(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 665 071**

(51) Int. CI.:

C12N 15/10 (2006.01)
C12N 15/70 (2006.01)
C12N 15/81 (2006.01)
C12N 15/85 (2006.01)
C12N 15/86 (2006.01)
C40B 40/08 (2006.01)
C40B 50/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **15.02.2013 PCT/US2013/026505**

(87) Fecha y número de publicación internacional: **22.08.2013 WO13123442**

(96) Fecha de presentación y número de la solicitud europea: **15.02.2013 E 13706397 (0)**

(97) Fecha y número de publicación de la concesión europea: **17.01.2018 EP 2814959**

(54) Título: **Composiciones y métodos para identificar mutaciones de manera precisa**

(30) Prioridad:

17.02.2012 US 201261600535 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.04.2018

(73) Titular/es:

FRED HUTCHINSON CANCER RESEARCH CENTER (100.0%)
1100 Fairview Avenue North
Seattle, WA 98109, US

(72) Inventor/es:

BIELAS, JASON, H. y
BERTOUT, JESSICA, A.

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Composiciones y métodos para identificar mutaciones de manera precisa

Antecedentes

1. Campo técnico

La presente divulgación se refiere a composiciones y métodos para detectar con precisión mutaciones utilizando la secuenciación y, más particularmente, marcando de manera única moléculas de ácidos nucleicos de doble cadena de manera que los datos de secuencia obtenidos para una cadena sentido pueden vincularse a datos de secuencia obtenidos a partir de la cadena anti-sentido cuando se obtiene a través de métodos de secuenciación masivamente paralelos.

2. Descripción de la Técnica Relacionada

La detección de mutaciones espontáneas (*p. ej.*, sustituciones, inserciones, deleciones, duplicaciones) o incluso mutaciones inducidas, que ocurren aleatoriamente a lo largo de un genoma puede ser un desafío, debido a que estos eventos mutacionales son raros y pueden existir en una o sólo en unas pocas copias de ADN. La forma más directa de detectar mutaciones es por secuenciación, pero los métodos de secuenciación disponibles no son lo suficientemente sensibles como para detectar mutaciones raras. Por ejemplo, las mutaciones que surgen *de novo* en el ADN mitocondrial (ADNmt) generalmente sólo estarán presentes en una sola copia de ADNmt, lo que significa que estas mutaciones no se encuentran fácilmente, ya que una mutación debe estar presente hasta en un 10-25% de una población de moléculas a detectar por secuenciación (Jones *et al.*, *Proc. Nat'l. Acad. Sci. U.S.A. 105*: 4283-88, 2008). Como otro ejemplo, se ha estimado que la frecuencia de mutación somática espontánea en ADN genómico es tan baja como 1×10^{-8} y $2{,}1 \times 10^{-6}$ en tejidos humanos normales y cancerosos, respectivamente (Bielas *et al.*, *Proc. Nat'l. Acad. Sci. U.S.A. 103*: 18238-42, 2008).

Una mejora en la secuenciación ha sido tomar moléculas de ADN individuales y amplificar el número de cada una de las moléculas mediante, por ejemplo, la reacción en cadena de la polimerasa (PCR) y PCR digital. De hecho, la secuenciación paralela masiva representa una forma particularmente potente de PCR digital porque se pueden analizar una por una millones de moléculas de ADN molde. Sin embargo, la amplificación de moléculas de ADN individuales antes o durante la secuenciación por PCR y/o amplificación puente padece la tasa de errores inherente de las polimerasas empleadas para la amplificación, y las mutaciones espurias generadas durante la amplificación pueden identificarse erróneamente como mutaciones espontáneas del ácido nucleico original (endógeno no amplificado). De manera similar, los moldes de ADN dañados durante la preparación (*ex vivo*) pueden amplificarse y puntuarse incorrectamente como mutaciones mediante técnicas de secuenciación paralela masiva. Nuevamente, utilizando ADNmt como un ejemplo, las frecuencias de mutación determinadas experimentalmente dependen fuertemente de la precisión del ensayo particular que se esté utilizando (Kraytsberg *et al.*, *Methods 46*:269-73, 2008) - estas discrepancias sugieren que la frecuencia de mutación espontánea del ADNmt está ya sea por debajo o muy cerca del límite de detección de estas tecnologías. La secuenciación paralela masiva generalmente no se puede utilizar para detectar variantes raras debido a la alta tasa de errores asociada con el proceso de secuenciación - un proceso que utiliza amplificación de puente y secuenciación por síntesis ha mostrado una tasa de errores que oscila entre aproximadamente 0,06% y 1% que depende de diversos factores, incluida la longitud de lectura. algoritmos de llamada de bases y el tipo de variantes detectadas (véase Kinde *et al., Proc. Nat'l. Acad. Sci. U.S.A. 108*:9530-5, 2011).

El documento DE 10 2008 025656 A1 describe un método para la determinación cuantitativa de ácidos nucleicos en una muestra, en particular para la determinación cuantitativa de transcritos de genes tales como ARNm, ADNc, ARNmicro, ARN no codificante, y para la provisión de marcadores para llevar a cabo los métodos analíticos.

El documento WO 2009/036525 A2 describe herramientas y métodos para uso en ensayos genéticos que indican técnicas de secuenciación de alto rendimiento, incluyendo un método de PCR multiplex, en el que los cebadores respectivos para amplificar los diferentes amplicones son físicamente aislados uno de otro.

El documento WO 2005/042759 A2 describe composiciones y métodos para el análisis de la expresión génica utilizando tecnología basada en micromatrices. Los métodos utilizan cebadores específicos para genes así como cebadores de amplificación universales durante la preparación de la muestra. Algunas realizaciones de la invención incorporan secuencias de códigos de barras en los productos amplificados.

El documento WO 98/44151 A1 describe un método para la amplificación de ácidos nucleicos para proporcionar moléculas de ácidos nucleicos amplificadas e inmovilizadas para usos tales como secuenciación, rastreo, diagnosis, síntesis de ácidos nucleicos in situ, vigilancia de la expresión génica y huella genética de ácidos nucleicos.

