(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 953 889**

(51) Int. CI.:

C12Q 1/6874 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **27.02.2020 PCT/IB2020/051702**

(87) Fecha y número de publicación internacional: **17.09.2020 WO20183280**

(96) Fecha de presentación y número de la solicitud europea: **27.02.2020 E 20710617 (0)**

(97) Fecha y número de publicación de la concesión europea: **07.06.2023 EP 3938541**

(54) Título: **Método para secuenciar una repetición directa**

(30) Prioridad:

14.03.2019 US 201962818527 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
16.11.2023

(73) Titular/es:

GENOME RESEARCH LIMITED (100.0%)
Wellcome Trust Genome Campus, Hinxton
Cambridge, Cambridgeshire CB10 1SA, GB

(72) Inventor/es:

OSBORNE, ROBERT

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

DESCRIPCIÓN

Método para secuenciar una repetición directa

ANTECEDENTES

Algunos métodos de secuenciación requieren comparar dos secuencias con una sola lectura de secuencia para determinar si existe una diferencia entre las secuencias. Sin embargo, tales métodos pueden ser difíciles de llevar a cabo debido a que el software que lleva a cabo esta tarea necesita identificar con precisión los comienzos y finales de las secuencias en una lectura de secuencia que deben compararse, extraer secuencias que deben compararse, y después llevar a cabo una alineación de esas secuencias. Estas etapas pueden ser un desafío para llevarlas a cabo automáticamente de forma consistente para todas las diferentes secuencias, composiciones de secuencias, y longitudes. Por ejemplo, la existencia de secuencias repetidas dentro de una lectura de secuencia puede provocar el deslizamiento de una alineación, lo que puede producir resultados erróneos.

La presente descripción proporciona una mejor manera alternativa de comparar secuencias con la misma lectura de secuencia.

SUMARIO

Se proporciona un método para secuenciar un molde que comprende una repetición directa, es decir, un molde que comprende una primera secuencia repetida y una segunda secuencia repetida que está en orientación directa con la primera repetición. En algunas realizaciones, el método puede comprender, en la misma reacción, hibridar un cebador con un primer sitio que está en dirección 5’ de la primera secuencia repetida, e hibridar un cebador con un segundo sitio que está en dirección 5’ de la segunda secuencia repetida. En estas realizaciones, el primer y el segundo sitio (es decir, los sitios a los que se unen el primer y el segundo cebador) deben estar en dirección 5’ de la primera y la segunda secuencia repetida, respectivamente (es decir, en dirección 3’ de los extremos 3’ de los cebadores), y equidistantes de la primera y la segunda secuencia repetida. El producto de hibridación producido por esta etapa contiene el molde con dos cebadores hibridados con él, ambos en dirección 5’ de una secuencia repetida mediante la misma distancia (por ejemplo, el mismo número de bases). A continuación, el método implica secuenciar el molde usando un método de secuenciación por síntesis (por ejemplo, usando terminadores de colorantes fluorescentes), para producir una lectura de secuencia que comprende una combinación de la primera y la segunda secuencia repetida, es decir, una lectura de secuencia que es esencialmente dos lecturas (una del primer cebador y la otra del segundo cebador) que se fusionan entre sí. Las diferencias entre la secuencia de la primera y la segunda repetición pueden identificarse como llamadas de bases de baja calidad.

En algunas realizaciones, dentro de cada molécula molde, la primera secuencia repetida y la segunda secuencia repetida se amplifican a partir de hebras opuestas de un fragmento bicatenario de ADN. En estas realizaciones, las secuencias de la primera y la segunda repetición deben ser idénticas excepto por las posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN o a errores que ocurren durante la amplificación. Por lo tanto, cualquier diferencia entre las hebras superior e inferior del fragmento bicatenario se puede identificar en la lectura de secuencia como una llamada de bases de “baja calidad”, es decir, una base que está asociada con datos subyacentes deficientes debido a que hay, en efecto, dos bases diferentes en una posición particular en la secuencia. Más detalladamente, dentro de cada molécula molde, la primera repetición puede amplificarse a partir de la una hebra de un fragmento bicatenario de ADN genómico, y la segunda repetición puede amplificarse a partir de la otra hebra del mismo fragmento de un fragmento bicatenario de ADN genómico. Dentro de una molécula, las secuencias de la primera y la segunda repetición suelen ser las mismas. Sin embargo, en los casos en que hay daño en la molécula original, las secuencias de la primera y la segunda repetición (dentro de una sola molécula) pueden diferir. Como tal, dentro de cada molécula de repetición, la primera y la segunda repetición son típicamente idénticas excepto por las posiciones que corresponden a (a) nucleótidos dañados en el fragmento bicatenario de ADN genómico del que se copiaron esas hebras, o (b) errores que ocurren durante amplificación de la molécula de repetición directa (por ejemplo, nucleótidos que se incorporaron incorrectamente o eliminaciones causadas por un evento de tartamudeo o deslizamiento durante la amplificación). Como tal, la primera y la segunda repetición son típicamente al menos en 95% idénticas en secuencia. Por lo tanto, las diferentes repeticiones en una molécula molde se pueden secuenciar usando dos cebadores (uno para cada repetición) al mismo tiempo para determinar si las repeticiones (que corresponden a la parte superior y al complemento de las hebras inferiores de un fragmento inicial de ADN genómico) difieren. Debido a que se usan dos cebadores, las secuencias de la primera y la segunda repetición se fusionan en la misma lectura de secuencia. Cualquier diferencia entre esas secuencias se puede observar como una llamada de bases de baja calidad debido a que los datos subyacentes para esa llamada de bases derivan esencialmente de dos bases (una base leída por el primer cebador y la otra base leída por el segundo cebador, en el que esas bases están a la misma distancia en dirección 3’ de los cebadores). Si hay una llamada de bases de baja calidad en una posición particular, entonces el método puede comprender excluir esa llamada de bases del análisis futuro. El método puede usarse para identificar nucleótidos dañados y errores de amplificación, así como errores de secuenciación (es decir, errores que surgen de la propia reacción de secuencia, no en el molde de secuenciación).

El método encuentra un uso particular en el análisis de muestras de ADN que contienen ADN dañado, muestras en las

que la cantidad de ADN es limitada, y/o muestras que contienen fragmentos que tienen una mutación de bajo número de copias (por ejemplo, una secuencia causada por una mutación que está presente en bajo número de copias con respecto a las secuencias que no contienen la mutación). Estas características a menudo están presentes en muestras de pacientes que se pueden obtener de forma no invasiva, por ejemplo muestras de tumores circulantes (ADNtc), que se pueden obtener de sangre periférica, o de forma invasiva, por ejemplo secciones de tejido. En algunas realizaciones, la muestra puede ser ADN obtenido de tejido embebido en parafina (es decir, una muestra FFPE). En tales muestras, las secuencias mutantes pueden estar presentes sólo en un número de copias muy limitado (por ejemplo, menos de 10, menos de 5 copias, o incluso 1 copia en un fondo de cientos o miles de copias de la secuencia de tipo salvaje). En estas situaciones, sin una forma efectiva de eliminar los errores generados por el daño del ADN, puede ser casi imposible identificar una verdadera variación de secuencia con una confianza significativa.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La invención se comprende mejor a partir de la siguiente descripción detallada cuando se lee junto con los dibujos adjuntos. Se enfatiza que, según la práctica común, las diversas características de los dibujos no están a escala. De hecho, las dimensiones de las diversas características se amplían o reducen arbitrariamente para mayor claridad. En los dibujos se incluyen las siguientes figuras.

La Fig. 1 ilustra esquemáticamente un molde de repetición directa que se ha obtenido a partir de un fragmento de ADN genómico bicatenario.

La Fig. 2 ilustra esquemáticamente dónde se hibridan el primer y el segundo cebador usados en el método con un molde de repetición directa.

La Fig. 3 ilustra esquemáticamente un ejemplo del método.

La Fig. 4 ilustra esquemáticamente un método ejemplar mediante el cual se puede producir una molécula de repetición directa.

La Fig. 5 ilustra esquemáticamente otro método ejemplar mediante el cual se puede producir una molécula de repetición directa.

DEFINICIONES

A menos que se defina lo contrario aquí, todos los términos técnicos y científicos usados aquí tienen el mismo significado que comúnmente entiende un experto en la técnica a la que pertenece esta invención. Aunque cualquier método y material similar o equivalente a los descritos aquí puede usarse en la práctica o ensayo de la presente invención, se describen los métodos y materiales preferidos.

Los intervalos numéricos incluyen los números que definen el intervalo. A menos que se indique lo contrario, los ácidos nucleicos se escriben de izquierda a derecha en orientación de 5' a 3'; las secuencias de aminoácidos se escriben de izquierda a derecha en orientación amino a carboxi, respectivamente.

Los encabezados proporcionados aquí no son limitaciones de los diversos aspectos o realizaciones de la invención. Por consiguiente, los términos definidos inmediatamente a continuación se definen más completamente por referencia a la memoria descriptiva como un todo.

A menos que se defina de otro modo, todos los términos técnicos y científicos usados aquí tienen el mismo significado que entienden comúnmente los expertos en la técnica a la que pertenece esta invención. Singleton, et al., DICTIONARY OF MICROBIOLOGY AND MOLECULAR BIOLOGY, 2D ED., John Wiley and Sons, Nueva York (1994), y Hale y Markham, THE HARPER COLLINS DICTIONARY OF BIOLOGY, Harper Perennial, N.Y. (1991) proporcionan a un experto el significado general de muchos de los términos usados aquí. Aun así, ciertos términos se definen a continuación en aras de la claridad y la facilidad de referencia.

Se observa además que las reivindicaciones pueden redactarse para excluir cualquier elemento opcional. Como tal, esta afirmación pretende servir como base antecedente para el uso de tal terminología exclusiva tal como "únicamente", "solamente", y similar, con respecto a la lectura de los elementos de la reivindicación, o el uso de una limitación "negativa".

El término "muestra", como se usa aquí, se refiere a un material o mezcla de materiales, que normalmente contiene uno o más analitos de interés. En una realización, el término, tal como se usa en su sentido más amplio, se refiere a cualquier material vegetal, animal, microbiano, o viral que contiene ADN genómico, tal como, por ejemplo, tejido o fluido aislado de un individuo (incluyendo, sin limitación, plasma, suero, líquido cefalorraquídeo), linfa, lágrimas, saliva, y secciones de tejido), o de constituyentes de cultivos celulares in vitro, así como muestras del medioambiente.

La expresión "muestra de ácido nucleico", como se usa aquí, denota una muestra que contiene ácidos nucleicos. Las muestras de ácido nucleico usadas aquí pueden ser complejas por cuanto contienen múltiples moléculas diferentes que

contienen secuencias. Las muestras de ADN genómico de un mamífero (por ejemplo, un ratón o un ser humano) son tipos de muestras complejas. Las muestras complejas pueden tener más de alrededor de 10^4, 10^5, 10^6 o 10^7, 10^8, 10^9 o 10^{10} moléculas de ácido nucleico diferentes. Una diana de ADN puede tener su origen en cualquier fuente, tal como el ADN genómico, o un constructo de ADN artificial. Aquí puede emplearse cualquier muestra que contenga ácidos nucleicos, por ejemplo ADN genómico de células de cultivo de tejido o una muestra de tejido.

El término "mezcla", como se usa aquí, se refiere a una combinación de elementos, que están intercalados y no en ningún orden particular. Una mezcla es heterogénea y no separable espacialmente en sus diferentes constituyentes. Los ejemplos de mezclas de elementos incluyen varios elementos diferentes que se disuelven en la misma disolución acuosa y varios elementos diferentes unidos a un soporte sólido en posiciones aleatorias (es decir, sin ningún orden particular). Una mezcla no es direccionable. Para ilustrar con un ejemplo, una matriz de polinucleótidos unidos a la superficie separados espacialmente, como se conoce comúnmente en la técnica, no es una mezcla de polinucleótidos unidos a la superficie ya que las especies de polinucleótidos unidos a la superficie son espacialmente distintas, y la matriz es direccionable.

El término "nucleótido" pretende incluir aquellos restos que pueden copiarse usando una polimerasa. Los nucleótidos contienen no solo las bases conocidas de purina y pirimidina, sino también otras bases heterocíclicas que se han modificado, por ejemplo bases "dañadas" que se han oxidado o desadenilado por ejemplo. Dichas modificaciones incluyen purinas o pirimidinas metiladas, purinas o pirimidinas aciladas, ribosas alquiladas, u otros heterociclos. Además, el término "nucleótido" incluye aquellos restos que contienen hapteno o etiquetas fluorescentes, y pueden contener no solo azúcares de ribosa y de desoxirribosa convencionales, sino también otros azúcares. Los nucleósidos o nucleótidos modificados también incluyen modificaciones en el resto del azúcar, por ejemplo en el que uno o más de los grupos hidroxilo se reemplazan por átomos de halógeno o grupos alifáticos, o se funcionalizan como éteres, aminas, o similares.

