de ADN contiene aproximadamente 300.000 equivalentes de genoma humano haploide. Se pueden lograr mejoras en la secuenciación siempre que por lo menos algunos de los polinucleótidos duplicados o afines tengan identificadores únicos entre sí, es decir, que lleven etiquetas diferentes. Sin embargo, en ciertas realizaciones, el número de etiquetas usadas se selecciona de manera que haya por lo menos un 95% de posibilidades de que todas las moléculas duplicadas que comienzan en cualquier posición lleven identificadores únicos. Por ejemplo, en una muestra que comprende aproximadamente 10.000 equivalentes de genoma humano

haploide de ADN genómico fragmentado, por ejemplo, ADNc, se espera que z esté entre 2 y 8. Dicha población puede etiquetarse con entre aproximadamente 10 y 100 identificadores diferentes, por ejemplo, aproximadamente 2 identificadores, aproximadamente 4 identificadores, aproximadamente 9 identificadores, aproximadamente 16 identificadores, aproximadamente 25 identificadores, aproximadamente 36 identificadores diferentes, aproximadamente 49 identificadores diferentes, aproximadamente 64identificadores diferentes, aproximadamente 81 identificadores diferentes, o aproximadamente 100 identificadores diferentes.

Los códigos de barras de ácidos nucleicos que tienen secuencias identificables que incluyen códigos de barras moleculares, pueden usarse para etiquetar. Por ejemplo, una pluralidad de códigos de barras de ADN puede comprender varios números de secuencias de nucleótidos. Pueden usarse una pluralidad de códigos de barras de ADN que tienen 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 o más secuencias identificables de nucleótidos. Cuando se une a un único extremo de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 o más identificadores diferentes. Alternativamente, cuando se unen a ambos extremos de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 o más identificadores diferentes (que es ^2 de cuando el código de barras de ADN está unido a sólo 1 extremo de un polinucleótido). En un ejemplo, se puede usar una pluralidad de códigos de barras de ADN que tienen 6, 7, 8, 9 ó 10 secuencias identificables de nucleótidos. Cuando se unen a ambos extremos de un polinucleótido, producen 36, 49, 64, 81 ó 100 posibles identificadores diferentes, respectivamente. En un ejemplo particular, la pluralidad de códigos de barras de ADN puede comprender 8 secuencias identificables de nucleótidos. Cuando se une a un solo extremo de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 8 identificadores diferentes. Alternativamente, cuando se unen a ambos extremos de un polinucleótido, la pluralidad de códigos de barras de ADN puede producir 64 identificadores diferentes. Las muestras etiquetadas 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, ADNcf.

Un polinucleótido puede identificarse de manera única de varias maneras. Un polinucleótido puede identificarse de manera única mediante un código de barras de ADN único. Por ejemplo, dos polinucleótidos cualquiera en una muestra está unidos a dos códigos de barras de ADN diferentes. Alternativamente, un polinucleótido puede identificarse de manera única mediante la combinación de un código de barras de ADN y una o más secuencias endógenas del polinucleótido. Por ejemplo, dos polinucleótidos cualquiera en una muestra puede estar unidos al mismo código de barras de ADN, pero los dos polinucleótidos pueden identificarse todavía mediante diferentes secuencias endógenas. La secuencia endógena puede estar en un extremo de un polinucleótido. Por ejemplo, la secuencia endógena puede ser adyacente (por ejemplo, base entre) al código de barras de ADN unido. En algunos casos, la secuencia endógena puede ser de por lo menos 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, ó 100 bases de longitud. Preferiblemente, la secuencia endógena es una secuencia terminal del fragmento/polinucleótidos a ser analizados. La secuencia endógena puede ser la longitud de la secuencia. Por ejemplo, una pluralidad de códigos de barras de ADN que comprenden 8 códigos de barras de ADN diferentes puede unirse a ambos extremos de cada polinucleótido en una muestra. Cada polinucleótido en la muestra puede identificarse mediante la combinación de códigos de barras de ADN y una secuencia endógena de aproximadamente 10 pares de bases en un extremo del polinucleótido. Sin estar limitados por la teoría, la secuencia endógena de un polinucleótido también puede ser la secuencia completa del polinucleótido.

También se divulgan en la presente composiciones de polinucleótidos etiquetados. El polinucleótido etiquetado puede ser de cadena sencilla. Alternativamente, el polinucleótido etiquetado puede ser de cadena doble (por ejemplo, polinucleótidos etiquetados dúplex). Por consiguiente, esta divulgación también proporciona composiciones de polinucleótidos etiquetados dúplex. Los polinucleótidos pueden comprender cualquier tipo de ácidos nucleicos (ADN y/o ARN). Los polinucleótidos comprenden cualquier tipo de ADN divulgado en la presente. Por ejemplo, los polinucleótidos pueden comprender ADN, por ejemplo, ADN fragmentado o ADNcf. Un conjunto de polinucleótidos en la composición que mapean una posición base mapeable en un genoma puede etiquetarse de manera no única, es decir, el número de identificadores diferentes puede ser por lo menos 2 o inferior al número de polinucleótidos que mapean la posición base mapeable. El número de identificadores diferentes puede ser por lo menos 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 y menos que el número de polinucleótidos que mapean la posición base mapeable.

En algunos casos, como una composición va de aproximadamente 1 ng a aproximadamente 10 µg o más, se puede usar un conjunto más grande de códigos de barras moleculares diferentes. Por ejemplo, se pueden usar entre 5 y 100 adaptadores de bibliotecas diferentes para etiquetar polinucleótidos en una muestra de ADNcf.

Los sistemas y métodos descritos en la presente pueden usarse en aplicaciones que implican la asignación de códigos de barras moleculares. Los códigos de barras moleculares pueden asignarse a cualquier tipo de polinucleótido divulgado. Por ejemplo, los códigos de barras moleculares pueden asignarse a polinucleótidos libres de células (por ejemplo, ADNcfs). A menudo, un identificador divulgado en la presente puede ser un oligonucleótido de código de barras que se usa para etiquetar el polinucleótido. El identificador de código de barras puede ser un

oligonucleótido de ácido nucleico (por ejemplo, un oligonucleótido de ADN). El identificador de código de barras puede ser de cadena sencilla. Alternativamente, el identificador de código de barras puede ser de cadena doble. El identificador de código de barras puede unirse a polinucleótidos usando cualquier método divulgado en la presente. Por ejemplo, el identificador de código de barras puede unirse al polinucleótido por ligación usando un enzima. El identificador de código de barras también puede incorporarse en el polinucleótido mediante PCR. En otros casos, la reacción puede comprender la adición de un isótopo de metal, ya sea directamente al analito o mediante una sonda etiquetada con el isótopo. Generalmente, la asignación de identificadores únicos o no únicos o códigos de barras moleculares en las reacciones de esta divulgación puede seguir métodos y sistemas descritos por, por ejemplo, las solicitudes de patente U.S 2001/0053519, 2003/0152490, y la patente U.S Nº 6.582.908.

Los identificadores o códigos de barras moleculares usados en la presente pueden ser completamente endógenos por lo que se puede realizar la ligación circular de fragmentos individuales seguido por cizallamiento aleatorio o amplificación dirigida. En este caso, la combinación de un nuevo punto de inicio y de finalización de la molécula y el punto de ligación intramolecular original puede formar un identificador específico.

Los identificadores o códigos de barras moleculares usados en la presente pueden comprender cualquier tipo de oligonucleótidos. En algunos casos, los identificadores pueden ser oligonucleótidos de secuencia predeterminados, aleatorios o semi-aleatorios. Los identificadores pueden ser códigos de barras. Por ejemplo, se puede usar una pluralidad de códigos de barras de modo que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. Alternativamente, puede usarse una pluralidad de códigos de barras de manera que cada código de barras sea único para cualquier otro código de barras en la pluralidad. Los códigos de barras pueden comprender secuencias específicas (por ejemplo, secuencias predeterminadas) que pueden rastrearse individualmente. Además, los códigos de barras pueden unirse (por ejemplo, por ligación) a moléculas individuales de manera que la combinación del código de barras y la secuencia a la que se puede ligar crea una secuencia específica que puede rastrearse individualmente. Como se describe en la presente, la detección de códigos de barras en combinación con datos de secuencia de las partes inicial (inicio) y/o final (finalización) 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 lectura de secuencia individual también puede usarse para asignar una identidad única a dicha molécula. Como se describe en la presente, los fragmentos de una cadena individual de ácido nucleico a la que se ha asignado una identidad única, pueden permitir de ese modo la identificación posterior de los fragmentos de la cadena inicial. De esta forma, los polinucleótidos en la muestra pueden etiquetarse de manera única o sustancialmente única. Una etiqueta dúplex puede incluir una secuencia de nucleótidos degenerada o semi-degenerada, por ejemplo, una secuencia degenerada aleatoria. La secuencia de nucleótidos puede comprender cualquier número de nucleótidos. Por ejemplo, la secuencia de nucleótidos puede comprender 1 (si se usa un nucleótido no natural), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 o más nucleótidos. En un ejemplo particular, la secuencia puede comprender 7 nucleótidos. En otro ejemplo, la secuencia puede comprender 8 nucleótidos. La secuencia también puede comprender 9 nucleótidos. La secuencia puede comprender 10 nucleótidos.

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

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

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

C. <u>Adaptadores</u>

Se puede sintetizar una biblioteca de moléculas de polinucleótidos para su uso en la secuenciación. Por ejemplo, puede hacerse una biblioteca de polinucleótidos que comprende una pluralidad de moléculas de polinucleótidos que son cada uno menores de o iguales a 100, 90, 80, 70, 60, 50, 45, 40 o 35 bases de ácidos

nucleicos (o nucleótidos) de longitud. Una pluralidad de moléculas de polinucleótidos puede ser cada una menor o igual a 35 bases de ácido nucleico de longitud. Una pluralidad de moléculas de polinucleótidos puede ser cada una menor o igual que 30 bases de ácidos nucleicos de longitud. Una pluralidad de las moléculas de polinucleótidos también puede ser inferior o igual a 250, 200, 150, 100 ó 50 bases de ácidos nucleicos. Adicionalmente, la pluralidad de moléculas de polinucleótidos también puede ser inferior o igual a 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, ó 10 bases de ácidos nucleicos.

Una biblioteca de polinucleótidos que comprende una pluralidad de moléculas de polinucleótidos también puede tener secuencias de códigos de barras moleculares distintas (unas respecto a las otras) (o códigos de barras moleculares) con respecto a por lo menos 4 bases de ácidos nucleicos. Una secuencia de código de barras molecular (también "código de barras" o "identificador") es una secuencia de nucleótidos que distingue a un polinucleótido de otro. En otras realizaciones, las moléculas de polinucleótidos también pueden tener diferentes secuencias de códigos de barras con respecto a 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 o más bases de ácido nucleico.

Una biblioteca de polinucleótidos que comprende una pluralidad de moléculas de polinucleótidos también puede tener una pluralidad de secuencias de códigos de barras diferentes. Por ejemplo, una pluralidad de moléculas de polinucleótidos puede tener por lo menos 4 secuencias de códigos de barras moleculares diferentes. En algunos casos, la pluralidad de moléculas de polinucleótidos tiene de 2-100, 4-50, 4-30, 4-20 o 4-10 secuencias de códigos de barras moleculares diferentes. La pluralidad de moléculas de polinucleótidos también puede tener otros intervalos de diferentes secuencias de códigos de barras, como, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, o 22-25 secuencias de códigos de barras diferentes. En otros casos, una pluralidad de moléculas de polinucleótidos puede tener por lo menos 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 , 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 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, 95, 96, 97, 98, 99, ó 100 secuencias de códigos de barras diferentes más. En un ejemplo particular, la pluralidad de adaptadores de bibliotecas comprende por lo menos 8 secuencias diferentes.

La localización de las diferentes secuencias de códigos de barras puede variar dentro de la pluralidad de polinucleótidos. Por ejemplo, las diferentes secuencias de códigos de barras pueden estar dentro de 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, ó 2 bases de ácidos nucleicos de un extremo terminal de una respectiva de la pluralidad de moléculas de polinucleótidos. En un ejemplo, una pluralidad de moléculas de polinucleótidos tiene secuencias de códigos de barras distintas que están dentro de 10 bases de ácidos nucleicos desde el extremo terminal. En otro ejemplo, una pluralidad de moléculas de polinucleótidos tiene secuencias de códigos de barras distintas que están dentro de 5 ó 1 bases de ácidos nucleicos desde el extremo terminal. En otros casos, las secuencias de códigos de barras distintas pueden estar en el extremo terminal de una respectiva de la pluralidad de moléculas de polinucleótidos. Otras variaciones incluyen que las distintas secuencias de códigos de barras moleculares puedan estar dentro de 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, ó 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, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, o más bases de ácidos nucleicos de un extremo terminal de una respectiva de la pluralidad de moléculas de polinucleótidos.

El extremo terminal de la pluralidad de moléculas de polinucleótidos puede adaptarse para ligación a una molécula de ácido nucleico objetivo. Por ejemplo, el extremo del terminal puede ser un extremo romo. En algunos otros casos, el extremo terminal está adaptado para hibridación con una secuencia complementaria de una molécula de ácido nucleico objetivos.

Una biblioteca de polinucleótidos que comprende una pluralidad de moléculas de polinucleótidos puede tener también una distancia de edición de por lo menos 1. En algunos casos, la distancia de edición es con respecto a bases individuales de la pluralidad de moléculas de polinucleótidos. En otros casos, la pluralidad de moléculas de polinucleótidos puede tener una distancia de edición de por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 o más. La distancia de edición puede ser una distancia de Hamming.

En algunos casos, la pluralidad de polinucleótidos no contiene adaptadores de secuenciación. Un adaptador de secuencia puede ser un polinucleótido que comprende una secuencia que hibrida con uno o más

adaptadores de secuenciación o cebadores. Un adaptador de secuenciación puede comprender además una secuencia que hibrida con un soporte sólido, por ejemplo, una secuencia de células de flujo. El término "secuencia de células de flujo" y sus equivalentes gramaticales como se usa en la presente, se refiere a una secuencia que permite la hibridación a un sustrato, por ejemplo, por medio de un cebador unido al sustrato. El sustrato puede ser una perla o una superficie plana. En algunas realizaciones, una secuencia de células de flujo puede permitir que un polinucleótido se una a una célula de flujo o superficie (por ejemplo, superficie de una perla, por ejemplo, una célula de flujo Illumina).

Cuando una pluralidad de moléculas de polinucleótidos no contiene adaptadores de secuenciación o cebadores, cada molécula de polinucleótidos de la pluralidad no contiene una secuencia de ácido nucleico u otra fracción que está adaptada para permitir la secuenciación de una molécula de ácido nucleico objetivo con un enfoque de secuenciación dado, como Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, Gnu-Bio, Ion Torrent, Oxford Nanopore o Genia. En algunos ejemplos, cuando una pluralidad de moléculas de polinucleótidos no contiene adaptadores de secuenciación o cebadores, la pluralidad de moléculas de polinucleótidos no contiene secuencias de células de flujo. Por ejemplo, la pluralidad de moléculas de polinucleótidos no se puede enlazar con las células de flujo, como se usa en secuenciadores de células de flujo Illumina. Sin embargo, estas secuencias de células de flujo, si se desea, pueden añadirse a la pluralidad de moléculas de polinucleótidos por métodos como amplificación por PCR o ligación. En este punto, se pueden usar secuenciadores de células de flujo Illumina. Alternativamente, cuando la pluralidad de moléculas de polinucleótidos no contiene adaptadores de secuenciación o cebadores, la pluralidad de moléculas de polinucleótidos no contiene adaptadores con forma de horquilla o adaptadores para generar giros de horquilla en una molécula de ácido nucleico objetivo, como los adaptadores Pacific Bioscience SMRTbell™. Sin embargo, estos adaptadores con forma de horquilla, si se desea, pueden añadirse a la pluralidad de moléculas de polinucleótidos por métodos como amplificación por PCR o ligación. La pluralidad de moléculas de polinucleótidos puede ser circular o lineal.

Una pluralidad de moléculas de polinucleótidos puede ser de cadena doble. En algunos casos, la pluralidad de moléculas de polinucleótidos puede ser de cadena sencilla, o puede comprender regiones hibridadas y no hibridadas. Una pluralidad de moléculas de polinucleótidos puede ser moléculas de polinucleótidos no naturales.

Los adaptadores pueden ser moléculas de polinucleótidos. Las moléculas de polinucleótidos pueden tener forma de Y, forma de burbuja o forma de horquilla. Un adaptador de horquilla puede contener un sitio(s) de restricción o una base que contiene uracilo. Los adaptadores pueden comprender una parte complementaria y una parte no complementaria. La parte no complementaria puede tener una distancia de edición (por ejemplo, distancia de Hamming). Por ejemplo, la distancia de edición puede ser de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19, por lo menos 20, por lo menos 21, por lo menos 22, por lo menos 23, por lo menos 24, por lo menos 25, por lo menos 26, por lo menos 27, por lo menos 28, por lo menos 29 o por lo menos 30. La parte complementaria del adaptador puede comprender secuencias que se seleccionan para permitir y/o promover la ligación con un polinucleótido, por ejemplo, una secuencia para permitir y/o promover la ligación con un polinucleótido a un rendimiento alto.

Una pluralidad de las moléculas de polinucleótidos como se divulgan en la presente puede purificarse. En algunos casos, una pluralidad de moléculas de polinucleótidos como se divulgan en la presente pueden ser moléculas de polinucleótidos aisladas. En otros casos, una pluralidad de moléculas de polinucleótidos como se divulgan en la presente pueden ser moléculas de polinucleótidos purificadas y aisladas.

En ciertos aspectos, cada una de la pluralidad de moléculas de polinucleótidos tiene forma de Y o forma de horquilla. Cada una de la pluralidad de las moléculas de polinucleótidos puede comprender un código de barras diferente. El código de barras diferente puede ser un aleatorizador en la parte complementaria (por ejemplo, parte de cadena doble) del adaptador con forma de Y o con forma de horquilla. Alternativamente, el código de barras diferente puede estar en una cadena de la parte no complementaria (por ejemplo, uno de los brazos con forma de Y). Como se ha tratado anteriormente, el código de barras diferente puede ser por lo menos de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 , 21, 22, 23, 24, 25 o más (o cualquier longitud como se describe a lo largo de la presente) bases de ácidos nucleicos, por ejemplo, 7 bases. El código de barras puede ser secuencias contiguas o no contiguas, como se describe anteriormente. La pluralidad de moléculas de polinucleótidos es de 10 bases de ácidos nucleicos a 35 bases de ácidos nucleicos (o cualquier longitud como se describe anteriormente) de longitud. Además, la pluralidad de moléculas de polinucleótidos puede comprender una distancia de edición (como se ha descrito anteriormente), que es una distancia de Hamming. Una pluralidad de moléculas de polinucleótidos puede tener secuencias de códigos de barras distintas que están dentro de 10 bases de ácidos nucleicos desde el extremo terminal.

En otro aspecto, una pluralidad de moléculas de polinucleótidos pueden ser adaptadores de secuenciación. Un adaptador de secuenciación puede comprender una secuencia que hibrida con uno o más cebadores de secuenciación. Un adaptador de secuenciación puede comprender además una secuencia que

hibrida con un soporte sólido, por ejemplo, una secuencia de células de flujo. Por ejemplo, un adaptador de secuencia puede ser un adaptador de células de flujo. Los adaptadores de secuenciación pueden unirse a uno o a ambos extremos de un fragmento de polinucleótido. En otro ejemplo, un adaptador de secuenciación puede tener forma de horquilla. Por ejemplo, el adaptador con forma de horquilla puede comprender una parte de cadena doble complementaria y una parte de giro, donde la parte de cadena doble puede unirse (por ejemplo, ligarse) a un polinucleótido de cadena doble. Los adaptadores de secuenciación con forma de horquilla pueden unirse a ambos extremos de un fragmento de polinucleótido para generar una molécula circular, que puede secuenciarse múltiples veces. Un adaptador de secuenciación puede ser de hasta 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 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, 95, 96, 97, 98, 99, 100, o más bases de extremo a extremo. Por ejemplo, un adaptador de secuenciación puede ser de hasta 70 bases de extremo a extremo. El adaptador de secuenciación puede comprender de 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases de extremo a extremo. En un ejemplo particular, el adaptador de secuenciación puede comprender de 20-30 bases de extremo a extremo. En otro ejemplo, el adaptador de secuenciación puede comprender de 50-60 bases de extremo a extremo. Un adaptador de secuencia puede comprender uno o más códigos de barras. Por ejemplo, un adaptador de secuencia puede comprender un código de barras de muestra. El código de barras de muestra puede comprender una secuencia predeterminada. Los códigos de barras de muestra pueden usarse para identificar la fuente de los polinucleótidos. El código de barras de muestra puede ser de por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 2 , 23, 24, 25 o más (o cualquier longitud como se describe a lo largo de) bases de ácidos nucleicos, por ejemplo, por lo menos 8 bases. El código de barras puede ser de secuencias contiguas o no contiguas, como se ha descrito con anterioridad.

La pluralidad de moléculas de polinucleótidos como se describe en la presente puede usarse como adaptadores. Los adaptadores pueden comprender uno o más identificadores. Un adaptador puede comprender un identificador con una secuencia aleatoria. Alternativamente, un adaptador puede comprender un identificador con secuencias predeterminadas. Algunos adaptadores pueden comprender un identificador con una secuencia aleatoria y otro identificador con una secuencia predeterminada. Los adaptadores que comprenden identificadores pueden ser adaptadores de cadena doble o adaptadores de cadena sencilla. Los adaptadores que comprenden identificadores pueden ser adaptadores con forma de Y. Un adaptador con forma de Y puede comprender uno o más identificadores con una secuencia aleatoria. El uno o más identificadores pueden estar en la parte híbrida y/o en la parte no hibridada del adaptador con forma de Y. Un adaptador con forma de Y puede comprender uno o más identificadores con una secuencia predeterminada. El uno o más identificadores con secuencia predeterminada puede estar en la parte hibridada y/o en la parte no hibridada del adaptador con forma de Y. Un adaptador con forma de Y puede comprender uno o más identificadores con una secuencia aleatoria y uno o más identificadores con una secuencia predeterminada. Por ejemplo, el uno o más identificadores con una secuencia aleatoria puede estar en la parte hibridada del adaptador con forma de Y y/o la parte no hibridada del adaptador con forma de Y. El uno o más identificadores con una secuencia predeterminada puede estar en la parte hibridada del adaptador con forma de Y y/o la parte no hibridada del adaptador con forma de Y. En un ejemplo particular, un adaptador con forma de Y puede comprender un identificador con una secuencia aleatoria en su parte hibridada y un identificador con una secuencia predeterminada en su parte no hibridada. Los identificadores pueden ser de cualquier longitud descrita en la presente. Por ejemplo, un adaptador con forma de Y puede comprender un identificador con una secuencia aleatoria de 7 nucleótidos en su parte hibridada y un identificador con una secuencia predeterminada de 8 nucleótidos en su parte no hibridada.

Un adaptador puede incluir una parte de cadena doble con un código de barras molecular y por lo menos una o dos partes de cadena sencilla. Por ejemplo, el adaptador puede tener forma de Y e incluir una parte de cadena doble y dos partes de cadena sencilla. Las partes de cadena sencilla pueden incluir secuencias que no son complementarias entre sí.

