

19



OFICINA ESPAÑOLA DE  
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 868 074**

51 Int. Cl.:

**C12Q 1/68** (2008.01)

**C12Q 1/6855** (2008.01)

12

## TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **13.04.2018** **PCT/US2018/027632**

87 Fecha y número de publicación internacional: **18.10.2018** **WO18191702**

96 Fecha de presentación y número de la solicitud europea: **13.04.2018** **E 18783709 (1)**

97 Fecha y número de publicación de la concesión europea: **10.03.2021** **EP 3610032**

54 Título: **Métodos para unir adaptadores a ácidos nucleicos de muestra**

30 Prioridad:

**14.04.2017 US 201762485769 P**  
**14.04.2017 WO PCT/US2017/027809**  
**18.04.2017 US 201762486663 P**  
**08.06.2017 US 201762517145 P**

45 Fecha de publicación y mención en BOPI de la traducción de la patente:  
**21.10.2021**

73 Titular/es:

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

72 Inventor/es:

**KENNEDY, ANDREW y**  
**MORTIMER, STEFANIE ANN WARD**

74 Agente/Representante:

**IZQUIERDO BLANCO, María Alicia**

ES 2 868 074 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

## DESCRIPCIÓN

Métodos para unir adaptadores a ácidos nucleicos de muestra

## 5 ANTECEDENTES

El cáncer es una de las principales causas de enfermedad en todo el mundo. Cada año, a decenas de millones de personas se les diagnostica cáncer en todo el mundo, y más de la mitad finalmente muere a causa de él. En muchos países, el cáncer es la segunda causa de muerte más común después de las enfermedades cardiovasculares. La detección temprana se asocia con mejores resultados para muchos cánceres.

Los cánceres se detectan a menudo mediante biopsias de tumores seguidas de análisis de células, marcadores o ADN extraído de las células. Pero más recientemente se ha propuesto que los cánceres también pueden detectarse a partir de ácidos nucleicos libres de células en fluidos corporales, como sangre u orina (ver, por ejemplo, Siravegna et al., Nature Reviews 2017). Tales pruebas tienen la ventaja de que no son invasivas y pueden realizarse sin identificar una sospecha de células cancerosas para realizar una biopsia. Sin embargo, la cantidad de ácidos nucleicos en los fluidos corporales es muy baja. Por lo tanto, tales análisis requieren métodos eficientes para convertir ADN libre de células nativas en fluidos corporales en formas susceptibles de análisis.

La preparación de moléculas de ADN a partir de muestras de pacientes para su análisis suele implicar en primer lugar la reparación de los salientes de cadena sencilla para permitir la ligación a los adaptadores para su amplificación y secuenciación. La reparación puede efectuarse digiriendo la cadena que sobresale o extendiendo la cadena opuesta para producir un extremo romo seguido de fosforilación de los extremos 5' y ligadura del extremo romo a los adaptadores. Alternativamente, después de un extremo romo, los extremos romos pueden tener una cola A con una polimerasa Taq. Los fragmentos de cola A se aparean y se ligan con adaptadores que incluyen una cola T de un único nucleótido en un extremo 3'. Esta configuración favorece la ligación deseada entre el adaptador y la molécula de ADN, pero la eficiencia de conversión general de las moléculas de ADN en una muestra en moléculas que pueden secuenciarse puede ser todavía inaceptablemente baja para muestras en las que solo están disponibles pequeñas cantidades de ácidos nucleicos.

La WO 2015/044262 describe composiciones enzimáticas para la reparación, adenilación y fosforilación del extremo del ADN. La WO 2016/135300 divulga métodos y kits para mejorar la eficiencia de la construcción de bibliotecas de secuenciación de próxima generación. Wu et. Al. (Nat Methods 9, i-ii, 2012) describe métodos para la creación más rápida de bibliotecas de secuenciación de próxima generación.

## 35 SUMARIO

La invención proporciona un método de preparación de ácidos nucleicos para análisis que comprende; (a) ácidos nucleicos de cadena doble de extremo romo con salientes de cadena sencilla en una muestra por la acción de una o más enzimas que proporcionan una actividad de polimerasa 5'-3' y una actividad de lectura de prueba 3'-5', y cuatro tipos de nucleótidos estándar, en donde los salientes de cadena sencilla con extremos 5' sirven como plantillas para la extensión de una cadena complementaria por la actividad de polimerasa y los salientes de cadena sencilla con extremos 3' son digeridos por la actividad de lectura de prueba produciendo ácidos nucleicos de extremos romos; (b) sin separar los ácidos nucleicos de extremos romos de otros componentes de la muestra, poner cola en los extremos de los ácidos nucleicos de extremos romos mediante la acción de una polimerasa sin una función de lectura de prueba 3'-5', que realiza una adición no dirigida por plantilla de un nucleótido a los extremos 3' de ácidos nucleicos de extremos romos, en donde A se añade preferentemente a G preferentemente a C o T; (c) aparear los ácidos nucleicos del paso (b) con adaptadores por lo menos parcialmente de cadena doble con un saliente T de un solo nucleótido en un extremo 3', y adaptadores por lo menos parcialmente de cadena doble con un saliente C de un solo nucleótido en un extremo 3'; y (d) ligar los ácidos nucleicos a los adaptadores. Opcionalmente, el método comprende además desnaturalizar una o más enzimas después del paso (a). Opcionalmente, el método comprende además poner en contacto la muestra con una o más enzimas, los cuatro tipos de nucleótidos estándar y la polimerasa sin una función de lectura de prueba 3'-5'. Opcionalmente, la muestra se pone en contacto con la una o más enzimas, los cuatro tipos de nucleótidos estándar y la polimerasa sin una función de lectura de prueba 3'-5' juntos. Opcionalmente, el paso (b) se realiza a una temperatura más alta que el paso (a). Opcionalmente, el paso (a) se realiza a temperatura ambiente y el paso (b) a una temperatura superior a 60 C. Opcionalmente, la una o más enzimas son una polimerasa con actividad de polimerasa 5'-3' y actividad de lectura de prueba 3'-5'. Opcionalmente, la polimerasa sin una función de lectura de prueba 3'-5' es una polimerasa termoestable y el método comprende además aumentar la temperatura de la muestra después del paso (a) para inactivar la polimerasa con actividad de polimerasa 5'-3' y actividad de lectura de prueba 3'-5'. Opcionalmente, el método comprende además (e) amplificar los ácidos nucleicos ligados a los adaptadores; y (f) analizar los ácidos nucleicos.

Opcionalmente, el método comprende además poner en contacto la muestra con adaptadores de extremos romos por lo menos parcialmente de cadena doble, que se ligan con ácidos nucleicos de cadena doble de extremos romos que no se han sometido a la adición no dirigida por plantilla de un nucleótido a los extremos 3' en el paso de

ligación. Opcionalmente, la primera polimerasa es la polimerasa T4 o el fragmento grande de Klenow. Opcionalmente, la segunda polimerasa es una polimerasa Taq. Opcionalmente, por lo menos los pasos (a)-(e) se realizan en un único tubo. Opcionalmente, en donde los pasos (a)-(f) o (a) a (g) se realizan en un único tubo. Opcionalmente, la relación molar de adaptadores por lo menos parcialmente de cadena doble con un solo nucleótido T a un solo nucleótido C es de 4:1 a 2:1, preferiblemente de 3:1. Opcionalmente, la relación molar de adaptadores de extremos romos a adaptadores de cola es de 1:5 a 1:500, preferiblemente de 1:10 a 1:100. Opcionalmente, por lo menos el 70% de los ácidos nucleicos de cadena doble de la muestra se unen a adaptadores. Opcionalmente, se analiza por lo menos el 70% de los ácidos nucleicos de cadena doble disponibles en la muestra. Opcionalmente, el paso (f) comprende secuenciar los ácidos nucleicos ligados a los adaptadores. Opcionalmente, la secuenciación

secuencia un nucleótido que formó un saliente en el paso (c) o (d).  
La invención proporciona además un método para convertir ADN de cadena doble en ADN etiquetado con adaptador que comprende: (a) poner en contacto una población de moléculas de ADN de cadena doble con una población de adaptadores por lo menos parcialmente de cadena doble, en donde: (i) la población de moléculas de ADN de cadena doble comprende moléculas de ADN que comprenden un saliente de un A único nucleótido y moléculas de ADN que comprenden un saliente G de un solo nucleótido, y en donde los salientes A de un solo nucleótido son más abundantes (por ejemplo, 10 veces, 100 veces, 1000 veces) que los salientes G de un solo nucleótido en la población, y (ii) la población de adaptadores por lo menos parcialmente de cadena doble comprende adaptadores que comprenden un saliente T de un solo nucleótido y adaptadores que comprenden un saliente C de un solo nucleótido; y (b) ligar los adaptadores a las moléculas de ADN, en donde la ligación produce ADN etiquetado con adaptador.