Breve sumario

La invención proporciona un método para detectar una mutación verdadera en una molécula de ácido nucleico, que comprende:

amplificar un banco de ácidos nucleicos de doble cadena, en donde el banco de ácidos nucleicos de doble cadena comprende una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos de doble cadena, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-Y-X^b (en orden 5' a 3'), en donde:

(a) X^a comprende un primer código;

(b) Y comprende una molécula de ácido nucleico diana, y

(c) X^b comprende un segundo código,

en donde cada una de la pluralidad de moléculas de ácidos nucleicos diana está asociada con un par único de primero y segundo códigos de doble cadena, en donde cada uno de la pluralidad de códigos comprende una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos, en donde se amplifican cada una de las cadenas de la pluralidad de moléculas de ácidos nucleicos diana y de la pluralidad de códigos de doble cadena;

secuenciar cada una de las cadenas amplificadas de la pluralidad de moléculas de ácidos nucleicos diana y de la pluralidad de códigos para obtener lecturas de secuenciación para la pluralidad de moléculas de ácidos nucleicos diana y la pluralidad de códigos, y de sus complementos inversos;

agrupar las lecturas de secuenciación de moléculas de ácidos nucleicos que comprenden pares de códigos idénticos en familias de lecturas de secuenciación, y

detectar la mutación verdadera a lo largo de una tasa de fondo de mutaciones de artefactos, comprendiendo dicha detección identificar como mutación verdadera una mutación presente sustancialmente en todas las lecturas en una familia de lecturas de secuenciación.

En una realización del método, las familias de lecturas de secuenciación comprenden los complementos inversos de las lecturas de secuenciación de moléculas de ácidos nucleicos diana que comprenden pares de códigos idénticos.

En otra realización, el método comprende generar secuencias consenso para las familias de lecturas de secuenciación. En una realización adicional, la generación de la secuencia consenso comprende eliminar por computación mutaciones que surgen durante la preparación del banco o durante la secuenciación.

En otra realización, la secuencia de doble cadena del código X^a para cada una de las moléculas de ácidos nucleicos diana es diferente de la secuencia de doble cadena del código X^b. En una realización adicional, ninguna de las secuencias de doble cadena del código X^a es la misma que la secuencia de doble cadena de cualquier otro código X^a, ninguna de las secuencias de doble cadena del código X^b es la misma que la secuencia de doble cadena de cualquier otro código X^b, y ninguna de las secuencias de doble cadena del código X^a y del código X^b son las mismas.

En otra realización, la secuencia de doble cadena del código X^a es idéntica a la secuencia de doble cadena del código X^b para una o más moléculas de ácidos nucleicos diana, con la condición de que el código de doble cadena para cada uno de los ácidos nucleicos diana sea diferente.

En otra realización, la detección de la mutación verdadera comprende secuenciar la pluralidad de moléculas de ácidos nucleicos con una tasa de errores que oscila entre aproximadamente 10^{-6} y aproximadamente 10^{-8}.

En otra realización, la detección de la mutación verdadera comprende secuenciar simultáneamente una pluralidad de moléculas de ácidos nucleicos diana diferentes con una tasa de errores de 5 x 10^{-6} o menos, 10^{-6} o menos, 5 x 10^{-7} o menos, 10^{-7} o menos, 5 x 10^{-8} o menos o 10^{-8} o menos.

En otra realización, la detección de la mutación verdadera comprende secuenciar una molécula de ácido nucleico diana sencilla en profundidad con una tasa de errores de 5 x 10^{-7} o menos, 10^{-7} o menos, 5 x 10^{-8} o menos o 10^{-8} o menos.

En otra realización, el primer y/o segundo códigos son códigos al azar. En otra realización, el primer y/o segundo códigos son códigos catalogados. En una realización adicional, el primer y/o segundo códigos son códigos al azar catalogados.

En otra realización, las lecturas de secuenciación no cubren la secuencia completa de una molécula de ácido nucleico diana de doble cadena. En una realización adicional, el método comprende enlazar lecturas de secuenciación obtenidas de un extremo de la molécula diana de doble cadena con lecturas de secuenciación obtenidas del extremo opuesto o de la segunda cadena de la misma molécula diana de doble cadena.

En otra realización, la pluralidad de moléculas de ácidos nucleicos diana comprende una molécula de ácido nucleico diana derivada de una célula tumoral circulante (CTC), un ADN mitocondrial de tumor circulante (ADNctmt), o un ADN viral.

En otra realización, la pluralidad de códigos tiene cada uno el mismo número de nucleótidos y comprende una longitud de aproximadamente 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 nucleótidos.

En otra realización, la molécula de ácido nucleico diana de Y comprende de aproximadamente 10 nucleótidos a aproximadamente 10.000 nucleótidos o de aproximadamente 100 nucleótidos a aproximadamente 1.000 nucleótidos.

En otra realización, la amplificación es mediante amplificación de puente, amplificación en emulsión, amplificación por nanoesferas o amplificación por PCR.

En otra realización, la secuenciación es secuenciación por síntesis, pirosecuenciación, secuenciación de colorante-terminador reversible o secuenciación de polonias.

La presente divulgación proporciona un banco de moléculas de ácidos nucleicos de doble cadena que incluye una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-X^b-Y, X^b-X^a-Y, Y-X^a-X^b, Y-X^b-X^a, X^a-Y-X^b o X^b-Y-X^a (en orden de 5' a 3'), en donde (a) X^a comprende un primer código al azar, (b) Y comprende una molécula de ácido nucleico diana y (c) X^b comprende un segundo código al azar. Además, cada uno de la pluralidad de códigos al azar comprende una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos (o aproximadamente 5 nucleótidos a aproximadamente 10 nucleótidos, o una longitud de aproximadamente 6, aproximadamente 7, aproximadamente 8, aproximadamente 9, aproximadamente 10, aproximadamente 11, aproximadamente 12, aproximadamente 13, aproximadamente 14, aproximadamente 15, aproximadamente 16, aproximadamente 17, aproximadamente 18, aproximadamente 19 o aproximadamente 20 nucleótidos).

En determinadas realizaciones, las secuencias de doble cadena de los códigos X^a y X^b son las mismas (*p. ej.*, $X^a = X^b$) para una o más moléculas de ácidos nucleicos diana, con la condición de que cada una de las moléculas de ácidos nucleicos diana no tenga la misma secuencia de código de doble cadena que cualquier otra molécula de ácido nucleico diana de este tipo. En determinadas otras realizaciones, la secuencia de doble cadena del código X^a para cada una de las moléculas de ácidos nucleicos diana es diferente de la secuencia de doble cadena del código X^b. En realizaciones adicionales, el banco de ácidos nucleicos de doble cadena está contenido en un vector autorreplicante tal como un plásmido, cósmido, YAC o vector viral.