Las expresiones "ácido nucleico" y "polinucleótido" se usan indistintamente aquí para describir un polímero de cualquier longitud, por ejemplo mayor que alrededor de 2 bases, mayor que alrededor de 10 bases, mayor que alrededor de 100 bases, mayor que alrededor de 500 bases, mayor que 1000 bases, mayor que 10.000 bases, mayor que 100.000 bases, mayor que alrededor de 1.000.000, hasta alrededor de 10^{10} o más bases compuestas de nucleótidos, por ejemplo desoxirribonucleótidos o ribonucleótidos, y puede producirse enzimática o sintéticamente (por ejemplo, PNA como se describe en la patente de EE. UU. n.º 5.948.902, y las referencias citadas allí) que se puede hibridar con ácidos nucleicos de origen natural en una forma específica de secuencia análoga a la de dos ácidos nucleicos de origen natural, por ejemplo puede participar en interacciones de emparejamiento de bases de Watson-Crick. Los nucleótidos de origen natural incluyen guanina, citosina, adenina, timina, uracilo (G, C, A, T y U respectivamente). El ADN y el ARN tienen una cadena principal de azúcar de desoxirribosa y ribosa, respectivamente, mientras que la cadena principal del PNA está compuesta por unidades repetidas de N-(2-aminoetil)-glicina unidas por enlaces peptídicos. En PNA, diversas bases de purina y pirimidina están unidas a la cadena principal por enlaces metilenocarbonilo. Un ácido nucleico bloqueado (LNA), a menudo denominado ARN inaccesible, es un nucleótido de ARN modificado. El resto de ribosa de un nucleótido de LNA se modifica con un puente adicional que conecta el oxígeno 2' y el carbono 4'. El puente "bloquea" la ribosa en la conformación 3'-endo (Norte), que a menudo se encuentra en los dúplex en forma de A. Los nucleótidos de LNA se pueden mezclar con restos de ADN o ARN en el oligonucleótido siempre que se desee. La expresión "ácido nucleico no estructurado", o "UNA", es un ácido nucleico que contiene nucleótidos no naturales que se unen entre sí con estabilidad reducida. Por ejemplo, un ácido nucleico no estructurado puede contener un resto G' y un resto C', en el que estos restos corresponden a formas de origen no natural, es decir, análogos, de G y C que emparejan las bases entre sí con estabilidad reducida, pero retiene una capacidad de emparejar las bases con restos C y G de origen natural, respectivamente. El ácido nucleico no estructurado se describe en el documento US20050233340.

El término "oligonucleótido", como se usa aquí, denota un multímero monocatenario de nucleótido de alrededor de 2 a 200 nucleótidos, hasta 500 nucleótidos de longitud. Los oligonucleótidos pueden ser sintéticos o pueden obtenerse enzimáticamente, y en algunas realizaciones, tienen una longitud de 30 a 150 nucleótidos. Los oligonucleótidos pueden contener monómeros de ribonucleótidos (es decir, pueden ser oligorribonucleótidos) o monómeros de desoxirribonucleótidos, o tanto monómeros de ribonucleótidos como monómeros de desoxirribonucleótidos. Un oligonucleótido puede tener, por ejemplo, 10 a 20, 21 a 30, 31 a 40, 41 a 50, 51 a 60, 61 a 70, 71 a 80, 80 a 100, 100 a 150, o 150 a 200 nucleótidos de longitud.

"Cebador" significa un oligonucleótido, ya sea natural o sintético, que es capaz, al formar un dúplex con un molde polinucleotídico, de actuar como un punto de inicio de la síntesis de ácido nucleico y de extenderse desde su extremo 3' a lo largo del molde para que se forme un dúplex extendido. La secuencia de nucleótidos añadidos durante el proceso de extensión está determinada por la secuencia del polinucleótido molde. Por lo general, los cebadores se extienden mediante una ADN polimerasa. Los cebadores son generalmente de una longitud compatible con su uso en la síntesis de productos de extensión de cebadores, y normalmente están en el intervalo de 8 a 100 nucleótidos de longitud, tal como 10 a 75, 15 a 60, 15 a 40, 18 a 30, 20 a 40, 21 a 50, 22 a 45, 25 a 40, etc. Los cebadores típicos pueden estar en el intervalo de entre 10-50 nucleótidos de longitud, tal como 15-45, 18-40, 20-30, 21-25, etc., y cualquier longitud entre los intervalos señalados. En algunas realizaciones, los cebadores no suelen tener más de alrededor de 10, 12, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 55, 60, 65 o 70 nucleótidos de longitud. En algunas realizaciones, un cebador se puede activar antes de la extensión del cebador. Por ejemplo, algunos cebadores tienen un bloque 3' y una base de ARN interna. La base de ARN se puede eliminar mediante ARNasaH u otro tratamiento, produciendo así un grupo hidroxilo 3' que se puede extender. Existen otros métodos para activar cebadores.

Los cebadores suelen ser monocatenarios para lograr la máxima eficiencia en la amplificación, pero alternativamente pueden ser bicatenarios o parcialmente bicatenarios. Si es bicatenario, el cebador se trata generalmente primero para separar sus hebras antes de usarlo para preparar productos de extensión. Esta etapa de desnaturalización se ve afectada normalmente por el calor, pero alternativamente se puede llevar a cabo usando álcali, seguido de neutralización. También se incluyen en esta definición los cebadores de intercambio de punto de apoyo, como se describe en Zhang et al (Nature Chemistry 2012 4: 208-214).

Por lo tanto, un "cebador" es complementario a un molde, y forma complejos mediante enlaces de hidrógeno o hibridación con el molde para dar un complejo cebador/molde para el inicio de la síntesis por una polimerasa, que se extiende mediante la adición de bases enlazadas covalentemente unidas en su extremo 3' complementario al molde en el proceso de síntesis del ADN.

El término "hibridación" o "hibrida" se refiere a un proceso en el que una región de una hebra de ácido nucleico se hibrida a y forma un dúplex estable, ya sea un homodúplex o un heterodúplex, en condiciones normales de hibridación con una segunda hebra de ácido nucleico complementaria, y no forma un dúplex estable con moléculas de ácido nucleico no relacionadas en las mismas condiciones normales de hibridación. La formación de un dúplex se logra hibridando dos regiones de hebras de ácido nucleico complementarias en una reacción de hibridación. La reacción de hibridación se puede realizar para que sea altamente específica mediante el ajuste de las condiciones de hibridación (a menudo denominadas rigurosidad de hibridación) bajo las cuales tiene lugar la reacción de hibridación, de modo que dos hebras de ácido nucleico no formaran un dúplex estable, por ejemplo un dúplex que retiene una región de bicatenaridad en condiciones de rigurosidad normales, a menos que las dos hebras de ácido nucleico contengan un cierto número de nucleótidos en secuencias específicas que son sustancial o completamente complementarias. Las "condiciones de hibridación normal o de rigurosidad normal" se determinan fácilmente para cualquier reacción de hibridación dada. Véase, por ejemplo, Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, Inc., Nueva York, o Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press. Como se usa aquí, el término "hibridar" o "hibridación" se refiere a cualquier proceso mediante el cual una hebra de ácido nucleico se une a una hebra complementaria a través del emparejamiento de bases.

Se considera que un ácido nucleico es "hibridable selectivamente" con una secuencia de ácido nucleico de referencia si las dos secuencias se hibridan específicamente entre sí en condiciones de lavado e hibridación de rigurosidad moderada a alta. Las condiciones de hibridación de rigurosidad moderada y alta son conocidas (véase, por ejemplo, Ausubel, et al., Short Protocols in Molecular Biology, 3ª ed., Wiley & Sons 1995 y Sambrook et al., Molecular Cloning: A Laboratory Manual, tercera edición, 2001 Cold Spring Harbor, N.Y.). Un ejemplo de condiciones de rigurosidad alta incluye la hibridación a alrededor de 42ºC en formamida al 50%, SSC 5X, disolución de Denhardt 5X, SDS al 0,5%, y 100 µg/ml de ADN portador desnaturalizado, seguido de lavado dos veces en SSC 2X y SDS al 0,5% a temperatura ambiente, y dos tiempos adicionales en SSC 0,1 X y SDS al 0,5% a 42ºC.

El término "amplificar", como se usa aquí, se refiere al proceso de sintetizar moléculas de ácido nucleico que son complementarias a una o ambas hebras de un ácido nucleico molde. Amplificar una molécula de ácido nucleico puede incluir desnaturalizar el ácido nucleico molde, hibridar cebadores al ácido nucleico molde a una temperatura que está por debajo de las temperaturas de fusión de los cebadores, y alargar enzimáticamente a partir de los cebadores para generar un producto de amplificación. Cada una de las etapas de desnaturalización, hibridación y elongación se puede llevar a cabo una o más veces. En ciertos casos, las etapas de desnaturalización, hibridación y elongación se llevan a cabo múltiples veces, de modo que la cantidad de producto de amplificación aumenta, a menudo de manera exponencial, aunque los presentes métodos no requieren amplificación exponencial. La amplificación normalmente requiere la presencia de trifosfatos de desoxirribonucleósido, una enzima ADN polimerasa, y un amortiguador y/o cofactores apropiados para una actividad óptima de la enzima polimerasa. La expresión "producto de amplificación" se refiere a los ácidos nucleicos, que se producen a partir del proceso de amplificación como se define aquí.

Los términos "determinar", "medir", "evaluar", "confirmar", "ensayar" y "analizar" se usan indistintamente aquí para referirse a cualquier forma de medida, e incluyen determinar si un elemento está presente o no. Estos términos incluyen determinaciones tanto cuantitativas como cualitativas. La evaluación puede ser relativa o absoluta. "Evaluar la presencia de" incluye determinar la cantidad de algo presente, así como determinar si está presente o ausente.

El término "ligar", como se usa aquí, se refiere a la unión catalizada enzimáticamente del nucleótido terminal en el extremo 5' de una primera molécula de ADN al nucleótido terminal en el extremo 3' de una segunda molécula de ADN.

Una "pluralidad" contiene al menos 2 miembros. En ciertos casos, una pluralidad puede tener al menos 2, al menos 5, al menos 10, al menos 100, al menos 100, al menos 10.000, al menos 100.000, al menos 10^6, al menos 10^7, al menos 10^8, o al menos 10^9 o más miembros.

Un "sitio de unión de oligonucleótidos" se refiere a un sitio con el que se hibrida un oligonucleótido en un polinucleótido diana. Si un oligonucleótido "proporciona" un sitio de unión para un cebador, entonces el cebador puede hibridarse con ese oligonucleótido o su complemento.

El término “hebra”, como se usa aquí, se refiere a un ácido nucleico formado por nucleótidos unidos covalentemente mediante enlaces covalentes, por ejemplo enlaces fosfodiéster. En una célula, el ADN generalmente existe en forma bicatenaria, y como tal, tiene dos hebras complementarias de ácido nucleico denominadas aquí como las hebras “Watson” (o “superior”) y “Crick” (o “inferior”). En ciertos casos, las hebras complementarias de una región cromosómica pueden denominarse hebras “más” y “menos”, las hebras “primera” y “segunda”, las hebras “codificantes” y “no codificantes”, las hebras “superior” e “inferior”, o las hebras “sentido” y “antisentido”. La asignación de una hebra como una hebra de Watson o Crick es arbitraria, y no implica ninguna orientación, función o estructura particular.

El término “extender”, como se usa aquí, se refiere a la extensión de un cebador mediante la adición de nucleótidos usando una polimerasa. Si se extiende un cebador que se hibrida con un ácido nucleico, el ácido nucleico actúa como molde para la reacción de extensión.

El término “secuenciar”, como se usa aquí, se refiere a un método por el cual se obtiene la identidad de al menos 10 nucleótidos consecutivos (por ejemplo, la identidad de al menos 20, al menos 50, al menos 100, o al menos 200 o más nucleótidos consecutivos) de un polinucleótido.

Las expresiones “secuenciación de próxima generación” o “secuenciación de alto rendimiento”, como se usan aquí, se refieren a las denominadas plataformas de secuenciación por síntesis o secuenciación por ligación en paralelo empleadas actualmente por Illumina, Life Technologies, y Roche, etc. Los métodos de secuenciación de próxima generación también pueden incluir métodos de secuenciación de nanoporos tal como el comercializado por Oxford Nanopore Technologies, métodos basados en detección electrónica tal como la tecnología Ion Torrent comercializada por Life Technologies, o métodos basados en fluorescencia de una sola molécula tal como el comercializado por Pacific Biosciences.