El adaptador puede incluir un extremo terminal que tiene una secuencia que se selecciona para permitir que el adaptador sea ligue eficientemente (por ejemplo, con una eficiencia de por lo menos aproximadamente el 20%, 30%, 40%, 50%) o se acople de otra manera a un polinucleótido. En algunos ejemplos, los nucleótidos terminales en una parte de cadena doble de un adaptador se seleccionan de una combinación de purinas y pirimidinas para proporcionar una ligación eficiente.

En algunos ejemplos, un conjunto de adaptadores de bibliotecas comprende una pluralidad de moléculas de polinucleótidos (adaptadores de biblioteca) con códigos de barras moleculares. Los adaptadores de biblioteca son menores o iguales a 80, 70, 60, 50, 45 o 40 bases de nucleótidos de longitud. Los códigos de barras moleculares pueden ser de por lo menos 4 bases de nucleótidos de longitud, pero pueden ser de 4 a 20 bases de nucleótidos de longitud. Los códigos de barras moleculares pueden ser diferentes entre sí y tener una distancia de edición de por lo menos 1, 2, 3, 4 o 5 entre uno y otro. Los códigos de barras moleculares están localizados por lo menos 1, 2, 3, 4, 5, 10 o 20 bases de nucleótidos alejadas desde un extremo terminal de sus respectivos adaptadores de biblioteca. En algunos casos, la por lo menos una base terminal es idéntica en todos los adaptadores de bibliotecas.

Los adaptadores de biblioteca pueden ser idénticos pero para los códigos de barras moleculares. Por ejemplo, los adaptadores de biblioteca pueden tener secuencias idénticas pero difieren solo con respecto a las secuencias de nucleótidos de los códigos de barras moleculares.

Cada uno de los adaptadores de bibliotecas puede tener una parte de cadena doble y por lo menos una parte de cadena sencilla. Por "parte de cadena sencilla" se entiende un área de no complementariedad o un saliente. En algunos casos, cada uno de los adaptadores de biblioteca tiene una parte de cadena doble y dos partes de cadena sencilla. La parte de cadena doble puede tener un código de barras molecular. En algunos casos, el código de barras molecular es un aleatorizador. Cada uno de los adaptadores de biblioteca puede incluir además un código de barras de identificación de cadenas en una parte de cadena sencilla. El código de barras de identificación de cadenas puede incluir por lo menos 4 bases de nucleótidos, en algunos casos de 4 a 20 bases de nucleótidos.

En algunos ejemplos, cada uno de los adaptadores de bibliotecas tiene una parte de cadena doble con un código de barras molecular y dos partes de cadena sencilla. Las partes de cadena sencilla pueden no hibridar entre sí. Las partes de cadena sencilla pueden no ser completamente complementarias entre sí.

Los adaptadores de bibliotecas pueden tener una secuencia de nucleótidos terminales en una parte de cadena doble que son los mismos. La secuencia de nucleótidos terminales puede ser de por lo menos 2, 3, 4, 5 ó 6 bases de nucleótidos de longitud. Por ejemplo, una cadena de una parte de cadena doble del adaptador de bibliotecas puede tener la secuencia ACTT, TCGC o TACC en el extremo terminal, mientras que la otra cadena puede tener una secuencia complementaria. En algunos casos, dicha secuencia se selecciona para optimizar la eficiencia a la que los adaptadores de las bibliotecas se unen a los polinucleótidos objetivo. Dichas secuencias pueden seleccionarse para optimizar una interacción de enlace entre los extremos de los adaptadores de bibliotecas y los polinucleótidos objetivo.

En algunos casos, ninguno de los adaptadores de bibliotecas contiene un motivo de identificación de muestras (o código de barras molecular de muestras). Dicho motivo de identificación de muestras puede proporcionarse a través de adaptadores de secuenciación. Un motivo de identificación de muestras puede incluir un secuenciador de por lo menos 4, 5, 6, 7, 8, 9, 10, 20, 30, ó 40 bases de nucleótidos que permite la identificación de moléculas de polinucleótidos de una muestra dada a partir de moléculas de polinucleótidos de otras muestras. Por ejemplo, esto puede permitir que se secuencien moléculas de polinucleótidos de dos sujetos en el mismo conjunto y se identifiquen posteriormente las lecturas de secuencia para los sujetos.

Un motivo secuenciador incluye la secuencia(s) de nucleótidos necesarias para acoplar un adaptador de bibliotecas a un sistema de secuenciación y secuenciar un polinucleótido objetivo acoplado al adaptador de bibliotecas. El motivo secuenciador puede incluir una secuencia que es complementaria a una secuencia de células de flujo y una secuencia (secuencia de iniciación de secuenciación) que puede hibridarse selectivamente a un cebador (o secuencia de cebado) para su uso en la secuenciación. Por ejemplo, dicha secuencia de inicio de secuenciación puede ser complementaria a un cebador que se emplea para su uso en secuencia por síntesis (por ejemplo, Illumina). Dicho cebador puede ser incluido en un adaptador de secuenciación. Una secuencia de iniciación de secuenciación puede ser un sitio de hibridación de cebador.

En algunos casos, ninguno de los adaptadores de bibliotecas contiene un motivo secuenciador completo. Los adaptadores de bibliotecas pueden contener motivos secuenciadores parciales o nulos. En algunos casos, los adaptadores de bibliotecas incluyen una secuencia de iniciación de la secuenciación. Los adaptadores de bibliotecas pueden incluir una secuencia de iniciación de la secuenciación pero no secuencia de células de flujo. La secuencia de iniciación de la secuenciación puede ser complementaria a un cebador para la secuenciación. El cebador puede ser un cebador específico de la secuencia o un cebador universal. Tales secuencias de iniciación de la secuenciación pueden situarse en partes de cadena sencilla de los adaptadores de bibliotecas. Como alternativa, tales secuencias de iniciación de la secuenciación pueden ser sitios de cebado (por ejemplo, acodamientos o muescas) para permitir que una polimerasa se acople con los adaptadores de bibliotecas durante la secuenciación.

En algunos casos, se proporcionan motivos secuenciadores parciales o completos mediante adaptadores de secuenciación. Un adaptador de secuenciación puede incluir un código de barras molecular de muestra y un motivo secuenciador. Los adaptadores de secuenciación pueden proporcionarse en un conjunto que está separado de los adaptadores de bibliotecas. Los adaptadores de secuenciación en un conjunto dado pueden ser idénticos - es decir, contienen el mismo código de barras de muestra y motivo del secuenciador.

Los adaptadores de secuenciación pueden incluir motivos de identificación de muestras y motivos secuenciadores. Los motivos secuenciadores pueden incluir cebadores que son complementarios a una secuencia de iniciación de secuenciación. En algunos casos, los motivos del secuenciador también incluyen secuencias de células de flujo u otras secuencias que permiten que un polinucleótido se configure o disponga de una manera que permita que el polinucleótido sea secuenciado por un secuenciador.

Los adaptadores de bibliotecas y los adaptadores de secuenciación pueden ser cada uno adaptadores

parciales, es decir, contienen parte pero no todas las secuencias necesarias para permitir la secuenciación mediante una plataforma de secuenciación. Juntos proporcionan adaptadores completos. Por ejemplo, los adaptadores de bibliotecas pueden incluir motivos secuenciadores parciales o nulos, pero tales motivos secuenciadores se proporcionan mediante adaptadores de secuenciación.

Las **FIGs. 9A-9C** ilustran esquemáticamente un método para etiquetar una molécula de polinucleótido objetivo con adaptadores de bibliotecas. La **FIG. 9A** muestra un adaptador de bibliotecas como un adaptador parcial que contiene un sitio de hibridación del cebador en una de las cadenas y un código de barras molecular hacia otro extremo. El sitio de hibridación del cebador puede ser una secuencia de inicio de secuenciación para la secuenciación posterior. El adaptador de bibliotecas es menor o igual a 80 bases de nucleótidos de longitud. En la **FIG. 9B**, los adaptadores de bibliotecas se ligan en ambos extremos de la molécula de polinucleótidos objetivo para proporcionar una molécula de polinucleótido objetivo etiquetada. La molécula de polinucleótidos objetivo etiquetada puede someterse a amplificación de ácidos nucleicos para generar copias del objetivo. Después, en la **FIG. 9C**, se proporcionan adaptadores de secuenciación que contienen motivos secuenciadores y se hibridan con la molécula de polinucleótidos objetivo etiquetada. Los adaptadores de secuenciación contienen motivos de identificación de muestras. Los adaptadores de secuenciación pueden contener secuencias para permitir la secuenciación del objetivo etiquetado con un secuenciador dado.

D. <u>Secuenciación</u>

Los polinucleótidos etiquetados pueden secuenciarse para generar lecturas de secuencia (por ejemplo, como se muestra en el paso (106), **FIG. 1**). Por ejemplo, puede secuenciarse un polinucleótido dúplex etiquetado. Las lecturas de secuencia pueden generarse a partir de solo una cadena de un polinucleótido dúplex etiquetado. Alternativamente, ambas cadenas de un polinucleótido dúplex etiquetado pueden generar lecturas de secuencia. Las dos cadenas del polinucleótido dúplex etiquetado pueden comprender las mismas etiquetas. Alternativamente, las dos cadenas del polinucleótido dúplex etiquetado pueden comprender diferentes etiquetas. Cuando las dos cadenas del polinucleótido dúplex etiquetado se etiquetan de manera diferente, las lecturas de secuencia generadas de una cadena (por ejemplo, una cadena de Watson) pueden distinguirse de las lecturas de secuencia generadas otras cadenas (por ejemplo, una cadena de Crick). La secuenciación puede implicar la generación de múltiples lecturas de secuencia para cada molécula. Esto tiene lugar, por ejemplo, como resultado de la amplificación de cadenas de polinucleótidos individuales durante el proceso de secuenciación, por ejemplo, por PCR.

Los métodos descritos en la presente pueden comprender amplificación de polinucleótidos. La amplificación de polinucleótidos puede dar como resultado la incorporación de nucleótidos en una molécula de ácido nucleico o cebador formando de este modo una nueva molécula de ácido nucleico complementaria a un ácido nucleico plantilla. La molécula de polinucleótidos recién formada y su plantilla pueden usarse como plantillas para sintetizar polinucleótidos adicionales. Los polinucleótidos que se están amplificando pueden ser cualquier ácido nucleico, por ejemplo, ácidos desoxirribonucleicos, incluyendo ADN genómicos, ADNcs (ADN complementario), ADNcfs, y ADNs tumorales circulantes (ADNcts). Los polinucleótidos que se están amplificando también pueden ser ARNs. Como se usa en la presente, una reacción de amplificación puede comprender muchas rondas de replicación de ADN. Las reacciones de amplificación de ADN pueden incluir, por ejemplo, reacción en cadena de polimerasa (PCR). Una reacción de PCR puede comprender 2-100 "ciclos" de desnaturalización, apareamiento, y síntesis de una molécula de ADN. Por ejemplo, se pueden realizar 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, ó 15-20 ciclos durante el paso de amplificación. La condición de la PCR se puede optimizar en función del contenido de GC de las secuencias, incluidos los cebadores.

Las técnicas de amplificación de ácidos nucleicos pueden usarse con los ensayos descritos en la presente. Algunas técnicas de amplificación son metodologías de PCR que pueden incluir, pero no se limitan a, PCR en solución y PCR in situ. Por ejemplo, la amplificación puede comprender una amplificación basada en PCR. Alternativamente, la amplificación puede comprender una amplificación no basada en PCR. La amplificación del ácido nucleico plantilla puede comprender el uso de una o más polimerasas. Por ejemplo, la polimerasa puede ser una ADN polimerasa o una ARN polimerasa. En algunos casos, la amplificación de alta fidelidad se realiza como con el uso de polimerasa de alta fidelidad (por ejemplo, ADN Polimerasa Phusion® High-Fidelity) o protocolos de PCR. En algunos casos, la polimerasa puede ser una polimerasa de alta fidelidad. Por ejemplo, la polimerasa puede ser ADN polimerasa KAPA HiFi. La polimerasa también puede ser ADN polimerasa Phusion. La polimerasa puede usarse bajo condiciones de reacción que reducen o minimizan los sesgos de la amplificación, por ejemplo, debido a la longitud del fragmento, contenido de GC, etc.

La amplificación de una única cadena de un polinucleótido por PCR generará copias tanto de esa cadena como de su complemento. Durante la secuenciación, tanto la cadena como su complemento generarán lecturas de secuencia. Sin embargo, las lecturas de secuencia generadas del complemento de, por ejemplo, la cadena de Watson, pueden identificarse como tales ya que llevan el complemento de la parte de la etiqueta dúplex que etiquetó la cadena de Watson original. Por el contrario, una lectura de secuencia generada a partir de una cadena de Crick o su producto de amplificación llevará la parte de la etiqueta dúplex que etiquetó la cadena original de Crick. De esta

manera, una lectura de secuencia generada de un producto amplificado de un complemento de la cadena de Watson puede distinguirse de una lectura de secuencia del complemento generada de un producto de amplificación de la cadena de Crick de la molécula original.

Todos los polinucleótidos amplificados pueden enviarse a un dispositivo de secuenciación para secuenciación. Alternativamente, un muestreo, o subconjunto, de todos los polinucleótidos amplificados se envía a un dispositivo de secuenciación para secuenciación. Con respecto a cualquier polinucleótido de cadena doble original, puede haber tres resultados con respecto a la secuenciación. Primero, las lecturas de secuencia pueden generarse a partir de ambas cadenas complementarias de la molécula original (es decir, de tanto la cadena de Watson como de la cadena de Crick). Segundo, las lecturas de secuencia pueden generarse a partir de solo una de las dos cadenas complementarias (es decir, o de la cadena de Watson o de la cadena de Crick, pero no de ambas). Tercero, no se puede generar lectura de secuencia a partir de ninguna de las dos cadenas complementarias. Por consiguiente, contar las lecturas de secuencia únicas que mapean un locus genético subestimará el número de polinucleótidos de cadena doble en el mapeo de muestra original al locus. En la presente, se describen métodos para estimar los polinucleótidos no vistos y no contados.

El método de secuenciación puede ser una secuenciación masivamente paralela, es decir, secuenciación simultánea (o en rápida sucesión) de por lo menos 100, 1000, 10.000, 100.000, 1 millón, 10 millones, 100 millones o 1 billón de moléculas de polinucleótidos. Los métodos de secuenciación pueden incluir, pero no están limitados a: secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de moléculas individuales, secuenciación de nanoporos, secuenciación de semiconductores, secuenciación por ligación, secuenciación por hibridación, RNA-Seq (Illumina), Expresión Génica Digitaln (Helicos), secuenciación de próxima generación, Secuenciación de Moléculas Individuales por Síntesis (SMSS) (Helicos), secuenciación masivamente paralela, Matriz de Moléculas Individuales Clonales (Solexa), secuenciación shotgun, Secuenciación de Maxam-Gilbert o Sanger, paseo con cebadores, secuenciación usando plataformas PacBio, SOLiD, Ion Torrent o Nanopore y cualquier otro método de secuenciación conocido en la técnica.

Por ejemplo, pueden amplificarse polinucleótidos etiquetados dúplex mediante, por ejemplo, PCR (ver, por ejemplo, **FIG. 4A** los polinucleótidos marcados dúplex se denominan mm' y nn'). En la **Fig. 4A**, la cadena del polinucleótido dúplex que incluye la secuencia m lleva etiquetas de secuencia w e y, mientras que la cadena del polinucleótido dúplex que incluye la secuencia m ' lleva etiquetas de secuencia x y z. De manera similar, la cadena del polinucleótido dúplex que incluye la secuencia n lleva etiquetas de secuencia a y c, mientras que la cadena del polinucleótido dúplex que incluye la secuencia n 'lleva etiquetas de secuencia b y d. Durante la amplificación, cada cadena produce ella misma y su secuencia complementaria. Sin embargo, por ejemplo, una progenie de amplificación de la cadena original m que incluye la secuencia complementaria, m ', es distinguible de una progenie de amplificación de la cadena original m' ya que la progenie de la cadena original m tendrá la secuencia 5'-y'm'w'-3' y la progenie de la cadena m' original una cadena tendrá la secuencia 5'-zm'x-3'. La **FIG. 4B** muestra la amplificación con más detalle. Durante la amplificación, pueden introducirse errores en la progenie de amplificación, representados por puntos. La progenie de la aplicación se muestrea para secuenciación, de modo que no todas las cadenas producen lecturas de secuencia, lo que da como resultado las lecturas de secuencia indicadas. Como las lecturas de secuencia pueden venir de cualquiera de una cadena o su complemento, ambas secuencias y secuencias de complemento se incluirán en el conjunto de lecturas de secuencia. Debe observarse que es posible que un polinucleótido lleve la misma etiqueta en cada extremo. Por tanto, para una etiqueta "a" y un polinucleótido "m", una primera cadena podría etiquetarse a-m-a', y el complemento podría etiquetarse a-m'-a.

E. <u>Determinación de lecturas de secuencia de consenso</u>

Los métodos divulgados en la presente pueden comprender determinar lecturas de secuencia de consenso en lecturas de secuencia (por ejemplo, como se muestra en el paso (108), **FIG. 1**), como por reducción o redundancia de rastreo. La secuenciación de polinucleótidos amplificados puede producir lecturas de los varios productos de amplificación del mismo polinucleótido original, referidas como "lecturas redundantes". Identificando lecturas redundantes, se pueden determinar moléculas únicas en la muestra original. Si las moléculas en una muestra están etiquetadas de forma única, las lecturas generadas a partir de la amplificación de una única molécula original individual pueden identificar en base a su distinto código de barras. Ignorando los códigos de barras, las lecturas de moléculas originales únicas pueden determinarse en base a las secuencias al principio y al final de una lectura, opcionalmente en combinación con la longitud de la lectura. En ciertos casos, sin embargo, se puede esperar que una muestra tenga una pluralidad de moléculas originales que tienen las mismas secuencias de inicio finalización y la misma longitud. Sin códigos de barras, estas moléculas son difíciles de distinguir entre sí. Sin embargo, si una colección de polinucleótidos no está etiquetada de forma única (es decir, una molécula original comparte el mismo identificador con por lo menos otra molécula original), combinar la información de un código de barras con la secuencia de inicio/finalización y/o la longitud de polinucleótido aumenta significativamente la probabilidad de que cualquier lectura de secuencia pueda rastrearse de vuelta a un polinucleótido original. Esto se debe a que, en parte, incluso sin un etiquetado único, es improbable que dos polinucleótidos originales que tiene la misma secuencia de inicio/finalización y longitud también se etiqueten con el mismo identificador.

F. Colapso

El colapso permite la reducción del ruido (es decir, el fondo) que se genera en cada paso del proceso. Los métodos divulgados en la presente pueden comprender colapsar, por ejemplo, generar una secuencia consenso comparando múltiples lecturas de secuencia. Por ejemplo, las lecturas de secuencia generadas a partir de un único polinucleótido original pueden usarse para generar una secuencia de consenso de ese polinucleótido original. Las rondas iterativas de amplificación pueden introducir errores en los polinucleótidos de la progenie. También, la secuenciación puede no realizarse típicamente con fidelidad perfecta por lo que también se introducen errores de secuenciación en esta etapa. Sin embargo, la comparación de lecturas de secuencia de moléculas derivadas de una única molécula original, incluyendo las que tienen variantes de secuencia, puede analizarse para determinar la secuencia original, o "de consenso". Esto se puede hacer filogenéticamente. Las secuencias de consenso pueden generarse a partir de familias de lecturas de secuencia por cualquiera de una variedad de métodos. Dichos métodos incluyen, por ejemplo, métodos lineales o no lineales para construir secuencias de consenso (como votación (por ejemplo, votación sesgada), promedio, estadística, máxima a posteriori o detección de probabilidad máxima, programación dinámica, Bayesiano, Markov oculto o métodos de máquina con vector de soporte, etc.) derivados de la teoría de la comunicación digital, la teoría de la información o la bioinformática. Por ejemplo, si todas o la mayoría de las lecturas de secuencia que se remontan a una molécula original llevan la misma variante de secuencia, esa variante probablemente existía en la molécula original. Por otro lado, si existe una variante de secuencia en un subconjunto de lecturas de secuencia redundante, esa variante puede haberse introducido durante la amplificación/secuenciación y representa un artefacto que no existe en el original. Además, si solo las lecturas de secuencia derivadas de la cadena de Watson o Crick de un polinucleótido original contienen la variante, la variante puede haberse introducido por daño de un solo lado de ADN, error de PCR de primer ciclo o mediante polinucleótidos contaminantes que se amplificaron a partir de una muestra diferente.

Después de que los fragmentos se han amplificado y las secuencias de los fragmentos amplificados se han leído y alineado, los fragmentos se someten a una tipificación de base, por ejemplo, determinando para cada locus el nucleótido más probable. Sin embargo, variaciones en el número de fragmentos amplificados y fragmentos amplificados no vistos (por ejemplo, aquellos que no se han leído sus secuencias; las razones podrían ser demasiadas como errores de amplificación, errores de secuencia de lectura, demasiado largos, demasiado cortos, estar troceado, etc.) pueden introducir errores en las tipificaciones base. Si hay demasiados fragmentos amplificados no vistos con respecto a los fragmentos amplificados vistos (fragmentos amplificados que realmente se leen), la fiabilidad de las tipificaciones base puede disminuirse.

Por lo tanto, en la presente, se divulga un método para corregir el número de fragmentos no vistos en tipificaciones base. Por ejemplo, cuando la base llama al locus A (un locus arbitrario), primero se asume que hay N fragmentos amplificados. Las lecturas de secuencia pueden venir de dos tipos de fragmentos: fragmentos de cadena doble y fragmentos de cadena sencilla. Por lo tanto, asignamos N1, N2 y N3 como los números de cadenas dobles, cadenas sencillas y fragmentos no vistos, respectivamente. Por lo tanto, N=N1+N2+N3 (N1 y N2 son conocidos de las lecturas de secuencia, y N y N3 son desconocidos). Si la fórmula se resuelve para N (o N3), entonces se deducirá N3 (o N).

Se usa probabilidad para estimar N. Por ejemplo, asignamos "p" a la probabilidad de haber detectado (o haber leído) un nucleótido del locus A en una lectura de secuencia de una cadena sencilla.

Para lecturas de secuencia de cadenas dobles, la tipificación de nucleótidos desde un fragmento amplificado de cadena doble tiene una probabilidad de p * p = p^2, viendo que todas las cadenas dobles N1 tienen la siguiente ecuación: N1=N*(p^2).

Para lecturas de secuencia de cadena sencilla. Asumiendo que se ve uno de las 2 cadenas, y la otra no se ve, la probabilidad de ver una cadena es "p", pero la probabilidad de perder la otra cadena es (1-p). Además, al no distinguir la fuente de cadena sencilla del cebador 5 y la fuente del cebador 3, hay un factor de 2. Por lo tanto, la tipificación de nucleótidos de un fragmento amplificado de cadena sencilla tiene una probabilidad 2XpX(1-p). Por tanto, el ver todas las cadenas sencillas N2, tiene la ecuación siguiente: N2 = Nx2xpx(1-p).