La invención proporciona un método para obtener una secuencia de ácidos nucleicos o detectar con precisión una mutación verdadera en una molécula de ácido nucleico amplificando cada una de las cadenas del banco de ácidos nucleicos de doble cadena antes mencionado en el que se amplifican una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos, y se secuencia cada una de las cadenas de la pluralidad de moléculas de ácidos nucleicos diana y la pluralidad de códigos al azar. La secuenciación se puede realizar utilizando métodos de secuenciación masivamente paralelos. En determinados casos, la secuencia de una cadena de una molécula de ácido nucleico diana asociada con el primer código al azar alineada con la secuencia de la cadena complementaria asociada con el segundo código al azar da como resultado una tasa de errores de secuenciación mensurable que oscila entre aproximadamente 10^{-6} y aproximadamente 10^{-8}.

Breve descripción de los dibujos

La Figura 1 es una ilustración diagramática de un vector ilustrativo de la presente divulgación útil para generar un banco de ácidos nucleicos de doble cadena.

La Figura 2 es una ilustración diagramática de un vector ilustrativo de la presente divulgación, en el que se incluyen secuencias de adaptadores y son útiles, por ejemplo, para métodos de amplificación de puente antes de la secuenciación.

Las Figuras 3A y 3B muestran características de un banco de códigos y la detección de mutaciones verdaderas. (A) Los datos generados en una sola secuencia de nueva generación ejecutada en MiSeq® demuestran una amplia cobertura y diversidad en el código de siete pares de bases en un banco de vectores, en donde el vector utilizado se ilustra en la Figura 2. (B) Cypher Seq elimina errores introducidos durante la preparación y secuenciación del banco. Las moléculas de ácidos nucleicos diana se ligaron en un banco de vectores de códigos que contiene códigos de doble cadena, dobles previamente catalogados. Las secuencias diana se amplificaron y secuenciaron. Todas las lecturas de secuenciación que tienen pares de códigos idénticos, junto con sus complementos inversos, se agruparon en familias. La comparación de secuencias de familias permitió la generación de una secuencia consenso en la que 'mutaciones' (errores) que surgen durante la preparación del banco (círculo en blanco) y durante la secuenciación (círculo gris y triángulo) fueron eliminados por computación. En general, las mutaciones que están presentes en todas o casi todas las lecturas (diamante negro) del mismo código y su complemento inverso se cuentan como verdaderas mutaciones.

Las Figuras 4A y 4B muestran que el sistema de códigos puede distinguir mutaciones verdaderas de mutaciones de artefactos. (A) Se ligó el exón 4 TP53 de tipo salvaje en un banco de vectores Cypher Seq y se secuenció en el instrumento Illumina MiSeq® con una profundidad de más de un millón. Las secuencias se compararon luego con la

secuencia de TP53 de tipo salvaje. Las sustituciones detectadas se trazaron antes (A) y después de la corrección (B) con Cypher Seq.

Descripción detallada

La presente divulgación proporciona un banco de ácidos nucleicos de doble cadena en el que moléculas de ácidos nucleicos diana incluyen dobles códigos (es decir, códigos de barras o etiquetas identificadoras de origen), una en cada extremo (igual o diferente), de modo que la secuenciación de cada una de las cadenas complementarias puede estar conectada o enlazada a la molécula original. El código único en cada una de las cadenas enlaza cada una de las cadenas con su cadena complementaria original (*p. ej.*, antes de cualquier amplificación), de modo que cada una de las secuencias emparejadas sirve como su propio control interno. En otras palabras, etiquetando de manera única moléculas de ácidos nucleicos de doble cadena, los datos de secuencia obtenidos de una cadena de una única molécula de ácido nucleico pueden enlazarse específicamente a datos de secuencia obtenidos de la cadena complementaria de esa misma molécula de ácido nucleico de doble cadena. Además, los datos de secuencia obtenidos de un extremo de una molécula de ácido nucleico diana de doble cadena se pueden enlazar específicamente a los datos de secuencia obtenidos del extremo opuesto de esa misma molécula de ácido nucleico diana de doble cadena (por ejemplo, si no es posible obtener datos de la secuencia a través del fragmento de la molécula de ácido nucleico entera del banco).

Las composiciones y los métodos de esta divulgación permiten a una persona de experiencia ordinaria en la técnica distinguir con mayor precisión mutaciones verdaderas (es decir, mutaciones *in vivo* que surgen de forma natural) de una molécula de ácido nucleico a partir de "mutaciones" de artefactos (es decir, mutaciones o errores *ex vivo*) de una molécula de ácido nucleico que puede surgir por diversas razones, tales como un error de amplificación aguas abajo, un error de secuenciación o daño físico o químico. Por ejemplo, si en la molécula de ácido nucleico de doble cadena original pre-existía una mutación antes del aislamiento, la amplificación o la secuenciación, entonces una mutación de transición de adenina (A) a guanina (G) identificada en una cadena se complementará con una transición de timina (T) a cisteína (C) en la otra cadena. Por el contrario, es extremadamente improbable que "mutaciones" de artefactos que surgen más tarde en una cadena de ADN individual (separada) debido a errores de la polimerasa durante el aislamiento, la amplificación o la secuenciación tengan un cambio de base coincidente en la cadena complementaria. El enfoque de esta divulgación proporciona composiciones y métodos para distinguir errores sistemáticos (*p. ej.*, errores de fidelidad de lectura de la polimerasa) y errores biológicos (*p. ej.*, daño químico u otro) de mutaciones verdaderas o polimorfismos de un solo nucleótido (SNP) conocidos o recientemente identificados.