La expresión “secuencia de código de barras” o “código de barras molecular”, como se usa aquí, se refiere a una secuencia única de nucleótidos que se puede usar para a) identificar y/o rastrear la fuente de un polinucleótido en una reacción, b) contar cuántas veces se secuencia una molécula inicial, y c) emparejar lecturas de secuencias de diferentes hebras de la misma molécula. Las secuencias de códigos de barras pueden variar mucho en tamaño y composición; las siguientes referencias proporcionan una guía para seleccionar conjuntos de secuencias de códigos de barras apropiados para realizaciones particulares: Casbon (Nuc. Acids Res. 2011, 22 e81), Brenner, patente de EE.UU. n.º 5.635.400; Brenner et al., Proc. Natl. Acad. Sci., 97: 1665-1670 (2000); Shoemaker et al., Nature Genetics, 14: 450-456 (1996); Morris et al., publicación de patente europea 0799897A1; Wallace, patente de EE.UU. n.º 5.981.179; y similares. En realizaciones particulares, una secuencia de código de barras puede tener una longitud en el intervalo de 2 a 36 nucleótidos, o de 6 a 30 nucleótidos, o de 8 a 20 nucleótidos.

En algunos casos, un código de barras puede contener una “región de base degenerada” o “DBR”, en el que las expresiones “región de base degenerada” y “DBR” se refieren a un tipo de código de barras molecular que tiene una complejidad suficiente para ayudar a distinguir entre fragmentos a los que se ha añadido la DBR. En algunos casos, sustancialmente cada fragmento etiquetado puede tener una secuencia de DBR diferente. En estas realizaciones, puede usarse una DBR de alta complejidad (por ejemplo, una que esté compuesta por al menos 10.000 o 100.000, o más secuencias). En otras realizaciones, algunos fragmentos pueden etiquetarse con la misma secuencia de DBR, pero esos fragmentos aún pueden distinguirse por la combinación de i. la secuencia de DBR, ii. la secuencia del fragmento, iii. la secuencia de los extremos del fragmento, y/o iv. el sitio de inserción de la DBR en el fragmento. En algunas realizaciones, al menos 95%, por ejemplo al menos 96%, al menos 97%, al menos 98%, al menos 99%, o al menos 99,5% de los polinucleótidos diana se asocian con una secuencia de DBR diferente. En algunas realizaciones, una DBR puede comprender uno o más (por ejemplo, al menos 2, al menos 3, al menos 4, al menos 5, o 5 a 30 o más) nucleótidos seleccionados de R, Y, S, W, K, M, B, D, H, V, N (según lo define el código IUPAC). En algunos casos, un código de barras bicatenario se puede obtener construyendo un oligonucleótido que contiene una secuencia degenerada (por ejemplo, un oligonucleótido que tiene una serie de 2-10 o más “N”), y copiando entonces el complemento del código de barras en la otra hebra, como se describe más abajo.

Los oligonucleótidos que contienen una secuencia variable, por ejemplo una DBR, se pueden preparar obteniendo una serie de oligonucleótidos por separado, mezclando juntos los oligonucleótidos, y amplificándolos en masa. En otras palabras, la población de oligonucleótidos que contiene una secuencia variable se puede obtener como un solo oligonucleótido que contiene posiciones degeneradas (es decir, posiciones que contienen más de un tipo de nucleótido). Alternativamente, dicha población de oligonucleótidos se puede obtener fabricándolos individualmente o usando una matriz de los oligonucleótidos usando métodos de síntesis in situ, escindiendo los oligonucleótidos del sustrato, y opcionalmente amplificándolos. Los ejemplos de tales métodos se describen, por ejemplo, en Cleary et al. (Nature Methods 2004 1: 241-248) y LeProust et al. (Nucleic Acids Research 2010 38: 2522-2540).

En algunos casos, un código de barras puede corregir errores. Las descripciones de secuencias de identificación de errores (o corrección de errores) ejemplares se pueden encontrar en la bibliografía (por ejemplo, se describen en las publicaciones de solicitud de patente de EE. UU. US2010/0323348 y US2009/0105959). Los códigos con corrección de errores pueden ser necesarios para cuantificar números absolutos de moléculas. Muchos informes en la bibliografía usan códigos que se desarrollaron originalmente para la corrección de errores de sistemas binarios (códigos de Hamming, códigos de Reed Solomon, etc.), o los aplican a sistemas cuaternarios (por ejemplo, códigos de Hamming cuaternarios; véase Generalized DNA barcode design based on Hamming codes, Bystrykh 2012 PLoS One. 2012 7: e36852).

En algunas realizaciones, un código de barras se puede usar adicionalmente para determinar el número de moléculas polinucleotídicas diana iniciales que se han analizado, es decir, para “contar” el número de moléculas polinucleotídicas diana iniciales que se han analizado. La amplificación mediante PCR de moléculas que se han etiquetado con un código de barras puede dar como resultado múltiples subpoblaciones de productos que están clonalmente relacionados en el sentido de que cada una de las diferentes subpoblaciones se amplifica a partir de una única molécula etiquetada. Como sería evidente, aunque puede haber varios miles o millones o más de moléculas en cualquiera de las subpoblaciones de productos de PCR clonalmente relacionadas, y el número de moléculas diana en esas subpoblaciones clonalmente relacionadas puede variar enormemente, el número de moléculas etiquetadas en la primera etapa del método se puede estimar contando el número de secuencias de DBR asociadas con una secuencia diana que está representada en la población de productos de PCR. Este número es útil debido a que, en ciertas realizaciones, la población de productos de PCR obtenida usando este método puede secuenciarse para producir una pluralidad de secuencias. Se puede contar el número de secuencias de código de barras diferentes que están asociadas con las secuencias de un polinucleótido diana, y este número se puede usar (junto con, por ejemplo, la secuencia del fragmento, la secuencia de los extremos del fragmento, y/o el sitio de inserción de la DBR en el fragmento) para estimar el número de moléculas de ácido nucleico molde iniciales que se han secuenciado. Tales etiquetas también pueden ser útiles para corregir errores de secuenciación.

Las expresiones “secuencia identificadora de muestra” o “índice de muestra” se refieren a un tipo de código de barras que se puede añadir a un polinucleótido diana, en el que la secuencia identifica la fuente del polinucleótido diana (es decir, la muestra de la que deriva el polinucleótido diana). En uso, cada muestra se etiqueta con una secuencia identificadora de muestra diferente (por ejemplo, se añade una secuencia a cada muestra, en el que las diferentes muestras se añaden a diferentes secuencias), y las muestras etiquetadas se reúnen. Después de secuenciar la muestra reunida, la secuencia identificadora de muestra se puede usar para identificar la fuente de las secuencias.

El término “adaptador” se refiere a un ácido nucleico que se puede unir a al menos una hebra de una molécula de ADN bicatenario. El término “adaptador” se refiere a moléculas que son al menos parcialmente bicatenarias. Un adaptador puede tener una longitud de 20 a 150 bases, por ejemplo 40 a 120 bases, aunque se contemplan adaptadores fuera de este intervalo.

La expresión “etiquetado con adaptador”, como se usa aquí, se refiere a un ácido nucleico que se ha etiquetado mediante, es decir, unido covalentemente con, un adaptador. Un adaptador se puede unir a un extremo 5’ y/o un extremo 3’ de una molécula de ácido nucleico.

La expresión “ADN etiquetado”, como se usa aquí, se refiere a moléculas de ADN que tienen una secuencia adaptadora añadida, es decir, una “etiqueta” de origen sintético. Una secuencia adaptadora se puede añadir (es decir, “anexar”) mediante ligación.

El término “complejidad” se refiere al número total de secuencias diferentes en una población. Por ejemplo, si una población tiene 4 secuencias diferentes, entonces esa población tiene una complejidad de 4. Una población puede tener una complejidad de al menos 4, al menos 8, al menos 16, al menos 100, al menos 1.000, al menos 10.000, o al menos 100.000 o más, dependiendo del resultado deseado.

La expresión “de la fórmula” significa que las moléculas individuales en una población están descritas por, es decir, abarcadas por, la fórmula.

Ciertos polinucleótidos descritos aquí pueden representarse mediante una fórmula. A menos que se indique lo contrario, los polinucleótidos definidos por una fórmula están orientados en la dirección 5’ a 3’. Los componentes de la fórmula se refieren a secuencias separadamente definibles de nucleótidos dentro de un polinucleótido, en la que, a menos que esté implícito a partir del contexto, las secuencias están unidas entre sí covalentemente de modo que un polinucleótido descrito por una fórmula es una sola molécula. En algunos casos, los componentes de la fórmula se encuentran inmediatamente adyacentes entre sí en una sola molécula. A menos que se indique lo contrario o esté implícito a partir del contexto, una región definida por una fórmula puede tener secuencias adicionales, un sitio de unión del cebador, un código de barras molecular, un promotor, o un espaciador, etc., en su extremo 3’, su extremo 5’, o tanto el extremo 3’ como 5’. Como sería evidente, las diversas secuencias componentes de un polinucleótido pueden tener independientemente cualquier longitud deseada siempre que sean capaces de llevar a cabo la función deseada (por ejemplo, hibridación con otra secuencia). Por ejemplo, las diversas secuencias componentes de un polinucleótido pueden tener independientemente una longitud en el intervalo de 8-80 nucleótidos, por ejemplo 10-50 nucleótidos o 12-30 nucleótidos.

La expresión “hebras opuestas”, como se usa aquí, se refiere a las hebras superior e inferior, en la que las hebras son complementarias entre sí, excepto por los nucleótidos dañados.

La expresión “variación potencial de secuencia”, como se usa aquí, se refiere a una variación de secuencia, por ejemplo una sustitución, eliminación, inserción, o reordenamiento de uno o más nucleótidos en una secuencia con respecto a otra.

La expresión “error de amplificación” se refiere a una base mal incorporada, o una eliminación/inserción causada por el tartamudeo de la polimerasa. El tartamudeo ocurre generalmente en secuencias repetidas, por ejemplo repeticiones cortas

en tándem (STR) o repeticiones de microsatélites, y se supone que se debe a errores de copia o deslizamiento por la polimerasa.

La expresión “enriquecimiento diana”, como se usa aquí, se refiere a un método en el que las secuencias seleccionadas se separan de otras secuencias en una muestra. Esto puede realizarse por hibridación con una sonda, por ejemplo hibridando un oligonucleótido biotinilado con la muestra para producir dúplex entre el oligonucleótido y la secuencia diana, inmovilizando los dúplex a través del grupo biotina, lavando los dúplex inmovilizados, y después liberando las secuencias diana de los oligonucleótidos. Alternativamente, una secuencia seleccionada puede enriquecerse amplificando esa secuencia, por ejemplo mediante PCR usando uno o más cebadores que se hibridan con un sitio que está próximo a la secuencia diana.

Las expresiones “variante minoritaria” y “variación de secuencia”, como se usan aquí, son una variante que está presente con una frecuencia menor que 50%, con respecto a otras moléculas en la muestra. En algunos casos, una variante minoritaria puede ser un primer alelo de una secuencia diana polimórfica, en la que, en una muestra, la relación de moléculas que contienen el primer alelo de la secuencia diana polimórfica en comparación con las moléculas que contienen otros alelos de la secuencia diana polimórfica es 1:5 o menos, 1:10 o menos, 1:100 o menos, 1:1.000 o menos, 1:10.000 o menos, 1:100.000 o menos, o 1:1.000.000 o menos.

La expresión “secuenciación dúplex” se refiere a un método en el que se obtienen secuencias para ambas hebras de una molécula bicatenaria de ADN genómico. En la secuenciación dúplex, las secuencias derivadas de la hebra superior de la molécula bicatenaria de ADN genómico son distinguibles de las secuencias derivadas de la hebra inferior de esa molécula, de tal manera que se pueden comparar las secuencias para las hebras superior e inferior de la misma molécula bicatenaria de ADN genómico.

La expresión “repetición directa” se refiere a una molécula que contiene dos copias de secuencias casi idénticas, es decir, secuencias que tienen la misma longitud y que son al menos 95% idénticas en la secuencia nucleotídica. El término “distancia”, como se usa aquí, depende del método de secuenciación por síntesis que se utilice para la secuenciación. Por ejemplo, en los métodos que se basan en terminadores de cadena reversibles, la distancia entre el extremo 3’ de un cebador y un nucleótido en dirección 3’ puede definirse por el número de bases. En los métodos de semiconductores o pirosecuenciación, la distancia entre el extremo 3’ de un cebador y un nucleótido en dirección 3’ se puede definir por el número de flujos debido que, en esos métodos, se pueden añadir varios nucleótidos en un solo flujo. Por lo tanto, “equidistante” puede significar el mismo número de nucleótidos si se usa un método de secuenciación basado en un terminador de cadena reversible, o el mismo número de flujos si se usa un método de secuenciación basado en semiconductores o pirosecuenciación.

Para facilitar la referencia, el complemento inverso de una secuencia se puede indicar mediante el símbolo de apóstrofo (“ ’ ”). Por ejemplo, el complemento inverso de una secuencia denominada “W” puede denominarse “ W’ ”.

Otras definiciones de términos pueden aparecer a lo largo de la memoria descriptiva.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

Antes de que se describa la presente invención, debe entenderse que esta invención no se limita a las realizaciones particulares descritas, ya que tales pueden, por supuesto, variar. También debe entenderse que la terminología utilizada aquí tiene el propósito de describir realizaciones particulares solamente, y no pretende ser limitativa, ya que el alcance de la presente invención estará limitado únicamente por las reivindicaciones adjuntas.