"p" también es desconocido. Para resolver p, se usa la relación de N1 a N2 para resolver para "p":

$$R = \frac{N1}{N2} = \frac{Np^2}{2Np(1-p)} = \frac{p^2}{2p(1-p)} = \frac{p}{2(1-p)}$$

Una vez que se encuentra "p", se puede encontrar N. Después de que se encuentra N, se puede encontrar N3 = N-N1--N2.

Además la proporción de cadenas emparejadas frente a cadenas desemparejadas (que es una medida

post- colapso), hay información útil en la profundidad de lectura pre-colapso en cada locus. Esta información puede usarse para mejorar adicionalmente la tipificación para el recuento total de moléculas y/o aumentar la confianza de las variantes de tipificación.

Por ejemplo, la **FIG. 4C** demuestra lecturas de secuencia corregidas para secuencias complementarias. Las secuencias generadas a partir de una cadena de Watson original o una cadena de Crick original pueden diferenciarse en base a sus etiquetas dúplex. Las secuencias generadas a partir de la misma cadena original pueden agruparse. El examen de las secuencias puede permitir inferir la secuencia de la cadena original (la "secuencia de consenso"). En este caso, por ejemplo, la variante de la secuencia en la molécula nn' se incluye en la secuencia de consenso porque se incluyó en cada secuencia leída mientras que otras variantes se ve que son errores parásitos. Después de colapsar las secuencias, las parejas de polinucleótidos originales se pueden identificar en base a sus secuencias complementarias y etiquetas dúplex.

La **FIG. 5** demuestra confianza aumentada en la detección de variantes de secuencia emparejando lecturas de las cadenas de Watson y Crick. La secuencia nn' puede incluir una variante de secuencia indicada por un punto. En algunos casos, la secuencia pp' no incluye una variante de secuencia. La amplificación, secuenciación, reducción de redundancia y emparejamiento pueden dar como resultado tanto cadenas de Watson como de Crick de la misma molécula original incluyendo la variante de secuencia. Por el contrario, como resultado de errores introducidos durante la amplificación y el muestreo durante la secuenciación, la secuencia de consenso de la cadena p de Watson puede contener una variante de secuencia, mientras que la secuencia de consenso de la cadena p de Crick no lo hace. Es menos probable que la amplificación y la secuenciación introduzcan la misma variante en ambas cadenas (secuencia nn ') de un dúplex que en una cadena (secuencia pp'). Por lo tanto, es más probable que la variante en la secuencia pp 'sea un artefacto, y es más probable que la variante en la secuencia nn' exista en la molécula original.

Los métodos divulgados en la presente pueden usarse para corregir errores resultado de experimentos, por ejemplo, PCR, amplificación y/o secuenciación. Por ejemplo, dicho método puede comprender unir uno o más adaptadores de cadena doble a ambos extremos de un polinucleótido de cadena doble, proporcionando de este modo un polinucleótido de cadena doble etiquetado; amplificar el polinucleótido etiquetado de cadena doble; secuenciar ambas cadenas del polinucleótido etiquetado; comparar la secuencia de una cadena con su complemento para determinar cualquier error introducido durante la secuenciación; y corregir errores en la secuencia en base a (d). Los adaptadores usados en este método pueden ser cualquier adaptador divulgado en la presente, por ejemplo, adaptadores con forma de Y. El adaptador puede comprender cualquier código de barras (por ejemplo, códigos de barras distintos) divulgados en la presente.

G. <u>Mapeo</u>

Las lecturas de secuencia o las secuencias consenso pueden mapearse para uno o más loci genéticos seleccionados (por ejemplo, como se muestra en el paso (110), **FIG. 1**). Un locus genético puede ser, por ejemplo, una posición de nucleótido específica en el genoma, una secuencia de nucleótidos (por ejemplo, un marco de lectura abierto), un fragmento de un cromosoma, un cromosoma completo, o un genoma completo. Un locus genético puede ser un locus polimórfico. El locus polimórfico puede ser un locus en el que existe una variación de secuencia en la población y/o existe en un sujeto y/o una muestra. Un locus polimórfico puede generarse mediante dos o más secuencias distintas que coexisten en la misma localización del genoma. Las distintas secuencias pueden diferir entre sí por una o más sustituciones de nucleótidos, una deleción/inserción y/o una duplicación de cualquier número de nucleótidos, generalmente un número relativamente pequeño de nucleótidos, como menos de 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, ó 1 nucleótido(s), entre otros. Puede crearse un locus polimórfico mediante una única posición de nucleótido que varía dentro de la población, por ejemplo, una variación de nucleótido único (SNV) o un polimorfismo de nucleótido único (SNP).

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

En los métodos divulgados en la presente, el colapso puede realizarse antes o después del mapeo. En algunos aspectos, el colapso puede realizarse antes del mapeo. Por ejemplo, las lecturas de secuencia pueden agruparse en familias en base a sus etiquetas y una o más secuencias endógenas, sin tener en cuenta dónde se lee el mapa en el genoma. Entonces, los miembros de una familia pueden colapsarse en una secuencia de consenso. La secuencia de consenso puede generarse usando cualquier método de colapso descrito en la presente. Entonces la secuencia de consenso puede mapearse a localizaciones en el genoma. Las lecturas mapeadas a un locus pueden cuantificarse (por ejemplo, contarse). También puede determinarse el porcentaje de lecturas que llevan una mutación en un locus. Alternativamente, el colapso puede realizarse después del mapeo. Por ejemplo, todas las lecturas pueden mapearse primero al genoma. Luego las lecturas pueden agruparse en familias

en base a sus etiquetas y una o más secuencias endógenas. Como las lecturas se han mapeado al genoma, se pueden determinar las bases de consenso para cada familia en cada locus. En otros aspectos, se puede generar una secuencia de consenso para una cadena de una molécula de ADN (por ejemplo, para una cadena de Watson o una cadena de Crick). El mapeo puede realizarse antes o después de que se determine la secuencia de consenso para una cadena de la molécula de ADN. Se puede determinar el número de dobletes y singletes. Estos números pueden usarse para calcular moléculas no visibles. Por ejemplo, las moléculas no visibles pueden calcularse usando la ecuación siguiente: N=D+S+U; D=Np(2), S=N2pq, donde p=1-q, donde p es la probabilidad de ver; q es la probabilidad de perder una cadena.

H. Agrupamiento

Los métodos divulgados en la presente también pueden comprender lecturas de secuencia de agrupamiento. Las lecturas de secuencia se pueden agrupar en base a varios tipos de secuencias, por ejemplo, secuencias de una etiqueta de oligonucleótido (por ejemplo, un código de barras), secuencia de fragmentos de polinucleótido, o cualquier combinación. Por ejemplo, como se muestra en el paso (112) (**FIG. 1**), las lecturas de secuencia pueden agruparse como sigue: Lecturas de secuencia generadas a partir de una cadena de "Watson" y las generadas a partir de una cadena de "Crick" de un polinucleótido de cadena doble en la muestra son identificables en base a las etiquetas dúplex que llevan. De esta manera, una lectura de secuencia o secuencia de consenso de una cadena de Watson de un polinucleótido dúplex puede emparejarse con una lectura de secuencia o lectura de consenso de su cadena complementaria de Crick. Las lecturas de secuencia emparejadas se conocen como "Par".

Las lecturas de secuencia para las cuales no se puede encontrar lectura de secuencia correspondiente a una cadena complementaria entre las lecturas de secuencia se denominan "Singletes".

Los polinucleótidos de cadena doble para los que no se ha generado una lectura de secuencia para ninguna de las dos cadenas complementarias se denominan moléculas "invisibles".

I. Cuantificación

Los métodos divulgados en la presente también comprenden cuantificar lecturas de secuencia. Por ejemplo, como se muestra en el paso (114) (**FIG. 1**), se cuantifican, Pares y Singletes que mapean para un locus genético seleccionado, o a cada uno de una pluralidad de loci genéticos seleccionados, por ejemplo, contados.

La cuantificación puede comprender estimar el número de polinucleótidos en la muestra (por ejemplo, pares de polinucleótidos, Singletes de polinucleótidos o polinucleótidos no vistos). Por ejemplo, como se muestra en el paso (116) (**FIG. 1**), se estima el número de polinucleótidos de cadena doble en la muestra para la que no se generaron lecturas de secuencia (polinucleótidos "no vistos"). La probabilidad de que un polinucleótido de cadena doble no genere lecturas de secuencia puede determinarse en base al número relativo de Pares y Singletes en cualquier locus. Usando esta probabilidad, se puede estimar el número de polinucleótidos no vistos.

En el paso (118), una estimación del número total de polinucleótidos de cadena doble en una muestra que mapea para un locus seleccionado es la suma del número de Pares, el número de Singletes y el número de moléculas no vistas que mapean para el locus.

El número de moléculas originales no vistas en una muestra puede estimarse en base al número relativo de Pares y Singletes (**FIG. 2**). Con referencia a la **FIG. 2**, como ejemplo, se registran los recuentos para un locus genómico particular, Locus A, donde 1000 moléculas están emparejadas y 1000 moléculas están desemparejadas. Asumiendo una probabilidad uniforme, p, para que una cadena de Watson o Crick individual pueda pasar el proceso posterior a la conversión, se puede calcular la proporción de moléculas que no logran pasar por el proceso (no visibles) como sigue: Sea R = proporción de moléculas emparejadas a no emparejadas = 1, entonces $R=1=p^2/(2p(1-p))$. Esto implica que p=2/3 y que la cantidad de moléculas perdidas es igual a $(1-p)^2=1/9$. Por tanto, en este ejemplo, aproximadamente el 11% de las moléculas convertidas se pierden y nunca se detectan. Considerando otro locus genómico, Locus B, en la misma muestra donde 1440 moléculas están emparejadas y 720 están desemparejadas. Usando el mismo método, podemos inferir que el número de moléculas que se pierden es solo el 4%. Comparando las dos áreas, se puede suponer que el Locus A tenía 2000 moléculas únicas en comparación con 2160 moléculas en el Locus B - una diferencia de casi el 8%. Sin embargo, añadiendo correctamente en las moléculas perdidas en cada región, concluimos que hay 2000/(8/9)=2250 moléculas en el Locus A y 2160/.96=2250 moléculas en el Locus B. Por tanto, los recuentos en ambas regiones son realmente iguales. Esta corrección y, por tanto una sensibilidad mucho mayor puede alcanzarse convirtiendo las moléculas de ácidos nucleicos de cadena doble originales y manteniendo un seguimiento bioinformático de todas las que están emparejadas y desemparejadas al final del proceso. De manera similar, puede usarse el mismo procedimiento para inferir variaciones de números de copias verdaderas en regiones que parecen tener recuentos similares de moléculas únicas observadas. Al tomar en consideración el número de moléculas no visibles en las dos o más regiones, se vuelve aparente la variación del número de copias.

Adicionalmente a usar distribución binomial, otros métodos para estimar el número de moléculas no vistas incluyen distribuciones exponenciales, beta, gamma o empíricas basadas en la redundancia de lecturas de secuencia observadas. En el último caso, la distribución de recuentos de lectura para moléculas emparejadas y desemparejadas puede derivarse de dicha redundancia para inferir la distribución subyacente de moléculas de polinucleótidos originales en un locus particular. Esto puede llevar a menudo a una mejor estimación del número de moléculas no vistas.

J. <u>Detección de CNV</u>

Los métodos divulgados en la presente también comprender detectar CNV. Por ejemplo, como se muestran en el paso (120) (**FIG. 1**), una vez se ha determinado el número total de polinucleótidos que mapean para un locus, este número puede usarse en métodos estándar de determinación de CNV en el locus. Puede normalizarse una medida cuantitativa frente a una estándar. La estándar puede ser una cantidad de cualquier polinucleótido. En un método, puede estandarizarse una medida cuantitativa en un locus de prueba frente a una medida cuantitativa de polinucleótidos que mapean para un locus de control en el genoma, como den de número de copias conocidas. Las medidas cuantitativas pueden compararse frente a la cantidad de ácido nucleico en cualquier muestra divulgada en la presente. Por ejemplo, en otro método, la medida cuantitativa puede compararse frente a la cantidad de ácido nucleico en la muestra original. Por ejemplo, si la muestra original contenía 10.000 equivalentes de genes haploides, la medida cuantitativa puede compararse frente a una medida esperada para diploides. En otro método, la medida cuantitativa puede normalizarse frente a una medida de una muestra de control, y pueden compararse medidas normalizadas en diferentes loci.

En algunos casos, en los que se desea un análisis de la variación del número de copias, los datos de secuencia pueden: 1) alinearse con un genoma de referencia; 2) filtrarse y mapearse; 3) particionarse en ventanas o compartimentos de secuencia; 4) lecturas de cobertura contadas para cada ventana; 5) las lecturas de cobertura pueden luego normalizarse usando un algoritmo de modelado estocástico o estadístico; 6) y puede generarse un archivo de salida que refleje los estados discretos de número de copias en varias posiciones en el genoma. En otros casos, en los que se desea un análisis de mutaciones raras, los datos de secuencia pueden 1) alinearse con un genoma de referencia; 2) filtrarse y mapearse; 3) frecuencia de las bases de variantes calculada en base a lecturas de cobertura para esa base específica; 4) frecuencia de las bases variantes normalizada usando un algoritmo de modelado estocástico, estadístico o probabilístico; 5) y puede generarse un archivo de salida que refleje los estados de mutaciones en varias posiciones en el genoma.

Después de que se han determinado las proporciones de cobertura de lecturas de secuencia, se puede aplicar opcionalmente un algoritmo de modelado estocástico para convertir las proporciones normalizadas para cada región ventana en estados de número de copias discretos. 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 enrejado, decodificación de Viterbi, maximización de expectativas, metodologías de filtrado de Kalman o redes neuronales.

Los métodos divulgados en la presente pueden comprender detectar SNVs, CNVs, inserciones, deleciones y/o reordenamientos en una región específica en un genoma. La región genómica específica puede comprender una secuencia en un gen, como ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RBI, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, o NTRK1

En algunos casos, el método usa un panel que comprende exones de uno o más genes. El panel puede comprender también intrones de uno o más genes. El panel también puede comprender exones e intrones de uno o más genes. El uno o más genes pueden ser los divulgados. El panel puede comprender aproximadamente 80.000 bases que cubren un panel de genes. El panel puede comprender aproximadamente 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, o más bases.

En algunos aspectos, el número de copias de un gen puede reflejarse en la frecuencia de una forma genética del gen en una muestra. Por ejemplo, en un individuo sano, no se refleja ninguna variación del número de copias en una variante en un gen en un cromosoma (por ejemplo, heterocigosidad) que se detecta en aproximadamente el 50% de las moléculas detectadas en una muestra. También, en un individuo sano, la duplicación de un gen que lleva una variante puede reflejarse en la variante que se detecta en aproximadamente el 66% de las moléculas detectadas en una muestra. En consecuencia, si la carga tumoral en una muestra de ADN es del 10%, la frecuencia de una mutación somática en un gen en un cromosoma de células cancerosas, sin CNV, puede ser de aproximadamente el 5%. Lo contrario puede ser cierto en el caso de aneuploidía.

Los métodos divulgados en la presente se pueden usar para determinar si es más probable que una variante de secuencia esté presente en el nivel de la línea germinal o sea el resultado de una mutación celular somática, por ejemplo, en una célula cancerosa. Por ejemplo, una variante de secuencia en un gen detectado en niveles posiblemente compatibles con la heterocigosidad en la línea germinal es más probablemente que sea el producto de una mutación somática si también se detecta CNV en ese gen. En algunos casos, en la medida en que esperamos que una duplicación génica en la línea germinal lleve una variante compatible con la dosis genética (por ejemplo, 66% para trisomía en un locus), la amplificación del gen de detección con una dosis variante de secuencia que se desvía significativamente de esta cantidad esperada indica es más probable que la CNV se presente como resultado de una mutación de células somáticas.

Los métodos divulgados en la presente también pueden usarse para inferir la heterogeneidad tumoral en una situación en la que las variantes de secuencia en dos genes se detectan a diferentes frecuencias. Por ejemplo, la heterogeneidad tumoral puede inferirse cuando se detectan dos genes a diferentes frecuencias, pero sus números de copias son relativamente iguales. Alternativamente, puede inferirse la homogeneidad del tumor cuando la diferencia de frecuencia entre dos variantes de secuencia es consistente con la diferencia en el número de copias para los dos genes. Por tanto, por ejemplo, si se detecta una variante de EGFR al 11% y se detecta una variante de KRAS al 5% y no se detecta CNV en estos genes, la diferencia en la frecuencia es probable que refleje heterogeneidad tumoral (por ejemplo, todas las células tumorales llevan un mutante EGFR y la mitad de las células tumorales también llevan un mutante KRAS). Alternativamente, si el gen EGFR que lleva el mutante se detecta a 2 veces el número de copias normal, una interpretación es una población homogénea de células tumorales, cada célula llevando un mutante en los genes EGFR y KRASA, pero en la que el gen KRAS está duplicado

En respuesta a quimioterapia, una forma de tumor dominante puede eventualmente ceder a través de la selección darwiniana a células cancerosas que llevan mutantes que vuelven al cáncer no sensible al régimen de terapia. La aparición de estos mutantes de resistencia puede retrasarse mediante los métodos de esta divulgación. En una realización de este método, un sujeto se somete a uno o más ciclos de terapia pulsátil, cada ciclo de terapia pulsátil comprendiendo un primer período durante el cual se administra un fármaco a una primera cantidad y un segundo ciclo durante el cual el fármaco se administra a una segunda cantidad reducida. El primer período puede caracterizarse por una carga tumoral detectada por encima de un primer nivel clínico. El segundo período puede caracterizarse por una carga tumoral detectada por debajo de un segundo nivel clínico. El primer y segundo niveles clínicos pueden ser diferentes en diferentes ciclos de terapia pulsátil. Por ejemplo, el primer nivel clínico puede ser menor en ciclos sucesivos. Una pluralidad de ciclos puede incluir por lo menos 2, 3, 4, 5, 6, 7, 8 o más ciclos. Por ejemplo, el mutante BRAF V600E puede detectarse en polinucleótidos de una célula de enfermedad en una cantidad que indica una carga tumoral del 5% en ADNcf. La quimioterapia puede comenzar con dabrafenib. Pruebas posteriores pueden mostrar que la cantidad del mutante BRAF en el ADNc desciende por debajo del 0,5% o hasta niveles indetectables. En este punto, la terapia con dabrafenib puede detenerse o reducirse significativamente. Pruebas posteriores adicionales pueden descubrir que el ADN que lleva la mutación BRAF ha aumentado al 2.5% de polinucleótidos en ADNcf. En este punto, la terapia con dabrafenib puede reiniciarse, por ejemplo, al mismo nivel que el tratamiento inicial. Pruebas posteriores pueden descubrir que el ADN que lleva la mutación BRAF ha disminuido al 0,5% de polinucleótidos en ADNcf. De nuevo, la terapia con dabrafenib puede detenerse o reducirse. El ciclo puede repetirse una variedad de veces.

Una intervención terapéutica también puede cambiarse tras la detección del aumento de una forma mutante resistente a un fármaco original. Por ejemplo, los cánceres con la mutación EGFR L858R responden a terapia con erlotinib. Sin embargo, los cánceres con la mutación EGFR T790M son resistentes al erlotinib. Sin embargo, son sensibles a ruxolitinib. Un método de esta divulgación implica monitorizar cambios en el perfil tumoral y cambiar una intervención terapéutica cuando una variante genética asociada con la resistencia a fármacos aumenta a un nivel clínico predeterminado.

Los métodos divulgados pueden comprender un método de detectar la heterogeneidad de células de enfermedad a partir de una muestra que comprende polinucleótidos de células somáticas y células de enfermedad, el método comprendiendo: a) cuantificar polinucleótidos en la muestra que lleva una variante de secuencia en cada uno de una pluralidad de loci genéticos; b) determinar la CNV en cada uno de la pluralidad de loci genéticos; cantidades relativas diferentes de moléculas de enfermedad en un locus, donde la CNV indica una dosis genética de un locus en los polinucleótidos de la célula de enfermedad; c) determinar una medida relativa de la cantidad de polinucleótidos que llevan una variante de secuencia en un locus por dosis genética en el locus para cada uno de una pluralidad de loci; y d) comparar las medidas relativas en cada uno de la pluralidad de loci, en donde diferentes medidas relativas indican heterogeneidad tumoral. En los métodos divulgados en la presente, la dosis genética puede determinarse en base a una molécula total. Por ejemplo, si hay IX moléculas totales en un primer locus y 1.2X moléculas mapeadas a un segundo locus, entonces la dosis genética es 1.2. Las variantes en este locus pueden dividirse por 1.2. En algunos aspectos, el método divulgado en la presente puede usarse para detectar cualquier heterogeneidad de la célula de enfermedad, por ejemplo, heterogeneidad de la célula tumoral. Los métodos pueden usarse para detectar heterogeneidad de células de enfermedad de una muestra que comprende cualquier tipo de polinucleótidos, por ejemplo, ADNcf, ADN genómico, ADNc o ADNct. En los métodos, la cuantificación puede

comprender, por ejemplo, determinar el número o la cantidad relativa de los polinucleótidos. La determinación de CNV puede comprender mapear y normalizar diferentes cantidades relativas de moléculas totales a un locus.

En otro aspecto, en respuesta a quimioterapia, una forma de tumor dominante puede eventualmente ceder a través de la selección darwiniana a células cancerosas que portan mutantes que vuelven al cáncer insensible al régimen de terapia. La aparición de estos mutantes de resistencia puede retrasarse mediante los métodos divulgados en este momento. Los métodos divulgados en la presente pueden comprender un método que comprende: a) someter a un sujeto a uno o más ciclos de terapia pulsátil, comprendiendo cada ciclo de terapia pulsátil (i) un primer período durante el cual se administra un fármaco en una primera cantidad y (ii) un segundo período durante el cual el fármaco se administra en una segunda cantidad reducida; en donde (A) el primer período se caracteriza por una carga tumoral detectada por encima de un primer nivel clínico; y (B) el segundo período se caracteriza por una carga tumoral detectada por debajo de un segundo nivel clínico.

K. <u>Detección de Variantes de Secuencia</u>

Los sistemas y métodos descritos en la presente pueden usarse para detectar variantes de secuencia, por ejemplo, SNVs. Por ejemplo, una variante de secuencia puede detectarse a partir de secuencias de consenso de múltiples lecturas de secuencia, por ejemplo, de por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19, por lo menos 20, por lo menos 21, por lo menos 22, por lo menos 23, por lo menos 24, por lo menos 25, por lo menos 26, por lo menos 27, por lo menos 28, por lo menos 29, por lo menos 30, por lo menos 31, por lo menos 32, por lo menos 33, por lo menos 34, por lo menos 35, por lo menos 36, por lo menos 37, por lo menos 38, por lo menos 39, por lo menos 40, por lo menos 41, por lo menos 42, por lo menos 43, por lo menos 44, por lo menos 45, por lo menos 46, por lo menos 47, por lo menos 48, por lo menos 49, por lo menos 50, por lo menos 51, por lo menos 52, por lo menos 53, por lo menos 54, por lo menos 55, por lo menos 56, por lo menos 57, por lo menos 58, por lo menos 59, por lo menos 60, por lo menos 61, por lo menos 62, por lo menos 63, por lo menos 64, por lo menos 65, por lo menos 66, por lo menos 67, por lo menos 68, por lo menos 69, por lo menos 70, por lo menos 71, por lo menos 72, por lo menos 73, por lo menos 74, por lo menos 75, por lo menos 76, por lo menos 77, por lo menos 78, por lo menos 79, por lo menos 80, por lo menos 81, por lo menos 82, por lo menos 83, por lo menos 84, por lo menos 85, por lo menos 86, por lo menos 87, por lo menos 88, por lo menos 89, por lo menos 90, por lo menos 91, por lo menos 92, por lo menos 93, por lo menos 94, por lo menos 95, por lo menos 96, por lo menos 97, por lo menos 98, por lo menos 99, por lo menos 100, por lo menos 200, por lo menos 300, por lo menos 400, por lo menos 500, por lo menos 600, por lo menos 700, por lo menos 800, por lo menos 900, por lo menos 1000, por lo menos 2000, por lo menos 3000, por lo menos 4000, por lo menos 5000, por lo menos 6000, por lo menos 7000, por lo menos 8000, por lo menos 9000, por lo menos 10000 o más lecturas de secuencia.