En determinadas realizaciones, los dos códigos en cada una de las moléculas diana tienen secuencias que son distintas entre sí y, por lo tanto, proporcionan un único par de identificadores, en los que un código identifica (o está asociado con) un primer extremo de una molécula de ácido nucleico diana y el segundo código identifica (o está asociado con) el otro extremo de la molécula de ácido nucleico diana. En determinadas otras realizaciones, los dos códigos en cada una de las moléculas diana tienen la misma secuencia y, por lo tanto, proporcionan un identificador único para cada una de las cadenas de la molécula de ácido nucleico diana. Cada una de las cadenas del banco de ácidos nucleicos de doble cadena (*p. ej.*, ADN genómico, ADNc) puede amplificarse y secuenciarse utilizando, por ejemplo, tecnologías de secuenciación de próxima generación (tales como PCR en emulsión o amplificación de puente combinada con pirosecuenciación o secuenciación por síntesis, o similares). La información de la secuencia de cada una de las cadenas complementarias de una primera molécula de ácido nucleico de doble cadena se puede enlazar y comparar (p. ej., computacionalmente "decontextua") debido a los códigos únicos asociados con cada uno de los extremos o una de las cadenas de esa molécula de ácido nucleico de doble cadena particular. En otras palabras, cada uno de los fragmentos de molécula de ácido nucleico de doble cadena original encontrado en un banco de moléculas se puede reconstruir individualmente debido a la presencia de un código de barras único asociado o un par de secuencias de código de barras (etiqueta identificadora) en cada uno de los fragmentos o cada una de las cadenas diana.

A modo de antecedentes, cualquier mutación espontánea o inducida estará presente en ambas cadenas de una molécula de ADN nativa, de doble cadena, genómica. Por lo tanto, dicho molde de ADN mutante amplificado utilizando la PCR dará como resultado un producto de la PCR en el que el 100% de las moléculas producidas por PCR incluyen la mutación. En contraposición con una mutación espontánea original, un cambio debido al error de la polimerasa sólo aparecerá en una cadena de la molécula de ADN molde inicial (mientras que la otra cadena no tendrá la mutación de artefacto). Si todas las cadenas de ADN en una reacción PCR se copian de manera igualmente eficaz, entonces cualquier error de polimerasa que surja del primer ciclo de PCR probablemente se encontrará en al menos el 25% del producto de la PCR total. Pero las moléculas o cadenas de ADN no se copian eficazmente de la misma manera, por lo que las secuencias de ADN amplificadas a partir de la cadena que incorporó una base de nucleótidos errónea durante la amplificación inicial podrían constituir más o menos del 25% de la población de secuencias de ADN amplificadas dependiendo de la eficacia de la amplificación, pero aún mucho menos del 100%. De forma similar, cualquier error de polimerasa que se produzca en ciclos posteriores de PCR generalmente representará una proporción incluso menor de productos de la PCR (es decir, 12,5% para el segundo ciclo, 6,25% para el tercero, etc.) que contiene una "mutación". Las mutaciones inducidas por la PCR pueden deberse a errores de la polimerasa o debido a que la polimerasa pasa por alto los nucleótidos dañados, dando con

ello como resultado un error (*véase*, *p. ej.*, Bielas y Loeb, *Nat. Methods 2*: 285-90, 2005). Por ejemplo, un cambio común en el ADN es la desaminación de la citosina, que es reconocida por Taq polimerasa como un uracilo y resulta en una mutación por transición de citosina a timina (Zheng *et al., Mutat. Res. 599*:11-20, 2006) - es decir, una alteración en la secuencia de ADN original se puede detectar cuando el ADN dañado es secuenciado, pero un cambio de este tipo puede o no ser reconocido como un error de la reacción de secuenciación o debido a un daño que surge *ex vivo* (*p. ej.*, durante o después del aislamiento de ácido nucleico).

Debido a artefactos y alteraciones potenciales de las moléculas de ácidos nucleicos que surgen del aislamiento, la amplificación y la secuenciación, la identificación precisa de mutaciones del ADN somático verdadero es difícil cuando se secuencian moléculas de ácidos nucleicos amplificadas. En consecuencia, se confunde la evaluación de si determinadas mutaciones están relacionadas con, o son un biomarcador para diversos estados patológicos (*p. ej.*, cáncer) o envejecimiento.

La secuenciación de próxima generación ha abierto la puerta a la secuenciación de múltiples copias de una molécula amplificada de ácido nucleico única - a la que se alude como secuenciación profunda. El pensamiento sobre la secuenciación profunda es que si un nucleótido particular de una molécula de ácido nucleico se secuencia múltiples veces, entonces uno puede identificar más fácilmente variantes o mutaciones de secuencias raras. De hecho, sin embargo, el proceso de amplificación y secuenciación tiene una tasa de errores inherente (que puede variar dependiendo de la calidad del ADN, de la pureza, la concentración (*p. ej.*, densidad del racimo) u otras condiciones), por lo que no importa cuántas veces se secuencia la molécula de ácido, una persona experta en la técnica no puede distinguir un artefacto de error de polimerasa de una mutación verdadera (especialmente mutaciones raras).

Si bien la secuenciación de muchas moléculas de ADN diferentes colectivamente es ventajosa en términos de costo y tiempo, el precio de esta eficiencia y conveniencia es que diversos errores de PCR complican el análisis de mutaciones, siempre que su frecuencia sea equiparable a la de las mutaciones que surgen *in vivo* - en otras palabras, las mutaciones genuinas *in vivo* serán esencialmente indistinguibles de los cambios que son artefactos de PCR o errores de secuenciación.

Por lo tanto, la presente invención proporciona métodos para identificar mutaciones presentes antes de la amplificación o secuenciación de un banco de ácidos nucleicos de doble cadena, en donde las moléculas diana incluyen un único código de doble cadena o códigos duales (es decir, códigos de barras o etiquetas identificadoras), una en cada uno de los extremos, de modo que la secuenciación de cada una de las cadenas complementarias se puede conectar de nuevo a la molécula original. En determinadas realizaciones, el método potencia la sensibilidad del método de secuenciación, de modo que la tasa de errores es 5×10^{-6}, 10^{-6}, 5×10^{-7}, 10^{-7}, 5×10^{-8}, 10^{8} o menos cuando se secuencian muchas moléculas diferentes de ácido nucleico diana simultáneamente o de manera que la tasa de errores es de 5×10^{-7}, 10^{-7}, 5×10^{-8}, 10^{-8} o menos cuando se secuencia una única molécula de ácido nucleico diana en profundidad.

Antes de exponer esta descripción con más detalle, puede ser útil para una comprensión de la misma proporcionar definiciones de determinados términos y expresiones a utilizar en esta memoria. Se recogen definiciones adicionales a lo largo de esta divulgación.