Cuando se proporciona un intervalo de valores, se entiende que cada valor intermedio, a la décima parte de la unidad del límite inferior, a menos que el contexto dicte claramente lo contrario, entre los límites superior e inferior de ese intervalo también se describe específicamente. Cada intervalo más pequeño entre cualquier valor establecido o valor intermedio en un intervalo establecido y cualquier otro valor establecido o intermedio en ese intervalo establecido está abarcado dentro de la invención. Los límites superior e inferior de estos intervalos más pequeños pueden incluirse o excluirse independientemente en el intervalo, y cada intervalo, en el que cualquiera de los límites, ninguno de ellos, o ambos están incluidos en los intervalos más pequeños, también está abarcado dentro de la invención, sujeto a cualquier límite específicamente excluido en el intervalo establecido. Cuando el intervalo establecido incluye uno o ambos de los límites, los intervalos que excluyen cualquiera o ambos de los límites incluidos también están incluidos en la invención.

A menos que se defina de otro modo, todos los términos técnicos y científicos usados aquí tienen el mismo significado que comúnmente entiende alguien de pericia normal en la técnica a la que pertenece esta invención. Aunque cualesquiera métodos y materiales similares o equivalentes a los descritos aquí se pueden usar en la práctica o ensayo de la presente invención, ahora se describen algunos métodos y materiales potenciales y preferidos.

Debe señalarse que como se usa aquí y en las reivindicaciones adjuntas, las formas singulares “un”, “una” y “el/la” incluyen referentes en plural a menos que el contexto dicte claramente lo contrario. Así, por ejemplo, la referencia a “un ácido nucleico” incluye una pluralidad de tales ácidos nucleicos, y la referencia a “el compuesto” incluye la referencia a uno o más compuestos y equivalentes de los mismos conocidos por los expertos en la técnica, y así sucesivamente.

La práctica de la presente invención puede emplear, a menos que se indique lo contrario, técnicas convencionales y descripciones de química orgánica, tecnología de polímeros, biología molecular (incluyendo técnicas recombinantes), biología celular, bioquímica, e inmunología, que están dentro de la pericia de la técnica. Tales técnicas convencionales incluyen síntesis de matrices poliméricas, hibridación, ligación, y detección de hibridación usando marcador. Se pueden obtener ilustraciones específicas de técnicas adecuadas con referencia al ejemplo aquí más abajo. Sin embargo, también pueden usarse, por supuesto, otros procedimientos convencionales equivalentes. Tales técnicas y descripciones convencionales se pueden encontrar en manuales de laboratorio estándar tales como Genome Analysis: A Laboratory Manual Series (Vols. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular cloning: A Laboratory Manual (todos de Cold Spring Harbor Laboratory Press), Stryer, L. (1995) Biochemistry (4ª ed.) Freeman, Nueva York, Gait, “Oligonucleotide Synthesis: A Practical Approach” 1984, IRL Press, Londres, Nelson y Cox (2000), Lehninger, A., Principles of Biochemistry 3ª ed., W. H. Freeman Pub., Nueva York, N.Y. y Berg et al. (2002) Biochemistry, 5ª ed., W. H. Freeman Pub., Nueva York, N.Y.

Las publicaciones referidas aquí se proporcionan únicamente para su descripción antes de la fecha de presentación de la presente solicitud. Nada de lo aquí contenido debe interpretarse como una admisión de que la presente invención no tiene derecho a ser anterior a dicha publicación en virtud de una invención anterior. Además, las fechas de publicación proporcionadas pueden diferir de las fechas de publicación reales, por lo que es posible que sea necesario confirmarlas de forma independiente.

Se proporciona aquí, entre otras, una manera de secuenciar un molde que tiene una repetición directa, es decir, un molde que comprende una primera secuencia repetida y una segunda secuencia repetida, en el que la primera y la segunda secuencia repetida están en una repetición directa y son idénticas o casi idénticas. En algunas realizaciones dentro de cada molécula molde, la primera secuencia repetida y la segunda secuencia repetida pueden amplificarse a partir de hebras opuestas de un fragmento bicatenario de ADN. En realizaciones en las que el fragmento de ADN es ADN genómico bicatenario (por ejemplo, ADN genómico eucariota, que puede aislarse de una biopsia de tejido o puede ser ADN libre de células (ADNlc), ADN genómico microbiano, o ADN genómico viral), las secuencias de las repeticiones pueden ser idénticas excepto por las posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN o a errores que ocurren durante la amplificación. Un ejemplo de tal repetición directa se ilustra en la Fig. 1. Como se muestra, dentro de cada molécula de repetición, la primera repetición y la segunda repetición se amplifican a partir de hebras opuestas de un fragmento de ADN genómico bicatenario, por ejemplo ADN genómico. La primera repetición tiene la misma secuencia o una muy similar a una hebra (la hebra superior del fragmento, por ejemplo) de un fragmento de ADN genómico bicatenario, mientras que la segunda repetición tiene la misma secuencia o una muy similar a la del complemento inverso de la otra hebra del fragmento (por ejemplo, la hebra inferior del fragmento). En realizaciones, en las que el fragmento es ADN genómico, la primera y la segunda secuencia repetida deben ser idénticas excepto por los nucleótidos que corresponden a (es decir, están en una posición que corresponde a la posición de) nucleótidos dañados en el fragmento de ADN genómico bicatenario o a errores que han ocurrido durante la amplificación. En otras realizaciones, el fragmento bicatenario se puede obtener sintéticamente o derivar de un plásmido bicatenario, por ejemplo.

Un “nucleótido dañado” se refiere a cualquier derivado de adenina, citosina, guanina, y timina que se ha alterado de una manera que le permita emparejarse con una base diferente. En el ADN no dañado, la base A se empareja con la base T, y la base C con la G. Sin embargo, algunas bases se pueden oxidar, alquilar o desaminar de una manera que efectúe el emparejamiento de bases. Por ejemplo, la 7,8-dihidro-8-oxoguanina (8-oxo-dG) es un derivado de la guanina que empareja la base con adenina en lugar de citosina. Este derivado provoca una transversión de G a T después de la replicación.

La desaminación de la citosina produce uracilo, que puede emparejar la base con adenina, lo que lleva a un cambio de C a T después de la replicación. Se conocen otros ejemplos o nucleótidos dañados que son capaces de emparejarse incorrectamente.

Dentro de una molécula molde de repetición directa, las secuencias de la primera y la segunda repetición tienen longitudes idénticas, y son al menos 95% idénticas (por ejemplo, al menos 95% idénticas, al menos 96% idénticas, al menos 97% idénticas, al menos 98% idénticas, al menos 99% idénticas, o 100% idénticas, dependiendo, por ejemplo, de la extensión del daño en el ADN en el fragmento de ADN genómico bicatenario y/o errores de amplificación), y deberían ser idénticas, con la excepción de nucleótidos que corresponden a nucleótidos dañados y a errores de amplificación. Como se muestra, las moléculas pueden tener una longitud unitaria de 1, lo que significa que sólo hay una copia de la primera repetición y una copia de la segunda repetición en cada molécula. Las moléculas molde pueden ser monocatenarias o bicatenarias. Sin embargo, como se apreciaría, el molde está en su forma monocatenaria cuando se está secuenciando. La secuencia de la primera y la segunda repetición puede tener una longitud de al menos 50 nucleótidos, y en algunas realizaciones puede estar en el intervalo de 50 nucleótidos a 2 kb de longitud, por ejemplo 50-500 nt o 50-300 nt. En algunas realizaciones, el molde de repetición directa puede estar en una muestra que contiene otros moldes de repetición directa. Dentro de la población, la complejidad y la mediana de la longitud de la secuencia de la primera repetición pueden variar y pueden ser aproximadamente iguales a la complejidad y la mediana de la longitud de la secuencia de la segunda repetición, ya que esas secuencias son casi idénticas. En la población, la primera repetición y la segunda repetición pueden tener cada una una complejidad de al menos 10^3, por ejemplo al menos 10^4, al menos 10^5, al menos 10^6, al menos 10^7, al menos 10^8, al menos 10^9, o al menos 10^{10}, por ejemplo, lo que significa que en la población, la primera repetición

y la segunda repetición están representadas cada una por al menos 10^3 secuencias diferentes. Las longitudes de la primera y la segunda repetición pueden depender de las longitudes de los fragmentos de ADN en la muestra a partir de la cual se obtienen las moléculas. En algunas realizaciones, los fragmentos pueden tener una mediana de tamaño de no más de 2 kb de longitud (por ejemplo, en el intervalo de 50 pb a 2 kb, por ejemplo 75 pb a 1,5 kb, 100 pb a 1 kb, 100 pb a 500 pb). Las longitudes del fragmento pueden adaptarse a la plataforma de secuenciación que se esté usando. A continuación se describirán con mayor detalle ejemplos de cómo pueden obtenerse estas moléculas.

Por lo tanto, en cualquier realización, la molécula de repetición directa puede obtenerse copiando un fragmento bicatenario de ADN para producir la molécula de repetición directa, en la que la primera y la segunda repetición de la molécula de repetición directa se van a amplificar a partir de hebras opuestas del fragmento bicaternario de ADN.

Como se indicó anteriormente, en algunas realizaciones, el método puede comprender, en la misma reacción, hibridar un cebador con un primer sitio que está en dirección 5’ de la primera secuencia repetida, e hibridar un cebador con un segundo sitio que está en dirección 5’ de la segunda secuencia repetida. En estas realizaciones, los sitios primero y segundo (es decir, los sitios a los que se unen los cebadores primero y segundo, respectivamente) están en dirección 5’ de las secuencias repetidas primera y segunda, respectivamente, y equidistantes de las secuencias repetidas primera y segunda. Esto se ilustra en la Fig. 2. Como se ilustra, el primer cebador se une a un sitio que está en dirección 5’ de (es decir, 3’ hacia) la primera repetición, mientras que el segundo cebador se une a un sitio que está en dirección 5’ de (es decir, 3’ hacia) la segunda repetición, en la que las distancias entre los cebadores y sus respectivas repeticiones son las mismas. Ilustrado con un ejemplo, si el extremo 3’ del primer cebador se hibrida con un nucleótido que está en dirección 5’ de (es decir, 3’ hacia) la primera repetición mediante n bases (en el que n está en el intervalo de, por ejemplo, 5 a 30), entonces el extremo 3’ del segundo cebador se hibrida con un nucleótido que está en dirección 5’ de (es decir, 3’ hacia) la segunda repetición mediante n bases. Mientras que la distancia entre los sitios de unión del cebador y las repeticiones se puede definir por el número de bases para algunos métodos de secuenciación (por ejemplo, el método de secuenciación del terminador de colorante de Illumina), la distancia se puede definir por “flujos” en otros métodos (por ejemplo, Ion Torrent, o métodos de pirosecuenciación).

Después de la hibridación de los cebadores, el método puede comprender someter el producto de hibridación a una reacción de secuenciación de tipo secuenciación por síntesis para producir una lectura de secuencia que comprende una combinación de la primera y la segunda secuencia repetida, lo que significa que las secuencias se fusionan en una sola. En algunas realizaciones, los métodos de secuenciación por síntesis son aquellos que implican extender un cebador usando un molde, y detectar qué nucleótido se añade en cada posición. Los métodos de secuenciación por síntesis incluyen, pero no se limitan a, el método de terminador de colorante reversible de Illumina, el método de Ion Torrent de Thermo (que detecta iones a medida que son liberados por la ADN polimerasa), y la pirosecuenciación, aunque se conocen otros. En el enfoque de terminador de colorante reversible, la secuencia de un molde se determina usando química de terminadores reversibles (Turcatti et al., Nucleic Acids Res. 2008 36:e25). En cada ciclo de secuenciación, se añade en una reacción de extensión del cebador moldeado un único nucleótido bloqueado en 3’, marcado fluorescentemente. Después de la incorporación, la identidad del marcador fluorescente añadido se detecta mediante imágenes fluorescentes. En cada ronda, los marcadores y terminadores se eliminan químicamente para preparar el producto de extensión del cebador para el siguiente ciclo. En Bentley, más arriba, puede encontrarse una descripción más detallada del proceso.

Como se señaló anteriormente, la lectura de secuencia producida usando este método será una combinación de la primera y la segunda secuencia repetida, en la que el término “combinación” significa que las secuencias de la primera y la segunda repetición se fusionan, superponen o combinan en una. A modo de ejemplo, si la secuencia de la primera repetición es GATCGGATCGA (SEQ ID NO: 1) y la secuencia de la segunda repetición es GATCGGATCGA (SEQ ID NO: 1), entonces la lectura de secuencia contendrá sólo una copia de la secuencia GATCGGATCGA (SEQ ID NO: 1), en la que parte de la señal usada para generar la lectura de secuencia se genera por la extensión del primer cebador, y parte de la señal usada para generar la lectura de secuencia se genera por la extensión del segundo cebador en la misma reacción.