Una secuencia consenso puede ser de lecturas de secuencia de un polinucleótido de cadena sencilla. Una secuencia de consenso también puede ser de lecturas de secuencia de una cadena de un polinucleótido de cadena doble (por ejemplo, lecturas de emparejamiento). En un método ejemplar, las lecturas de emparejamiento permiten identificar con confianza aumentada la existencia de una variante de secuencia en una molécula. Por ejemplo, si ambas cadenas de una pareja incluyen la misma variante, se puede estar razonablemente seguro de que la variante existía en la molécula original, como la posibilidad de que la misma variante se introduzca en ambas cadenas durante amplificación/secuenciación es rara. Por el contrario, si solo una cadena de una Pareja incluye la variante de secuencia, es más probable que sea un artefacto. De manera similar, la confianza de que un Singlete que lleva una variante de secuencia existiese en la molécula original es menor que la confianza si la variante existía en un Dúplex, ya que hay una mayor probabilidad de que la variante se pueda introducir una vez durante la amplificación/secuenciación.

Otros métodos de detección de la variación del número de copias y la detección de variantes de secuencia se describen enWO2014149134, WO2014039556.

Las lecturas de secuencia pueden colapsarse para generar una secuencia de consenso, que puede mapearse a una secuencia de referencia para identificar variantes genéticas, tales como CNV o SNV. Como alternativa, las lecturas de secuencia se mapean antes o incluso sin mapeo. En tal caso, las lecturas de secuencia pueden mapearse individualmente a la referencia para identificar una CNV o SNV.

La **FIG. 3** muestra una secuencia de referencia que codifica un Locus A genético. Los polinucleótidos en la FIG. 3 pueden tener forma de Y o pueden tener otras formas, como de horquilla.

En algunos casos, se puede determinar una variante de SNV o de nucleótido múltiple (MNV) mediante múltiples lecturas de secuencia en un locus dado (por ejemplo, base de nucleótidos) alineando las lecturas de secuencia que se corresponden con ese locus. Después, se mapea una pluralidad de bases de nucleótidos secuenciales de por lo menos un subconjunto de las lecturas de secuencia con la referencia a una SNV o MNV en una molécula de polinucleótidos o parte del mismo que se corresponde con las lecturas. La pluralidad de bases de

nucleótidos secuenciales puede abarcar una localización real, inferida o sospechada de SNV o MNV. La pluralidad de bases de nucleótidos secuenciales puede abarcar por lo menos 3, 4, 5, 6, 7, 8, 9, ó 10 bases de nucleótidos.

L. Detección/Cuantificación de Ácidos Nucleicos

Los métodos descritos en la presente pueden usarse para etiquetas fragmentos de ácidos nucleicos, como ácido desoxirribonucleico (ADN), con una eficacia extremadamente alta. Este etiquetado eficiente permite a una persona detectar de forma eficiente y precisa ADN raro en poblaciones heterogéneas de los fragmentos de ADN original (como en ADNcf). Un polinucleótido raro (por ejemplo, ADN raro) puede ser un polinucleótido que comprende una variante genética que tiene lugar en una población de polinucleótidos a una frecuencia de menos del 10%, 5%, 4%, 3%, 2%, 1%, o 0,1 % Un ADN raro puede ser un polinucleótido con una propiedad detectable a una concentración de menos del 50%, 25%, 10%, 5%, 1%, o 0,1%.

El etiquetado puede tener lugar en una reacción individual. En algunos casos, se pueden realizar dos o más reacciones y agruparlas juntas. Etiquetar cada fragmento de ADN original en una sola reacción puede dar como resultado en el etiquetado de manera que más del 50% (por ejemplo, 60%, 70%, 80%, 90%, 95% o 99%) de los fragmentos de ADN original estén etiquetados en ambos extremos con etiquetas que comprenden códigos de barras moleculares, proporcionando de este modo fragmentos de ADN etiquetados. El etiquetado también puede resultar en más del 30%, 35%, 40%, 45%, 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%, 95%, 96%, 97%, 98%, ó 99% de los fragmentos de ADN original etiquetados en ambos extremos con etiquetas que comprenden códigos de barras moleculares. El etiquetado también puede dar como resultado que el 100% de los fragmentos de ADN original estén etiquetados en ambos extremos con etiquetas que comprenden códigos de barras moleculares. El etiquetado también puede dar como resultado el etiquetado de un único extremo.

El etiquetado también puede tener lugar usando una cantidad en exceso de etiquetas en comparación con los fragmentos de ADN original. Por ejemplo, el exceso puede ser un exceso de por lo menos 5 veces. En otros casos el exceso puede ser por lo menos 1,25, 1,5, 1,75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 o más veces en exceso. El etiquetado puede comprender la unión de extremos romos o extremos adhesivos. El etiquetado también puede realizarse por PCR de hibridación. El etiquetado también puede realizarse en volúmenes de reacción baja, como 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 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, 95, 96, 97, 98, 99, ó 100 pico- y/o microlitros.

El método también puede incluir realizar una amplificación de alta fidelidad en los fragmentos de ADN etiquetados. Se puede usar cualquier ADN polimerasa de alta fidelidad. Por ejemplo, la polimerasa puede ser ADN polimerasa KAPA HiFi o ADN polimerasa Phusion.

Además, el método puede comprender enriquecer selectivamente un subconjunto de los fragmentos de ADN etiquetados. Por ejemplo, el enriquecimiento selectivo puede realizarse mediante técnicas de hibridación o amplificación. El enriquecimiento selectivo puede realizarse usando un soporte sólido (por ejemplo, perlas). El soporte sólido (por ejemplo, perlas) puede comprender sondas (por ejemplo, oligonucleótidos que hibridan específicamente con ciertas secuencias. Por ejemplo, las sondas pueden hibridar con ciertas regiones genómicas, por ejemplo, genes. En algunos casos, las regiones genómicas, por ejemplo, genes, pueden ser regiones asociadas con enfermedades, por ejemplo, cáncer. Después del enriquecimiento, el fragmento seleccionado puede unirse a cualquier adaptador de secuenciación divulgado en la presente. Por ejemplo, un adaptador de secuencia puede comprender una secuencia de células de flujo, un código de barras de muestra o ambos. En otro ejemplo, un adaptador de secuencia puede ser un adaptador con forma de horquilla y/o comprende un código de barras de la muestra. Además, los fragmentos resultantes pueden amplificarse y secuenciarse. En algunos casos, el adaptador no comprende una región de cebador de secuenciación.

El método puede incluir la secuenciación de una o ambas cadenas de los fragmentos de ADN. En un caso, ambas cadenas del fragmento de ADN se secuencian independientemente. Los fragmentos de ADN etiquetados, amplificados y/o enriquecidos selectivamente se secuencian para obtener lecturas de secuencia que comprenden información de secuencia de los códigos de barras moleculares y por lo menos una parte de los fragmentos de ADN original.

El método puede incluir reducir o rastrear la redundancia (como se ha descrito anteriormente) en las lecturas de secuencia para determinar lecturas de consenso que son representativas de cadenas sencillas de los fragmentos de ADN original. Por ejemplo, para reducir o rastrear la redundancia, el método puede incluir comparar lecturas de secuencia que tienen códigos de barras moleculares iguales o similares y el mismo extremo o similar de secuencias de fragmentos. El método puede comprender realizar un análisis filogenético en las lecturas de

secuencia que tienen los mismos o similares códigos de barras moleculares. Los códigos de barras moleculares pueden tener un código de barras con distancias de edición variables (incluyendo cualquier distancia de edición como se describe en la presente), por ejemplo, una distancia de edición de hasta 3. El extremo de las secuencias de fragmentos puede incluir secuencias de fragmentos que tienen una distancia de edición con distancias variables (incluyendo cualquier distancia de edición como se describe en la presente), por ejemplo, una distancia de edición de hasta 3.

El método puede comprender pre-clasificar las lecturas de secuencia de acuerdo con los códigos de barras moleculares y la información de secuencia. Por ejemplo, la pre-clasificación de las lecturas de secuencia de acuerdo con los códigos de barras moleculares y la información de secuencia puede realizarse de al menos un extremo de cada uno de los fragmentos de ADN original para crear pre-clasificaciones de lecturas de cadenas sencillas. El método puede comprender además en cada pre-clasificación, determinar una secuencia de un fragmento de ADN original dado entre los fragmentos de ADN original analizando las lecturas de secuencia.

En algunos casos, las lecturas de secuencia en cada pre-clasificación pueden colapsarse a una secuencia de consenso y posteriormente mapearse en un genoma. Como alternativa, las lecturas de secuencia pueden mapearse a un genoma antes de la pre-clasificación y posteriormente colapsarse a una secuencia de consenso.

El método puede comprender también lecturas de secuencia de clasificación en lecturas emparejadas y lecturas desemparejadas. Después de la clasificación, se puede cuantificar el número de lecturas emparejadas y lecturas desemparejadas que mapean para cada uno de los uno o más loci genéticos.

El método puede incluir cuantificar las lecturas de consenso para detectar y/o cuantificar el ADN raro, que se describen a lo largo de la presente. El método puede comprender detectar y/o cuantificar el ADN raro comparando un número de veces que aparece cada base en cada posición de un genoma representado por los fragmentos de ADN etiquetados, amplificados y/o enriquecidos.

El método puede comprender etiquetar los fragmentos de ADN original en una única reacción utilizando una biblioteca de etiquetas. La biblioteca puede incluir por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19, por lo menos 20, por lo menos 50, por lo menos 100, por lo menos 500, por lo menos 1000, por lo menos 5000, por lo menos 10000, o cualquier número de etiquetas como se describe a lo largo de la presente. Por ejemplo, la biblioteca de etiquetas puede incluir por lo menos 8 etiquetas. La biblioteca de etiquetas puede incluir 8 etiquetas (que pueden generar 64 combinaciones posibles diferentes). El método puede realizarse de tal manera que un alto porcentaje de fragmentos, por ejemplo, más del 50% (o cualquier porcentaje como se describe a lo largo de la presente) se etiqueten en ambos extremos, en donde cada una de las etiquetas comprende un código de barras molecular.

M. Procesamiento y/o Análisis de Ácidos Nucleicos

Los métodos descritos a lo largo de la presente pueden usarse para procesar y/o analizar una muestra de ácidos nucleicos de un sujeto. El método puede comprender exponer los fragmentos de polinucleótidos de la muestra de ácidos nucleicos a una pluralidad de moléculas de polinucleótidos para producir fragmentos de polinucleótidos etiquetados. La pluralidad de moléculas de polinucleótidos que puede usarse se describe a lo largo de la solicitud.

Por ejemplo, la pluralidad de moléculas de polinucleótidos puede ser cada una menor de o igual a 40 bases de ácidos nucleicos de longitud y tener secuencias de códigos de barras distintas con respecto a por lo menos 4 bases de ácidos nucleicos y una distancia de edición de por lo menos 1, en donde cada una de las secuencias de códigos de barras distintas está dentro de 20 bases de ácido nucleico de un extremo terminal de una respectiva de la pluralidad de moléculas de polinucleótidos, y en donde la pluralidad de moléculas de polinucleótidos no son adaptadores de secuenciación.

Los fragmentos de polinucleótidos marcados pueden someterse a reacciones de amplificación de ácidos nucleicos bajo condiciones que producen fragmentos de polinucleótidos amplificados como productos de la amplificación de los fragmentos de polinucleótidos etiquetados. Tras la amplificación, se determina la secuencia de nucleótidos de los fragmentos de polinucleótidos etiquetados amplificados. En algunos casos, las secuencias de nucleótidos de los fragmentos de polinucleótidos etiquetados amplificados se determinan sin el uso de reacción en cadena de polimerasa (PCR).

El método puede comprender analizar las secuencias de nucleótidos con un procesador informático programado para identificar una o más variantes genéticas en la muestra de nucleótidos del sujeto. Se puede identificar cualquier alteración genética, incluyendo pero no limitado a, cambio(s) de base, inserción(es), repetición(es), deleción(es), variación(es) del número de copia, modificación(es) epigenética(s), sitio(s) de unión al nucleosoma, cambio(s) de número de copia debido a origen(es) de replicación y transversión(es). Otras alteraciones

genéticas pueden incluir, pero no están limitadas a, una o más alteraciones genéticas asociadas a tumores.

Se puede sospechar que el sujeto de los métodos tenga una enfermedad. Por ejemplo, se puede sospechar que el sujeto tiene cáncer. El método puede comprender recoger una muestra de ácidos nucleicos de un sujeto. La muestra de ácidos nucleicos puede recogerse de sangre, plasma, suero, orina, saliva, excreciones mucosales, esputo, heces, fluido espinal cerebral, piel, cabello, sudor y/o lágrimas. La muestra de ácidos nucleicos puede ser una muestra de ácidos nucleicos libre de células. En algunos casos, la muestra de ácidos nucleicos se recoge de no más de 100 nanogramos (ng) de moléculas de polinucleótidos de cadena dobles del sujeto.

Los fragmentos de polinucleótidos pueden comprender moléculas de polinucleótidos de cadena doble. En algunos casos, la pluralidad de moléculas de polinucleótidos están acopladas a los fragmentos de polinucleótidos mediante ligación de extremos romos, ligación de extremos adhesivos, sondas de inversión molecular, reacción en cadena de polimerasa (PCR), PCR basada en ligación, PCR multiplexada, ligación de cadena sencilla, o circularización de cadena sencilla.

El método como se describe en la presente da como resultado un etiquetado de alta eficiencia de ácidos nucleicos. Por ejemplo, exponer los fragmentos de polinucleótidos de la muestra de ácidos nucleicos a la pluralidad de moléculas de polinucleótidos produce los fragmentos de polinucleótidos etiquetados con una eficiencia de conversión de por lo menos el 30%, por ejemplo, por lo menos el 50% (por ejemplo, 60%, 70%, 80%, 90%, 95%, ó 99%). Puede lograrse una eficiencia de conversión de por lo menos el 30%, 35%, 40%, 45%, 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%, 95%, 96%, 97%, 98%, ó 99%.

El método puede dar como resultado un fragmento de polinucleótidos etiquetado que comparte moléculas de polinucleótidos comunes. Por ejemplo, cualquiera de por lo menos el 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% ó 100% de los fragmentos de polinucleótidos etiquetados comparten una molécula de polinucleótidos común. El método puede comprender generar los fragmentos de polinucleótidos a partir de la muestra de ácidos nucleicos.

En algunos casos, la aplicación del método comprende amplificar los fragmentos de polinucleótidos etiquetados en presencia de cebadores correspondientes a una pluralidad de genes seleccionados del grupo que consiste de ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RBI, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, y NTRK1. Adicionalmente, puede amplificarse cualquier combinación de estos genes. Por ejemplo, pueden amplificarse 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, o los 54 de estos genes.

Los métodos descritos en la presente pueden comprender generar una pluralidad de lecturas de secuencia a partir de una pluralidad de moléculas de polinucleótidos. La pluralidad de moléculas de polinucleótidos puede cubrir loci genómicos de un genoma objetivo. Por ejemplo, los loci genómicos pueden corresponder a una pluralidad de genes como se enumera anteriormente. Además, los loci genómicos pueden ser cualquier combinación de estos genes. Cualquier locus genómico dado puede comprender por lo menos dos bases de ácidos nucleicos. Cualquier locus genómico dado también puede comprender una pluralidad de bases de ácidos nucleicos, por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, o más bases de ácidos nucleicos.

El método puede comprender agrupar con un procesador informático la pluralidad de lecturas de secuencia en familias. Cada una de las familias puede comprender lecturas de secuencia de uno de los polinucleótidos plantilla. Cada familia puede comprender lecturas de secuencia de solo uno de los polinucleótidos plantilla. Para cada una de las familias, las lecturas de secuencia pueden fusionarse para generar una secuencia de consenso. La agrupación puede comprender clasificar la pluralidad de lecturas de secuencia en familias identificando (i) códigos de barras moleculares distintos acoplados a la pluralidad de moléculas de polinucleótidos y (ii) similitudes entre la pluralidad de lecturas de secuencia, en donde cada familia incluye una pluralidad de secuencias de ácidos nucleicos que están asociados con una combinación distinta de códigos de barras moleculares y lecturas de secuencia similares o idénticas.

Una vez fusionada, puede designarse una secuencia de consenso en un locus genómico dado entre los loci genómicos. En cualquier loci genómico dado, puede determinarse cualquiera de los siguientes: i) variantes genéticas entre las designadas; ii) frecuencia de una alteración genética entre las designadas; iii) número total de designaciones; y iv) número total de alteraciones entre las designaciones. La designación puede comprender

designar por lo menos a una base de ácidos nucleicos en el locus genómico dado. La designación también puede comprender designar una pluralidad de bases de ácidos nucleicos en el locus genómico dado. En algunos casos, la designación puede comprender análisis filogenético, votación (por ejemplo, votación sesgada), ponderación, asignación de una probabilidad a cada lectura en el locus en una familia o designar la base con la probabilidad más alta. La secuencia de consenso puede generarse evaluando una medida cuantitativa o un nivel de significancia estadística para cada una de las lecturas de secuencia. Si se realiza una medida cuantitativa, el método puede comprender usar una distribución binomial, distribución exponencial, distribución beta o distribución empírica. Sin embargo, la frecuencia de la base en la localización particular también puede usarse para designar, por ejemplo, si el 51% o más de las lecturas es una "A" en la ubicación, entonces la base puede designarse "A" en esa particular ubicación. El método puede comprender además mapear una secuencia consenso en un genoma diana. la frecuencia de la base en la ubicación particular también se puede usar para designaciones, por ejemplo, si el 51% o más de las lecturas es una "A" en la localización, entonces la base puede designarse como una "A" en esa localización particular. El método puede comprender además mapear una secuencia de consenso en un genoma objetivo.

El método puede comprender además realizar designaciones de consenso en un locus genómico adicional entre los loci genómicos. El método puede comprender determinar una variación en el número de copias en uno de los locus genómicos dados y el locus genómico adicional en base a los recuentos en el locus genómico dado y el locus genómico adicional.

Los métodos descritos en la presente pueden comprender proporcionar moléculas de polinucleótidos plantilla y una biblioteca de moléculas de polinucleótidos adaptadores en un recipiente de reacción. Las moléculas de polinucleótidos adaptadores pueden tener de 2 a 1.000 secuencias de códigos de barras diferentes y en algunos casos no son adaptadores de secuenciación. Se describen a lo largo de la presente otras variaciones de las moléculas de polinucleótidos adaptadores que también pueden usarse en los métodos.

Las moléculas de polinucleótidos de los adaptadores pueden tener la misma etiqueta de muestra. Las moléculas de polinucleótidos adaptadores se pueden acoplar en ambos extremos de las moléculas de polinucleótidos plantilla. El método puede comprender el acoplamiento de las moléculas de polinucleótidos adaptadores a las moléculas de polinucleótidos plantilla con una eficiencia de por lo menos el 30%, por ejemplo, por lo menos el 50% (por ejemplo, 60%, 70%, 80%, 90%, 95%, o 99 %), etiquetando de este modo cada polinucleótido plantilla con una combinación de etiquetado que está entre de 4 a 1.000.000 de combinaciones de etiquetado diferentes, para producir moléculas de polinucleótidos etiquetadas. En algunos casos, la reacción puede tener lugar en un solo recipiente de reacción. La eficiencia del acoplamiento también puede ser de por lo menos el 30%, 35%, 40%, 45%, 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%, 95%, 96%, 97%, 98%, o 99%. El etiquetado puede ser etiquetado no único.

Las moléculas de polinucleótidos etiquetadas pueden someterse luego a una reacción de amplificación bajo condiciones que producirán moléculas de polinucleótidos amplificadas como productos de la amplificación de las moléculas de polinucleótidos etiquetadas. Las moléculas de polinucleótidos plantilla pueden ser de cadena doble. Además, las moléculas de polinucleótidos plantilla pueden ser de extremo romo. En algunos casos, la reacción de amplificación comprende amplificar no específicamente las moléculas de polinucleótidos etiquetadas. La reacción de amplificación puede comprender también el uso de un sitio de cebado para amplificar cada una de las moléculas de polinucleótidos etiquetadas. El sitio de cebado puede ser un cebador, por ejemplo, un cebador universal. El sitio de cebado también puede ser una muesca.

El método puede comprender también la secuenciación de las moléculas de polinucleótidos amplificadas. La secuenciación puede comprender (i) someter a las moléculas de polinucleótidos amplificadas a una reacción de amplificación adicional bajo condiciones que produzcan moléculas de polinucleótidos amplificadas adicionales como productos de la amplificación de las moléculas de polinucleótidos amplificadas, y/o (ii) secuenciar las moléculas de polinucleótidos amplificadas adicionales. La amplificación adicional puede realizarse en presencia de cebadores que comprenden secuencias de células de flujo, que producirán moléculas de polinucleótidos que son capaces de enlazar a una célula de flujo. La amplificación adicional también puede realizarse en presencia de cebadores que comprenden secuencias para adaptadores con forma de horquilla. Los adaptadores con forma de horquilla pueden unirse a ambos extremos de un fragmento de polinucleótido para generar una molécula circular, que puede secuenciarse múltiples veces. El método puede comprender además identificar variantes genéticas tras la secuenciación de las moléculas de polinucleótidos amplificadas.

El método puede comprender además separar moléculas de polinucleótidos que comprenden una o más secuencias dadas de las moléculas de polinucleótidos amplificadas, para producir moléculas de polinucleótidos enriquecidas. El método también puede comprender amplificar las moléculas de polinucleótidos enriquecidas con cebadores que comprenden las secuencias de las células de flujo. Esta amplificación con cebadores que comprenden secuencias de células de flujo producirá moléculas de polinucleótidos que son capaces de enlazar con una célula de flujo. La amplificación también puede realizarse en presencia de cebadores que comprenden

secuencias para adaptadores con forma de horquilla. Los adaptadores con forma de horquilla pueden unirse a ambos extremos de un fragmento de polinucleótido para generar una molécula circular, que puede secuenciarse múltiples veces.

Las secuencias de células de flujo o los adaptadores con forma de horquilla pueden añadirse por métodos de no amplificación como a través de la ligación de tales secuencias. Se pueden usar otras técnicas como métodos de hibridación, por ejemplo, salientes de nucleótidos.