En la presente descripción, se debe entender que cualquier intervalo de concentraciones, intervalo porcentual, intervalo de relaciones o intervalo de números enteros incluye el valor de cualquier número entero dentro del intervalo indicado y, cuando sea apropiado, fracciones del mismo (tal como una décima y una centésima de un número entero), a menos que se indique lo contrario. Además, debe entenderse que cualquier intervalo de números enumerado aquí en relación con cualquier característica física, tal como subunidades de polímero, tamaño o grosor, ha de entenderse que incluye cualquier número entero dentro del intervalo indicado, a menos que se indique lo contrario. Tal como se utiliza en esta memoria, el término "aproximadamente" y la expresión "consiste esencialmente en" significa ± 20% del intervalo, valor o estructura indicados, a menos que se indique lo contrario. Debe entenderse que los términos "un" y "una" tal como se utilizan en esta memoria se refieren a "uno o más" de los componentes enumerados. El uso de la alternativa (*p. ej.*, "o") debe entenderse como una, ambas o cualquier combinación de las alternativas. Tal como se utiliza en esta memoria, los términos "incluir", "tener" y "comprender" se utilizan de manera sinónima, términos y variantes que se pretende sean considerados no limitantes.

Tal como se utiliza en esta memoria, la expresión "código al azar" o el término "código" o la expresión "código de barras" o "etiqueta identificadora" y sus variantes se utilizan indistintamente y se refieren a una molécula de ácido nucleico que tiene una longitud que oscila entre aproximadamente 5 y aproximadamente 50 nucleótidos. En determinadas realizaciones, todos los nucleótidos del código no son idénticos (es decir, comprenden al menos dos nucleótidos diferentes) y opcionalmente no contienen tres nucleótidos contiguos que sean idénticos. En realizaciones adicionales, el código está comprendido entre aproximadamente 5 y aproximadamente 15 nucleótidos, entre aproximadamente 6 y aproximadamente 10 nucleótidos, y preferiblemente entre aproximadamente 7 y aproximadamente 12 nucleótidos. Generalmente, los códigos estarán ubicados en uno o en ambos extremos de una molécula diana, que se puede incorporar directamente en moléculas diana de interés o en un vector en el que posteriormente se añadirán las moléculas diana.

Tal como se utiliza en esta memoria, "moléculas de ácidos nucleicos diana " y variantes de las mismas se refieren a una pluralidad de moléculas de ácidos nucleicos de doble cadena que pueden ser fragmentos o moléculas más cortas generadas a partir de moléculas de ácidos nucleicos más largas, incluyendo las de muestras naturales (*p. ej.*, un genoma), o las moléculas de ácido nucleico diana pueden ser sintéticas (*p. ej.*, ADNc), recombinantes, o una combinación de las mismas. Fragmentos de ácidos nucleicos diana de moléculas más largas se pueden generar utilizando una diversidad de técnicas conocidas en la técnica tales como cizallamiento mecánico o escisión específica con endonucleasas de restricción.

Tal como se utiliza en esta memoria, un "banco de moléculas de ácidos nucleicos" y variantes de las mismas se refiere a un banco de moléculas o fragmentos de ácidos nucleicos. En determinadas realizaciones, el banco de moléculas o fragmentos de ácidos nucleicos se incorpora en un vector, que se puede transformar o transfectar en una célula huésped apropiada. Las moléculas de ácidos nucleicos diana de esta divulgación se pueden introducir en una diversidad de estructuras principales de vectores diferentes (tales como plásmidos, cósmidos, vectores virales o similares), de modo que la producción recombinante de un banco de moléculas de ácidos nucleicos puede mantenerse en una célula huésped de elección (tal como bacterias, levaduras, células de mamíferos o similares).

Por ejemplo, un banco de moléculas de ácidos nucleicos que representa el genoma completo se denomina una genoteca y a un banco de copias de ADN de ARN mensajero se la alude como un banco de ADN complementario (ADNc). Métodos para introducir bancos de moléculas de ácidos nucleicos en vectores son bien conocidos en la técnica (*véase, p. ej.*, Current Protocols in Molecular Biology, Ausubel *et al.*, Comps., Greene Publishing and Wiley-Interscience, Nueva York, 1995; Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, 2ª Ed., Cold Spring Harbor Laboratory Vols. 1-3, 1989; Methods in Enzymology, Vol. 152, Guide to Molecular Cloning Techniques, Berger y Kimmel, Comps., San Diego: Academic Press , Inc., 1987).

Dependiendo del tipo de banco a generar, los extremos de los fragmentos de ácido nucleico diana pueden tener colgantes o pueden estar "pulidos" (es decir, romos). Juntos, los fragmentos de moléculas de ácidos nucleicos diana pueden, por ejemplo, clonarse directamente en un vector de códigos para generar un banco de vectores, o pueden ligarse con adaptadores para generar, por ejemplo, polonias. Las moléculas de ácidos nucleicos diana, que son las moléculas de ácidos nucleicos de interés para la amplificación y secuenciación, pueden variar en tamaño desde unos pocos nucleótidos (*p. ej.*, 50) hasta muchos miles (*p. ej.*, 10.000). Preferiblemente, los fragmentos diana en el banco varían en tamaño desde aproximadamente 100 nucleótidos a aproximadamente 750 nucleótidos o aproximadamente 1.000 nucleótidos, o desde aproximadamente 150 nucleótidos a aproximadamente 250 nucleótidos o aproximadamente 500 nucleótidos.

Tal como se utiliza en esta memoria, un "sitio de cebado de moléculas de ácidos nucleicos" o “PS” y variantes del mismo son secuencias cortas de ácidos nucleicos conocidas contenidas en el vector. Una secuencia PS puede variar en longitud desde 5 nucleótidos a aproximadamente 50 nucleótidos de longitud, desde aproximadamente 10 nucleótidos a aproximadamente 30 nucleótidos, y preferiblemente es desde aproximadamente 15 nucleótidos a aproximadamente 20 nucleótidos de longitud. En determinadas realizaciones, una secuencia PS puede incluirse en uno o ambos extremos o ser una parte integral de las moléculas de ácido nucleico de códigos al azar, o puede estar incluida en uno o ambos extremos o ser una parte integral de una secuencia de adaptador, o puede estar incluida como parte del vector. Un cebador de moléculas de ácidos nucleicos que es complementario a un PS incluido en un banco de la presente divulgación se puede utilizar para iniciar una reacción de secuenciación.