Opcionalmente, (i) la población de moléculas de ADN de cadena doble comprende además por lo menos una de: moléculas de ADN que comprenden un saliente C de un solo nucleótido, moléculas de ADN que comprenden un saliente T de un solo nucleótido y un extremo romo, y (ii) la población de adaptadores de por lo menos parcialmente cadena doble comprende además por lo menos uno de: adaptadores que comprenden un saliente G de un solo nucleótido, adaptadores que comprenden un saliente A de un solo nucleótido y un extremo romo. Opcionalmente, los adaptadores por lo menos parcialmente de cadena doble comprenden un sitio de unión de cebadores de NGS ("secuenciación de próxima generación") y un código de barras de ADN. Opcionalmente, la población de adaptadores por lo menos parcialmente de cadena doble comprende una pluralidad de códigos de barras de ADN diferentes. Opcionalmente, el número de combinaciones de códigos de barras que pueden unirse a ambos extremos de una molécula de ADN de cadena doble es menor que el número de moléculas de ADN de cadena doble en la población, por ejemplo, entre 5 y 10.000 combinaciones diferentes. Opcionalmente, el método comprende además: (c) amplificar el ADN etiquetado con adaptador usando cebadores de amplificación que comprenden un código de barras de índice de muestra y una secuencia de nucleótidos adaptada para hibridar con un oligonucleótido inmovilizado en un soporte de celda de flujo. Opcionalmente, los adaptadores son adaptadores con forma de Y. Opcionalmente, la muestra es una muestra de fluido corporal, como sangre completa, suero o plasma. Opcionalmente, la población de ácidos nucleicos es una población de ácidos nucleicos libre de células. Opcionalmente, la muestra es de un sujeto sospechoso de tener cáncer. Opcionalmente, el paso de analizar detecta una variante somática o de línea germinal, una variación del número de copias, una variación de nucleótidos individuales (SNV), y fusión de indeles o genes.

La divulgación proporciona además una población de ácidos nucleicos adaptados producidos por el método de cualquier reivindicación anterior, la población comprendiendo una pluralidad de moléculas de ácido nucleico, cada una de las cuales comprende un fragmento de ácido nucleico flanqueado en ambos lados por un adaptador que incluye un código de barras con una pareja de bases A/T o G/C entre el fragmento de ácido nucleico y el adaptador. Opcionalmente, la pluralidad de moléculas de ácido nucleico es de por lo menos 100.000 moléculas. Opcionalmente, la proporción de parejas de bases A/T a parejas de bases G/C está entre 2:1 y 4:1. Opcionalmente, por lo menos el 99% de las moléculas de ácido nucleico de la población tienen un fragmento de ácido nucleico flanqueado por adaptadores con diferentes códigos de barras.

La divulgación proporciona además un kit que comprende un par de adaptadores por lo menos parcialmente de cadena doble con colas 3' de un solo nucleótido T y C respectivamente, que son idénticos entre sí excepto por las colas. Opcionalmente, los adaptadores son adaptadores con forma de Y que comprenden oligonucleótidos de las SEQ ID NO. 1 y 2, y 3 y 2. Opcionalmente, el kit comprende además una polimerasa T4 o un fragmento grande de Klenow, una polimerasa Taq y cuatro tipos de nucleótidos estándar.

#### BREVE DESCRIPCIÓN DE LA FIGURA

La Figura 1 muestra adaptadores de terminación roma, cola de extremo y unión a cola -T y -C con forma de Y de ADN de muestra.

#### DEFINICIONES

Un sujeto se refiere a un animal, como una especie de mamífero (preferiblemente un humano) o una

especie de ave (por ejemplo, un pájaro) u otro organismo, como una planta. Más específicamente, un sujeto puede ser un vertebrado, por ejemplo, un mamífero como un ratón, un primate, un simio o un humano. Los animales incluyen animales de granja, animales deportivos y mascotas. Un sujeto puede ser un individuo sano, un individuo que tiene o se sospecha que tiene una enfermedad o predisposición a la enfermedad, o un individuo que necesita terapia o se sospecha que necesita terapia.

Una variante genética se refiere a una alteración, variante o polimorfismo en una muestra de ácido nucleico o genoma de un sujeto. Tal alteración, variante o polimorfismo puede ser con respecto a un genoma de referencia, que puede ser un genoma de referencia del sujeto u otro individuo. Las variaciones incluyen una o más variaciones de un solo nucleótido (SNV), inserciones, deleciones, repeticiones, inserciones pequeñas, deleciones pequeñas, repeticiones pequeñas, uniones de variantes estructurales, repeticiones en tándem de longitud variable, y/o secuencias flanqueantes, las variantes del número de copias (CNV), transversiones y otros reordenamientos también son formas de variación genética. Una variación puede ser un cambio de base, inserción, deleción, repetición, variación del número de copias, transversión o una combinación de las mismas.

Un marcador de cáncer es una variante genética asociada con la presencia o el riesgo de desarrollar un cáncer. Un marcador de cáncer puede proporcionar una indicación de que un sujeto tiene cáncer o un mayor riesgo de desarrollar cáncer que un sujeto de la misma edad y género de la misma especie. Un marcador de cáncer puede ser o no causante de cáncer.

Una etiqueta de ácido nucleico es un ácido nucleico corto (por ejemplo, de menos de 100, 50 o 10 nucleótidos de longitud), habitualmente de secuencia artificial y habitualmente ADN, que se usa para marcar ácidos nucleicos de muestra para distinguir los ácidos nucleicos de diferentes muestras (por ejemplo, que representa un índice de muestras), de diferentes tipos o que hayan sido sometidos a procesamiento diferente. Las etiquetas pueden ser de cadena sencilla o doble. Las etiquetas nucleicas pueden decodificarse para revelar información como la muestra de origen, la forma o el procesamiento de un ácido nucleico. Las etiquetas pueden usarse para permitir la agrupación y el procesamiento paralelo de múltiples ácidos nucleicos que llevan diferentes etiquetas con los ácidos nucleicos que se deconvolucionan posteriormente mediante la lectura de las etiquetas. A las etiquetas también puede hacerse referencia como identificadores moleculares o códigos de barras.

Los adaptadores son ácidos nucleicos cortos (por ejemplo, de menos de 500, 100 o 50 nucleótidos de largo y típicamente ADN) habitualmente por lo menos parcialmente de cadena doble para el enlace a uno o ambos extremos de una molécula de ácido nucleico de muestra. Los adaptadores pueden incluir sitios de unión de cebadores para permitir la amplificación de una molécula de ácido nucleico de muestra flanqueada por adaptadores en ambos extremos, y/o un sitio de unión de cebadores de secuenciación, incluyendo sitios de unión de cebadores para secuencias de próxima generación. Los adaptadores también pueden incluir sitios de unión para sondas de captura, como un oligonucleótido unido a un soporte de celda de flujo. Los adaptadores también pueden incluir una etiqueta como se ha descrito anteriormente. Las etiquetas están preferiblemente en posición con respecto al cebador y los sitios de unión del cebador de secuenciación, de tal manera que se incluye una etiqueta en los amplicones y las lecturas de secuenciación de una muestra de ácido nucleico. Pueden enlazarse adaptadores iguales o diferentes a los extremos respectivos de una molécula de muestra. Algunas veces, el mismo adaptador está enlazado a los extremos respectivos, excepto que la etiqueta es diferente. Un adaptador preferido es un adaptador con forma de Y en el que un extremo es un extremo romo o con cola como se describe en la presente, para unirse a un ácido nucleico de muestra, que también tiene un extremo romo o con cola con un nucleótido complementario. Otro adaptador preferido es un adaptador con forma de campana, de igual manera con un extremo romo o con cola para unirse a un ácido nucleico a ser analizado.

Los cuatro tipos de nucleótidos estándar se refieren a A, C, G, T para desoxirribonucleótidos y A, C, T y U para ribonucleótidos.

## DESCRIPCIÓN DETALLADA

### 1. General

La preparación de muestras para plataformas de secuenciación de próxima generación a menudo sigue un protocolo similar. Las muestras contienen típicamente fragmentos de ácido nucleico de cadena doble con salientes de cadena sencilla. Tales fragmentos pueden tener extremos romos y ligarse directamente a adaptadores. Pero tales ligaciones también dan como resultado subproductos en los que los adaptadores o fragmentos forman concatémicos. La formación de tales subproductos puede reducirse mediante un procedimiento alternativo en el que a los fragmentos de extremos romos se les añade cola A y se ligan a adaptadores con cola T. Los kits comerciales que realizan la reparación de extremos y se les añaden colas en un solo tubo son simples de usar y rápidos y pueden usarse con adaptadores disponibles comercialmente. (Por ejemplo, NEBNext Ultra II (New England Biolabs, Ipswich, MA.). Sin embargo, el uso de kits no optimizados para añadir colas A puede dar como resultado la adición de colas con otros nucleótidos, como G, T y C. El resultado de la adición de colas ineficiente es la ligación ineficiente de adaptadores y bibliotecas de baja complejidad.

La invención proporciona métodos mejorados para preparar ácidos nucleicos de cadena doble (preferiblemente ADN) con salientes de cadena sencilla para amplificación y análisis posterior, particularmente secuenciación. Se ha descubierto que poner en contacto ácidos nucleicos de cadena doble de extremos romos con Taq en presencia de los cuatro tipos de nucleótidos estándar da como resultado una adición dirigida sin plantilla de un solo nucleótido a los extremos 3' del ácido nucleico, de tal manera que se añade A seguido con mayor frecuencia por G seguido de C y T. Aunque la inclusión de moléculas de ácido nucleico adicionales aumenta el potencial de reacciones secundarias fuera del objetivo, se ha descubierto que la proporción de la adición de colas G individuales es suficientemente alta con respecto a la adición de colas A individuales para que la eficiencia de ligación de moléculas de ácidos nucleicos en una muestra a adaptadores puede aumentarse significativamente incluyendo una mezcla personalizada de adaptadores con adición de cola no solo con T (como en los métodos anteriores sino también con C, tales adaptadores se aparean respectivamente con los extremos 3' de moléculas de ADN con colas A y G. La eficiencia de ligación puede aumentarse aun más incluyendo también adaptadores con extremos romos (es decir, son colas con cualquier nucleótido) para ligar las moléculas de ácidos nucleicos de extremos romos en la muestra que no han experimentado adición de colas con ningún nucleótido.