El método puede realizarse sin alicuotar las moléculas de polinucleótidos etiquetadas. Por ejemplo, una vez que se hace la molécula de polinucleótido etiquetada, la amplificación y la secuenciación pueden tener lugar en el mismo tubo sin ninguna preparación adicional.

Los métodos descritos en la presente pueden ser útiles para detectar variaciones de nucleótido único (SNV), variaciones de número de copias (CNV), inserciones, deleciones y/o reordenamientos. En algunos casos, las SNV, CNV, inserciones, eliminaciones y/o reordenamientos, pueden asociarse con enfermedades, por ejemplo, cáncer.

N. Monitorización de un Estado del Paciente

Los métodos divulgados en la presente también pueden usarse para monitorizar el estado de la enfermedad de un paciente. La enfermedad de un sujeto puede monitorizarse a lo largo del tiempo para determinar una progresión de la enfermedad (por ejemplo, regresión). Los marcadores indicativos de la enfermedad pueden monitorizarse en una muestra biológica del sujeto, como una muestra de ADN libre de células.

Por ejemplo, la monitorización del estado de cáncer de un sujeto puede comprender (a) determinar una cantidad de uno o más de SNVs o números de copias de una pluralidad de genes (por ejemplo, en un exón), (b) repetir dicha determinación en diferentes puntos en el tiempo, y (c) determinar si hay una diferencia en el número de SNVs, nivel de SNVs, número o nivel de reordenaciones genómicas, o números de copias entre (a) y (b). Los genes se pueden seleccionar del grupo que consiste en ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, y NTRK1. Los genes pueden seleccionarse de cualquiera de 5, 10, 15, 20, 30, 40, 50, o todos los genes en este grupo.

O. Sensibilidad y Especificidad

Los métodos divulgados en la presente pueden usarse para detectar polinucleótidos de cáncer en una muestra, y cáncer en un sujeto, con altas medidas de conformidad, por ejemplo, alta sensibilidad y/o especificidad. Por ejemplo, tales métodos pueden detectar polinucleótidos de cáncer (por ejemplo, ADN raro) en una muestra a una concentración que es menor del 5%, 1%, 0,5%, 0,1%, 0,05% o 0,01%, a una especificidad de por lo menos el 99%, 99,9%, 99,99%, 99,999%, 99,9999%, ó 99,99999%. Tales polinucleótidos pueden ser indicativos de cáncer u otra enfermedad. Además, tales métodos pueden detectar polinucleótidos de cáncer en una muestra con un valor predictivo positivo de por lo menos el 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99,9%, 99,99%, 99,999%, o 99,9999%.

Los sujetos identificados como positivos en una prueba que son en realidad positivos se denominan verdaderos positivos (TP). Los sujetos identificados como positivos en una prueba que en realidad son negativos son referidos como falsos positivos (FP). Los sujetos identificados como negativos en una prueba que en realidad son negativos son referidos como verdaderos negativos (TN). Los sujetos identificados como negativos en una prueba que son en realidad positivos son referidos como falsos negativos (FN). La sensibilidad es el porcentaje de positivos reales identificados en una prueba como positivos. Esto incluye, por ejemplo, situaciones en las que uno debería haber encontrado una variante genética de cáncer y lo hizo. (Sensibilidad = TP/(TP+FN).) La especificidad es el porcentaje de negativos reales identificados en una prueba como negativos. Esto incluye, por ejemplo, situaciones en las que no se debería haber encontrado una variante genética de cáncer y no lo hizo. La especificidad puede calcularse usando la siguiente ecuación: Especificidad = TN/(TN+FP). El valor predictivo positivo (PPV) se puede medir por el porcentaje de sujetos que dan positivo que son verdaderos positivos. El PPV puede calcularse usando la siguiente ecuación: PPV = TP/(TP+FP). El valor predictivo positivo puede aumentarse aumentando la sensibilidad (por ejemplo, la probabilidad de que se detecte un positivo real) y/o la especificidad (por ejemplo, la probabilidad de no confundir un negativo real con un positivo).

Las bajas tasas de conversión de polinucleótidos en polinucleótidos etiquetados con adaptadores pueden comprometer la sensibilidad ya que disminuye la posibilidad de convertir, y por lo tanto detectar, objetivos de polinucleótidos raros. El ruido en una prueba puede comprometer la especificidad ya que aumenta el número de

falsos positivos detectados en una prueba. Tanto la baja tasa de conversión como el ruido comprometen el valor predictivo positivo, ya que disminuyen el porcentaje de positivos verdaderos y aumentan el porcentaje de falsos positivos.

Los métodos divulgados en la presente pueden alcanzar altos niveles de conformidad, por ejemplo, sensibilidad y especificidad, llevando a valores predictivos positivos altos. Los métodos para aumentar la sensibilidad incluyen conversión de polinucleótidos de alta eficacia en polinucleótidos etiquetados con adaptadores en una muestra. Los métodos para aumentar la especificidad incluyen reducir los errores de secuenciación, por ejemplo, por rastreo molecular.

Los métodos de la presente divulgación pueden usarse para detectar la variación genética en material genético de partida inicial no etiquetado de manera única (por ejemplo, ADN raro) a una concentración que es menor del 5%, 1%, 0,5%, 0,1%, 0,05. %, o 0.01%, a una especificidad de por lo menos el 99%, 99,9%, 99,99%, 99,999%, 99,9999% o 99,99999%. En algunos aspectos, los métodos pueden comprender además convertir polinucleótidos en el material de partida inicial con una eficiencia 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 70%, por lo menos el 80%) o por lo menos el 90%. Las lecturas de secuencias de polinucleótidos etiquetados pueden rastrearse posteriormente para generar secuencias de consenso para polinucleótidos con una tasa de error de no más del 2%, 1%, 0,1% o 0,01%.

2. <u>Métodos de Agrupamiento</u>

En la presente se divulgan métodos para detectar la variación del número de copias y/o las variantes de secuencia en uno o más loci genéticos en una muestra de prueba. Una realización se muestra en la **FIG. 8**. Típicamente, detectar la variación del número de copias implica determinar una medida cuantitativa (por ejemplo, un número absoluto o relativo) de mapeo de polinucleótidos a un locus genético de interés en un genoma de una muestra de prueba, y comparar ese número con una medida cuantitativa de mapeo de polinucleótidos a ese locus en una muestra de control. En ciertos métodos, la medida cuantitativa se determina comparando el número de moléculas en la muestra de prueba que mapea para un locus de interés con un número de moléculas en el mapeo de la muestra de prueba a una secuencia de referencia, por ejemplo, una secuencia que se espera esté presente en número de ploidías de tipo salvaje. En algunos ejemplos, la secuencia de referencia es HG19, estructura 37 o estructura 38. La comparación podría implicar, por ejemplo, determinar una proporción. Luego, esta medida se compara con una medida similar determinada en una muestra de control. Así, por ejemplo, si una muestra de prueba tiene una proporción de 1,5:1 para el locus de interés frente al locus de referencia, y una muestra de control tiene una proporción de 1:1 para los mismos loci, se puede concluir que la muestra de prueba muestra poliploidía en el locus de interés.

Cuando la muestra de prueba y la muestra de control se analizan por separado, el flujo de trabajo puede introducir distorsiones entre los números finales en las muestras de control y de prueba.

En un método divulgado en la presente (por ejemplo, diagrama de flujo 800), se proporcionan polinucleótidos a partir de una muestra de prueba y de control (802). Los polinucleótidos en una muestra de prueba y los que están en una muestra de control se etiquetan con etiquetas que identifican a los polinucleótidos como originarios de la muestra de prueba o control (una etiqueta fuente). (804.) La etiqueta puede ser, por ejemplo, una secuencia de polinucleótidos o código de barras que identifique inequívocamente la fuente.

Los polinucleótidos en cada una de las muestras de control y prueba también pueden etiquetarse con etiquetas identificadoras que serán llevadas por toda la progenie de amplificación de un polinucleótido. La información de las secuencias de inicio y finalización de un polinucleótido y las etiquetas identificadoras pueden identificar lecturas de secuencia de polinucleótidos amplificados de una molécula inicial original. Cada molécula puede etiquetarse de manera única en comparación con otras moléculas en la muestra. Alternativamente, cada molécula no necesita ser etiquetada de manera única en comparación con otras moléculas en la muestra. Es decir, el número de secuencias identificadoras diferentes puede ser menor que el número de moléculas en la muestra. Combinando información de identificadores con información de secuencia de inicio/finalización, la probabilidad de confundir dos moléculas que tienen la misma secuencia de inicio/finalización disminuye significativamente.

El número de identificadores diferentes usados para etiquetar un ácido nucleico (por ejemplo, ADNcf) puede depender del número de equivalentes de genoma haploide diferentes. Se pueden usar diferentes identificadores para etiquetar por lo menos 2, por lo menos 10, por lo menos 100, por lo menos 200, por lo menos 300, por lo menos 400, por lo menos 500, por lo menos 600, por lo menos 700, por lo menos 800, por lo menos 900, por lo menos 1.000, por lo menos 2.000, por lo menos 3.000, por lo por lo menos 4.000, por lo menos 5.000, por lo menos 6.000, por lo menos 7.000, por lo menos 8.000, por lo menos 9.000, por lo menos 10.000 o más equivalentes de genoma haploide diferentes. Por consiguiente, el número de identificadores diferentes usados para etiquetar una muestra de ácido nucleico, por ejemplo, ADN libre de células de 500 a 10.000 equivalentes de genoma haploide diferentes y estar entre cualquiera de 1, 2, 3, 4 y 5 y no más de 100, 90, 80, 70, 60, 50, 40 ó 30. Por ejemplo, el número de

identificadores diferentes usados para etiquetar una muestra de ácido nucleico de 500 a 10.000 equivalentes de genoma haploide diferentes puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 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, 95, 96, 97, 98, 99, 100 o menos.

Los polinucleótidos se pueden etiquetar por ligación de adaptadores que comprenden las etiquetas o identificadores antes de la amplificación. La ligación puede realizarse usando un enzima, por ejemplo, una ligasa. Por ejemplo, el etiquetado puede realizarse usando una ADN ligasa. La ADN ligasa puede ser una ADN ligasa de T4, una ADN ligasa de E. coli y/o una ligasa de mamífero. La ligasa de mamífero puede ser ADN ligasa I, ADN ligasa III o ADN ligasa IV. La ligasa también puede ser una ligasa termoestable. Las etiquetas pueden ligarse a un extremo romo de un polinucleótido (ligación de extremos romos). Alternativamente, las etiquetas pueden ligarse a un extremo adhesivo de un polinucleótido (ligación del extremo adhesivo). Los polinucleótidos pueden etiquetarse mediante ligación de extremos romos usando adaptadores (por ejemplo, adaptadores que tienen extremos bifurcados). Se puede lograr alta eficiencia de ligación usando un exceso alto de adaptadores (por ejemplo, más de 1.5X, más de 2X, más de 3X, más de 4X, más de 5X, más de 6X, más de 7X, más de 8X, más de 9X, más de 10X, más de 11X, más de 12X, más de 13X, más de 14X, más de 15X, más de 20X, más de 25X, más de 30X, más de 35X, más de 40X, más de 45X, más de 50X, más de 55X, más de 60X, más de 65X, más de 70X, más de 75X, más de 80X, más de 85X, más de 90X, más de 95X, o más de 100).

Una vez etiquetados con etiquetas que identifican la fuente de polinucleótidos, pueden agruparse los polinucleótidos de diferentes fuentes (por ejemplo, muestras diferentes). Después del agrupamiento, los polinucleótidos de diferentes fuentes (por ejemplo, muestras diferentes) pueden distinguirse por una medición usando las etiquetas, incluyendo cualquier proceso de medición cuantitativa. Por ejemplo como se muestra en (806) (**FIG. 8**), pueden agruparse los polinucleótidos de la muestra de control y la muestra de control. Las moléculas agrupadas pueden someterse a secuenciación (808) y flujo de trabajo bioinformático. Ambas se someterán a las mismas variaciones en el proceso y, por lo tanto, se reduce cualquier sesgo diferencial. Como las moléculas que se originan de las muestras de control y de prueba están etiquetadas de manera diferente, pueden distinguirse en cualquier proceso de medición cuantitativa.

La cantidad relativa de muestra de control y de prueba agrupada puede variarse. La cantidad de muestra de control puede ser la misma que la cantidad de muestra de prueba. La cantidad de muestra de control también puede ser mayor que la cantidad de muestra de prueba. Alternativamente, la cantidad de muestra de control puede ser más pequeña que la cantidad de muestra de prueba. Cuanto menor sea la cantidad relativa de una muestra al total, menos etiquetas de identificación se necesitarán en el proceso de etiquetado original. Puede seleccionarse un número para reducir a niveles aceptables la probabilidad de que dos moléculas iniciales que tienen las mismas secuencias de inicio/finalización lleven la misma etiqueta de identificación. Esta probabilidad puede ser inferior al 10%, inferior al 1%, inferior al 0,1% o inferior al 0,01%. La probabilidad puede ser inferior al 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, o 1%.

Los métodos divulgados en la presente también pueden comprender lecturas de secuencia de agrupamiento. Por ejemplo, el flujo de trabajo bioinformático puede incluir lecturas de secuencia de agrupamiento producidas a partir de la progenie de una molécula inicial única, como se muestra en (810) (**FIG. 8**). Esto puede implicar cualquiera de los métodos de reducción de la redundancia descritos en la presente. Las moléculas procedentes de muestras de prueba y control pueden diferenciarse en función de las etiquetas de origen que llevan (812). Las moléculas que mapean para un locus objetivo se cuantifican para tanto las moléculas originadas de prueba como originadas de control (812). Esto puede incluir los métodos de normalización tratados en la presente, por ejemplo, en los que los números en un locus objetivo están normalizados frente a los números en un locus de referencia.

Las cantidades normalizadas (o brutas) en un locus objetivo de muestras de prueba y control se comparan para determinar la presencia de variación del número de copias (814).

3. <u>Sistemas de Control Informáticos</u>

La presente divulgación proporciona sistemas de control informáticos que se programan para implementar los métodos de divulgación. La **FIG.** 6 muestra un sistema informático 1501 que está programado o configurado de otra manera 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 de muestras y el análisis de muestras, incluyendo la secuenciación de ácidos nucleicos. El sistema informático 1501 puede ser un dispositivo electrónico de un usuario o un sistema informático que está localizado de manera remota con respecto al dispositivo electrónico. El dispositivo electrónico puede ser un dispositivo electrónico móvil.

El sistema informático 1501 incluye una unidad central de procesamiento (CPU, también "procesador" y

"procesador informático" en la presente) 1505, que puede ser un procesador de núcleo único o multi-núcleo, o una pluralidad de procesadores para procesamiento en paralelo. El sistema informático 1501 también incluye memoria o localización de memoria 1510 (por ejemplo, memoria de acceso aleatorio, memoria de solo lectura, memoria flash), la unidad de almacenamiento electrónico 1515 (por ejemplo, disco duro), interfaz de comunicación 1520 (por ejemplo, adaptador de red) para comunicarse con uno o más sistemas distintos, y dispositivos periféricos 1525, como caché, otra memoria, adaptadores de almacenamiento de datos y/o 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 un depósito de datos) para almacenar datos. El sistema informático 1501 puede estar acoplado operativamente a una red informática ("red") 1530 con la ayuda de la interfaz de comunicación 1520. La red 1530 puede ser la Internet, una internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. La red 1530 en algunos casos es una red de telecomunicaciones y/o datos. La red 1530 puede incluir uno o más servidores informáticos, que pueden habilitar la computación distribuida, como computación en la nube. La red 1530, en algunos casos con la ayuda del sistema informático 1501, puede implementar una red de pares, que puede permitir que los dispositivos acoplados al sistema informático 1501 se comporten como un cliente o un servidor.

La CPU 1505 puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden incorporarse en un programa o software. Las instrucciones pueden almacenarse en una localización de memoria, como la memoria 1510. Las instrucciones pueden dirigirse a la CPU 1505, que posteriormente puede programar o configurar de otra manera la CPU 1505 para implementar los métodos de la presente divulgación. Ejemplos de operaciones realizadas por la CPU 1505 pueden incluir recuperación, decodificación, ejecución y reescritura.

La CPU 1505 puede ser parte de un circuito, como un circuito integrado. Pueden incluirse en el circuito uno o más de otros componentes del sistema 1501. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC).

La unidad de almacenamiento 1515 puede almacenar archivos, como controladores, bibliotecas y programas guardados. 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 localizadas en un servidor remoto que está en comunicación con el sistema informático 1501 a través de una intranet o de Internet.

El sistema informático 1501 puede comunicarse con uno o más sistemas informáticos remotos a través de la red 1530. Por ejemplo, el sistema informático 1501 puede comunicarse con un sistema informático remoto de un usuario (por ejemplo, un operador). Ejemplos de sistemas informáticos remotos incluyen ordenadores personales (por ejemplo, PC portátil), tableta o tableta PC (por ejemplo, iPad de Apple®, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, iPhone de Apple®, dispositivo habilitado para Android, Blackberry®), o asistentes digitales personales. El usuario puede acceder al sistema informático 1501 a través de la red 1530.

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

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

Los aspectos de los sistemas y métodos proporcionados en la presente, como el sistema informático 1501, pueden incorporarse en la programación. Varios aspectos de la tecnología pueden considerarse como "productos" o "artículos de fabricación" típicamente en forma de código ejecutable por máquina (o procesador) y/o datos asociados que se llevan a cabo o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina puede almacenarse en una unidad de almacenamiento electrónico, dicha memoria (por ejemplo, memoria de solo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir cualquiera o la totalidad de la memoria tangible de los ordenadores, procesadores o similares, o módulos asociados de los mismos, como varias memorias semiconductoras, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación del software. Todo o partes del software puede comunicarse a veces a través de Internet o varias otras redes de

telecomunicación. 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 huésped a la plataforma informática de un servidor de aplicaciones. Por lo tanto, otro tipo de medio que puede admitir los elementos de software incluye ondas ópticas, eléctricas y electromagnéticas, como las usadas en interfaces físicas entre dispositivos locales, a través de redes terrestres de cable y ópticas y en varios enlaces aéreos. Los elementos físicos que llevan tales ondas, como enlaces por cable o inalámbricos, enlaces ópticos o similares, también se pueden considerar como medios que llevan el software. Como se usa en la presente, a menos que esté restringido a medios de "almacenamiento" no transitorios, tangibles, los términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participe en la provisión de instrucciones a un procesador para su ejecución.

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

El sistema informático 1501 puede incluir o estar en comunicación con una pantalla electrónica 1535 que comprende una interfaz de usuario (UI) 1540. La UI puede permitir a un usuario establecer varias condiciones para los métodos descritos en la presente, por ejemplo, PCR o condiciones de secuenciación. Los ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en la web.

Los métodos y sistemas de la presente divulgación pueden implementarse por medio de uno o más algoritmos. Se puede implementar un algoritmo por medio de software tras la ejecución por la unidad central de procesamiento 1505. El algoritmo puede, por ejemplo, procesar las lecturas para generar una secuencia de consecuencia.

La **FIG. 7** ilustra esquemáticamente otro sistema para analizar una muestra que comprende ácidos nucleicos de un sujeto. El sistema incluye un secuenciador, software bioinformático y conexión a Internet para el análisis de informes mediante, por ejemplo, un dispositivo manual o un ordenador de sobremesa.

Se divulga en la presente un sistema para analizar una molécula de ácidos nucleicos objetivo de un sujeto, que comprende: una interfaz de comunicación que recibe lecturas de secuencias de ácidos nucleicos para una pluralidad de moléculas de polinucleótidos que cubren loci genómicos de un genoma objetivo; memoria informática que almacena las lecturas de secuencias de ácidos nucleicos para la pluralidad de moléculas de polinucleótidos recibidas por la interfaz de comunicación; y un procesador informático acoplado operativamente a la interfaz de comunicación y la memoria y programado para (i) agrupar la pluralidad de lecturas de secuencia en familias, en donde cada familia comprende lecturas de secuencia de uno de los polinucleótidos plantilla, (ii) para cada una de las familias, fusionar lecturas de secuencias para generar una secuencia de consenso, (iii) designar la secuencia de consenso en un locus genómico dado entre los loci genómicos, y (iv) detectar en el locus genómico dado cualquiera de las variantes genéticas entre las designaciones, la frecuencia de una alteración genética entre las designaciones, el número total de designaciones; y número total de alteraciones entre las designaciones, en donde los loci genómicos corresponden a una pluralidad de genes seleccionados del grupo que consiste de ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RBI, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, y NTRK1. Las diferentes variaciones de cada componente del sistema se describen a lo largo de la divulgación dentro de los métodos y composiciones. Estos componentes individuales y variaciones de los mismos también son aplicables en este sistema.

4. Kits

Kits que comprenden las composiciones como se describe en la presente. Los kits pueden ser útiles para realizar los métodos como se describe en la presente. En la presente se divulga un kit que comprende una pluralidad de sondas de oligonucleótidos que hibridan selectivamente con por lo menos 5, 6, 7, 8, 9, 10, 20, 30, 40 o todos los genes seleccionados del grupo que consiste de ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RBI, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, y NTRK1. El número de genes a los que las sondas de oligonucleótidos pueden hibridar selectivamente puede variar. Por ejemplo, el número de genes puede comprender 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, o 54. El kit puede incluir un recipiente que incluye la pluralidad de sondas de oligonucleótidos e instrucciones para realizar cualquiera de los métodos descritos en la presente.

Las sondas de oligonucleótidos pueden hibridar selectivamente a regiones de exones de los genes, por ejemplo, de por lo menos 5 genes. En algunos casos, las sondas de oligonucleótidos pueden hibridar selectivamente a por lo menos 30 exones de los genes, por ejemplo, de por lo menos 5 genes. En algunos casos, las sondas múltiples pueden hibridar selectivamente a cada uno de los por lo menos 30 exones. Las sondas que hibridan con cada exón pueden tener secuencias que se superponen con al menos 1 otra sonda. En algunas realizaciones, las oligosondas pueden hibridar selectivamente a regiones no codificantes de genes divulgados en la presente, por ejemplo, regiones intrónicas de los genes. Las oligosondas también pueden hibridar selectivamente a regiones de genes que comprenden tanto regiones exónicas como intrónicas de los genes divulgados en la presente.

Se pueden dirigir cualquier número de exones a las sondas de oligonucleótidos. Por ejemplo, pueden dirigirse por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290,, 295, 300, 400, 500, 600, 700, 800, 900, 1.000, o más, exones.