Por ejemplo, si un código al azar solo tiene un PS aguas arriba (5') del código, entonces se puede utilizar un cebador complementario al PS para cebar una reacción de secuenciación para obtener la secuencia del código al azar y algo de una secuencia de una molécula de ácido nucleico diana clonada aguas abajo de la cifra. En otro ejemplo, si un código al azar tiene un primer PS aguas arriba (5') y un segundo PS aguas abajo (3') del código, entonces puede utilizarse un cebador complementario del primer PS para cebar una reacción de secuenciación para obtener la secuencia del código al azar, el segundo PS y algo de la secuencia de una molécula de ácido nucleico diana clonada aguas abajo del segundo PS. Por el contrario, puede utilizarse un cebador complementario al segundo PS para cebar una reacción de secuenciación para obtener directamente la secuencia de la molécula de ácido nucleico diana clonada aguas abajo del segundo PS. En este último caso se obtendrá más información de la secuencia de la molécula diana, ya que la reacción de secuenciación que comienza desde el segundo PS puede extenderse más allá en la molécula diana que lo que lo hace la reacción que tiene que extenderse a través tanto del código como de la molécula diana.

Tal como se utiliza en esta memoria, "secuenciación de nueva generación" se refiere a métodos de secuenciación de alto rendimiento que permiten la secuenciación de miles o millones de moléculas en paralelo. Ejemplos de métodos de secuenciación de nueva generación incluyen secuenciación por síntesis, secuenciación por ligamiento, secuenciación por hibridación, secuenciación de polonias y pirosecuenciación. Al unir cebadores a un sustrato sólido y una secuencia complementaria a una molécula de ácido nucleico, se puede hibridar una molécula de ácido nucleico al sustrato sólido a través del cebador y luego se pueden generar múltiples copias en una zona discreta sobre el sustrato sólido utilizando polimerasa para amplificar (a estas agrupaciones se las alude a veces como colonias de polimerasa o polonias). Por consiguiente, durante el proceso de secuenciación, un nucleótido en una posición particular se puede secuenciar múltiples veces (*p. ej.*, cientos o miles de veces) - a esta profundidad de cobertura se la alude como “secuenciación profunda”.

Tal como se utiliza en esta memoria, "llamada de bases" se refiere a la conversión en computación de datos en bruto o procesados de un instrumento de secuenciación en puntuaciones de calidad y luego secuencias reales. Por ejemplo, muchas de las plataformas de secuenciación utilizan cámaras de detección óptica y de dispositivo acoplado a carga (CCD) para generar imágenes de información de intensidad (es decir, la información de intensidad indica qué nucleótido está en qué posición de una molécula de ácido nucleico), por lo que la llamada de bases se refiere generalmente al análisis de imagen en computación que convierte los datos de intensidad en secuencias y puntuaciones de calidad. Otro ejemplo es la tecnología de secuenciación por torrente de iones, que emplea una tecnología en propiedad de detección de iones semiconductores para detectar la liberación de iones hidrógeno durante la incorporación de bases de nucleótidos en reacciones de secuenciación que tienen lugar en una matriz de alta densidad de pocillos micro-mecanizados. Existen otros ejemplos de métodos conocidos en la técnica que se pueden emplear para la secuenciación simultánea de grandes números de moléculas de nucleótidos. Se describen diversos métodos de llamada de bases en, por ejemplo, Niedringhaus *et al*. (*Anal. Chem. 83*:4327, 2011).

En la siguiente descripción, se establecen determinados detalles específicos con el fin de proporcionar una comprensión exhaustiva de diversas realizaciones de esta divulgación. Sin embargo, después de revisar esta divulgación, un experto en la técnica entenderá que la invención puede ponerse en práctica sin muchos de estos detalles. En otros casos, tecnologías de secuenciación de nueva generación emergentes, así como métodos de secuenciación de nueva generación bien conocidos o ampliamente disponibles (*p. ej.*, secuenciación de terminación de la cadena, secuenciación de colorante y terminador, secuenciación de colorante y terminador reversible, secuenciación por síntesis, secuenciación por ligamiento, secuenciación por hibridación, secuenciación de polonias, pirosecuenciación, secuenciación de semiconductores de iones, secuenciación de nanoesferas, secuenciación de nanoporos, secuenciación de molécula única, secuenciación FRET, secuenciación de bases pesada y secuenciación de microfluidos), no se han descrito todas en detalle para evitar oscurecer innecesariamente las descripciones de las realizaciones de la presente divulgación. Las descripciones de algunos de estos métodos pueden encontrarse, por ejemplo, en las Publicaciones PCT Nºs WO 98/44151, WO 00/18957 y WO 2006/08413; y las Publicaciones de Solicitudes de Patente de EE.UU. Nºs 6.143.496, 6833246 y 7.754.429; y las Publicaciones de Solicitudes de Patente de EE.UU. 2010/0227329 y US 2009/0099041.

Se describen diversas realizaciones de la presente divulgación para fines de ilustración en el contexto del uso con vectores que contienen un banco de fragmentos de ácidos nucleicos (*p. ej.*, banco genómico o de ADNc). Sin embargo, como apreciarán los expertos en la técnica al revisar esta divulgación, el uso con otros bancos de ácidos nucleicos o métodos para producir un banco de fragmentos de ácidos nucleicos también pueden ser adecuados.

En determinadas realizaciones, un banco de ácidos nucleicos de doble cadena comprende una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-Y-X^b (en orden de 5' a 3'), en donde (a) X^a comprende un primer código al azar, (b) Y comprende una molécula de ácido nucleico diana y (c) X^b comprende un segundo código al azar; en donde cada uno de la pluralidad de códigos al azar comprende una longitud de aproximadamente 5 nucleótidos a aproximadamente 50 nucleótidos. En determinadas realizaciones, la secuencia de doble cadena del código X^a para cada una de las moléculas de ácidos nucleicos diana es diferente de la secuencia de doble cadena del código X^b. En determinadas otras realizaciones, el código X^a de doble cadena es idéntico al código X^b para una o más moléculas de ácidos nucleicos diana, con la condición de que el código doble cadena para cada una de las moléculas de ácido nucleico diana sea diferente.