## 2. Muestras

Una muestra puede ser cualquier muestra biológica aislada de un sujeto. Las muestras pueden incluir tejidos corporales, como tumores sólidos conocidos o sospechosos, sangre completa, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejido, líquido cefalorraquídeo líquido sinovial, líquido linfático, líquido de ascitis, líquido intersticial o extracelular, el líquido en los espacios entre las células, incluyendo el líquido crevicular gingival, médula ósea, derrames pleurales, líquido cefalorraquídeo, saliva, moco, esputo, semen, sudor, orina. Las muestras son preferiblemente fluidos corporales, particularmente sangre y fracciones de la misma, y orina. Una muestra puede estar en la forma originalmente aislada de un sujeto o puede haber sido sometida a un procesamiento adicional para eliminar o añadir componentes, como células, o enriquecer un componente con respecto a otro. Por tanto, un fluido corporal preferido para análisis es plasma o suero que contiene ácidos nucleicos libres de células.

El volumen de plasma puede depender de la profundidad de lectura deseada para las regiones secuenciadas. Los volúmenes ejemplares son 0,4-40 ml, 5-20 ml, 10-20 ml. Por ejemplo, el volumen puede ser 0,5 ml, 1 ml, 5 ml 10 ml, 20 ml, 30 ml o 40 ml. Un volumen de plasma muestreado puede ser, por ejemplo, de 5 a 20 ml.

Una muestra puede comprender varias cantidades de ácido nucleico que contienen equivalentes de genoma. Por ejemplo, una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 equivalentes de genoma humano haploide y, en el caso de ADN libre de células, aproximadamente 200 billones de moléculas de ácidos nucleicos individuales. De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes de genoma humano haploide y, en el caso de ADN libre de células, aproximadamente 600 billones de moléculas individuales. Algunas muestras contienen 1-500, 2-100, 5-150 ng de ADN libre de células, por ejemplo, 5-30 ng o 10-150 ng de ADN libre de células.

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

Las cantidades ejemplares de ácidos nucleicos libres de células en una muestra antes de la amplificación varían de aproximadamente 1 fg a aproximadamente 1 ug, por ejemplo, de 1 pg a 200 ng, de 1 ng a 100 ng, de 10 ng a 1000 ng. Por ejemplo, la cantidad puede ser de hasta aproximadamente 600 ng, de hasta aproximadamente 500 ng, de hasta aproximadamente 400 ng, de hasta aproximadamente 300 ng, de hasta aproximadamente 200 ng, de hasta aproximadamente 100 ng, de hasta aproximadamente 50 ng, o de hasta aproximadamente 20 ng de moléculas de ácidos nucleicos libres de células. La cantidad puede ser por lo menos 1 fg, por lo menos 10 fg, por lo menos 100 fg, por lo menos 1 pg, por lo menos 10 pg, por lo menos 100 pg, por lo menos 1 ng, por lo menos 10 ng, por lo menos 100 ng, por lo menos por lo menos 150 ng, o por lo menos 200 ng de moléculas de ácidos nucleicos libres de células. La cantidad puede ser de hasta 1 femtogramo (fg), 10 fg, 100 fg, 1 picogramo (pg), 10 pg, 100 pg, 1 ng, 10 ng, 100 ng, 150 ng o 200 ng de moléculas de ácidos nucleicos libres de células. El método puede comprender la obtención de 1 femtogramo (fg) a 200 ng.

Una muestra ejemplar es de 5-10 ml de sangre completa, plasma o suero, que incluye aproximadamente 30 ng de ADN o aproximadamente 10.000 equivalentes de genoma haploide.

Los ácidos nucleicos libres de células son ácidos nucleicos que no están contenidos dentro de una célula o que no están unidos de otro modo a ella o, en otras palabras, los ácidos nucleicos que quedan en una muestra de eliminar células intactas. Los ácidos nucleicos libres de células incluyen ADN, ARN e híbridos de los mismos,

incluyendo ADN genómico, ADN mitocondrial, ARNip, ARNmi, ARN circulante (ARNc), ARNt, ARNr, ARN nucleolar pequeño (ARNsno), ARN que interactúa con Piwi (ARNpi) ARN largo no codificante (ARNc largo), o fragmentos de cualquiera de estos. Los ácidos nucleicos libres de células pueden ser de cadena doble, de cadena sencilla o un híbrido de los mismos. Las moléculas de ADN de cadena doble, por lo menos algunas de las cuales tienen salientes de cadena sencilla, son una forma preferida de ADN libre de células para cualquier método divulgado en la presente. Un ácido nucleico libre de células puede liberarse en el fluido corporal a través de procesos de secreción o muerte celular, por ejemplo, necrosis celular y apoptosis. Algunos ácidos nucleicos libres de células se liberan en los fluidos corporales de las células cancerosas, por ejemplo, el ADN tumoral circulante (ADNct). Otros se liberan de células sanas.

Un ácido nucleico libre de células puede tener una o más modificaciones epigenéticas, por ejemplo, un ácido nucleico libre de células puede acetilarse, metilarse, ubiquitinarse, fosforilarse, sumoizarse, ribosilarse y/o citrulinarse.

Los ácidos nucleicos libres de células tienen una distribución de tamaño de aproximadamente 100-500 nucleótidos, particularmente de 110 a aproximadamente 230 nucleótidos, con un modo de aproximadamente 168 nucleótidos y un segundo pico secundario en un intervalo de 240 a 440 nucleótidos.

Los ácidos nucleicos libres de células pueden aislarse de los fluidos corporales mediante un paso de división en el que los ácidos nucleicos libres de células, tal como se encuentran en solución, se separan de las células intactas y otros componentes no solubles del fluido corporal. La división puede incluir técnicas como centrifugación o filtración. Alternativamente, las células en los fluidos corporales pueden lisarse y los ácidos nucleicos celulares y libres de células pueden procesarse juntos. Generalmente, después de la adición de tampones y pasos de lavado, los ácidos nucleicos pueden precipitarse con un alcohol. Pueden usarse pasos de limpieza adicionales, como columnas a base de sílice, para eliminar contaminantes o sales. Pueden añadirse ácidos nucleicos portadores a granel no específicos, por ejemplo, a lo largo de la reacción para optimizar ciertos aspectos del procedimiento, como el rendimiento.

Después de dicho procesamiento, las muestras pueden incluir varias formas de ácido nucleico, incluyendo ADN de cadena doble, ADN de cadena sencilla y ARN de cadena sencilla. Opcionalmente, el ADN y el ARN de cadena sencilla pueden convertirse en formas de cadena doble para que se incluyan en los pasos de procesamiento y análisis posteriores.

### 3. Enlace de moléculas de ácidos nucleicos de muestra a adaptadores

El ácido nucleico presente en una muestra con o sin procesamiento previo como se ha descrito anteriormente contiene típicamente una porción sustancial de moléculas en forma de moléculas parcialmente de cadena doble con salientes de cadena sencilla. Tales moléculas pueden convertirse en moléculas de cadena doble de extremos romos mediante el tratamiento con una o más enzimas para proporcionar una polimerasa 5'-3' y una exonucleasa 3'-5' (o función de lectura de prueba), en presencia de los cuatro tipos de nucleótidos estándar como se muestra en la Fig. 1, parte superior. Esta combinación de actividades puede extender las cadenas con un extremo 3' empotrado para que terminen al ras con el extremo de 5' de la cadena opuesta (en otras palabras, generando un extremo romo) o puede digerir cadenas con salientes 3' para que también estén al ras con el extremo 5' de la cadena opuesta. Opcionalmente, ambas actividades pueden ser conferidas por una única polimerasa. Preferiblemente, la polimerasa es sensible al calor, de tal manera que su actividad puede interrumpirse cuando se eleva la temperatura. El fragmento grande de Klenow y la polimerasa T4 son ejemplos de polimerasa adecuada.

Las una o más enzimas que confieren la actividad de polimerasa 5'-3' y una exonucleasa 3'-5' se desnaturalizan preferiblemente aumentando la temperatura o de otro modo. Por ejemplo, la desnaturalización puede efectuarse elevando la temperatura, por ejemplo, a 75° -80° C. Luego, las muestras se activan mediante una polimerasa que carece de función de lectura de prueba (Fig. 1, parte media). Esta polimerasa es preferiblemente termoestable de tal manera que permanezca activa a la temperatura elevada. Las polimerasas Taq, fragmento grande de Bst y Tth son ejemplos de tal polimerasa. La segunda polimerasa efectúa una adición sin plantilla de un solo nucleótido a los extremos 3' de los ácidos nucleicos de extremos romos. Aunque la mezcla de la reacción contiene típicamente cantidades molares iguales de cada uno de los cuatro tipos de nucleótidos estándar del paso anterior, los cuatro tipos de nucleótidos no se añaden a los extremos 3' en proporciones iguales. Más bien, A se añade con mayor frecuencia, seguido de G seguido de C y T.

Después de la adición de cola de las moléculas de muestra, y con o sin la purificación posterior de las moléculas de muestra con cola, las moléculas de la muestra con cola se ponen en contacto con adaptadores con cola con nucleótidos T y C complementarios en un extremo de los adaptadores (Fig. 1, parte inferior). Los adaptadores se forman típicamente mediante síntesis separada e hibridación de sus cadenas respectivas. Las colas de T y C adicionales pueden, por tanto, añadirse como un nucleótido extra en la síntesis de una de las cadenas. Típicamente, los adaptadores con adición de cola con G y A no se incluyen porque, aunque estos adaptadores pueden aparearse con moléculas de muestra con adición de cola con C y T, respectivamente, también se aparean

con otros adaptadores. Las moléculas adaptadoras y las moléculas de muestra que portan nucleótidos complementarios (es decir, T-A y C-G) en sus extremos 3' se aparean y pueden ligarse entre sí. El porcentaje de adaptadores con cola C con respecto a los adaptadores con cola T varía de aproximadamente el 5-40% en moles, por ejemplo, 10-35%, 15-25%, 20-35%, 25-35% o aproximadamente el 30%. Debido a que la adición no dirigida por plantillas de un solo nucleótido a los extremos 3' de las moléculas de muestra no se completa, la muestra también contiene algunas moléculas de muestra de extremos romos sin adición de colas. Estas moléculas pueden recuperarse suministrando también a la muestra adaptadores que tengan un, y preferiblemente solo un, extremo romo. Los adaptadores de extremos romos se suministran habitualmente en una proporción molar del 0,2-20%, o 0,5-15% o 1-10% de adaptadores con adaptadores con cola T y C. Pueden proporcionarse adaptadores de extremos romos al mismo tiempo, antes o después de los adaptadores con cola T y C. Los adaptadores de extremos romos ligados con moléculas de muestra de extremos romos de nuevo da como resultado moléculas de muestra flanqueadas en ambos lados por adaptadores. Estas moléculas carecen de las parejas de nucleótidos A-T o C-G entre la muestra y los adaptadores presentes cuando las moléculas de muestra con cola se ligan a adaptadores con cola.