El kit puede comprender por lo menos 4, 5, 6, 7 u 8 adaptadores de bibliotecas diferentes que tienen distintos códigos de barras moleculares distintos y códigos de barras de muestra idénticos. Los adaptadores de bibliotecas pueden no ser adaptadores de secuenciación. Por ejemplo, los adaptadores de bibliotecas no incluyen secuencias de células de flujo o secuencias que permiten la formación de giros de horquilla para la secuenciación. Las diferentes variaciones y combinaciones de códigos de barras moleculares y códigos de barras de muestra se describen a lo largo de la presente y son aplicables al kit. Además, en algunos casos, los adaptadores no son adaptadores de secuenciación. Adicionalmente, los adaptadores provistos con el kit pueden comprender también adaptadores de secuenciación. Un adaptador de secuenciación puede comprender una secuencia que hibrida con uno o más cebadores de secuenciación. Un adaptador de secuenciación puede comprender además una secuencia que se hibrida con un soporte sólido, por ejemplo, una secuencia de células de flujo. Por ejemplo, un adaptador de secuencia puede ser un adaptador de células de flujo. Los adaptadores de secuenciación pueden unirse a uno o ambos extremos de un fragmento de polinucleótido. En algunos casos, el kit puede comprender por lo menos 8 adaptadores de bibliotecas diferentes que tienen códigos de barras moleculares distintos y códigos de barras de muestra idénticos. Los adaptadores de bibliotecas pueden no ser adaptadores de secuenciación. El kit puede incluir además un adaptador de secuenciación que tiene una primera secuencia que hibrida selectivamente a los adaptadores de biblioteca y una segunda secuencia que hibrida selectivamente a una secuencia de células de flujo. En otro ejemplo, un adaptador de secuenciación puede tener forma de horquilla. Por ejemplo, el adaptador con forma de horquilla puede comprender una parte de cadena doble complementaria y una parte de giro, donde la parte de cadena doble se puede unir (por ejemplo, ligar) a un polinucleótido de cadena doble. Los adaptadores de secuenciación con forma de horquilla pueden unirse a ambos extremos de un fragmento de polinucleótido para generar una molécula circular, que puede secuenciarse múltiples veces. Un adaptador de secuenciación puede ser de hasta 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 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, 95, 96, 97, 98, 99, 100,o más bases de extremo a extremo. El adaptador de secuenciación puede comprender 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases de extremo a extremo. En un ejemplo particular, el adaptador de secuenciación puede comprender 20-30 bases de extremo a extremo. En otro ejemplo, el adaptador de secuenciación puede comprender 50-60 bases de extremo a extremo. Un adaptador de secuenciación puede comprender uno o más códigos de barras. Por ejemplo, un adaptador de secuenciación puede comprender un código de barras de muestra. El código de barras de muestra puede comprender una secuencia predeterminada. Los códigos de barras de muestra se pueden usar para identificar la fuente de los polinucleótidos. El código de barras de muestra puede ser de por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, o más (o cualquier longitud como se describe a lo largo de la presente) bases de ácidos nucleicos, por ejemplo, por lo menos 8 bases. El código de barras puede ser de secuencias

contiguas o no contiguas, como se ha descrito con anterioridad.

Los adaptadores de bibliotecas pueden tener extremos romos y forma de Y, y pueden tener menos de o igual a 40 bases de ácidos nucleicos de longitud. Pueden encontrarse otras variaciones de la misma en la presente y son aplicables al kit.

EJEMPLOS

Ejemplo 1. Métodos para detección de variación del número de copias

Recogida de sangre

Se recogen muestras de sangre de 10-30 ml a temperatura ambiente. Las muestras se centrifugan para eliminar células. El plasma se recoge tras la centrifugación.

Extracción de ADNcf

La muestra se somete a digestión con proteinasa K. El ADN se precipita con isopropanol. El ADN se captura en una columna de purificación de ADN (por ejemplo, un QIAamp DNA Blood Mini Kit) y se eluye en una solución de 100 µl. Los ADN por debajo de 500 pb se seleccionan con captura de perlas magnéticas Ampure SPRI (PEG/sal). La producción resultante se suspende en 30 µl de H_2O. Se verifica la distribución por tamaño (pico principal = 166 nucleótidos, pico menor = 330 nucleótidos) y se cuantifica. 5 ng de ADN extraído contienen aproximadamente 1700 equivalentes de genoma haploide ("HGE"). La correlación general entre la cantidad de ADN y HGE es la siguiente: 3 pg DNA = 1 HGE; 3 ng DNA = 1K HGE; 3 mg DNA = 1M HGE; 10 pg DNA = 3 HE; 10 µg DNA = 3K HGE; 10 mg DNA = 3M HGE

Preparación de biblioteca de "Molécula Única"

Se realiza etiquetado de ADN de alta eficacia (> 80%) se realiza por reparación y ligación de extremos romos con 8 octómeros diferentes (es decir, 64 combinaciones) con adaptadores de horquilla sobrecargados. Se usan 2,5 ng de ADN (es decir, aproximadamente 800 HGE) como material de partida. Cada adaptador de horquilla comprende una secuencia aleatoria en su parte no complementaria. Ambos extremos de cada fragmento de ADN se unen con adaptadores de horquilla. Cada fragmento etiquetado puede identificarse por la secuencia aleatoria en los adaptadores de horquilla y una secuencia endógena 10 p en el fragmento.

El ADN marcado se amplifica por 10 ciclos de PCR para producir aproximadamente 1 -7 µg de ADN que contienen aproximadamente 500 copias de cada uno de los 800 HGE en el material de partida.

Pueden realizarse optimización del tampón, optimización de la polimerasa y reducción del ciclo para optimizar las reacciones de PCR. El sesgo de amplificación, por ejemplo, el sesgo no específico, el sesgo GC y/o el sesgo de tamaño también se reducen por optimización. El ruido(s) (por ejemplo, errores introducidos por la polimerasa) se reducen mediante el uso de polimerasas de alta fidelidad.

La Biblioteca puede prepararse usando métodos Verniata o Sequenom.

Las secuencias se pueden enriquecer como sigue: los ADNs con regiones de interés (ROI) se capturan usando perlas etiquetadas con biotina con sonda a ROIs. Los ROIs se amplifican con 12 ciclos de PCR para generar una amplificación de 2000 veces. El ADN resultante se desnaturaliza y diluye a 8 pM y se carga en un secuenciador Illumina.

Secuenciación masivamente paralela

Se usa del 0,1 al 1% de la muestra (aproximadamente 100 pg) para la secuenciación.

Bioinformática digital

Las lecturas de secuencia se agrupan en familias, con aproximadamente 10 lecturas de secuencia en cada familia. Las familias se colapsan en secuencias de consenso por votación (por ejemplo, votación sesgada) de cada posición en una familia. Se designa una base para la secuencia de consenso si 8 ó 9 miembros los confirman. No se designa una base para una secuencia de consenso si no más del 60% de los miembros lo confirman.

Las secuencias de consenso resultantes se mapean para un genoma de referencia. Cada base en una secuencia de consenso está cubierta por aproximadamente 3000 familias diferentes. Se calcula un puntaje de calidad para cada secuencia y las secuencias se filtran en base a sus puntuaciones de calidad.

La variación de secuencia se detecta contando la distribución de bases en cada locus. Si el 98% de las lecturas tienen la misma base (homocigotos) y el 2% tienen una base diferente, es probable que el locus tenga una variante de secuencia, presumiblemente de ADN del cáncer.

La CNV se detecta contando el número total de secuencias (bases) que mapean para un locus y comparando con un locus de control. Para aumentar la detección de CNV, el análisis de CNV se realiza en regiones específicas, incluyendo las regiones en los genes ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RBI, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, o NTRK1

Ejemplo 2. Método Para Corregir la Designación de Bases Determinando el Número Total de Molécula No vistas en una Muestra

Después de que los fragmentos se han amplificado y las secuencias de fragmentos amplificados se han leído y alineado, los fragmentos se someten a una designación de bases. Las variaciones en el número de fragmentos amplificados y fragmentos amplificados no vistos pueden introducir errores en la designación de bases. Estas variaciones se corrigen al calcular el número de fragmentos amplificados no vistos.

Cuando se designan las bases para el locus A (un locus arbitrario), se asume en primer lugar que hay N fragmentos amplificados. Las lecturas de secuencia pueden provenir de dos tipos de fragmentos: fragmentos de cadena doble y fragmentos de cadena sencilla. Lo siguiente es un ejemplo teórico del cálculo del número total de moléculas no vistas en una muestra.

N es el número total de moléculas en la muestra.
Suponiendo que 1000 es el número de dúplex detectados.
Suponiendo que 500 es el número de moléculas de cadena sencilla detectadas.
P es la probabilidad de ver una cadena.
Q es la probabilidad de no detectar una cadena
Donde Q = 1 -P.

1000 = NP(2).
500 = N2PQ.

$$1000 / P(2) = N.$$

$$500 \div 2 PQ = N.$$

$$1000 / P(2) = 500 \div 2PQ.$$

$$1000 * 2 PQ = 500 P(2).$$

2000 PQ = 500 P(2).
2000 Q = 500 P.
2000 (1-P) = 500P

$$2000-2000 P = 500P.$$

$$2000 = 500P + 2000 P.$$

2000 = 2500 P.

$$2000 \div 2500 = P.$$

0.8 = P.

$$1000/ P(2) = N.$$

$$1000 \div 0.64 = N.$$

1562 = N.
Número de fragmentos no vistos = 62.

Ejemplo 3. Identificación de variantes genéticas en variantes somáticas asociadas con el cáncer en un paciente

Se usa un ensayo para analizar un panel de genes para identificar variantes genéticas en variantes somáticas asociadas con el cáncer con alta sensibilidad.

Se extrae ADN libre de células del plasma de un paciente y se amplifica por PCR. Las variantes genéticas se analizan por secuenciación masivamente paralela de los genes objetivo amplificados. Para un conjunto de genes, se secuencian todos los exones, ya que dicha cobertura de secuenciación ha demostrado tener utilidad clínica (Tabla 1). Para otro conjunto de genes, la cobertura de secuenciación incluyó aquellos exones que se había informado previamente tenían una mutación somática (Tabla 2). El alelo mutante detectable mínimo (límite de detección) depende de la concentración de ADN libre de células de la muestra del paciente, que varió en menos de 10 a más de 1.000 equivalentes genómicos por ml de sangre periférica. La amplificación puede no detectarse en muestras con cantidades menores de ADN libre de células y/o amplificación de copias de genes de bajo nivel. Ciertas características de las muestras o variantes dieron como resultado una sensibilidad analítica reducida, como una calidad de la muestra baja o una recolección inadecuada.

El porcentaje de variantes genéticas encontradas en el ADN libre de células que circula en la sangre está relacionado con la biología tumoral única de este paciente. Factores que afectaron a la cantidad/porcentajes de variantes genéticas detectadas en el ADN libre de células circulante en la sangre incluyen crecimiento tumoral, conversión, tamaño, heterogeneidad, vascularización, progresión de la enfermedad o tratamiento. La Tabla 3 anota el porcentaje, o frecuencia de alelos, de ADN libre de células circulante alterado (% de ADNcf detectado en este paciente. Algunas de las variantes genéticas detectadas se enumeran en orden descendente por % de ADNcf.

Se detectan variantes genéticas en el ADN libre de células circulante aislado de la muestra de sangre de este paciente. Estas variantes genéticas son variantes somáticas asociadas con el cáncer, algunas de las cuales se han asociado con una respuesta clínica o aumentada o reducida para el tratamiento específico. Las "Alteraciones Menores" se definen como aquellas alteraciones detectadas a menos del 10% de la frecuencia de alelos de las "Alteraciones Mayores". Se anotan las frecuencias de alelos detectadas de estas alteraciones (Tabla 3) y los tratamientos asociados para este paciente.

Todos los genes enumerados en las Tablas 1 y 2 se analizan como parte de la prueba Guardant360™. La amplificación no se detecta para ERBB2, EGFR o MET en el ADN libre de células circulante aislado de la muestra de sangre de este paciente.

Los resultados de las pruebas del paciente que comprenden las variantes genéticas se enumeran en la Tabla 4.

Tabla 1. Genes en los que se secuencian todos los exones

GENES EN LOS QUE SE SECUENCIAN TODOS LOS EXONES			
ALK	< 0.1%	APC	< 0.1%
AR	< 0.1%	BRAF	< 0.1%
CDKN2A	< 0.1%	EGFR	< 0.1%
ERBB2	< 0.1%	FBXW7	< 0.1%
KRAS	< 0.1%	MET	< 0.1%
MYC	< 0.1%	NOTCH1	< 0.1%
NRAS	< 0.1%	PIK3CA	< 0.1%
PTEN	< 0.1%	PROC	< 0.1%
RB1	< 0.1%	TP53	< 0.1%

LOD: Limite de Detección. La frecuencia de alelos mutantes mínima detectable para este espécimen en la que se detectan el 80% de las variantes somáticas.

Tabla 2. Genes en los que se secuencian exones con una mutación somática informada anteriormente

GENES EN LOS QUE SE SECUENCIAN EXONES CON UNA MUTACIÓN SOMÁTICA INFORMADA ANTERIORMENTE			
ABL1	< 0.1%	AKT1	< 0.1%
ATM	< 0.1%	CDH1	< 0.1%
CSF1R	< 0.1%	CTNNB1	< 0.1%
ERBB4	< 0.1%	EZH2	< 0.1%
FGFR1	< 0.1%	FGFR2	< 0.1%
FGFR3	< 0.1%	FLT3	< 0.1%
GNA11	< 0.1%	GNAQ	< 0.1%
GNAS	< 0.1%	HNF1A	< 0.1%
HRAS	< 0.1%	IDH1	< 0.1%
IDH2	< 0.1%	JAK2	< 0.1%
JAK3	< 0.1%	KDR	< 0.1%
KIT	< 0.1%	MLH1	< 0.1%
MPL	< 0.1%	NPM1	< 0.1%
PDGFRA	< 0.1%	PTPN11	< 0.1%
RET	< 0.1%	SMAD4	< 0.1%
SMARCB1	< 0.1%	SMO	< 0.1%
SRC	< 0.1%	STK11	< 0.1%
TERT	< 0.1%	VHL	< 0.1%

LOD: Limite de Detección. La frecuencia de alelos mutantes mínima detectable para este espécimen en la que se detectan el 80% de las variantes somáticas.

Tabla 3. Frecuencia de alelos de ADN libre de células circulante alterado detectado en este paciente

Gen	% de ADNcf	0.01%----------0.1%----------1%----------10%----------100%
BRAF V600E	8.9%	
NRAS Q61K	6.2%	
JAK2 V617F	1.5%	

Leyenda:
% de ADNcf: Frecuencia de alelos de alteración genómica observada en el ADN libre de células circulante de este paciente
: ADN libre de células con alteraciones.
: ADN libre de células sin alteraciones.

Tabla 4. Alteraciones genómicas detectadas en genes seleccionados

Detectadas: 51 Alteraciones Genómicas

Gen	Cromosoma	Posición	Mutación (nt)	Mutación (AA)	Porcentaje	ID Cósmico	ID DBSNP
KRAS	12	25368462	C>T		100.0%		rs4362222
ALK	2	29416572	T>C	11461 V	100.0%		rs1670283
ALK	2	29444095	C>T		100.0%		rs1569156
ALK	2	29543663	T>C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A>T	P234P	100.0%		rs2246745
APC	5	112176756	T>A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C>G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G>A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A>G		100.0%		rs3125006
PDGFRA	4	55141055	A>G	P567P	100.0%		rs1873778
HRAS	11	534242	A>G	H27H	100.0%	COSM249860	rs12628
EGFR	7	55214348	C>T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G>C	P72R	99.8%		rs1042522
APC	5	112162854	T>C	Y486Y	55.0%		rs2229992
APC	5	112177171	G>A	P1960P	53.8%		rs465899
EGFR	7	55266417	T>C	T903T	53.6%		rs1140475
APC	5	112176325	G>A	G1678G	53.2%		rs42427
APC	5	112176559	T>G	S1756S	53.0%		rs866006
EGFR	7	55229255	G>A	R521K	53.0%		
MET	7		A>G	Q648Q	116397572		
APC	5	112175770	G>A	T1493T	52.7%		rs41115
EGFR	7	55249063	G>A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T>C		52.4%		rs111457 67
EGFR	7	55238874	T>A	T629T	52.0%		rs2227984

(continuación)

Detectadas: 51 Alteraciones Genómicas						
Gen	Cromosoma	Posición	Mutación (nt)	Mutación (AA)	Porcentaje	ID Cósmico	ID DBSNP
ERBB2	17	37879588	A>G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G>A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G>C	L9L	51.0%		rs4358080
APC	5	112164561	G>A	A545A	51.0%		rs351771
FLT3	13	28610183	A>G		50.8%		rs2491231
NOTCH1	9	139418260	A>G	N104N	50.5%		rs4489420
ALK	2	29444076	G>T		50.4%		rs1534545
PIK3CA	3	178917005	A>G		50.3%		rs3729674
NOTCH1	9	139412197	G>A		50.2%		rs9411208
ALK	2		A>G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A>C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G>A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C>T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T>C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G>T	G1125G	48.6%		rs3795850
NOTCH1	9	139410177	T>C		48.5%		rs3124603
RET	10	43613843	G>T	L769L	48.2%		rs1800861
EGFR	7	55214443	G>A		48.0%		rs7801956
ALK	2	29416366	G>C	D1529E	47.2%		rs1881421
EGFR	7	55238087	C>T		45.5%		rs10258429
RET	10	43615633	C>G	S904S	44.8%		rs1800863
BRAF	7	140453136	A>T	V600E	8.9%	COSM476	
NRAS	1	115256530	G>T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G>T	V617F	1.5%	COSM12600	rs77375493

Ejemplo 4. Determinación de límites específicos del paciente de detección de genes analizados por ensayos Guardant360™

Usando el método del Ejemplo 3, se detectan alteraciones genéticas en el ADN libre de células de un paciente. Las lecturas de secuencias de estos genes incluyen secuencias de exones y/o intrones.

Los límites de detección de la prueba se muestran en la Tabla 5. Los límites de los valores de detección dependen de la concentración de ADN libre de células y de la cobertura de secuenciación para cada gen.

Tabla 5. Límites de Detección de genes seleccionados en una paciente usando Guardant

Cobertura de Exones Completa y de Intrones Parcial					
APC	0.1%	AR *	0.2%	ARID1A	
BRAF *	0.1%	BRCA1		BRCA2	
CCND1 *		CCND2 *		CCNE1 *	
CDK4 *		CDK6 *		CDKN2A	0.1%
CDKN2B		EGFR *	< 0.1%	ERBB2 *	0.1%
FGFR1 *	< 0.1%	FGFR2 *	0.1%	HRAS	0.1%
KIT *	0.1%	KRAS *	0.1%	MET *	0.1%
MYC *	0.1%	NF1		NRAS	0.1%
PDGFRA *	0.1%	PIK3CA *	0.1%	PTEN	0.1%
RAF1 *		TP53	0.1%		
Exones Cubiertos con Mutaciones Somáticas Informadas					
AKT1	0.1%	ALK	< 0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	< 0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusiones					
ALK	< 0.1%	RET	0.1%	ROS1	
NTRK1					
LOD: Limite de Detección. La frecuencia de alelos mutantes mínima detectable para este espécimen en la que se detectan el 80% de las variantes somáticas. * indica genes de CNV					

Ejemplo 5. Corrección de Errores de Secuencia Comparando Secuencias de Watson y de Crick

El ADN libre de células de cadena doble se aísla del plasma de un paciente. Los fragmentos de ADN libre de células se etiquetan usando 16 adaptadores que contienen burbujas diferentes, cada uno de los cuales comprende un código de barras distintivo. Los adaptadores que contienen burbujas están unidos a ambos extremos de cada fragmento de ADN libre de células por ligación. Después de la ligación, cada fragmento de ADN libre de células puede identificarse distintivamente por la secuencia de los distintos códigos de barras y dos 20 pb de secuencias endógenas en cada extremo del fragmento de ADN libre de células.

Los fragmentos de ADN libre de células etiquetados se amplifican por PCR. Los fragmentos amplificados se enriquecen usando perlas que comprenden sondas de oligonucleótidos que enlazan específicamente a un grupo de genes asociados con el cáncer. Por lo tanto, los fragmentos de ADN libre de células del grupo de genes asociados con el cáncer se enriquecen selectivamente.

Los adaptadores de secuenciación, cada uno de los cuales comprende un sitio de enlace al cebador de secuenciación, un código de barras de muestra y una secuencia de flujo de células, están unidos a las moléculas de ADN enriquecido. Las moléculas resultantes se amplifican por PCR.

Se secuencian ambas cadenas de los fragmentos amplificados. Como cada adaptador que contiene burbujas comprende una parte no complementaria (por ejemplo, la burbuja), la secuencia de una cadena del adaptador que contiene burbujas es diferente de la secuencia de la otra cadena (complemento). Por lo tanto, las lecturas de secuencias de amplicones derivados de la cadena Watson de un ADN libre de células original pueden distinguirse de los amplicones de la cadena de Crick del ADN libre de células original por las secuencias adaptadoras que contienen burbujas unidas.

Las lecturas de secuencia de una cadena de un fragmento de ADN libre de células original se comparan con las lecturas de secuencia de la otra cadena del fragmento de ADN libre de células original. Si se produce una variante en solo las lecturas de secuencias de una cadena, pero no la otra cadena, del fragmento de ADN libre de células original, esta variante se identificará como un error (por ejemplo, resultado de un PCR y/o amplificación), en lugar de una variante genética verdadera.

Las lecturas de secuencia se agrupan en familias. Se corrigen los errores en las lecturas de secuencias. La secuencia de consenso de cada familia se genera por colapso.

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

REIVINDICACIONES

1. Un método para determinar una medida cuantitativa indicativa de una serie de moléculas de ácido desoxirribonucleico (ADN) de cadena doble individuales en una muestra, que comprende:

(a) determinar una medida cuantitativa de moléculas de ADN individuales para las que se detectan ambas cadenas;
(b) determinar una medida cuantitativa de moléculas de ADN individuales para las que solo se detecta una de las cadenas de ADN;
(c) inferir de (a) y (b) anteriores una medida cuantitativa de moléculas de ADN individuales para las cuales no se detectó ninguna cadena; y
(d) usar (a)-(c) para determinar la medida cuantitativa indicativa de una serie de moléculas de ADN de cadena doble individuales en la muestra;

en donde determinar una medida cuantitativa de moléculas de ADN individuales comprende: (i) etiquetar dichas moléculas de ADN con un conjunto de etiquetas dúplex que etiquetan de manera diferente cadenas complementarias de una molécula de ADN de cadena doble en dicha muestra para proporcionar cadenas etiquetadas; y (ii) secuenciar por lo menos algunas de dichas cadenas etiquetadas para producir un conjunto de lecturas de secuencia.

2. El método de la reivindicación 1, que comprende además detectar la variación del número de copias en dicha muestra determinando una medida cuantitativa normalizada determinada en el paso (d) en uno o más loci genéticos y determinando la variación del número de copias en base a la medida normalizada.

3. El método de la reivindicación 1 o la reivindicación 2, que comprende además clasificar lecturas de secuencia en lecturas emparejadas y lecturas no emparejadas, en donde (i) una lectura emparejada corresponde a lecturas de secuencia generadas a partir de una primera cadena etiquetada y una segunda cadena complementaria etiquetada de manera diferente derivada de un molécula de polinucleótidos de cadena doble, y (ii) una lectura no emparejada representa una primera cadena etiquetada que no tiene una segunda cadena complementaria etiquetada de manera diferente derivada de una molécula de polinucleótido de cadena doble representada entre dichas lecturas de secuencia en dicho conjunto de lecturas de secuencia.

4. El método de la reivindicación 3, que comprende además determinar medidas cuantitativas de (i) dichas lecturas emparejadas y (ii) dichas lecturas no emparejadas que mapean uno o más loci genéticos para determinar una medida cuantitativa del total de moléculas de ADN cadena doble en dicha muestra que mapea dichos uno o más loci genéticos en base a dicha medida cuantitativa de lecturas emparejadas y lecturas no emparejadas que mapean un locus.