Las diferencias en las secuencias de la primera y la segunda repetición se pueden identificar debido a que la señal subyacente correspondiente a la diferencia se mezclará (es decir, será una combinación de señales producidas por dos bases diferentes en esa posición). Las posiciones que tienen una señal mixta pueden identificarse debido a que están asociadas con una llamada de bases de baja calidad. Como tal, las diferencias en las secuencias de la primera y la segunda repetición pueden identificarse como posiciones que tienen una llamada de bases de baja calidad. En estas realizaciones, la lectura de secuencia comprende, para cada posición de la lectura de secuencia, una puntuación de calidad que indica la fiabilidad de la o las bases llamadas en esa posición. La llamada de bases es el proceso mediante el cual se infiere un orden de nucleótidos en un molde durante una reacción de secuenciación. Por ejemplo, las plataformas de secuenciación de próxima generación que usan terminadores reversibles marcados fluorescentemente tienen un color único para cada base. Estos se incorporan a la cadena complementaria del molde de ADN y se capturan con una cámara CCD sensible. Estas imágenes se procesan en señales que se usan para inferir el orden de los nucleótidos, también conocido como llamada de bases.

La precisión de la llamada de bases se puede medir de diferentes maneras. En algunas realizaciones, la precisión de la llamada de bases se puede medir usando una puntuación Q (puntuación de calidad de Phred), que es una métrica común

para evaluar la precisión de un experimento de secuenciación. Las puntuaciones Q se definen como relacionadas logarítmicamente con la probabilidad de error de llamada de bases, en las que Q = - 10 log P / log 10. En este sistema, si a una base se le asigna una puntuación Q de 40, esto equivale a la probabilidad de una llamada de bases incorrecta de 1 en 10.000 veces, o una precisión de llamada de bases del 99,99%; una puntuación Q menor que 10 significa que existe la probabilidad de una llamada incorrecta en 1 de 10 bases. Las puntuaciones Q más bajas pueden conducir a aumentos en las llamadas de variantes falsas positivas, y reducen la confianza general que un investigador tiene en sus datos de secuenciación. Los detalles de la llamada de bases y los métodos para calcular la calidad de una llamada de bases se describen en una variedad de publicaciones, que incluyen, por ejemplo, Ledergerber et al. (Brief Bioinform. 2011 12: 489-497), Whiteford et al. (Bioinformatics 2009 25: 2194-2199), Erlich (Nat. Methods. 2008 5: 679-682) y Kao et al. (Genome Res. 2009 19: 1884-95).

En algunas realizaciones, el método puede usarse para identificar posiciones que difieren en la primera y la segunda repetición. En estas realizaciones, una posición en la lectura de secuencia que no se llama o está asociada con una puntuación de baja calidad indica que las secuencias repetidas primera y segunda difieren en un nucleótido que corresponde a esa posición. A modo de ejemplo, si la secuencia de la primera repetición es GATCGGATCGA (SEQ ID NO: 1) y la secuencia de la segunda repetición es GATCGTATCGA (SEQ ID NO: 2), entonces la lectura de secuencia puede contener sólo una copia de la secuencia GATCGG[G/T]ATCGA (SEQ ID NO: 3), en la que “G/T” es una base que tiene una señal mixta y por lo tanto está asociada con una llamada de bases de mala calidad. En este ejemplo, la calidad de las llamadas de bases para las bases que no son G/T será alta, y la calidad de la llamada de bases para la base G/T será mala debido a que parte de la señal para esa posición, según se analiza mediante el algoritmo de la llamada de bases, se generará por extensión del primer cebador y parte de la señal se generará por extensión del segundo cebador, en la misma reacción.

Después de identificar una posición que tiene una llamada de bases de baja calidad (o, en algunos casos, una posición que no está llamada), el método puede comprender además analizar las señales subyacentes para esa posición para determinar las identidades de los nucleótidos en esa posición en la primera y la segunda repetición. Por ejemplo, en el ejemplo descrito en el párrafo anterior, las señales subyacentes (es decir, antes de la llamada de bases y denominadas datos de secuencia primaria) podrían analizarse para determinar que la posición contiene una mezcla de G y T, indicando de ese modo que la primera repetición contiene una G o una T en esa posición, y la segunda repetición contiene el otro nucleótido. Como tal, en cualquier realización, el método puede comprender leer una combinación de señales obtenidas mediante la extensión simultánea del primer y segundo cebador para producir datos de secuenciación primaria, procesar los datos de secuenciación primaria usando un algoritmo de llamada de bases para producir una lectura de secuencia compuesta por un secuencia de llamadas de bases, indicando cada llamada de bases asociada con una puntuación de calidad la fiabilidad de la llamada de bases; y generar la lectura de secuencia en base a las puntuaciones de calidad. Las puntuaciones de calidad permiten identificar las diferencias entre la primera y la segunda repetición.

En algunas realizaciones, los sitios primero y segundo en el molde (es decir, las secuencias a las que se unen los cebadores primero y segundo) son la misma secuencia. En estas realizaciones, se puede usar un solo cebador en el método, en el que el cebador se une a dos sitios en el molde. En realizaciones alternativas, los sitios primero y segundo en el molde (es decir, las secuencias a las que se unen los cebadores primero y segundo) pueden ser secuencias diferentes. En estas realizaciones, se pueden usar dos o más cebadores en el método, en el que el cebador se une a diferentes secuencias en el molde, uno en dirección 5’ de la primera repetición y el otro en la dirección 5’ de la segunda repetición.

En algunas realizaciones, el método puede implicar determinar cuántas hebras de la primera repetición se secuencian con respecto al número de hebras de la segunda repetición, o si se ha secuenciado un número suficiente de moléculas. Estas realizaciones pueden implementarse añadiendo una secuencia de calibración al molde, como se muestra en la Fig. 3. En estas realizaciones, el molde puede comprender: una primera secuencia calibradora que está presente entre el primer sitio y la primera repetición; y una segunda secuencia calibradora que está presente entre el segundo sitio y la segunda repetición, en el que la primera y la segunda secuencia calibradora tienen la misma longitud (por ejemplo, pueden tener dos, tres o cuatro bases de longitud o el mismo número de flujos de longitud, dependiendo del método de secuenciación usado) y tienen una secuencia diferente; y la lectura de secuencia de la etapa (b) incluye posiciones que corresponden a la primera y a la segunda secuencia calibradora. En estas realizaciones, las señales subyacentes correspondientes a la primera y a la segunda secuencia calibradora (antes de la llamada de bases) pueden examinarse para determinar cuántas hebras de la primera y de la segunda repetición se secuencian en la reacción. Asimismo, las señales subyacentes correspondientes a la primera y a la segunda secuencia calibradora (antes de la llamada de bases) pueden examinarse para determinar si se ha secuenciado un número suficiente de moléculas.

En muchos métodos de secuenciación por síntesis, las moléculas molde se amplifican clonalmente, y los productos de amplificación se secuencian de forma muy paralela. Tales métodos se repasan en, por ejemplo, Metzker et al. (Genome Res. 2005 15:1767-1776) y Bentley (Curr. Opin. Genet. Dev. 2006 16: 545-55). En la secuenciación de Illumina, los moldes se distribuyen en una celda de flujo y se inmovilizan en un soporte (normalmente, vidrio; véase Fedurco et al., Nucleic Acids Res. 2006 34:e22), en el que se amplifican *in situ* mediante PCR puente, que genera grupos de moldes idénticos (o “colonias”) en el soporte. Como tal, el presente método puede implementarse amplificando el molde sobre un sustrato mediante PCR puente para producir una colonia que comprende copias del molde, hibridando uno o más cebadores con la colonia, en el que un cebador se hibrida con un primer sitio que está en dirección 5’ de la primera secuencia repetida,

y un cebador se hibrida con un segundo sitio que está en dirección 5' de la segunda secuencia repetida, en el que los sitios primero y segundo están: en dirección 5' de la primera y de la segunda secuencia repetida, respectivamente, equidistantes de la primera y de la segunda secuencia repetida; y obteniendo la secuencia del molde mediante una reacción de secuenciación de tipo secuenciación por síntesis para producir una lectura de secuencia que comprende una combinación de la primera y de la segunda secuencia repetida. En algunas realizaciones (y como se ilustra en la Fig. 3), las hebras superior e inferior de los productos de amplificación por PCR puente se pueden secuenciar mediante el método de secuenciación de Illumina (que se denomina secuenciación de "extremos emparejados"). Como tal, en algunas realizaciones, la secuencia de una hebra superior de un producto de PCR puente se puede comparar con la secuencia de una hebra inferior de un producto de PCR puente. Las posiciones que están asociadas con una llamada de bases de baja calidad como resultado de una diferencia en la secuencia entre la primera y la segunda repetición deben tener una llamada de bases de baja calidad en ambas hebras. En algunas realizaciones, después de secuenciar ambas hebras del producto mediante secuenciación de extremos emparejados, se puede producir una secuencia consenso para la hebra superior del fragmento bicatenario inicial y una secuencia consenso para la hebra inferior del fragmento bicatenario inicial. Las bases de baja calidad se pueden enmascarar o integrar en un modelo en el que se tienen en cuenta las puntuaciones de calidad. Las secuencias que no están presentes en las hebras superior e inferior del fragmento bicatenario inicial pueden eliminarse de ese modo de análisis futuros.

La Fig. 3 ilustra un ejemplo del método. En este ejemplo, el molde es una molécula bicatenaria, y es necesario secuenciar una o ambas hebras (se muestra la secuenciación de la hebra inferior). En este ejemplo, el molde de repetición directa tiene secuencias de celdas de flujo (por ejemplo, las secuencias P5 y P7 de Illumina) en los extremos, y un sitio de unión al cebador entre la primera y la segunda repetición. Como se muestra, esta molécula se amplifica a partir de un fragmento bicatenario, en el que la primera y la segunda secuencia repetida (W* y W, o C* y C) se amplifican a partir de hebras opuestas de un fragmento bicatenario de ADN y son idénticas excepto por posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN o a errores que ocurren durante la amplificación. Como se muestra, el método puede implicar hibridar dos cebadores (designados como P_1 y P_2, que pueden ser iguales o diferentes) con el molde (después de que se haya amplificado). En esta realización, cada una de las repeticiones tiene una secuencia calibradora (denominada "clave 1" y "clave 2", que se puede usar para determinar el número relativo de copias de la primera y de la segunda repetición que se secuencian en una reacción. Como se muestra, la parte de la lectura de secuencia obtenida del cebador P_1 debe contener la clave 1 (TT), y la parte de la lectura de secuencia obtenida del cebador P_2 debe contener la clave 1 (AA). En este ejemplo, hay una diferencia en la secuencia en la primera y la segunda repetición, que puede identificarse como una llamada de bases con una baja calidad (como resultado de que el molde tiene un nucleótido mixto en esa posición).

En realizaciones en las que hay una secuencia no informativa inmediatamente en dirección 3' de un sitio de unión del cebador, los cebadores pueden extenderse pero no leerse para los primeros ciclos, permitiendo de ese modo la secuencia de las claves y/o repeticiones más rápidamente.

En algunas realizaciones, el molde de repetición directa puede tener diferentes secuencias no complementarias (Secuencias 1 y 2 en la Fig. 3) en al menos 10 nucleótidos (por ejemplo, al menos 10, 12 o 14 nucleótidos de longitud) que permiten que los fragmentos sean amplificados por un único par de cebadores: un primer cebador que se hibrida con una secuencia, y otro que se hibrida con el complemento de la otra secuencia. Estas secuencias pueden ser compatibles con la plataforma de secuenciación que se está usando. Estas secuencias no necesitan estar al final de una molécula, aunque, en muchas realizaciones, las secuencias están dentro de los 50 nt, por ejemplo dentro de los 30 nt del extremo de la molécula. Como será evidente, la molécula molde deberá tener una secuencia de unión entre la primera y la segunda repetición. La secuencia de unión debe ser de 10 nucleótidos (por ejemplo, 10 a 100 nt). El molde puede contener un código de barras molecular (por ejemplo, un identificador de muestra o de molécula) en cualquier posición (fuera de las repeticiones).