Los adaptadores usados en estas reacciones tienen preferiblemente uno y solo un extremo con adición de cola T o C o uno y solo un extremo romo para que puedan ligarse con moléculas de muestra en solo una orientación. Los adaptadores pueden ser, por ejemplo, adaptadores con forma de Y en los que un extremo tiene una cola o es romo y el otro extremo tiene dos cadenas individuales. Los adaptadores con forma de Y ejemplares tienen las siguientes secuencias con (6 bases) que indican una etiqueta. El oligonucleótido superior incluye una cola T de base única.

Adaptador universal:

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT (SEQ ID NO. 1).

Adaptador, índice 1-12: 5' GATCGGAAGAGCACACGTCTGAACTCCAGTCAC (6 bases)  
ATCTCGTATGCCGTCTTCTGCTTG (SEQ ID NO. 2)

Otro adaptador con forma de Y con una cola C tiene las secuencias:

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCC (SEQ ID NO. 3) y  
Adaptador, índice 1-12: 5' GATCGGAAGAGCACACGTCTGAACTCCAGTCAC (6 bases)  
ATCTCGTATGCCGTCTTCTGCTTG (SEQ ID NO. 2).

Para su uso en los presentes métodos pueden sintetizarse combinaciones personalizadas de tales oligonucleótidos incluyendo oligonucleótidos con colas tanto T como C.

Una versión truncada de estas secuencias adaptadoras ha sido descrita por Rohland et al., Genome Res. mayo 2012; 22(5): 939-946.

Los adaptadores también pueden tener forma de campana con un solo extremo, que tiene cola o es romo. Los adaptadores pueden incluir un sitio de unión a cebadores para amplificación, un sitio de unión para un cebador de secuenciación y/o una etiqueta de ácido nucleico con propósitos de identificación. Pueden usarse adaptadores iguales o diferentes en una única reacción.

Cuando los adaptadores incluyen una etiqueta de identificación y los ácidos nucleicos en una muestra se unen a los adaptadores en cada extremo, el número de combinaciones posibles de identificadores aumenta exponencialmente con el número de etiquetas únicas suministradas (es decir,  $n^n$  combinaciones, donde  $n$  es el número de etiquetas de identificación únicas). En algunos métodos, el número de combinaciones de etiquetas únicas es suficiente para que sea estadísticamente probable que todas o sustancialmente todas (por ejemplo, por lo menos el 90%) de las diferentes moléculas de ADN de cadena doble de la muestra reciban una combinación diferente de etiquetas. En algunos métodos, el número de combinaciones únicas de etiquetas de identificación es menor que el número de moléculas de ADN de cadena doble únicas en la muestra (por ejemplo, 5-10.000 combinaciones de etiquetas diferentes).

Un kit que proporciona las enzimas adecuadas para realizar los métodos anteriores es el kit de preparación de bibliotecas de ADN NEB Next® Ultra™ II para Illumina®. El kit proporciona los siguientes reactivos

Mezcla de enzimas preparativa NEBNext Ultra II End, Tampón de reacción preparativo NEBNext Ultra II End, potenciador de ligación NEBNext, mezcla maestra de ligación NEBNext Ultra II -20, mezcla maestra NEBNext® Ultra II Q5®.

Los extremos romos y la adición de colas de los ácidos nucleicos de muestra pueden realizarse en un solo tubo. Los ácidos nucleicos de extremos romos no necesitan separarse de la enzima o enzimas que realizan el

extremo romo antes de que se produzca la reacción de adición de cola. Opcionalmente, todas las enzimas, nucleótidos y otros reactivos se suministran juntos antes de que se produzca la reacción de extremo romo. Suministrar juntos significa que todos se introducen en la muestra lo suficientemente cerca en el tiempo de tal manera que todos estén presentes cuando se produce la incubación de la muestra para que se produzca un extremo romo. Opcionalmente, no se elimina nada de las muestras después de suministrar las enzimas, nucleótidos y otros reactivos por lo menos hasta que se hayan completado las incubaciones de los extremos romos y de la adición de colas de los extremos. A menudo, la reacción de adición de colas a los extremos se realiza a una temperatura más alta que la reacción de extremos romos. Por ejemplo, la reacción de extremos romos puede realizarse a temperatura ambiente en la que la polimerasa 5'-3' y la exonucleasa 3'-5' son activas y la polimerasa termoestable es inactiva o es mínimamente activa, y la reacción de adición de colas a los extremos realizarse a una temperatura elevada, como por encima de 60° C, cuando la polimerasa 5'-3' y la exonucleasa 3'-5' son inactivas y la polimerasa termoestable es activa.

La unión de adaptadores de cola T y C como se describe da como resultado una población de ácidos nucleicos adaptados, la población comprendiendo una pluralidad de moléculas de ácidos nucleicos, cada una de las cuales comprende un fragmento de ácido nucleico flanqueado en ambos lados por un adaptador que incluye un código de barras con una pareja de bases A/T o G/C entre el fragmento de ácido nucleico y el adaptador. La pluralidad de moléculas de ácidos nucleicos puede ser de por lo menos 10.000, 100.000 o 1.000.000 de moléculas. La proporción de parejas de bases A/T a parejas de bases G/C en las regiones de unión entre fragmentos y adaptadores flanqueantes depende de la proporción de adaptadores con cola T y C y estar, por ejemplo, entre 2:1 y 4:1. La mayoría de los ácidos nucleicos de la población están flanqueados por adaptadores con diferentes códigos de barras (por ejemplo, por lo menos el 99%). Si también se incluyen adaptadores de extremos romos, entonces la población incluye moléculas de ácidos nucleicos en un fragmento de ácido nucleico se une directamente a cualquiera o ambos extremos a un adaptador (es decir sin parejas A/T o G/C intermedias).

#### 4. Amplificación

Los ácidos nucleicos de muestra flanqueados por adaptadores pueden amplificarse mediante PCR y otros métodos de amplificación típicamente cebados a partir de cebadores que se unen a sitios de unión de cebadores en adaptadores que flanquean un ácido nucleico a amplificar. Los métodos de amplificación pueden implicar ciclos de extensión, desnaturalización y apareamiento que resultan del termociclado o pueden ser isotérmicos como en la amplificación mediada por transcripción. Otros métodos de amplificación incluyen la reacción en cadena de la ligasa, la amplificación por desplazamiento de cadena, la amplificación basada en la secuencia de ácido nucleico, y la replicación basada en secuencia autosostenida.

Preferiblemente, los presentes métodos dan como resultado que por lo menos un 75, 80, 85, 90 o 95% de ácidos nucleicos de cadena doble en la muestra se enlacen a adaptadores. Preferiblemente, el uso de adición de colas T y C aumenta el porcentaje de ácidos nucleicos de cadena doble en la muestra enlazada a adaptadores con respecto a los métodos de control realizados con adaptadores de cola T solos en por lo menos un 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10% (un aumento de rendimiento del 75% al 80% se considera un aumento del 5%). Preferiblemente, el uso de adición de colas T y C en combinación con adaptadores de extremos romos aumenta el porcentaje de ácidos nucleicos de cadena doble enlazados a adaptadores en por lo menos un 5, 10, 15, 20 o 25%. El porcentaje de ácidos nucleicos enlazados a los adaptadores puede determinarse mediante electroforesis en gel comparativa de la muestra original y la muestra procesada después de que se haya completado el enlace con los adaptadores.

Preferiblemente, los presentes métodos dan como resultado por lo menos un 75, 80, 85, 90 o 95% de las moléculas de cadena doble disponibles en la muestra que se está secuenciando. Preferiblemente, el uso de adición de colas T y C aumenta el porcentaje de ácidos nucleicos de cadena doble en la muestra que se está secuenciando con respecto a los métodos de control realizados con adaptadores con cola T solos en por lo menos un 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10%. Preferiblemente, el uso de adición de colas T y C en combinación con adaptadores de extremos romos aumenta el porcentaje de ácidos nucleicos de cadena doble en la muestra que se está secuenciando con respecto a los métodos de control realizados con adaptadores con cola T junto con por lo menos un 5, 10, 15, 20 o 25%. El porcentaje de ácidos nucleicos que se están secuenciando puede determinarse comparando el número de moléculas secuenciadas realmente en base al número que podría hacerse secuenciado en base a los ácidos nucleicos de entrada y las regiones del genoma objetivo para la secuenciación.

#### 5. Etiquetas

Pueden incorporarse o unirse de otro modo etiquetas que proporcionan identificadores moleculares o códigos de barras a los adaptadores mediante ligación, PCR de extensión por superposición, entre otros métodos. Generalmente, la asignación de identificadores únicos o no únicos, o códigos de barras moleculares en reacciones sigue métodos y sistemas descritos por Solicitudes de Patente de Estados Unidos 20010053519, 20030152490, 20110160078, y la Patente de Estados Unidos Nº 6.582.908 y la Patente de Estados Unidos Nº 7.537.898.