5. El método de cualquiera de las reivindicaciones 1 a 4, en el que las moléculas de ADN de cadena doble comprenden ADNcf.

6. El método de cualquiera de las reivindicaciones 1 a 5, en el que las etiquetas dúplex son etiquetas de cadena doble que tienen forma de Y con una porción hibridada en un extremo de la etiqueta y una porción no hibridada está en el extremo opuesto de la etiqueta.

7. El método de una cualquiera de las reivindicaciones 1 a 6, en el que las etiquetas dúplex contienen códigos de barras moleculares, opcionalmente en el que el etiquetado se produce en una reacción única y da como resultado que más del 50% de las moléculas de ADN se etiqueten en ambos extremos.

8. El método de la reivindicación 7, en el que las moléculas de ADN en la muestra están etiquetadas de manera no única, opcionalmente en el que el número de códigos de barras moleculares diferentes no es mayor de 100, 50, 40, 30, 20 o 10 códigos de barras moleculares.

9. El método de cualquiera de las reivindicaciones 7 a 8, en donde el método comprende además reducir o hacer un seguimiento de la redundancia en las lecturas de secuencia para determinar lecturas de consenso que son representativas de cadenas sencillas del ADN original, opcionalmente en donde el método para reducir o hacer un seguimiento de la redundancia comprende comparar lecturas de secuencia que tienen los mismos códigos de barras moleculares o unos similares y el mismo o similar final de secuencias.

10. El método de cualquiera de las reivindicaciones 7 a 9, en donde el método comprende además agrupar las lecturas de secuencia de acuerdo con los códigos de barras moleculares y la información de secuencia, opcionalmente en donde el agrupamiento se realiza desde por lo menos un extremo del ADN original para crear agrupaciones de lecturas de cadena sencilla.

11. El método de cualquiera de las reivindicaciones 1 a 10, en el que la muestra se deriva de sangre, plasma, suero, orina, saliva, excreciones de mucosa, esputo, heces, líquido cefalorraquídeo, piel, cabello, sudor y/o lágrimas.

12. El método de cualquiera de las reivindicaciones 1 a 11, en el que la muestra se deriva de un sujeto que se sospecha que tiene una enfermedad, opcionalmente en donde la enfermedad es cáncer.

13. El método de cualquiera de las reivindicaciones 1 a 12, en donde el método comprende además enriquecer selectivamente un subconjunto del ADN etiquetado, opcionalmente en donde el enriquecimiento selectivo se realiza mediante técnicas de hibridación o amplificación.

14. El método de la reivindicación 13, en el que el enriquecimiento selectivo se realiza por hibridación usando un soporte sólido, opcionalmente en donde el soporte sólido comprende sondas que hibridan específicamente con regiones genómicas asociadas con cáncer.

15. El método de cualquiera de las reivindicaciones 1 a 14, en donde el método comprende además analizar las secuencias de nucleótidos con un procesador informático programado para identificar una o más alteraciones genéticas en la muestra de nucleótidos de un sujeto, opcionalmente en donde la una o más alteraciones genéticas se selecciona de la lista que comprende cambio(s) de base, inserción(es), repetición(es), supresión(es), variación(es) del número de copias, modificación(es) epigenética, sitio(s) de unión de nucleosomas, cambio(s) de número de copias debido a origen(es) de replicación y transversión(es).

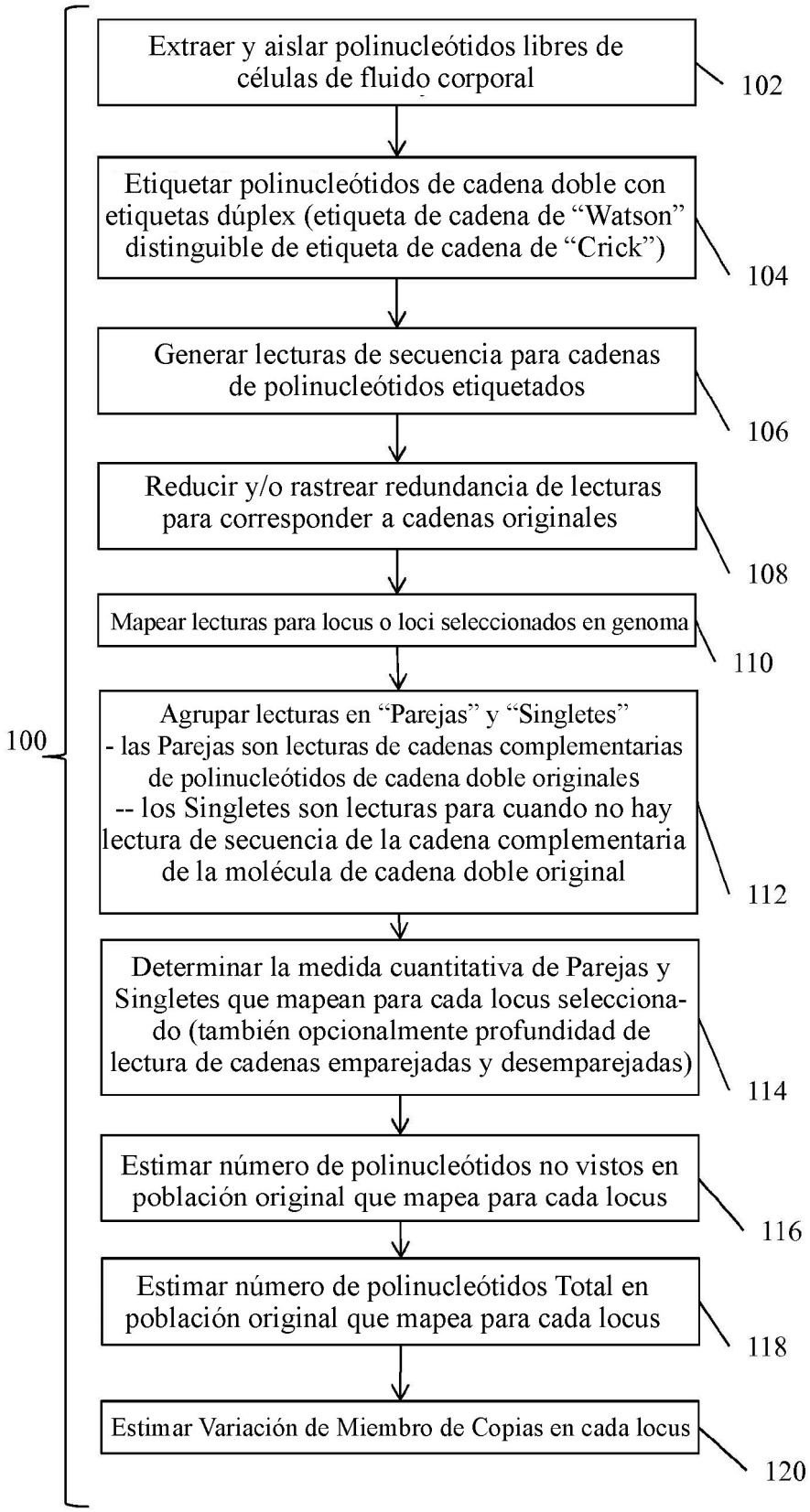
Extraer y aislar polinucleótidos libres de células de fluido corporal
102
Etiquetar polinucleótidos de cadena doble con etiquetas dúplex (etiqueta de cadena de "Watson" distinguible de etiqueta de cadena de "Crick")
104
Generar lecturas de secuencia para cadenas de polinucleótidos etiquetados
106
Reducir y/o rastrear redundancia de lecturas para corresponder a cadenas originales
108
Mapear lecturas para locus o loci seleccionados en genoma
110
Agrupar lecturas en "Parejas" y "Singletes"
- las Parejas son lecturas de cadenas complementarias de polinucleótidos de cadena doble originales
-- los Singletes son lecturas para cuando no hay lectura de secuencia de la cadena complementaria de la molécula de cadena doble original
112
Determinar la medida cuantitativa de Parejas y Singletes que mapean para cada locus seleccionado (también opcionalmente profundidad de lectura de cadenas emparejadas y desemparejadas)
114
Estimar número de polinucleótidos no vistos en población original que mapea para cada locus
116
Estimar número de polinucleótidos Total en población original que mapea para cada locus
118
Estimar Variación de Miembro de Copias en cada locus
120
100

Fig. 1

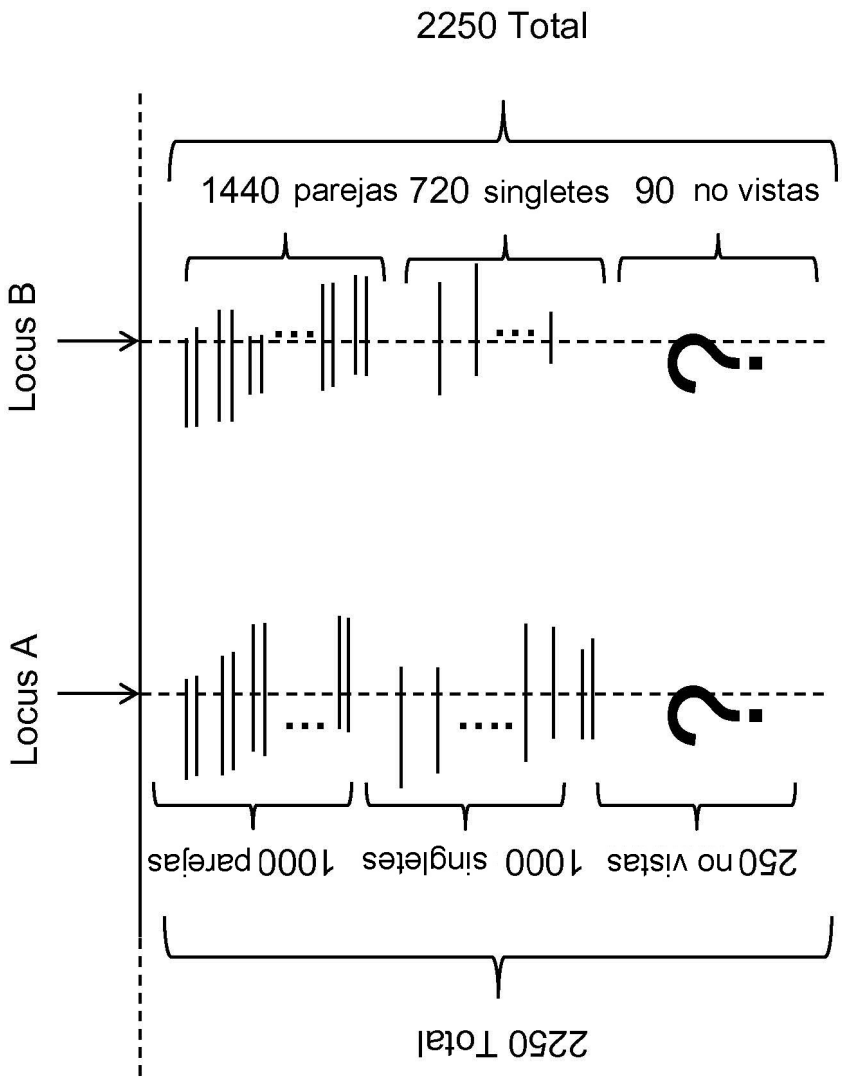
2250 Total
1440 parejas 720 singletes 90 no vistas
Locus B
Locus A
1000 parejas 1000 singletes 250 no vistas
2250 Total

Fig. 2

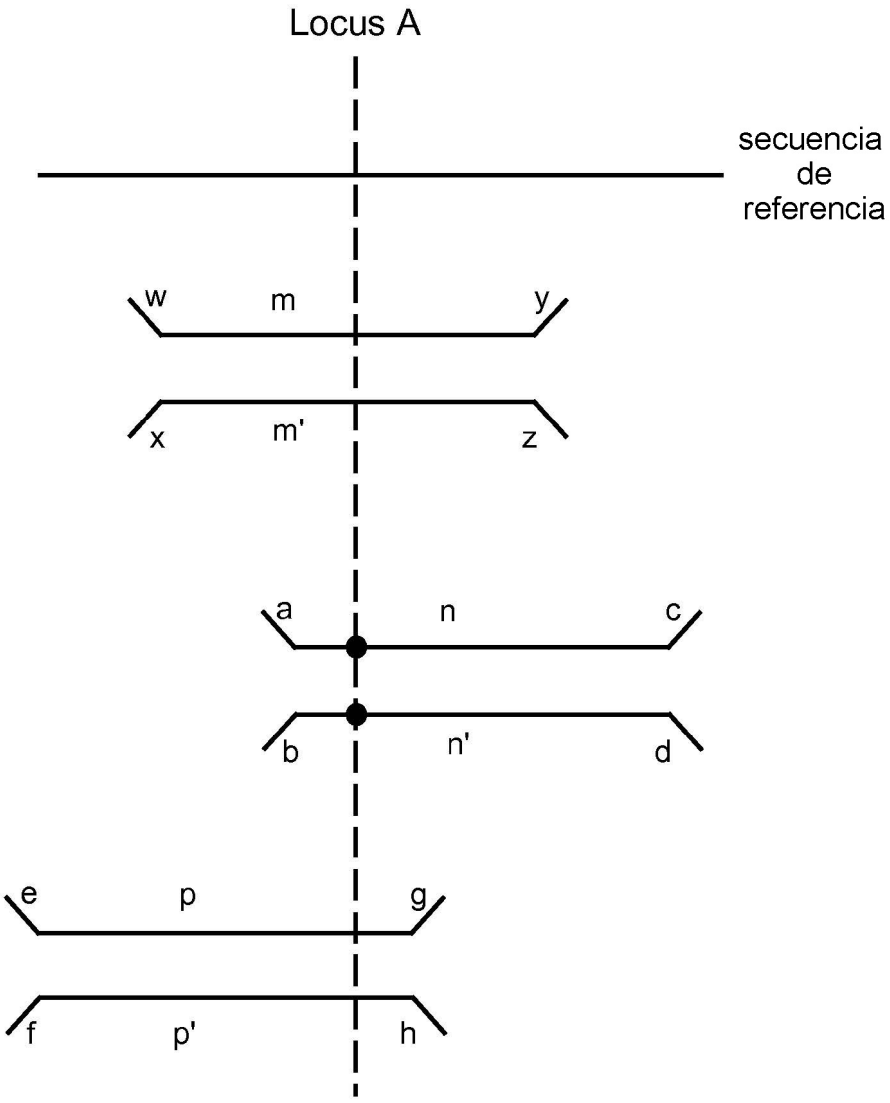

Fig. 3

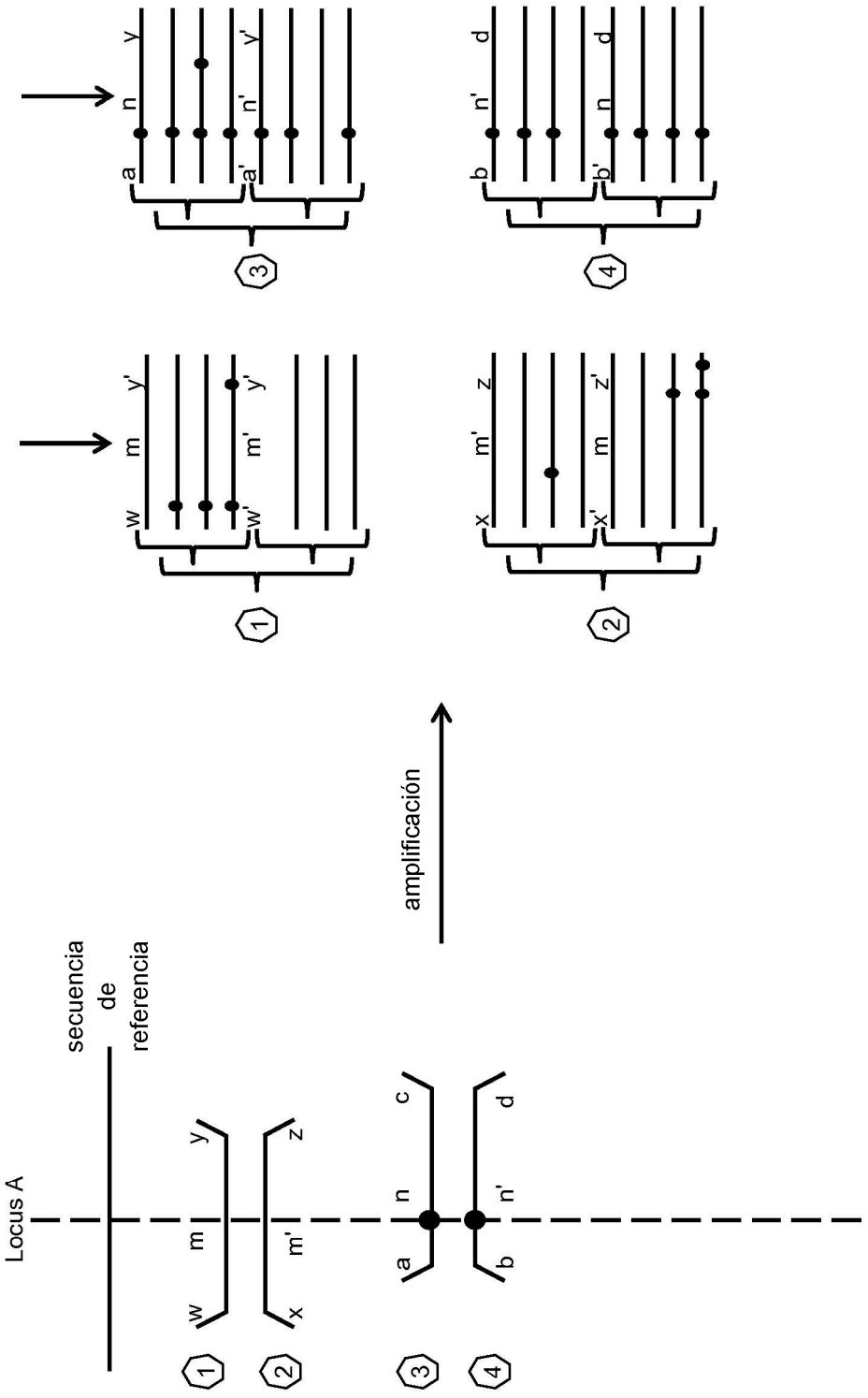
a
n
y
a'
n'
y'
3
b
n'
d
b'
n
d
4
w
m
y'
w'
m'
y'
1
x
m'
z
x'
m
z'
2
amplificación
secuencia
de
referencia
Locus A
w
m
y
x
m'
z
a
n
c
b
n'
d
1
2
3
4

Fig. 4A

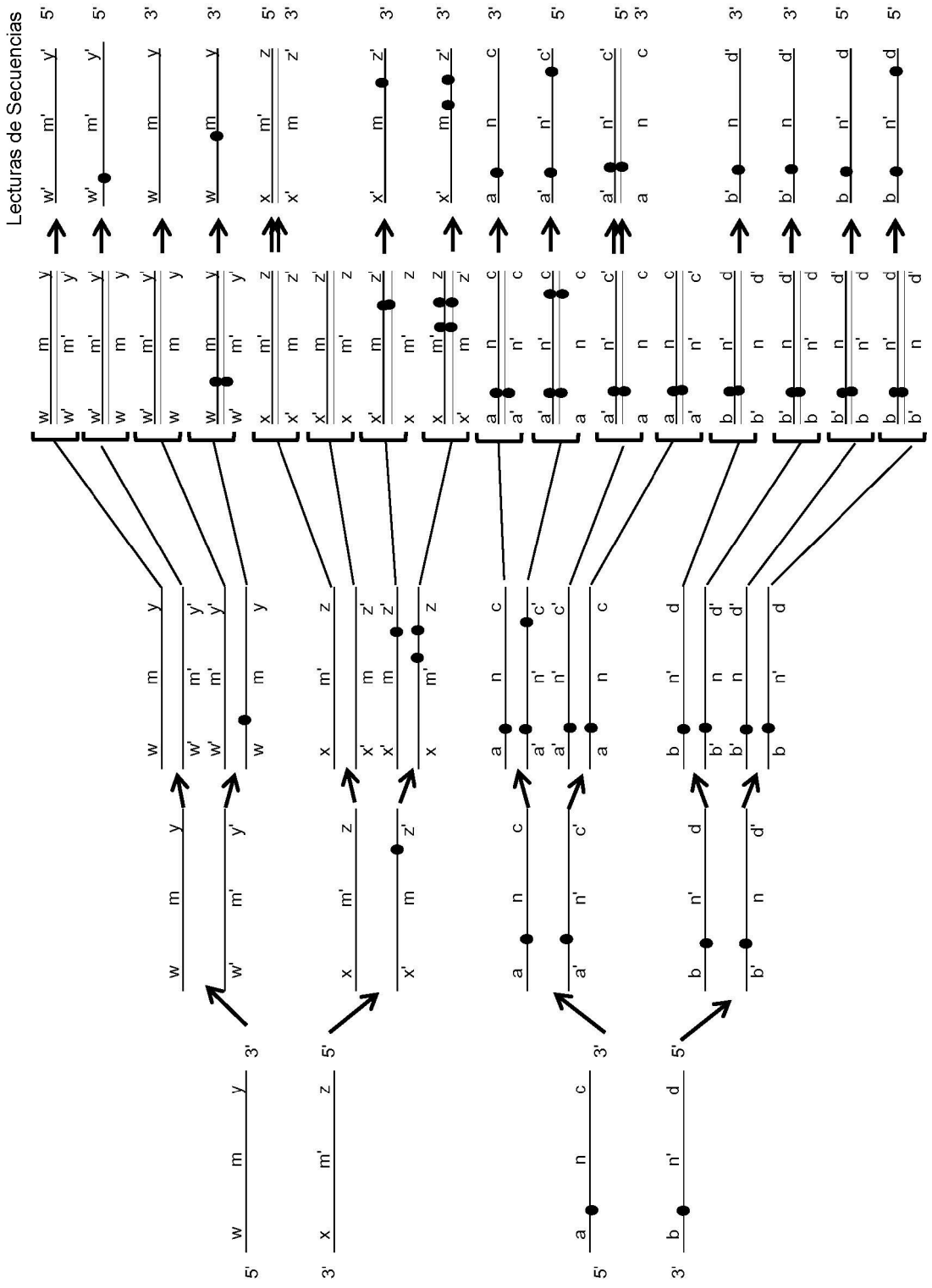

Fig. 4B

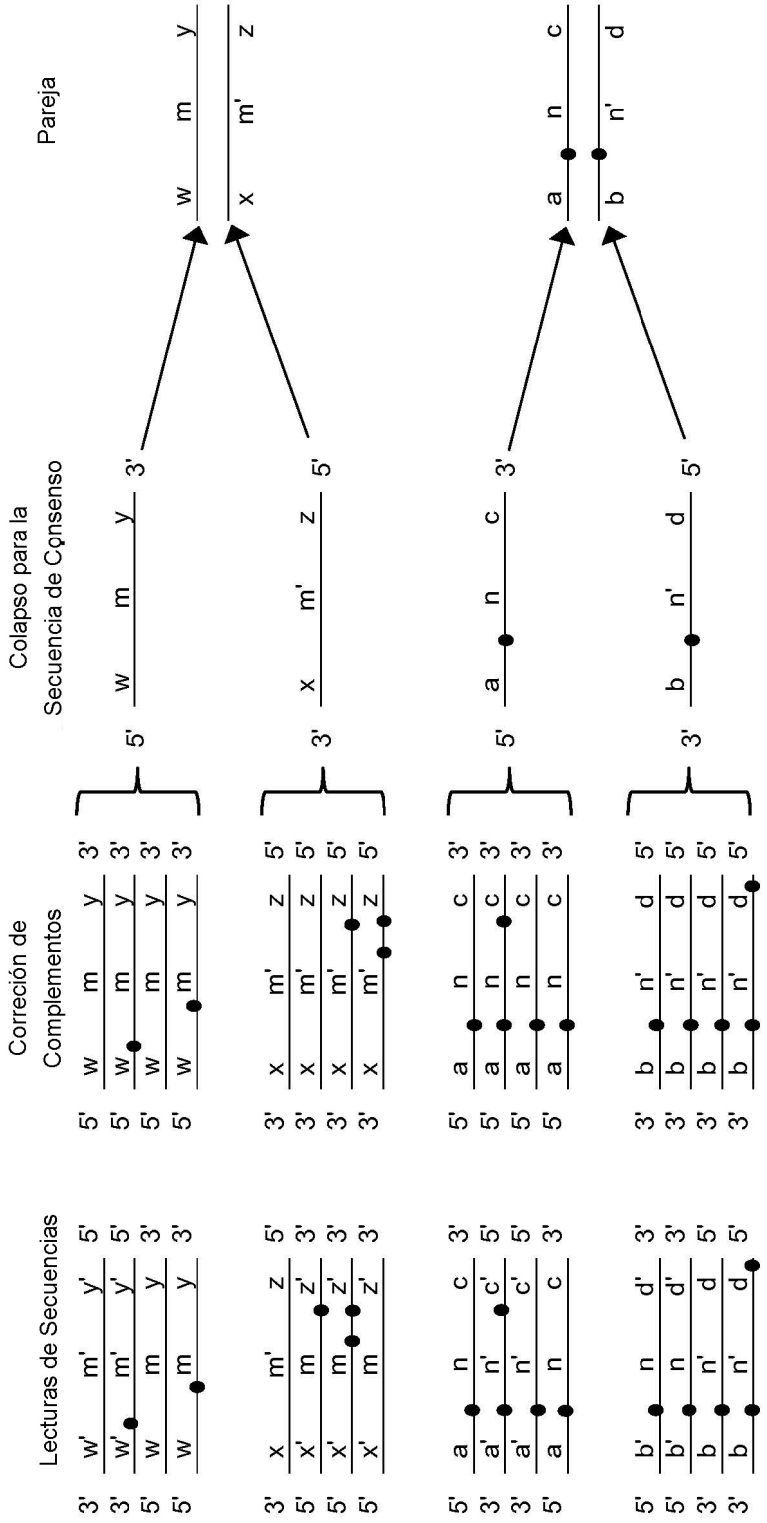
Lecturas de Secuencias
Correción de Complementos
Colapso para la Secuencia de Cqnsenso
Pareja