En realizaciones adicionales, la pluralidad o agrupación de códigos al azar utilizados en el banco de moléculas de ácidos nucleicos de doble cadena o el banco de vectores comprende de aproximadamente 5 nucleótidos a aproximadamente 40 nucleótidos, aproximadamente 5 nucleótidos a aproximadamente 30 nucleótidos, aproximadamente 6 nucleótidos a aproximadamente 30 nucleótidos, aproximadamente 6 nucleótidos a aproximadamente 20 nucleótidos, aproximadamente 6 nucleótidos a aproximadamente 10 nucleótidos, aproximadamente 6 nucleótidos a aproximadamente 8 nucleótidos, aproximadamente 7 nucleótidos a aproximadamente 9 o aproximadamente 10 nucleótidos, o aproximadamente 6, aproximadamente 7 o aproximadamente 8 nucleótidos. En determinadas realizaciones, un código tiene preferiblemente una longitud de aproximadamente 6, aproximadamente 7, aproximadamente 8, aproximadamente 9, aproximadamente 10, aproximadamente 11, aproximadamente 12, aproximadamente 13, aproximadamente 14, aproximadamente 15, aproximadamente 16, aproximadamente 17, aproximadamente 18, aproximadamente 19 o aproximadamente 20 nucleótidos. En determinadas realizaciones, un par de códigos al azar asociados con secuencias de ácidos nucleicos o vectores tendrán diferentes longitudes o tendrán la misma longitud. Por ejemplo, una molécula o vector de ácido nucleico diana puede tener un primer código al azar aguas arriba (5') de aproximadamente 6 nucleótidos de longitud y un segundo código aguas abajo (3') de aproximadamente 9 nucleótidos de longitud, o una molécula de ácido nucleico diana o un vector puede tener un primer código al azar aguas arriba (5') de aproximadamente 7 nucleótidos de longitud y un segundo código al azar aguas abajo (3') de aproximadamente 7 nucleótidos de longitud.

En determinadas realizaciones, tanto el código X^a como el código X^b comprenden cada uno 6 nucleótidos, 7 nucleótidos, 8 nucleótidos, 9 nucleótidos, 10 nucleótidos, 11 nucleótidos, 12 nucleótidos, 13 nucleótidos, 14 nucleótidos, 15 nucleótidos, 16 nucleótidos, 17 nucleótidos, 18 nucleótidos, 19 nucleótidos o 20 nucleótidos. En determinadas otras realizaciones, el código X^a comprende 6 nucleótidos y el código X^b comprende 7 nucleótidos u 8

nucleótidos; o el código X^a comprende 7 nucleótidos y el código X^b comprende 6 nucleótidos u 8 nucleótidos; o el código X^a comprende 8 nucleótidos y el código X^b comprende 6 nucleótidos o 7 nucleótidos; o el código X^a comprende 10 nucleótidos y el código X^b comprende 11 nucleótidos o 12 nucleótidos.

El número de nucleótidos contenidos en cada uno de los códigos al azar o códigos de barras regirá el número total de posibles códigos de barras disponibles para su uso en un banco. Los códigos de barras más cortos permiten un número menor de códigos únicos, que pueden ser útiles cuando se realiza una secuencia profunda de una o unas pocas secuencias de nucleótidos, mientras que códigos de barras más largos pueden ser deseables cuando se examina una población de moléculas de ácidos nucleicos tales como ADNcs o fragmentos genómicos. En determinadas realizaciones puede desearse la secuenciación múltiple cuando se fijan como objetivo moléculas de ácido nucleico específicas, regiones genómicas específicas, genomas más pequeños o un subconjunto de transcritos de ADNc. La secuenciación múltiplex implica amplificar dos o más muestras que se han agrupado en, por ejemplo, una sola pista de una célula de flujo para la amplificación de puente para aumentar exponencialmente el número de moléculas analizadas en una sola operación sin sacrificar tiempo o costo. En realizaciones relacionadas, se incluye una secuencia de índice único (que comprende una longitud que oscila entre aproximadamente 4 nucleótidos y aproximadamente 25 nucleótidos) específica para una muestra particular con cada uno de los bancos de vectores de doble código. Por ejemplo, si se combinan diez muestras diferentes en la preparación de la secuencia multiplex, entonces se utilizarán diez secuencias de índices diferentes de modo que se utilicen diez bancos de vectores de doble código en los que cada uno de los bancos tenga un único identificador de secuencia de índice único (pero cada uno de los banco tiene una pluralidad de códigos al azar).

Por ejemplo, un código de barras de 7 nucleótidos tendría una fórmula de 5'-NNNNNNN-3' (SEQ ID NO.:1), en donde N puede ser cualquier nucleótido que se produce de forma natural. Los cuatro nucleótidos que se producen de forma natural son A, T, C y G, por lo que el número total de posibles códigos al azar es 4^7, o 16,384 disposiciones al azar posibles (es decir, 16.384 códigos diferentes o únicos). Para códigos de barras de 6 y 8 nucleótidos, la cantidad de códigos al azar sería 4.096 y 65.536, respectivamente. En determinadas realizaciones de 6, 7 u 8 códigos de nucleótidos al azar, puede haber menos de la agrupación de 4.094, 16.384 o 65.536 códigos únicos, respectivamente, disponibles para su uso cuando se excluyen, por ejemplo, secuencias en las que todos los nucleótidos son idénticos (*p. ej.* , todas las A o todas las T o todas las C o todas las G) o cuando se excluyen las secuencias en las que tres nucleótidos contiguos son idénticos o cuando se excluyen ambos tipos de moléculas. Además, los primeros aproximadamente 5 nucleótidos a aproximadamente 20 nucleótidos de la secuencia de molécula de ácido nucleico diana se pueden utilizar como una etiqueta de identificador adicional junto con la secuencia de un código al azar asociado.

En aún otras realizaciones, un banco de ácidos nucleicos de doble cadena comprende una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-Y-X^b (en orden de 5' a 3'), en donde (a) X^a comprende un primer código al azar, (b) Y comprende una molécula de ácido nucleico diana y (c) X^b comprende un segundo código al azar, en donde cada uno de la pluralidad de códigos al azar tiene una longitud de aproximadamente 5 a aproximadamente 50 nucleótidos, y en donde (i) al menos dos de esos nucleótidos son diferentes en cada uno de los códigos o (ii) cada uno de los códigos no contiene tres nucleótidos contiguos que son idénticos. En determinadas realizaciones en las que cada uno de los códigos no contiene tres nucleótidos contiguos que son idénticos, el código X^a de doble cadena es idéntica al código X^b para una o más moléculas de ácidos nucleicos diana, con la condición de que el código de doble cadena para cada una de las moléculas de ácidos nucleicos diana es diferente.