Las etiquetas pueden enlazarse a los ácidos nucleicos de la muestra aleatoria o no aleatoriamente. En



algunos casos, se introducen en una proporción esperada de identificadores únicos a micropocillos. Por ejemplo, los identificadores únicos pueden cargarse de tal manera que se cargan más de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 identificadores únicos por muestra de genoma. En algunos casos, los identificadores únicos pueden cargarse de tal manera que se cargan menos de aproximadamente 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 identificadores únicos por muestra de genoma. En algunos casos, el número medio de identificadores únicos cargados por genoma de muestra es menor o mayor de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 identificadores únicos por muestra de genoma.

En algunos casos, los identificadores únicos pueden ser oligonucleótidos de secuencia predeterminada o aleatoria o semialeatoria. En otros casos, pueden usarse una pluralidad de códigos de barras de tal manera que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. En este ejemplo, los códigos de barras pueden ligarse a moléculas individuales de tal manera que la combinación del código de barras y la secuencia a la que puede ligarse crea una secuencia única que puede rastrearse individualmente. Como se describe en la presente, la detección de códigos de barras no únicos en combinación con datos de secuencia de porciones de inicio (comienzo) y final (parada) de lecturas de secuencia puede permitir la asignación de una identidad única a una molécula particular. La longitud, o el número de pares de bases, de una secuencia individual leída también puede usarse para asignar una identidad única a dicha molécula. Como se describe en la presente, los fragmentos de una cadena sencilla de ácidos nucleicos a los que se ha asignado una identidad única pueden permitir de este modo la identificación posterior de fragmentos de la cadena original.

Los polinucleótidos en una muestra pueden etiquetarse con un número suficiente de etiquetas diferentes para que exista una alta probabilidad (por ejemplo, por lo menos el 90%, por lo menos el 95%, por lo menos el 98%, por lo menos el 99%, por lo menos el 99,9% o por lo menos el 99,99%) de que todos los polinucleótidos que se mapean en una región genómica particular porten una etiqueta de identificación diferente (las moléculas dentro de la región están etiquetadas sustancialmente de forma única). La región genómica a la que se asignan los polinucleótidos puede ser, por ejemplo, (1) el panel completo de genes que se están secuenciando, (2) alguna parte de ese panel, como el mapeo dentro de un único gen, exón o intrón, (3) un coordenada de un solo nucleótido (por ejemplo, por lo menos un nucleótido en los mapas de polinucleótidos a la coordenada, por ejemplo, la posición de inicio, la posición de parada, el punto medio o cualquier punto entre ellos) o (4) un par particular de coordenadas de nucleótidos de inicio/parada (comienzo/fin). El número de identificadores diferentes (recuentos de etiquetas) necesarios para etiquetar polinucleótidos de manera sustancialmente única es una función de cuántas moléculas de polinucleótidos originales en la muestra se mapean en la región. Esto, a su vez, es una función de varios factores. Un factor es el número total de equivalentes del genoma haploide incluidos en el ensayo. Otro factor es el tamaño medio de las moléculas de polinucleótidos. Otro factor es la distribución de las moléculas en la región. Esto, a su vez, puede ser una función del patrón de escisión - puede esperarse que la escisión se produzca principalmente entre nucleosomas, de tal manera que se mapeen más polinucleótidos a través de una localización de nucleosoma que entre nucleosomas. Otro factor es la distribución de códigos de barras en el grupo y la eficiencia de ligación de códigos de barras individuales, causando potencialmente diferencias en la concentración efectiva de un código de barras frente a otro. Otro factor es el tamaño de la región dentro de la cual están confinadas las moléculas que se van a marcar de manera única (por ejemplo, el mismo inicio/parada o el mismo exón).

El identificador puede ser un código de barras único unido a un extremo de una molécula, o dos códigos de barras, cada uno unido a diferentes extremos de la molécula. Unir códigos de barras de manera independiente a ambos extremos de una molécula aumenta al cuadrado el número de identificadores posibles. En este caso, el número de códigos de barras diferentes se selecciona de tal manera que la combinación de códigos de barras en cada extremo de un polinucleótido particular tenga una alta probabilidad de ser única con respecto a otros polinucleótidos que mapean la misma región genómica seleccionada.

En ciertas realizaciones, el número de identificadores diferentes o combinaciones de códigos de barras (recuento de etiquetas) usados puede ser por lo menos cualquiera de 64, 100, 400, 900, 1400, 2500, 5625, 10.000, 14.400, 22.500 o 40.000 y no más de cualquiera de 90.000, 40.000, 22.500, 14.400 o 10.000. Por ejemplo, el número de identificadores o combinaciones de códigos de barras puede estar entre 64 y 90.000, entre 400 y 22.500, 400 y 14.400 o entre 900 y 14.400.

En una muestra que comprende ADN genómico fragmentado, por ejemplo, ADN libre de células (ADNcf), de una pluralidad de genomas, hay cierta probabilidad de que más de un polinucleótido de diferentes genomas tenga las mismas posiciones de inicio y finalización ("duplicados" o "cognados"). El número probable de duplicados que comienzan en cualquier posición es función del número de equivalentes del genoma haploide en una muestra y la distribución de los tamaños de los fragmentos. Por ejemplo, el ADNcf tiene un pico de fragmentos a aproximadamente 160 nucleótidos, y la mayoría de los fragmentos en este pico varían aproximadamente de 140 nucleótidos a 180 nucleótidos. En consecuencia, El ADNcf de un genoma de aproximadamente 3 billones de bases (por ejemplo, el genoma humano) puede estar compuesto por casi 20 millones ( $2 \times 10^7$ ) fragmentos de

polinucleótidos. Una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 equivalentes de genoma humano haploide. (De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes de genoma humano haploide.) Una muestra que contiene aproximadamente 10.000 ( $10^4$ ) equivalentes de genoma haploide de tal ADN pueden tener aproximadamente 200 billones ( $2 \times 10^{11}$ ) moléculas de polinucleótidos individuales. Se ha determinado empíricamente que en una muestra de aproximadamente 10.000 equivalentes de genoma haploide de ADN humano, hay aproximadamente 3 polinucleótidos duplicados que comienzan en cualquier posición dada. Por tanto, una colección de este tipo puede contener una diversidad de aproximadamente  $6 \times 10^{10}$ - $8 \times 10^{10}$  (aproximadamente 60 billones-80 billones, por ejemplo, aproximadamente 70 billones ( $7 \times 10^{10}$ )) moléculas de polinucleótidos secuenciadas de manera diferente.

La probabilidad de identificar correctamente las moléculas depende del número inicial de equivalentes del genoma, la distribución de longitud de las moléculas secuenciadas, la uniformidad de la secuencia y el número de etiquetas. El número puede calcularse usando una distribución de Poisson. Cuando el recuento de etiquetas es igual a uno, es decir, equivalente a no tener etiquetas únicas o no tener etiquetas. La Tabla 1 a continuación enumera la probabilidad de identificar correctamente una molécula como única asumiendo una distribución de tamaño libre de células típica como la anterior.

Tabla 1

Recuento de etiquetas	% de etiquetas identificadas correctamente de forma única
1000 equivalentes del genoma haploide humano	
1	96.9643
4	99.2290
9	99.6539
dieciséis	99.8064
25	99.8741
100	99.9685
3000 equivalentes del genoma haploide humano	
1	91.7233
4	97.8178
9	99.0198
16	99.4424
25	99.6412
100	99.9107

En este caso, tras secuenciar el ADN genómico, puede que no sea posible determinar qué lecturas de secuencia se derivan de qué moléculas originales. Este problema puede disminuirse etiquetando las moléculas originales con un número suficiente de identificadores únicos (por ejemplo, el recuento de etiquetas) de tal manera que exista la probabilidad de que dos moléculas duplicadas, es decir, moléculas que tienen las mismas posiciones de inicio y parada, tengan identificadores únicos diferentes, por lo que esas lecturas de secuencia son rastreables hasta moléculas originales particulares. Un enfoque para este problema es etiquetar de manera única todas o casi todas las moléculas originales diferentes de la muestra. Sin embargo, dependiendo del número de equivalentes de genes haploides y la distribución de los tamaños de los fragmentos en la muestra, esto puede requerir billones de identificadores únicos diferentes.