Fig. 4C

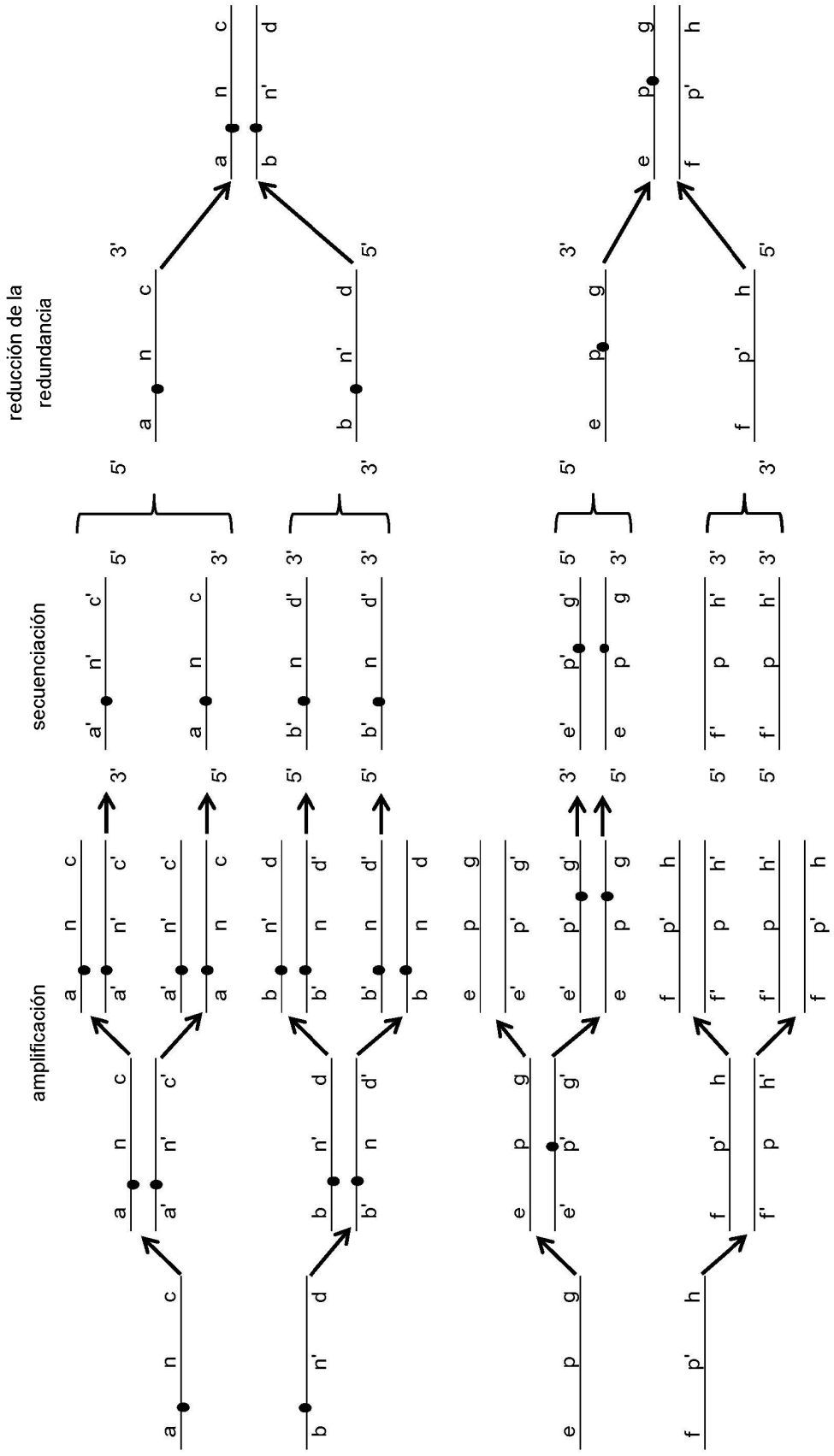
amplificación
secuenciación
reducción de la redundancia

Fig. 5

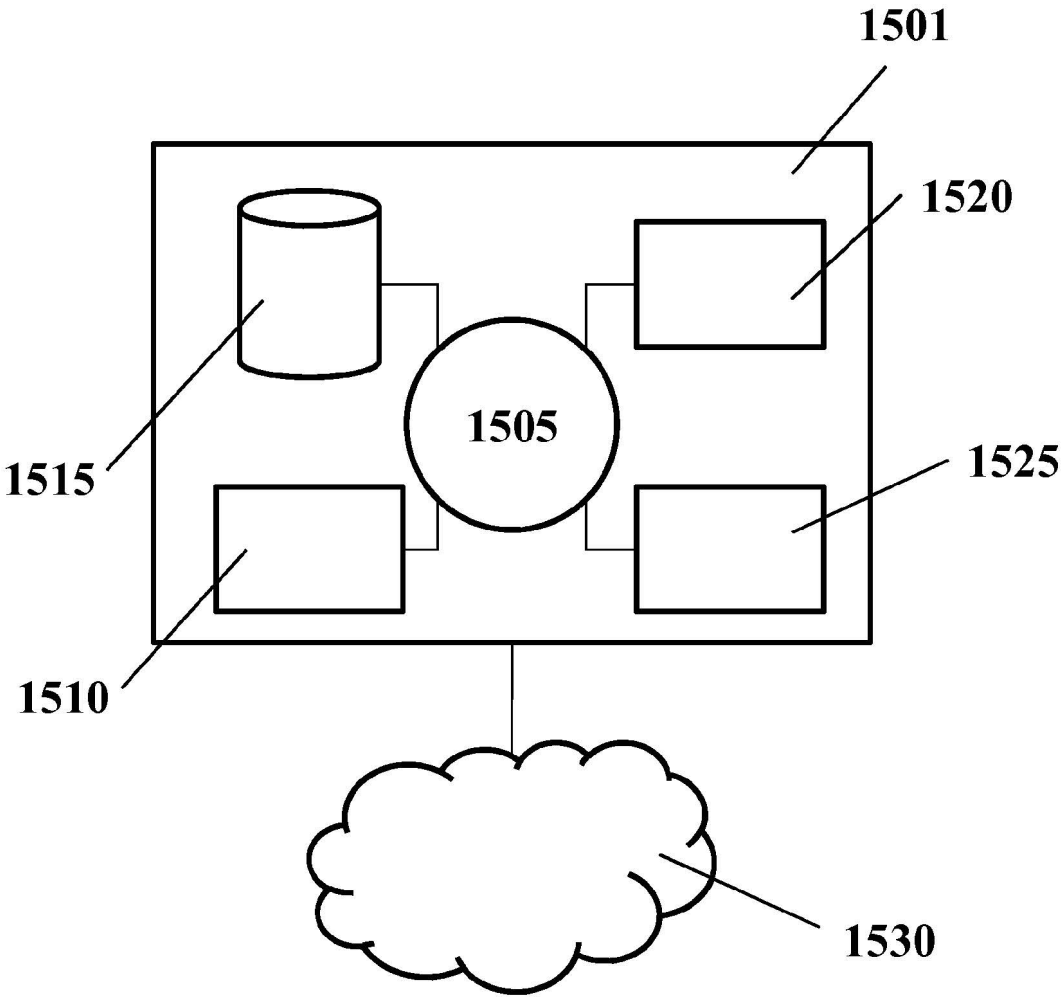

Fig. 6

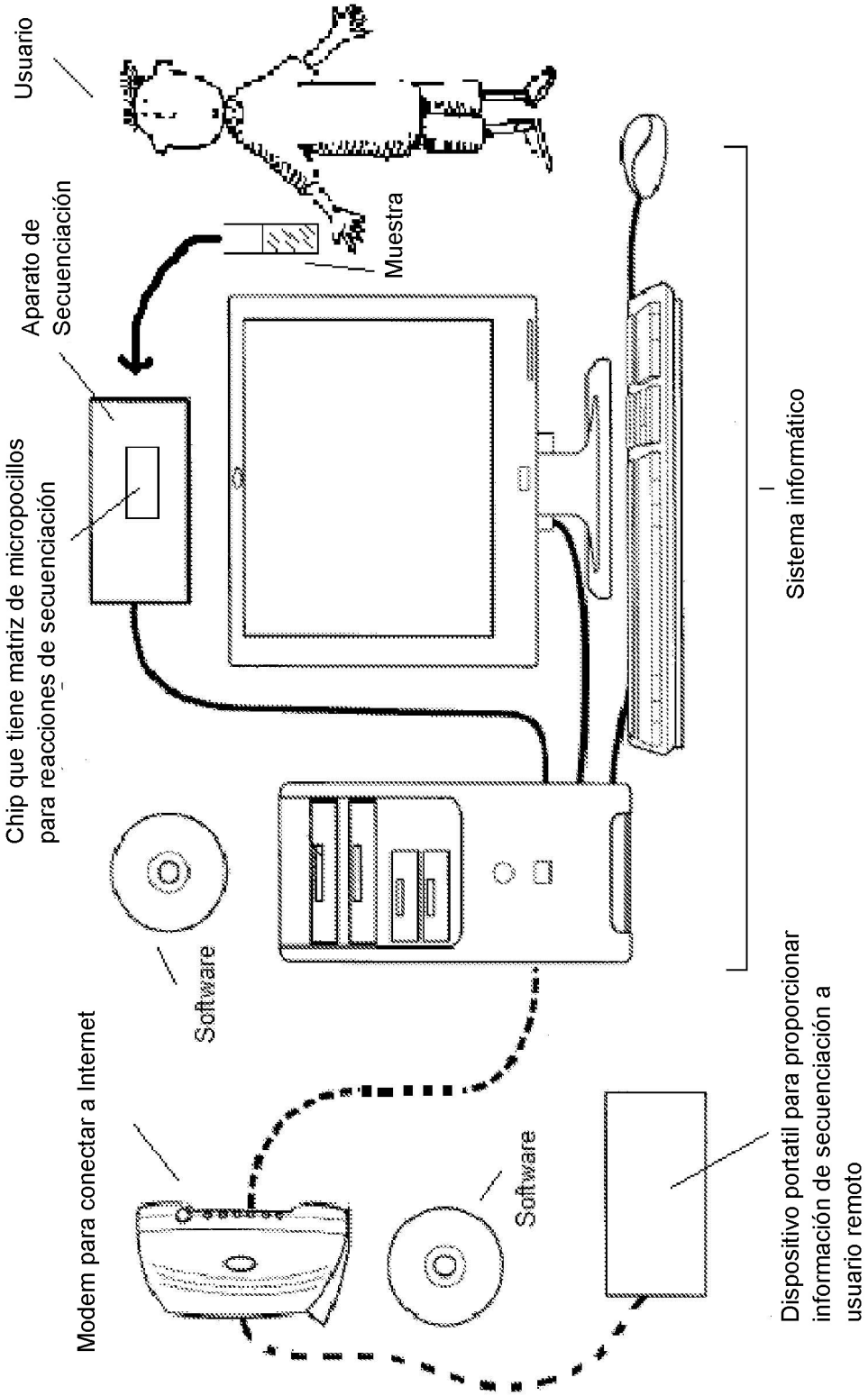
Usuario
Aparato de Secuenciación
Muestra
Chip que tiene matriz de micropocillos para reacciones de secuenciación
Sistema informático
Software
Modem para conectar a Internet
Software
Dispositivo portatil para proporcionar información de secuenciación a usuario remoto

Fig. 7

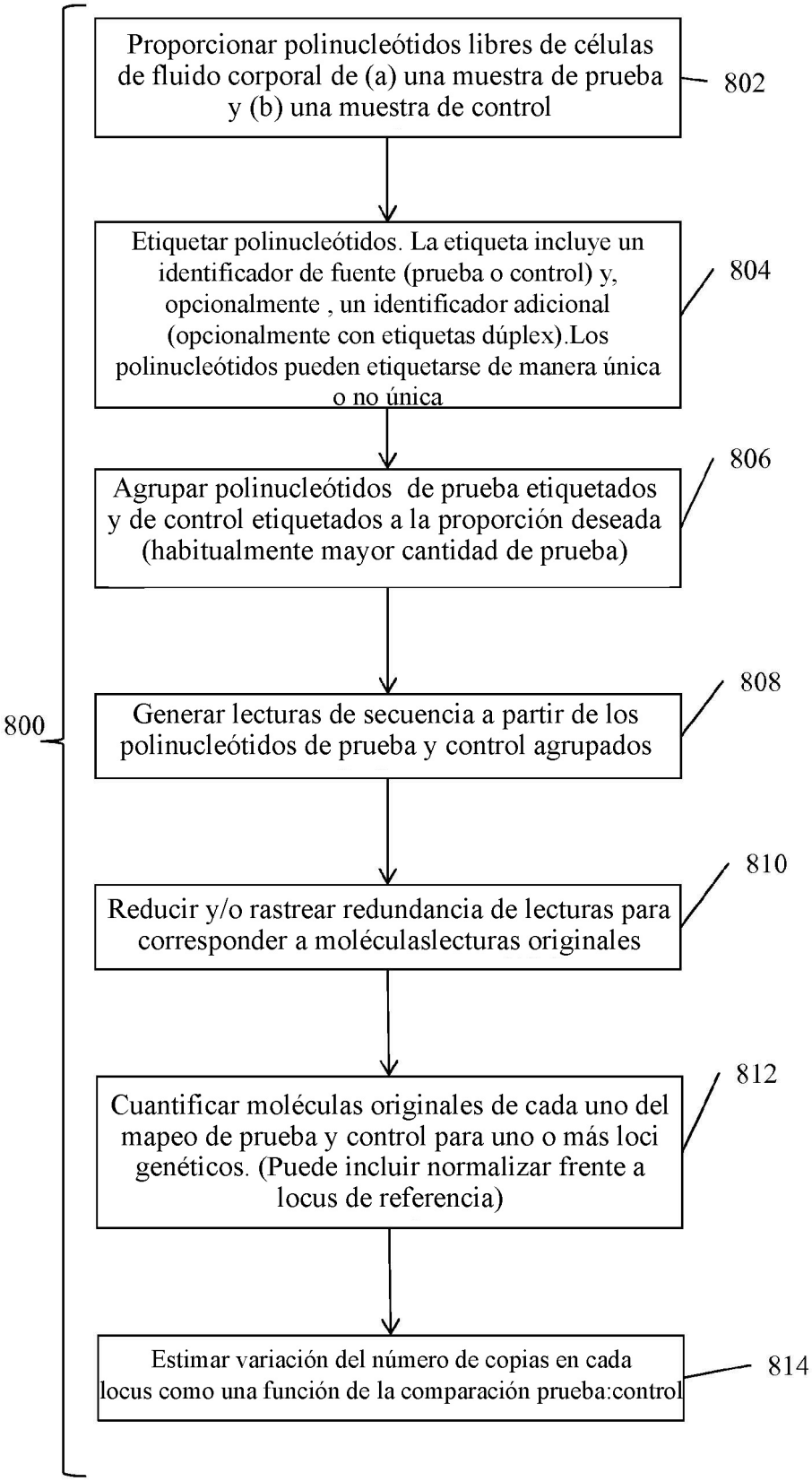
Proporcionar polinucleótidos libres de células de fluido corporal de (a) una muestra de prueba y (b) una muestra de control
802
Etiquetar polinucleótidos. La etiqueta incluye un identificador de fuente (prueba o control) y, opcionalmente , un identificador adicional (opcionalmente con etiquetas dúplex).Los polinucleótidos pueden etiquetarse de manera única o no única
804
Agrupar polinucleótidos de prueba etiquetados y de control etiquetados a la proporción deseada (habitualmente mayor cantidad de prueba)
806
Generar lecturas de secuencia a partir de los polinucleótidos de prueba y control agrupados
808
800
Reducir y/o rastrear redundancia de lecturas para corresponder a moléculaslecturas originales
810
Cuantificar moléculas originales de cada uno del mapeo de prueba y control para uno o más loci genéticos. (Puede incluir normalizar frente a locus de referencia)
812
Estimar variación del número de copias en cada locus como una función de la comparación prueba:control
814

Fig. 8

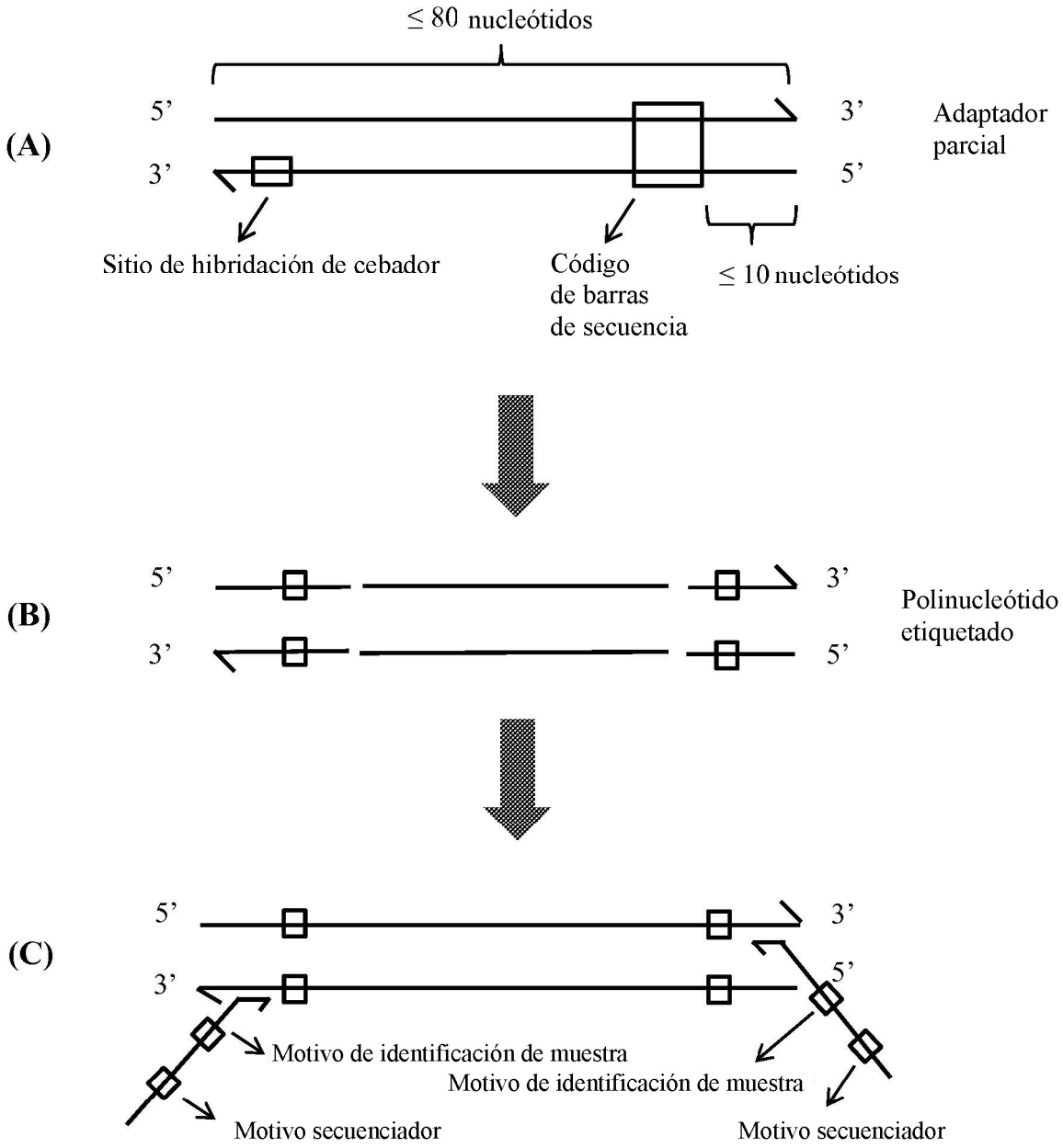
≤ 80 nucleótidos
(A)
5'
3'
3'
5'
Adaptador parcial
Sitio de hibridación de cebador
Código de barras de secuencia
≤ 10 nucleótidos
(B)
5'
3'
3'
5'
Polinucleótido etiquetado
(C)
5'
3'
3'
5'
Motivo de identificación de muestra
Motivo de identificación de muestra
Motivo secuenciador
Motivo secuenciador

Fig. 9