El método descrito anteriormente se puede emplear para analizar el ADN genómico de prácticamente cualquier organismo, incluyendo, pero sin limitarse a, plantas, animales (por ejemplo, reptiles, mamíferos, insectos, gusanos, peces, etc.), muestras de tejido, bacterias, hongos (por ejemplo, levadura), fagos, virus, tejido cadavérico, muestras arqueológicas/antiguas, etc. En ciertas realizaciones, el ADN genómico usado en el método puede derivar de un mamífero, en el que, en ciertas realizaciones, el mamífero es un ser humano. En realizaciones ejemplares, la muestra puede contener ADN genómico de una célula de mamífero, tal como una célula humana, de ratón, de rata, o de mono. La muestra puede estar hecha de células cultivadas o células de una muestra clínica, por ejemplo una biopsia de tejido, raspado o lavado, o células de una muestra forense (es decir, células de una muestra recogida en la escena del crimen). En realizaciones particulares, la muestra de ácido nucleico se puede obtener de una muestra biológica tal como células, tejidos, fluidos corporales, y heces. Los fluidos corporales de interés incluyen, pero no se limitan a, sangre, suero, plasma, saliva, moco, flema, líquido cefalorraquídeo, líquido pleural, lágrimas, líquido del conducto lácteo, linfa, esputo, líquido sinovial, orina, líquido amniótico, y semen. En realizaciones particulares, una muestra se puede obtener de un sujeto, por ejemplo un ser humano. En algunas realizaciones, la muestra comprende fragmentos de ADN genómico humano. En algunas realizaciones, la muestra se puede obtener de un paciente con cáncer. En algunas realizaciones, la muestra se puede obtener extrayendo ADN fragmentado de una muestra de paciente, por ejemplo una muestra de tejido fijado en formalina embebido en parafina. En algunas realizaciones, la muestra del paciente puede ser una muestra de ADN "circulante" libre de células procedente de un fluido corporal, por ejemplo sangre periférica, por ejemplo procedente de la sangre de un paciente o de una mujer embarazada. Los fragmentos de ADN usados en la etapa inicial del método deben ser ADN no

amplificado que no se ha desnaturalizado previamente.

El ADN en la muestra inicial puede obtenerse extrayendo ADN genómico de una muestra biológica, y después fragmentándolo. En algunas realizaciones, la fragmentación se puede realizar mecánicamente (por ejemplo, mediante sonicación, nebulización, o cizallamiento, etc.), o usando una enzima fragmentasa de ADN bicatenario "ADNbc" (New England Biolabs, Ipswich MA). En algunos de estos métodos (por ejemplo, los métodos mecánico y de fragmentasa), después de fragmentar el ADN, los extremos pueden pulirse y formar colas A antes de la ligación a uno o más adaptadores. Alternativamente, los extremos pueden pulirse y ligarse a adaptadores en una reacción de ligación de extremos romos. En otras realizaciones, el ADN en la muestra inicial ya puede estar fragmentado (por ejemplo, como es el caso de las muestras FFPE (embebidas en parafina fijadas con formalina), y ADN circulante libre de células (ADNlf), por ejemplo, ADNtc). Los fragmentos en la muestra inicial pueden tener una mediana de tamaño que está por debajo de 1 kb (por ejemplo, en el intervalo de 50 pb a 500 pb, u 80 pb a 400 pb), aunque se pueden usar fragmentos que tienen una mediana de tamaño fuera de este intervalo.

En algunas realizaciones, la cantidad de ADN en una muestra puede ser limitante. Por ejemplo, la muestra inicial de ADN fragmentado puede contener menos de 200 ng de ADN humano fragmentado, por ejemplo 1 pg a 20 pg, 10 pg a 200 ng, 100 pg a 200 ng, 1 ng a 200 ng, o 5 ng a 50 ng, o menos de 10.000 (por ejemplo, menos de 5.000, menos de 1.000, menos de 500, menos de 100, menos de 10, o menos de 1) equivalentes de genoma haploide, dependiendo del genoma.

En algunas realizaciones, los identificadores de muestra (es decir, una secuencia que identifica la muestra a la que se añade la secuencia, que puede identificar al paciente o un tejido, etc.) se pueden añadir a los polinucleótidos antes de la secuenciación, de modo que pueden multiplexarse múltiples muestras (por ejemplo al menos 2, al menos 4, al menos 8, al menos 16, al menos 48, al menos 96, o más). En estas realizaciones, el identificador de la muestra puede ligarse a los polinucleótidos iniciales como parte del adaptador asimétrico, o el identificador de la muestra puede ligarse a los polinucleótidos en las submuestras, antes o después de la amplificación de esos polinucleótidos. Como alternativa, la etiqueta se puede añadir mediante la extensión del cebador, es decir, usando un cebador que tiene un extremo 3' que se hibrida con una secuencia adaptadora, y una cola 5' que contiene el identificador de la muestra.

La población de moléculas de repetición directa se puede obtener de muchas maneras distintas. Estos métodos se basan en la creación de moléculas circulares, manteniendo la proximidad física entre las dos hebras de una molécula de ADN bicatenario, o aislando físicamente dos hebras de una molécula bicatenaria, durante las etapas de manipulación. Los métodos también se dividen en estrategias que requieren uno, o más, tipos de adaptadores. Estos métodos se pueden realizar fragmentando, puliendo, y añadiendo entonces colas a los extremos de los fragmentos antes de la ligación del adaptador. Alternativamente, se pueden usar transposasas para añadir secuencias adaptadoras. En algunas realizaciones, se pueden usar transposones estándar, pero después se pueden modificar para crear un adaptador en forma de Y usando reemplazo de oligonucleótidos (Grunenwald H, Baas B, Goryshin I, Zhang B, Adey A, Hu S, Shendure J, Caruccio N, Maffitt M 2011. Nextera PCR-free DNA library preparation for next-generation. [Presentación de póster, AGBT 2011]; Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, Kuersten S, Myers RM 2012. Transposase mediated construction of RNA-seq libraries. Genome Res 22: 134-141).

En algunas realizaciones, el molde de repetición directa puede obtenerse (a) ligando secuencias adaptadoras en ambos extremos de las hebras superior e inferior de una población de fragmentos de ADN genómico bicatenario para producir moléculas bicatenarias que comprenden (i) una hebra superior que comprende una secuencia 5' (por ejemplo, X) en el extremo 5', y una secuencia de unión (por ejemplo, J) en el extremo 3'; y (ii) una hebra inferior que comprende una secuencia 5' (por ejemplo, Y') en el extremo 5', y el complemento de la secuencia de unión (J') en el extremo 3'; y (b) extendiendo el extremo 3' de las hebras superiores (es decir, la hebra que contiene la secuencia X) usando la hebra inferior como molde, copiando así el complemento de la hebra inferior, así como las secuencias J e Y, en la misma molécula que la hebra superior para producir una molécula de repetición directa de fórmula: X-SUP-J-INF'-Y, en la que: (i) dentro de cada molécula de repetición, SUP e INF' se amplifican a partir de hebras opuestas de un fragmento del ADN genómico bicatenario e idénticas excepto por las posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN genómico o a errores de amplificación. En estas realizaciones, SUP e INF' varían en la población y tienen una mediana de longitud de al menos 50 nucleótidos, y X e Y son secuencias diferentes no complementarias de al menos 10 nucleótidos de longitud que no varían en la población; y J es una secuencia de unión. Los ejemplos de este método se muestran en las figuras, y se describen con mayor detalle más abajo.

En algunas realizaciones y como se muestra en las Fig. 4 y 5, una molécula de repetición directa se puede obtener ligando un solo adaptador en ambos extremos de las hebras superior e inferior de una población de fragmentos de ADN genómico bicatenario, de modo que las moléculas individuales están en un círculo covalentemente abierto y, en las moléculas individuales en la población, la secuencia X se añade al extremo 5' de las hebras superiores del fragmento, y la secuencia Y' se liga al 5' de las hebras inferiores de los fragmentos. Este método implica extender el extremo 3' de las hebras superiores (es decir, la hebra que contiene la secuencia X) usando la hebra inferior como molde, copiando así el complemento de la hebra inferior, así como la secuencia Y, en la misma molécula que la hebra superior. Tal molécula se puede amplificar usando cebadores que tienen un extremo 3' que es el mismo que o que se hibrida con la secuencia X e Y. Un ejemplo de tal método se ilustra en las Figs. 4 y 5, en las que la hebra superior de los fragmentos de ADN genómico se indica como "directa" e "inversa", respectivamente, y las secuencias X e Y' se indican como secuencias R1 y R2.

En algunas realizaciones, las moléculas de repetición directa pueden obtenerse ligando un solo adaptador en ambos extremos de las hebras superior e inferior de una población de fragmentos de ADN genómico bicatenario, de modo que las moléculas individuales estén en un círculo covalentemente cerrado, y, en las moléculas individuales de la población, la secuencia X se añade al extremo 5' de las hebras superiores del fragmento, y la secuencia Y' se liga al extremo 5' de las hebras inferiores de los fragmentos. Este método implica crear una o más muescas haciendo reaccionar, por ejemplo, un adaptador que contiene dUTP y una mezcla de UDG/endonucleasa IV, extendiendo el extremo 3' de las hebras superiores (es decir, la hebra que contiene la secuencia X) usando la hebra inferior como un molde, copiando así el complemento de la hebra inferior, así como la secuencia Y, en la misma molécula que la hebra superior. Tal molécula se puede amplificar usando cebadores que tienen un extremo 3' que es el mismo que o que se hibrida con la secuencia X e Y.

Un producto similar se puede obtener mediante PCR en emulsión, usando un enfoque de inmovilización, o amplificación de círculo rodante, métodos de adaptador único y métodos de más de 1 adaptador, como se describe en el documento WO2018229547.

En algunas realizaciones, el molde de repetición directa puede tener la fórmula X-SUP-J-INF'-Y, en la que (i) dentro de cada molécula de repetición, SUP e INF' se amplifican a partir de hebras opuestas de un fragmento bicatenario de ADN genómico y son idénticas excepto por las posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN genómico o a errores que ocurren durante la amplificación; (ii) SUP e INF' tienen una mediana de longitud de al menos 50 nucleótidos; (iii) X e Y son secuencias diferentes no complementarias de al menos 10 nucleótidos; y (iv) J es una secuencia de unión de, por ejemplo, al menos 10 nucleótidos de longitud. En algunas realizaciones, el molde de repetición directa puede tener una hebra de la fórmula X-(T)SUP(A)-J-(T)INF'(A)-Y, en la que (T) y (A) son los nucleótidos timina y adenina que están inmediatamente adyacentes a SUP e INF'. Tales moléculas pueden obtenerse, por ejemplo, (a) ligando secuencias adaptadoras en ambos extremos de las hebras superior e inferior de una población de fragmentos de ADN genómico bicatenario para producir moléculas bicatenarias que comprenden: (i) una hebra superior que comprende la secuencia X en el extremo 5', y la secuencia J en el extremo 3'; y (ii) una hebra inferior que comprende la secuencia Y' en el extremo 5', y la secuencia J' en el extremo 3'; y (b) extender el extremo 3' de las hebras superiores usando las hebras inferiores como molde, añadiendo así el complemento de las hebras inferiores y la secuencia Y en el extremo 3' de las hebras superiores. Este método se ilustra en las Figs. 4 y 5.

Kits

Esta descripción también proporciona un kit para practicar el método en cuestión, como se describe anteriormente. Los diversos componentes del kit pueden estar presentes en recipientes separados, o ciertos componentes compatibles pueden combinarse previamente en un solo recipiente, según se desee.

Además de los componentes mencionados anteriormente, los kits en cuestión pueden incluir además instrucciones para usar los componentes del kit para practicar los métodos en cuestión, es decir, para proporcionar instrucciones para el análisis de muestras. Las instrucciones para practicar los métodos en cuestión se guardan generalmente en un medio de registro adecuado. Por ejemplo, las instrucciones pueden estar impresas en un sustrato, tal como papel o plástico, etc. Como tal, las instrucciones pueden estar presentes en los kits como un prospecto, en el etiquetado del envase del kit o de los componentes del mismo (es decir, asociado con el embalaje o subembalaje), etc. En otras realizaciones, las instrucciones están presentes como un archivo de datos de almacenamiento electrónico presente en un medio de almacenamiento legible por computadora adecuado, por ejemplo CD-ROM, disquete, etc. En aún otras realizaciones, las instrucciones reales no están presentes en el kit, pero se proporcionan medios para obtener las instrucciones de una fuente remota, por ejemplo a través de Internet. Un ejemplo de esta realización es un kit que incluye una dirección web en la que se pueden ver las instrucciones y/o desde la cual se pueden descargar las instrucciones. Al igual que con las instrucciones, este medio para obtener las instrucciones se guarda en un sustrato adecuado.

Utilidad

Como será fácilmente manifiesto, el método descrito anteriormente se puede emplear para analizar cualquier tipo de muestra, incluyendo, pero sin limitarse a, muestras que contienen mutaciones heredables, muestras que contienen mutaciones somáticas, individuos con mosaicismo, hembras preñadas (en las que parte de la muestra contiene ADN de un feto en desarrollo), y muestras que contienen una mezcla de ADN de diferentes fuentes. En ciertas realizaciones, el método puede usarse para identificar una variante minoritaria que, en algunos casos, puede deberse a una mutación somática en una persona.