Este método puede resultar engorroso y caro. En algunos aspectos, se proporcionan en la presente métodos y composiciones en los que una población de polinucleótidos en una muestra de ADN genómico fragmentado se etiqueta con  $n$  identificadores únicos diferentes, en donde  $n$  es por lo menos 2 y no más de  $100.000 \times z$ , en donde  $z$  es una medida de tendencia central (por ejemplo, media, mediana, moda) de un número esperado de moléculas duplicadas que tienen las mismas posiciones de inicio y parada. En ciertas realizaciones,  $n$  es por lo menos cualquiera de  $2 \times z$ ,  $3 \times z$ ,  $4 \times z$ ,  $5 \times z$ ,  $6 \times z$ ,  $7 \times z$ ,  $8 \times z$ ,  $9 \times z$ ,  $10 \times z$ ,  $11 \times z$ ,  $12 \times z$ ,  $13 \times z$ ,  $14 \times z$ ,  $15 \times z$ ,  $16 \times z$ ,  $17 \times z$ ,  $18 \times z$ ,  $19 \times z$ ,  $20 \times z$  o  $100 \times z$  (por ejemplo, límite inferior). En otras realizaciones,  $n$  no es mayor de  $100.000 \times z$ ,  $10.000 \times z$ ,  $2000 \times z$ ,  $1000 \times z$ ,  $500 \times z$  o  $100 \times z$  (por ejemplo, límite superior). Por tanto,  $n$  puede variar entre cualquier combinación de estos límites superiores e inferiores. En ciertas realizaciones,  $n$  está entre  $100 \times z$  y  $1000 \times z$ ,  $5 \times z$  y  $15 \times z$ , entre  $8 \times z$  y  $12 \times z$ , o aproximadamente  $10 \times z$ . Por ejemplo, un equivalente de genoma humano haploide tiene aproximadamente 3 picogramos de ADN. Una muestra de aproximadamente 1 microgramo de ADN contiene aproximadamente 300.000 equivalentes del genoma haploide humano. El número  $n$  puede estar entre 15 y 45, entre 24 y 36, entre 64 y 2500, entre 625 y 31.000, o entre 900 y 4000. Pueden lograrse mejoras en la secuenciación siempre que por lo menos algunos de los polinucleótidos duplicados o cognados porten identificadores únicos, es decir, porten etiquetas

diferentes. Sin embargo, en ciertas realizaciones, el número de etiquetas usadas se selecciona de tal manera que haya por lo menos un 95% de posibilidades de que todas las moléculas duplicadas que comienzan en cualquier posición porten identificadores únicos. Por ejemplo, una muestra que comprende aproximadamente 10.000 equivalentes del genoma humano haploide de ADNcf puede etiquetarse con aproximadamente 36 identificadores únicos. Los identificadores únicos pueden comprender seis códigos de barras de ADN únicos. Unidos a ambos extremos de un polinucleótido, se producen 36 posibles identificadores únicos. Las muestras marcadas de esta manera pueden ser aquellas con un intervalo de aproximadamente 10 ng a cualquiera de aproximadamente 100 ng, aproximadamente 1 µg, aproximadamente 10 µg de polinucleótidos fragmentados, por ejemplo, ADN genómico, por ejemplo, ADNcf.

Por consiguiente, la presente divulgación también proporciona composiciones de polinucleótidos etiquetados. Los polinucleótidos pueden comprender ADN fragmentado, por ejemplo, ADNcf. Un conjunto de polinucleótidos en la composición que mapean una posición de base mapeable en un genoma puede etiquetarse de manera no única, es decir, el número de identificadores diferentes puede ser por lo menos 2 y menos que el número de polinucleótidos que mapean la posición base mapeable. Una composición de entre aproximadamente 10 ng y aproximadamente 10 µg (por ejemplo, cualquiera de aproximadamente 10 ng-1 µg, aproximadamente 10 ng-100 ng, aproximadamente 100 ng-10 µg, aproximadamente 100 ng-1 µg, aproximadamente 1 µg-10 µg) puede portar entre cualquiera de 2, 5, 10, 50 o 100 a cualquiera de 100, 1000, 10.000 o 100.000 identificadores diferentes. Por ejemplo, pueden usarse entre 5 y 100 o entre 100 y 4000 identificadores diferentes para marcar los polinucleótidos en dicha composición.

A los eventos en los que se mapean diferentes moléculas en la misma coordenada (en este caso que tienen las mismas posiciones de inicio/parada) y portan las mismas etiquetas, en lugar de diferentes, se hace referencia como "colisiones moleculares". En ciertos casos, el número real de colisiones moleculares puede ser mayor que el número de colisiones teóricas, calculado, por ejemplo, como anteriormente. Esto puede ser una función de la distribución desigual de moléculas a través de coordenadas, diferencias en la eficiencia de la ligación entre códigos de barras y otros factores. En este caso, pueden usarse métodos empíricos para determinar el número de códigos de barras necesarios para aproximarse al número de colisión teórico. En una realización, se proporciona en la presente un método para determinar un número de códigos de barras requeridos para disminuir las colisiones de códigos de barras para un equivalente de genoma haploide dado en base a la distribución de longitud de las moléculas secuenciadas y la uniformidad de la secuencia. El método comprende crear una pluralidad de grupos de moléculas de ácidos nucleicos; etiquetar cada grupo con un número incrementalmente creciente de códigos de barras; y determinar un número óptimo de códigos de barras que reduzca el número de colisiones de códigos de barras a un nivel teórico, por ejemplo, que podría deberse a diferencias en las concentraciones de códigos de barras afectivas debido a diferencias en la eficiencia de agrupación y ligación.

En una realización, puede determinarse empíricamente el número de identificadores necesarios para etiquetar de manera sustancialmente única el mapeo de polinucleótidos en una región. Por ejemplo, puede unirse un número seleccionado de identificadores diferentes a moléculas en una muestra, y puede contarse el número de identificadores diferentes para el mapeo de moléculas en la región. Si se usa un número insuficiente de identificadores, algunos polinucleótidos que mapean la región llevarán el mismo identificador. En ese caso, el número de identificadores contados será menor que el número de moléculas originales en la muestra. El número de identificadores diferentes usados puede aumentarse iterativamente para un tipo de muestra hasta que no se detecten identificadores adicionales que representen nuevas moléculas originales. Por ejemplo, en una primera iteración, pueden contarse cinco identificadores diferentes, que representan por lo menos cinco moléculas originales diferentes. En una segunda iteración, usando más códigos de barras, se cuentan siete identificadores diferentes, que representan por lo menos siete moléculas originales diferentes. En una tercera iteración, usando más códigos de barras, se cuentan 10 identificadores diferentes, que representan por lo menos diez moléculas originales diferentes. En una cuarta iteración, usando más códigos de barras, se cuentan de nuevo 10 identificadores diferentes. En este punto, no es probable que añadir más códigos de barras aumente el número de moléculas originales detectadas.

## 6. Secuenciación

Los ácidos nucleicos de muestra flanqueados por adaptadores con o sin amplificación previa pueden someterse a secuenciación. Los métodos de secuenciación incluyen, por ejemplo, secuenciación de Sanger, secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de una única molécula, secuenciación de nanoporos, secuenciación de semiconductores, secuenciación por ligación, secuenciación por hibridación, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Secuenciación de próxima generación, Secuenciación de una única molécula por síntesis (SMSS) (Helicos), Secuenciación masivamente paralela, Matriz de moléculas individuales clonales (Solexa), secuenciación de escopeta, Ion Torrent, nanoporos de Oxford, Roche Genia, secuenciación de Maxim-Gilbert, caminata de cebadores, secuenciación usando plataformas PacBio, SOLiD, Ion Torrent o Nanopore. Las reacciones de secuenciación pueden realizarse en una variedad de unidades de procesamiento de muestras, que pueden tener múltiples carriles, múltiples canales, múltiples pocillos u otro medio de procesamiento de múltiples conjuntos de muestras de manera sustancialmente simultánea. La unidad de procesamiento de muestras también puede incluir múltiples cámaras de muestras para permitir el procesamiento

de múltiples series simultáneamente.

Las reacciones de secuenciación pueden realizarse en uno más tipos de fragmentos que se sabe que contienen marcadores de cáncer de otra enfermedad. Las reacciones de secuenciación también pueden realizarse en cualquier fragmento de ácido nucleico presente en la muestra. Las reacciones de secuencia pueden proporcionar una cobertura de secuencia del genoma de por lo menos un 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99,9% o 100%. En otros casos, la cobertura de la secuencia del genoma puede ser inferior al 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95 %, 99%, 99,9% o 100%.

Pueden realizarse reacciones de secuenciación simultáneas usando secuenciación multiplex. En algunos casos, los ácidos nucleicos libres de células pueden secuenciarse con por lo menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. En otros casos, los polinucleótidos libres de células pueden secuenciarse con menos de 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. Las reacciones de secuenciación pueden realizarse de manera secuencial o simultánea. Pueden realizarse análisis de datos posteriores en todas o en parte de las reacciones de secuenciación. En algunos casos, el análisis de datos puede realizarse en por lo menos 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación. En otros casos, el análisis de datos puede realizarse en menos de 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100.000 reacciones de secuenciación.

El método de secuenciación puede ser secuenciación masivamente paralela, es decir, secuenciar simultáneamente (o en rápida sucesión) cualquiera de por lo menos 100, 1000, 10.000, 100.000, 1 millón, 10 millones, 100 millones o 1 billón de moléculas de ácido nucleico.

## 7. Análisis

Los presentes métodos pueden usarse para diagnosticar la presencia de afecciones, particularmente cáncer, en un sujeto, para caracterizar afecciones (por ejemplo, estadificación del cáncer o determinación de la heterogeneidad de un cáncer), monitorizar la respuesta al tratamiento de una afección, efectuar el pronóstico de riesgo de desarrollar una afección o el curso posterior de una afección.

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

Los tipos y la cantidad de cánceres que pueden detectarse pueden incluir cánceres de sangre, cánceres de cerebro, cánceres de pulmón, cánceres de piel, cánceres de nariz, cánceres de garganta, cánceres de hígado, cánceres de huesos, linfomas, cánceres de páncreas, cánceres de piel, cánceres de intestino, cánceres de recto, cánceres de tiroides, cánceres de vejiga, cánceres de riñón, cánceres de boca, cánceres de estómago, tumores en estado sólido, tumores heterogéneos, tumores homogéneos y similares.

Los cánceres pueden detectarse a partir de variaciones genéticas que incluyen mutaciones, mutaciones raras, indeles, variaciones en el número de copias, transversiones, translocaciones, inversiones, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones de genes, fusiones de cromosomas, truncamientos de genes, amplificación de genes, duplicaciones de genes, lesiones cromosómicas, lesiones de ADN, cambios anormales en las modificaciones químicas del ácido nucleico, cambios anormales en los patrones epigenéticos, cambios anormales en la infección por metilación del ácido nucleico y cáncer.

También pueden usarse datos genéticos para caracterizar una forma específica de cáncer. Los cánceres suelen ser heterogéneos tanto en tanto composición como en estadios. Los datos del perfil genético pueden permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes en el diagnóstico o tratamiento de ese subtipo específico. Esta información también puede proporcionar pistas al sujeto o al médico con respecto al pronóstico de un tipo específico de cáncer y permitirle al sujeto o al médico adaptar las opciones de tratamiento de acuerdo con el progreso de la enfermedad. Algunos cánceres progresan, volviéndose más agresivos y genéticamente inestables. Otros cánceres pueden permanecer benignos, inactivos o latentes. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad.

El presente análisis también es útil para determinar la eficacia de una opción de tratamiento particular. Las opciones de tratamiento con éxito pueden aumentar la cantidad de variación en el número de copias o las

mutaciones raras detectadas en la sangre del sujeto si el tratamiento tiene éxito, ya que más cánceres pueden morir y perder ADN. En otros ejemplos, esto puede no producirse. En otro ejemplo, quizás ciertas opciones de tratamiento puedan estar correlacionadas con perfiles genéticos de cánceres a lo largo del tiempo. Esta correlación puede resultar útil para seleccionar una terapia. Además, si se observa que un cáncer está en remisión después del tratamiento, los presentes métodos pueden usarse para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

Los presentes métodos también pueden usarse para detectar variaciones genéticas en afecciones distintas del cáncer. Las células inmunes, como las células B, pueden experimentar una rápida expansión clonal ante la presencia de ciertas enfermedades. Las expansiones clonales pueden monitorizarse usando la detección de la variación del número de copias y pueden monitorizarse ciertos estados inmunes. En este ejemplo, el análisis de la variación del número de copias puede realizarse a lo largo del tiempo para producir un perfil de cómo puede estar progresando una enfermedad en particular. La variación del número de copias o incluso la detección de mutaciones raras pueden usarse para determinar cómo está cambiando una población de patógenos durante el curso de la infección. Esto puede ser particularmente importante durante las infecciones crónicas, como las infecciones por VIH/SIDA o hepatitis, en las que los virus pueden cambiar el estado del ciclo de vida y/o mutar en formas más virulentas durante el curso de la infección. Los presentes métodos pueden usarse para determinar o hacer perfiles de rechazo de actividades del cuerpo del huésped, como las células inmunes que intentan destruir tejido trasplantado para monitorizar el estado del tejido trasplantado así como alterar el curso de tratamiento o prevención del rechazo.

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

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

Los presentes métodos pueden usarse para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades de origen fetal. Es decir, estas metodologías pueden emplearse en un sujeto embarazado para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades en un sujeto no nacido cuyo ADN y otros ácidos nucleicos pueden co-circular con moléculas maternas.

## 9. Kits

La divulgación también proporciona kits para la puesta en práctica de cualquiera de los métodos anteriores. Un kit ejemplar incluye un par de adaptadores por lo menos parcialmente de cadena doble con colas 3' de un solo nucleótido T y C respectivamente. Preferiblemente, los oligonucleótidos emparejados son idénticos excepto por las colas T y C. Opcionalmente, el kit está libre de adaptadores por lo menos parcialmente de cadena doble con colas 3' de un solo nucleótido A y G. Preferiblemente, los adaptadores tienen forma de Y, como adaptadores que comprenden oligonucleótidos de las SEQ ID NOS. 1 y 2, y 3 y 2. Los kits también pueden incluir enzimas para la puesta en práctica de los métodos, como polimerasa T4 o fragmento grande de Klenow y/o polimerasa Taq, y opcionalmente los cuatro tipos de nucleótidos estándar. Los kits también pueden incluir envases, folletos, CD o similares que proporcionan instrucciones para la puesta en práctica de los métodos reivindicados.

## Ejemplos

El uso de adaptadores con cola C y T contribuyó a aumentar la sensibilidad al capturar más moléculas en una muestra. Los adaptadores C se probaron en proporciones que varían de 0 a 1:2,75 (36%) con respecto a los adaptadores T, como se muestra en la Tabla 2 a continuación.

Tabla 2

Nº de Muestra	Entrada (ng)	Cola T (40 uM)	Cola C (40 uM)	% LIG
1	20	3,25	0,5	80%
2	20	3,25	0,5	77%
3	20	3,25	1	79%
4	20	3,25	1	80%
5	20	2,75	0,5	79%
6	20	2,75	0,5	77%
7	20	2,75	1	80%
8	20	2,75	1	78%
9	20	3,25	-	75%
10	20	3,25	-	75%

Todas las muestras en las que estaban presentes adaptadores de cola C mostraron un mayor rendimiento de ácidos nucleicos ligados a adaptadores (% de ligación) que las muestras en las que estaban ausentes las colas C. El mejor rendimiento fue para los cebadores con cola C a cola T en una proporción de 1:3,25 (aproximadamente el 30%), pero se obtuvieron rendimientos mejorados en proporciones de 0,5:3,25 (aproximadamente el 15%) a 1:2,75 (36%).

Después de secuenciar el ADN amplificado, se calculó la diversidad para cada preparación. La diversidad es el número de moléculas secuenciadas, calculada por: (tamaño medio de la molécula de ADN en pb) \* (Nº de moléculas únicas secuenciadas)/(tamaño de la región objetivo en pb). La diversidad fue generalmente mayor en las muestras en las que había adaptador de cola C. La secuenciación también indicó que la proporción de adaptadores con cola T a C incorporados fue de aproximadamente el 10%.

Si diferentes versiones de una secuencia están asociadas con un número de registro en diferentes momentos, se entiende la versión asociada con el número de registro en la fecha de presentación efectiva de esta solicitud. La fecha de presentación efectiva significa la primera entre la fecha de presentación real o la fecha de presentación de una solicitud con prioridad que se refiere al número de registro, si corresponde. De igual manera, si se publican diferentes versiones de una publicación, sitio web o similar en diferentes momentos, se entenderá la versión publicada más recientemente en la fecha de presentación efectiva de la solicitud, a menos que se indique lo contrario. Cualquier característica, paso, elemento, realización o aspecto de la invención puede usarse en combinación con cualquier otro a menos que se indique específicamente lo contrario. La presente invención se ha descrito con detalle a modo de ilustración y ejemplo con propósitos de claridad y comprensión.

#### LISTADO DE SECUENCIAS

<110> Guardant Health, Inc.  
Kennedy, Andrew  
Mortimer, Stefanie Ann Ward

<120> MÉTODOS PARA UNIR ADAPTADORES A ÁCIDOS NUCLEICOS DE MUESTRA

<130> 065777-512837

<150> US 62/485,769  
<151> 2017-04-14

<150> US 62/486,663  
<151> 2017-04-18

<150> US 62/517,145  
<151> 2017-06-08

<150> PCT/US2017/027809  
<151> 2017-04-14

<160> 3

<170> PatentIn versión 3.5

5      <210> 1  
       <211> 58  
       <212> DNA  
       <213> Secuencia artificial  
  
       <220>  
       <223> Sintetizado  
  
 10     <400> 1  
       aatgatacgg cgaccaccga gatctacact cttccctac acgacgtct tccgatct 58  
  
 15     <210> 2  
       <211> 63  
       <212> DNA  
       <213> Secuencia artificial  
  
       <220>  
       <223> Sintetizado  
 20  
       <220>  
       <221> característica\_misc  
       <222> (34)..(39)  
 25     <223> n es a, g, c, o t  
       <400> 2  
  
       gatcgggaaga gcacacgtct gaactccagt cacnnnnnna tctcgtatgc cgtcttctgc      60  
 30     ttg      63  
  
       <210> 3  
       <211> 58  
       <212> DNA  
 35     <213> Secuencia artificial  
  
       <220>  
       <223> Sintetizado  
  
 40     <400> 3  
       aatgatacgg cgaccaccga gatctacact cttccctac acgacgtct tccgatcc 58  
  
 45

## REIVINDICACIONES

1. Un método de preparación de ácidos nucleicos para análisis que comprende:

- 5 (a) hacer romos los extremos de ácidos nucleicos de cadena doble con salientes de cadena sencilla en una muestra por la acción de una o más enzimas que proporcionan una actividad de polimerasa 5'-3' y una actividad de lectura de prueba 3'-5', y cuatro tipos de nucleótidos estándar, en donde los salientes de cadena sencilla con extremos 5' sirven como plantillas para la extensión de una cadena complementaria por la actividad de polimerasa y los salientes de cadena sencilla con extremos 3' son digeridos por la actividad de lectura de prueba produciendo ácidos nucleicos de extremos romos;
- 10 (b) sin separar los ácidos nucleicos con extremos romos de otros componentes de la muestra, añadir colas a los extremos de los ácidos nucleicos de extremos romos mediante la acción de una polimerasa sin una función de lectura de prueba 3'-5', que realiza una adición no dirigida por plantilla de un nucleótido a los extremos 3' de ácidos nucleicos de extremos romos, en donde A se añade preferentemente a G preferentemente a C o T;
- 15 (c) aparear los ácidos nucleicos del paso (b) con adaptadores por lo menos parcialmente de cadena doble con un saliente T de un solo nucleótido en un extremo 3', y adaptadores por lo menos parcialmente de cadena doble con un saliente C de un solo nucleótido en un extremo 3'; y
- (d) ligar los ácidos nucleicos a los adaptadores.

20 2. El método de la reivindicación 1, que comprende además desnaturalizar una o más enzimas después del paso (a).

3. El método de la reivindicación 1 o 2, que comprende además poner en contacto la muestra con una o más enzimas, los cuatro tipos de nucleótidos estándar y la polimerasa sin una función de lectura de prueba 3'-5', opcionalmente en donde la muestra se pone en contacto con la una o más enzimas, los cuatro tipos de nucleótidos estándar y la polimerasa sin una función de lectura de prueba 3'-5' juntos.