En algunas realizaciones, el método puede emplearse para detectar una mutación oncogénica (que puede ser una mutación somática) en, por ejemplo, PIK3CA, NRAS, KRAS, JAK2, HRAS, FGFR3, FGFR1, EGFR, CDK4, BRAF, RET, PGDFRA, KIT o ERBB2, que puede estar asociada con cáncer de mama, melanoma, cáncer renal, cáncer de endometrio, cáncer de ovario, cáncer de páncreas, leucemia, cáncer colorrectal, cáncer de próstata, mesotelioma, glioma, meduloblastoma, policitemia, linfoma, sarcoma, o mieloma múltiple (véase, por ejemplo, Chial 2008 Proto-oncogenes to oncogenes to cancer. Nature Education 1:1). Otras mutaciones oncogénicas (que pueden ser mutaciones somáticas) de interés incluyen mutaciones en, por ejemplo, APC, AXIN2, CDH1, GPC3, CYLD, EXT1, EXT2, PTCH, SUFU, FH, SDHB, SDHC, SDHD, VHL, TP53, WT1, STK11/LKB1, PTEN, TSC1, TSC2, CDKN2A, CDK4, RB1, NF1, BMPR1A, MEN1,

SMAD4, BHD, HRPT2, NF2, MUTYH, ATM, BLM, BRCA1, BRCA2, FANCA, FANCC, FANCD2, FANCE, FANCF, FANCG, NBS1, RECQL4, WRN, MSH2, MLH1, MSH6, PMS2, XPA, XPC, ERCC2-5, DDB2 o MET, que pueden estar asociadas con cánceres de colon, tiroides, paratiroides, pituitaria, célula de los islotes, estómago, intestinal, embrionario, óseo, renal, de mama, de cerebro, de ovario, pancreático, uterino, ocular, de folículo piloso, sanguíneo, o de útero, pilotricomas, meduloblastomas, leiomiomas, paragangliomas, feocromocitomas, hamartomas, gliomas, fibromas, neuromas, linfomas, o melanomas. En algunas realizaciones, el método puede emplearse para detectar una mutación somática en genes que están implicados en el cáncer, por ejemplo CTNNB1, BCL2, TNFRSF6/FAS, BAX, FBXW7/CDC4, GLI, HPVE6, MDM2, NOTCH1, AKT2, FOXO1A, FOXO3A, CCND1, HPVE7, TAL1, TFE3, ABL1, ALK, EPHB2, FES, FGFR2, FLT3, FLT4, KRAS2, NTRK1, NTRK3, PDGFB, PDGFRB, EWSR1, RUNX1, SMAD2, TGFBR1, TGFBR2, BCL6, EVI1, HMGA2, HOXA9, HOXA11, HOXA13, HOXC13, HOXD11, HOXD13, HOX11, HOX11L2, MAP2K4, MLL, MYC, MYCN, MYCL1, PTNP1, PTNP11, RARA, SS18 (véase, por ejemplo, Vogelstein y Kinzler 2004 Cancer genes and the pathways they control. Nature Medicine 10:789-799). El método de realización se puede emplear para detectar cualquier mutación somática que esté implicada en el cáncer que esté catalogada por COSMIC (Catálogo de Mutaciones Somáticas en Cáncer), cuyos datos se pueden acceder en Internet.

Otras mutaciones de interés incluyen mutaciones en, por ejemplo, ARID1A, ARID1B SMARCA4, SMARCB1, SMARCE1, AKT1, ACTB/ACTG1, CHD7, ANKRD11, SETBP1, MLL2, ASXL1, que pueden estar al menos asociadas con síndromes raros tales como el síndrome de Coffin-Siris, síndrome de Proteus, síndrome de Baraitser-Winter, síndrome de CHARGE, síndrome de KBG, síndrome de Schinzel-Giedion, síndrome de Kabuki, o síndrome de Bohring-Opitz (véase, por ejemplo, Veltman y Brunner 2012 *De novo* mutations in human genetic disease. Nature Reviews Genetics 13:565-575). Por lo tanto, el método puede emplearse para detectar una mutación en esos genes.

En otras realizaciones, el método puede emplearse para detectar una mutación en genes que están implicados en una variedad de trastornos del neurodesarrollo, por ejemplo KAT6B, THRA, EZH2, SRCAP, CSF1R, TRPV3, DNMT1, EFTUD2, SMAD4, LIS1, DCX, que pueden estar asociados con el síndrome de Ohdo, hipotiroidismo, síndrome Genitopatelar, síndrome de Weaver, síndrome de puerto flotante, leucoencefalopatía hereditaria difusa con esferoides, síndrome de Olmsted, ADCA-DN (ataxia cerebelosa autosómica dominante, sordera y narcolepsia), disostosis mandibulofacial con microcefalia o síndrome de Myhre (véase, por ejemplo, Ku et al. (2012) A new paradigm emerges from study of *de novo* mutations in the context of neurodevelopmental disease. Molecular Psychiatry 18:141-153). El método también se puede emplear para detectar una mutación somática en genes que están implicados en una variedad de trastornos neurológicos y neurodegenerativos, por ejemplo SCN1A, MECP2, IKBKG/NEMO o PRNP (véase, por ejemplo, Poduri et al. (2014) Somatic mutation, genetic variation, and neurological disease. Science 341 (6141): 1237758).

En algunas realizaciones, una muestra se puede recoger de un paciente en una primera localización, por ejemplo en un entorno clínico tal como un hospital o en el consultorio de un médico, y la muestra se puede enviar a una segunda localización, por ejemplo un laboratorio en el que se procesa, y el método descrito anteriormente se lleva a cabo para generar un informe. Un “informe”, como se describe aquí, es un documento electrónico o tangible que incluye elementos del informe que proporcionan resultados de ensayos que pueden indicar la presencia y/o cantidad de variante o variantes minoritarias en la muestra. Una vez generado, el informe puede enviarse a otra localización (que puede ser la misma localización que la primera localización), en la que puede se puede interpretar por un profesional de la salud (por ejemplo, un médico especialista, un técnico de laboratorio, o un médico tal como un oncólogo, cirujano, patólogo, o virólogo), como parte de una decisión clínica.

El método se puede usar para analizar enfermedades asociadas con mutaciones, rechazo de trasplantes, y tiene aplicaciones en ensayos prenatales no invasivos.

REIVINDICACIONES

1. Un método para secuenciar un molde que comprende una primera secuencia repetida y una segunda secuencia repetida, en el que la primera y la segunda secuencia repetida están en una repetición directa y son idénticas o casi idénticas, que comprende:

(a) en la misma reacción, hibridar un cebador con un primer sitio que está en dirección 5' de la primera secuencia repetida, e hibridar un cebador con un segundo sitio que está en dirección 5' de la segunda secuencia repetida, en el que los sitios primero y segundo están:

(i) en dirección 5' de la primera y la segunda secuencia repetida, respectivamente, y

(ii) equidistantes de la primera y la segunda secuencia repetida; y

(b) someter el producto de hibridación de (a) a una reacción de secuenciación de tipo secuenciación por síntesis para producir una lectura de secuencia que comprende una combinación de la primera y la segunda secuencia repetida.

2. El método de la reivindicación 1, en el que dentro de cada molde, la primera secuencia repetida y la segunda secuencia repetida se amplifican a partir de hebras opuestas de un fragmento bicatenario de ADN y son idénticas excepto por las posiciones que corresponden a nucleótidos dañados en el fragmento bicatenario de ADN o a errores que ocurren durante la amplificación.

3. El método de la reivindicación 2, en el que el fragmento bicatenario de ADN es ADN genómico.

4. El método de la reivindicación 3, en el que el ADN genómico es ADN genómico eucariota, ADN aislado de una biopsia de tejido, ADN libre de células (ADNlc), ADN genómico microbiano, o ADN genómico viral.

5. El método de cualquier reivindicación anterior, en el que la lectura de secuencia de (b) comprende, para cada posición de la lectura de secuencia, una puntuación de calidad que indica la fiabilidad de la o las bases llamadas en esa posición.

6. El método de la reivindicación 5, en el que una posición en la lectura de secuencia que no está llamada o que está asociada con una puntuación de baja calidad indica que la primera y la segunda secuencia repetida difieren en un nucleótido que corresponde a esa posición.

7. El método de la reivindicación 6, que comprende además analizar datos de secuenciación primaria para una posición que tiene una puntuación de baja calidad para determinar las identidades de los nucleótidos en esa posición en la primera y la segunda repetición.

8. El método de cualquier reivindicación anterior, en el que la etapa (b) comprende:

(i) leer una combinación de señales obtenidas por extensión simultánea del primer y segundo cebador para producir datos de secuenciación primaria;

(ii) procesar los datos de secuenciación primaria usando un algoritmo de llamada de bases para producir una lectura de secuencia compuesta por una secuencia de llamadas de bases, indicando cada llamada de bases asociada con una puntuación de calidad la fiabilidad de la llamada de bases; y

(iii) generar la lectura de secuencia basada en (ii).

9. El método de cualquier reivindicación anterior, en el que la secuenciación por síntesis de la etapa (b) comprende extender simultáneamente el primer y el segundo cebador en presencia de terminadores de cadena reversibles.

10. El método de cualquier reivindicación anterior, en el que los sitios primero y segundos en el molde son la misma secuencia.

11. El método de cualquiera de las reivindicaciones 1-9, en el que los sitios primero y segundo en el molde son secuencias diferentes.

12. El método de cualquier reivindicación anterior, en el que el molde comprende:

(i) una primera secuencia calibradora que está presente entre el primer sitio y la primera repetición; y

(ii) una segunda secuencia calibradora que está presente entre el segundo sitio y la segunda repetición, en el que la primera y la segunda secuencia calibradora tienen la misma longitud y tienen una secuencia diferente; y

la lectura de secuencia de la etapa (b) incluye posiciones que corresponden a la primera y la segunda secuencia calibradora.

13. El método de la reivindicación 12, que comprende además analizar las señales correspondientes a la primera y a la segunda secuencia calibradora para determinar cuántas hebras de la primera y la segunda repetición se secuencian en la reacción, y opcionalmente que comprende además analizar las señales correspondientes a la primera y a la segunda secuencia calibradora para determinar si se ha secuenciado un número suficiente de moléculas.

14. El método de cualquier reivindicación anterior, en el que la primera y la segunda repetición tienen menos de 2.000 nucleótidos de longitud.

15. El método de cualquier reivindicación anterior, en el que el método se realiza:

amplificando el molde sobre un sustrato mediante PCR puente para producir una colonia que comprenda copias del molde;

hibridando uno o más cebadores con la colonia, en el que un cebador se hibrida con un primer sitio que está en dirección 5' de la primera secuencia repetida y un cebador se hibrida con un segundo sitio que está en dirección 5' de la segunda secuencia repetida, en el que los sitios primero y segundo están: en dirección 5' de la primera y la segunda secuencia repetida, respectivamente, y equidistantes de la primera y la segunda secuencia repetida; y

obteniendo la secuencia del molde mediante una reacción de secuenciación de tipo secuenciación por síntesis para producir una lectura de secuencia que comprende una combinación de la primera y la segunda secuencia repetida.

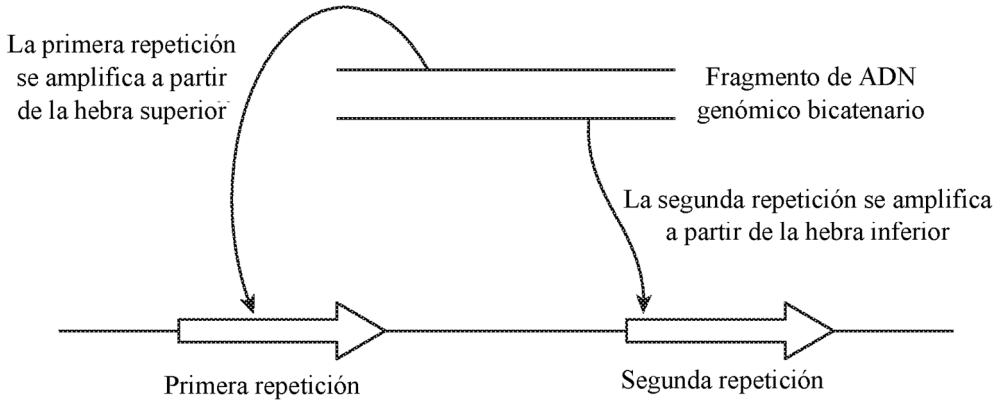
La primera repetición
se amplifica a partir
de la hebra superior
Fragmento de ADN
genómico bicatenario
La segunda repetición se amplifica
a partir de la hebra inferior
Primera repetición
Segunda repetición

FIG. 1

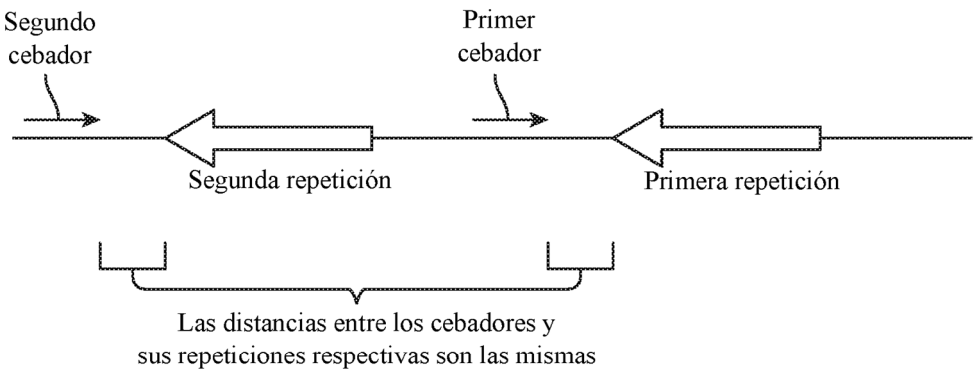
Segundo
cebador
Primer
cebador
Segunda repetición
Primera repetición
Las distancias entre los cebadores y
sus repeticiones respectivas son las mismas