25

4. El método de cualquier reivindicación anterior, en donde el paso (b) se realiza a una temperatura más alta que el paso (a), opcionalmente en donde el paso (a) se realiza a temperatura ambiente y el paso (b) a una temperatura superior a 60° C.

30

5. El método de cualquier reivindicación anterior, en donde:

- (i) la una o más enzimas son una polimerasa con actividad de polimerasa 5'-3' y actividad de lectura de prueba 3'-5', opcionalmente en donde la polimerasa con actividad de polimerasa 5'-3' y la actividad de lectura de prueba 3'-5' es polimerasa T4 o fragmento grande de Klenow; y/o
- 35 (ii) la polimerasa sin una función de lectura de prueba 3'-5' es una polimerasa termotable y el método comprende además aumentar la temperatura de la muestra después del paso (a) para inactivar la polimerasa con actividad de polimerasa 5'-3' y actividad de lectura de prueba 3'-5'.

6. El método de cualquier reivindicación anterior, que comprende además (e) amplificar los ácidos nucleicos ligados a los adaptadores; y (f) analizar los ácidos nucleicos, opcionalmente en donde:

40

- (i) los pasos (a)-(e) se realizan en un solo tubo;
- (ii) se analizan por lo menos el 70% de los ácidos nucleicos de cadena doble disponibles en la muestra;
- 45 (iii) el paso (f) comprende secuenciar los ácidos nucleicos ligados a los adaptadores, opcionalmente en donde la secuenciación secuenciar un nucleótido que formó un saliente en el paso (c) o (d);
- (iv) el paso de analizar detecta una variante somática o de línea germinal;
- (v) el paso de analizar detecta una variación del número de copias; o
- 50 (vi) el paso de analizar detecta una variación de un solo nucleótido (SNV).

7. El método de cualquier reivindicación anterior, que comprende además poner en contacto la muestra con adaptadores de extremos romos por lo menos parcialmente de cadena doble, que se ligan con ácidos nucleicos de cadena doble de extremos romos que no se han sometido a la adición dirigida sin plantilla de un nucleótido a los extremos 3' en el paso de ligación.

55

8. El método de cualquier reivindicación anterior, en donde:

- (i) la polimerasa sin una función de lectura de prueba 3'-5' es una polimerasa Taq;
- (ii) por lo menos los pasos (a)-(d) se realizan en un solo tubo;
- 60 (iii) para por lo menos los pasos (a)-(d) no se elimina ningún componente de la muestra;
- (iv) una relación molar de adaptadores por lo menos parcialmente de cadena doble con un solo nucleótido T a un solo nucleótido C es de 4:1 a 2:1, opcionalmente en donde la relación molar de adaptadores de extremos romos a adaptadores de cola es de 1:5 a 1:500; y/o
- 65 (v) por lo menos el 70% de los ácidos nucleicos de cadena doble de la muestra están unidos a adaptadores.



9. Un método para convertir ADN de cadena doble en ADN etiquetado con adaptador que comprende:

(a) poner en contacto una población de moléculas de ADN de cadena doble con una población de adaptadores por lo menos parcialmente de cadena doble, en donde:

(i) la población de moléculas de ADN de cadena doble comprende moléculas de ADN que comprenden un saliente A de un solo nucleótido y moléculas de ADN que comprenden un saliente G de un solo nucleótido, y en donde los salientes A de un solo nucleótido son más abundantes (por ejemplo, 10 veces, 100 veces, 1000 veces) que los salientes G de un solo nucleótido en la población, y  
(ii) la población de adaptadores por lo menos parcialmente de cadena doble comprende adaptadores que comprenden un saliente T de un solo nucleótido y adaptadores que comprenden un saliente C de un solo nucleótido; y

(b) ligar los adaptadores a las moléculas de ADN, en donde la ligación produce ADN etiquetado con adaptador.

10. El método de la reivindicación 9, en donde:

(i) la población de moléculas de ADN de cadena doble comprende además por lo menos uno de: moléculas de ADN que comprenden un saliente C de un solo nucleótido, moléculas de ADN que comprenden un saliente T de un solo nucleótido y un extremo romo, y  
(ii) la población de adaptadores por lo menos parcialmente de cadena doble comprende además por lo menos uno de: adaptadores que comprenden un saliente G de un solo nucleótido, adaptadores que comprenden un saliente A de un solo nucleótido y un extremo romo.

11. El método de la reivindicación 9 o 10, en donde los adaptadores por lo menos parcialmente de cadena doble comprenden un sitio de unión del cebador de NGS ("secuenciación de próxima generación") y un código de barras de ADN, que opcionalmente comprende además:

amplificar el ADN etiquetado con el adaptador usando cebadores de amplificación que comprenden un código de barras de índice de muestra y una secuencia de nucleótidos adaptada para hibridar con un oligonucleótido inmovilizado en un soporte de celda de flujo.

12. El método de una cualquiera de las reivindicaciones 9-11, en donde la población de los adaptadores por lo menos parcialmente de cadena doble comprende una pluralidad de códigos de barras de ADN diferentes, opcionalmente en donde el número de combinaciones de códigos de barras que pueden unirse a ambos extremos de una molécula de ADN de cadena doble es menor que el número de moléculas de ADN de cadena doble en la población, por ejemplo, entre 5 y 10.000 combinaciones diferentes.

13. El método de cualquiera de las reivindicaciones 9-12, en donde:

(i) los adaptadores son adaptadores con forma de Y; y/o  
(ii) la población de ácidos nucleicos es una población de ácidos nucleicos libres de células, preferiblemente ADN libre de células.

14. El método de cualquier reivindicación anterior, en donde la muestra es una muestra de fluido corporal, opcionalmente en donde la muestra es sangre completa, suero o plasma.

15. El método de cualquier reivindicación anterior, en donde la muestra es de un sujeto que se sospecha que tiene cáncer.

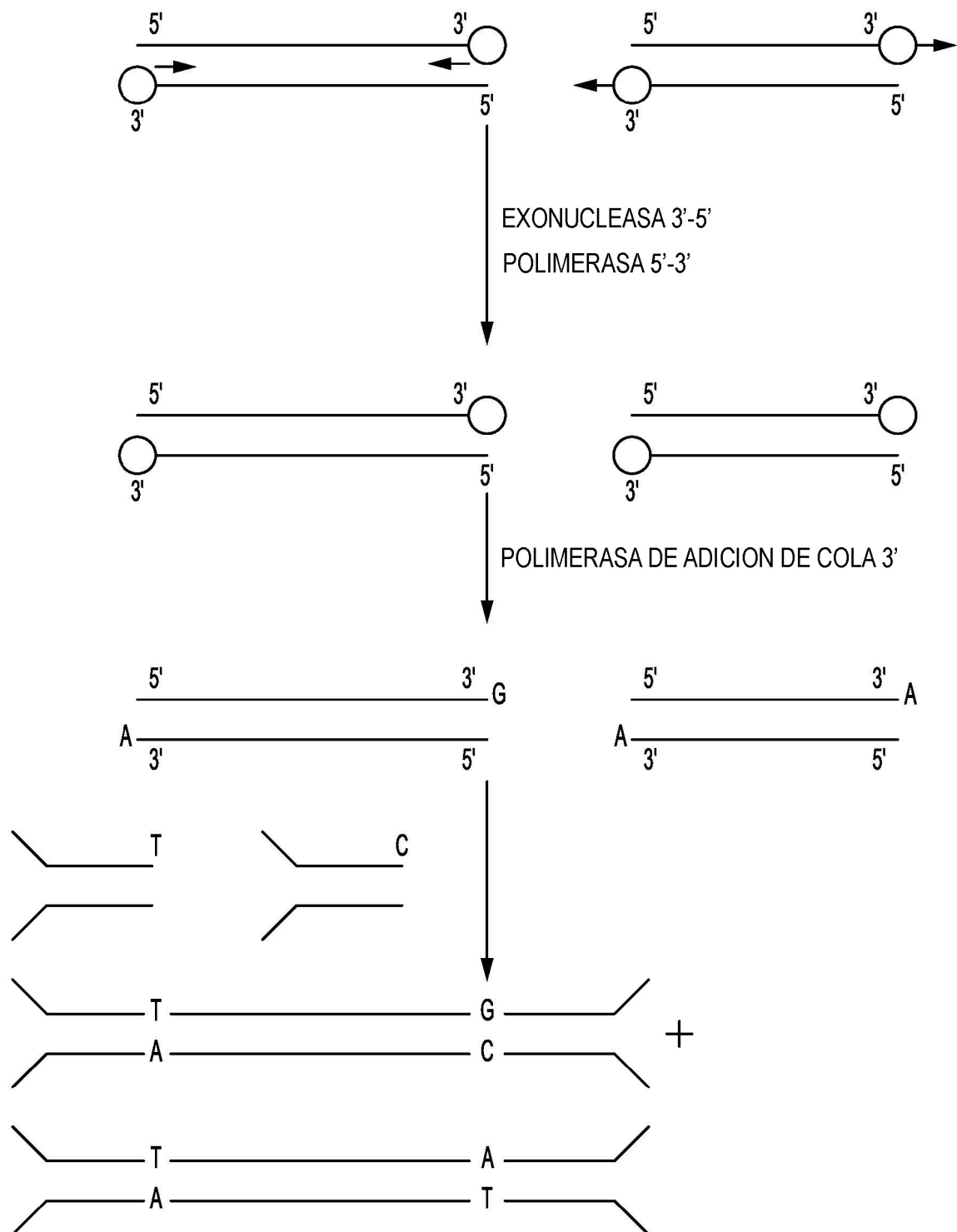


FIG. 1