FIG. 2

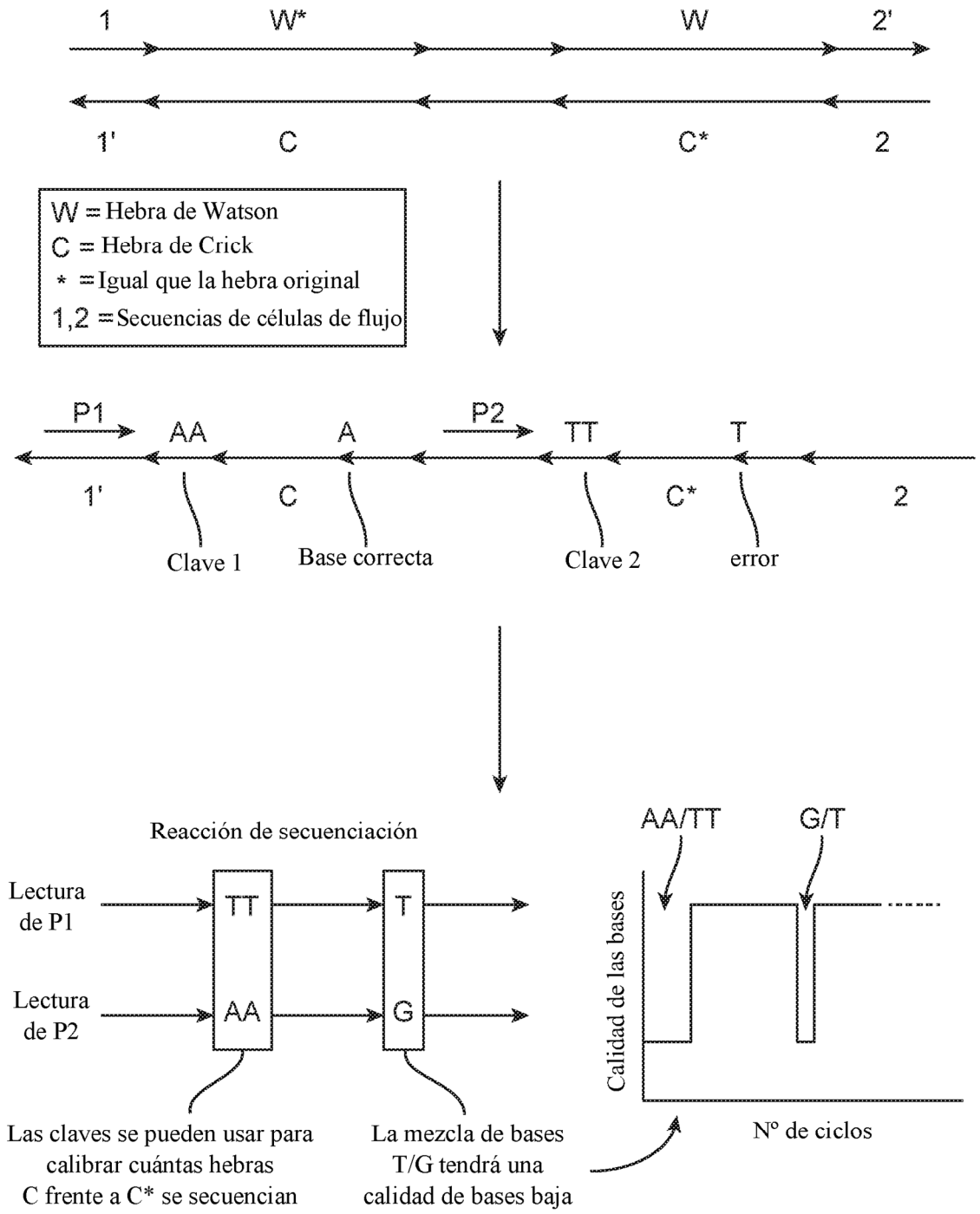
1
W*
W
2'
1'
C
C*
2
W = Hebra de Watson
C = Hebra de Crick
* = Igual que la hebra original
1,2 = Secuencias de células de flujo
P1
AA
A
P2
TT
T
1'
C
C*
2
Clave 1
Base correcta
Clave 2
error
Reacción de secuenciación
Lectura de P1
Lectura de P2
TT
AA
T
G
Las claves se pueden usar para calibrar cuántas hebras C frente a C* se secuencian
La mezcla de bases T/G tendrá una calidad de bases baja
AA/TT
G/T
Calidad de las bases
Nº de ciclos

FIG. 3

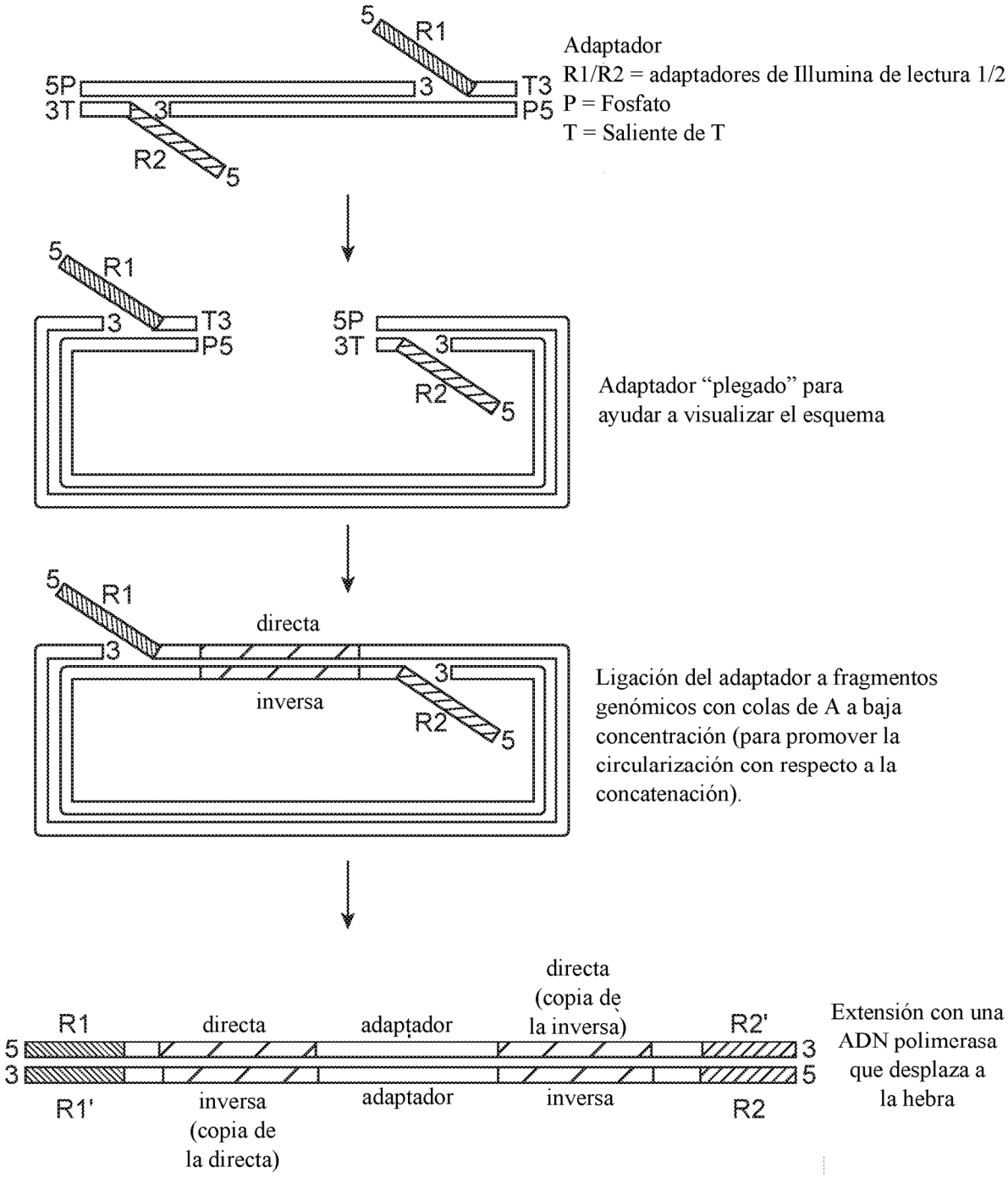
5
R1
5P
3
T3
3T
3
P5
R2
5
Adaptador
R1/R2 = adaptadores de Illumina de lectura 1/2
P = Fosfato
T = Saliente de T
Adaptador "plegado" para ayudar a visualizar el esquema
directa
inversa
Ligación del adaptador a fragmentos genómicos con colas de A a baja concentración (para promover la circularización con respecto a la concatenación).
directa (copia de la inversa)
adaptador
R2'
Extensión con una ADN polimerasa que desplaza a la hebra
R1'
inversa (copia de la directa)
adaptador
inversa
R2

FIG. 4

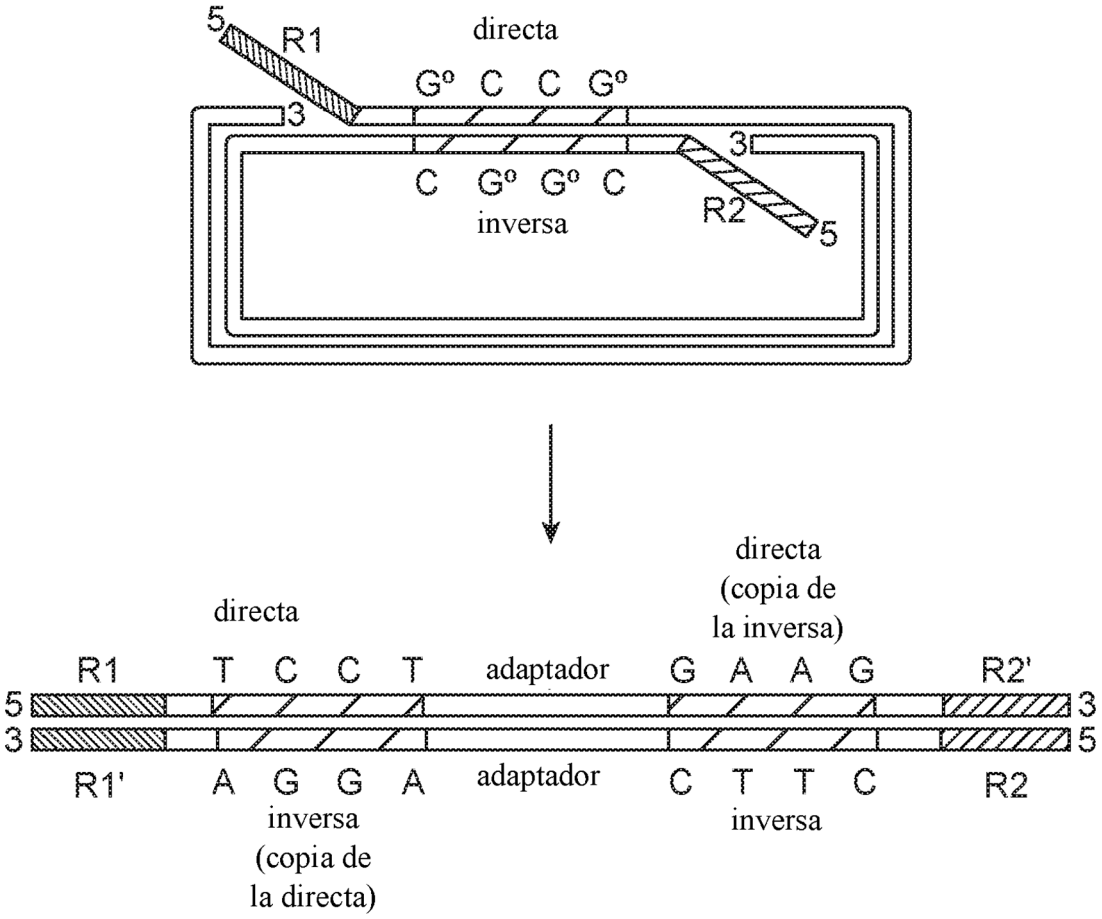
5
R1
directa
Gº C C Gº
3
3
C Gº Gº C
inversa
R2
5
directa
(copia de
la inversa)
directa
R1
T C C T
adaptador
G A A G
R2'
5
3
3
5
R1'
A G G A
adaptador
C T T C
R2
inversa
(copia de
la directa)
inversa

FIG. 5