En algunos casos, un banco de ácidos nucleicos de doble cadena comprende una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-X^b-Y, X^b-X^a-Y, Y-X^a-X^b, Y-X^b-X^a, X^a-Y, X^b-Y, Y-X^a o Y-X^b (en orden de 5' a 3'), en donde (a) X^a comprende un primer código al azar, (b) Y comprende una molécula de ácido nucleico diana y (c) X^b comprende un segundo código al azar, en donde cada uno de la pluralidad de códigos al azar tiene una longitud de aproximadamente 5 a aproximadamente 50 nucleótidos.

En cualquiera de las realizaciones descritas en esta memoria, un código X^a comprende, además, aproximadamente una secuencia de 5 nucleótidos a aproximadamente una secuencia de 20 nucleótidos de la molécula de ácido nucleico diana que está aguas abajo del código X^a, o un código X^b comprende, además, aproximadamente una secuencia de 5 nucleótidos a aproximadamente una secuencia de 20 nucleótidos de la molécula de ácido nucleico diana que está aguas arriba del código X^b, o un código X^a y un código X^b comprenden, además, aproximadamente una secuencia de 5 nucleótidos a aproximadamente una secuencia de 20 nucleótidos de la molécula de ácido nucleico diana que está aguas abajo o aguas arriba, respectivamente, de cada uno de los códigos.

En aún otras realizaciones adicionales, una primera molécula diana está asociada con y dispuesta entre un primer código al azar X^a y un segundo código al azar X^b, una segunda molécula diana está asociada con y dispuesta entre un tercer código al azar X^a y un cuarto código al azar X^b, y así sucesivamente, en donde las moléculas diana de un banco o de un banco de vectores tiene cada uno un código X^a único (es decir, ninguno de los códigos X^a tienen la misma secuencia) y cada uno tiene un código X^b único (es decir, ninguno de los códigos X^b tiene la misma secuencia), y en donde ninguno o solo una minoría de los códigos X^a y X^b tienen la misma secuencia.

Por ejemplo, si la longitud del código al azar es de 7 nucleótidos, entonces habrá un total de 16.384 códigos de barras diferentes disponibles como el primer código al azar X^a y el segundo código al azar X^b. En este caso, si una primera molécula de ácido nucleico diana se asocia y se dispone entre el código al azar X^a número 1 y el código al azar X^b número 2 y una segunda molécula de ácido nucleico diana está asociada y dispuesta entre el código al azar X^a número 16.383 y el código al azar X^b número 16.384, entonces una tercera molécula de ácido nucleico diana solo puede estar asociada y dispuesta entre cualquier par de números de códigos seleccionados de los números 3 a 16.382, y así sucesivamente para cada una de las moléculas de ácido nucleico diana de un banco hasta que se haya utilizado cada uno de los diferentes códigos al azar (que pueden ser o no todos los 16.382). En esta realización, cada una de las moléculas de ácido nucleico diana de un banco tendrá un par único de códigos que difieren de cada uno de los otros pares de códigos encontrados asociados con cada una de las otras moléculas de ácidos nucleicos diana del banco.

En cualquiera de las realizaciones descritas en esta memoria las secuencias del código al azar de una agrupación particular de códigos (p. ej., agrupaciones de 4.094, 16.384 o 65.536 códigos únicos) pueden utilizarse más de una vez. En realizaciones adicionales, cada una de las moléculas de ácidos nucleicos diana o un subconjunto de moléculas diana tiene un par diferente (único) de códigos. Por ejemplo, si una primera molécula diana está asociada con y se dispone entre el número de código al azar 1 y el número de código al azar 100, entonces una segunda molécula diana necesita estar flanqueada por un par doble diferente de códigos - tal como el número de código al azar 1 y el número de código al azar 65, o el número de código al azar 486 y el número de código al azar 100 - que puede ser cualquier combinación distinta de 1 y 100. En determinadas otras realizaciones, cada una de las moléculas de ácidos nucleicos diana o un subconjunto de moléculas diana tiene códigos idénticos en cada uno de los extremos de una o más moléculas de ácidos nucleicos diana, con la condición de que el código de doble cadena para cada una de las moléculas de ácidos nucleicos diana sea diferente. Por ejemplo, si una primera molécula diana está flanqueada por el número de código 10, entonces una segunda molécula diana con códigos idénticos en cada uno de los extremos tendrá que tener un código diferente - tal como el número de código al azar 555 o similar - que puede ser cualquier otro código distinto de 10. En aún otras realizaciones, las moléculas del ácido nucleico diana del banco de moléculas de ácidos nucleicos tendrán cada una códigos duales únicos X^a y X^b, en donde ninguno de los códigos X^a tiene la misma secuencia que cualquier otro código X^a, ninguno de los códigos X^b tiene la misma secuencia que cualquier otro código X^b y ninguno de los códigos X^a tiene la misma secuencia que cualquier código X^b. En aún realizaciones adicionales, las moléculas de ácidos nucleicos diana del banco de moléculas de ácidos nucleicos tendrán cada una un par único de códigos X^a-X^b, en donde ninguno de los códigos X^a o X^b tiene la misma secuencia. Una mezcla de cualquiera de las realizaciones mencionadas anteriormente puede constituir un banco de moléculas de ácido nucleico de esta divulgación.

En cualquiera de las realizaciones descritas en esta memoria, la pluralidad de moléculas de ácidos nucleicos diana, que juntas se utilizan para generar un banco de moléculas de ácidos nucleicos (o se utilizan para la inserción en un vector para generar un banco de vectores que contiene una pluralidad de moléculas de ácidos nucleicos diana) puede tener cada una longitud que oscila entre aproximadamente 10 nucleótidos y aproximadamente 10.000 nucleótidos, entre aproximadamente 50 nucleótidos y aproximadamente 5.000 nucleótidos, entre aproximadamente 100 nucleótidos y aproximadamente 1.000 nucleótidos o entre aproximadamente 150 nucleótidos y aproximadamente 750 nucleótidos, o entre aproximadamente 250 nucleótidos y aproximadamente 500 nucleótidos.

En cualquiera de las realizaciones descritas en esta memoria, la pluralidad de códigos al azar puede estar unida a un primer sitio de cebado de moléculas de ácidos nucleicos (PS1), unida a un segundo sitio de cebado de moléculas de ácidos nucleicos (PS2) o unida tanto a un primer como a un segundo sitio de cebado de moléculas de ácidos nucleicos. En determinadas realizaciones, una pluralidad de códigos al azar se pueden asociar con y disponer entre un primer sitio de cebado de moléculas de ácidos nucleicos (PS1) y un segundo sitio de cebado de moléculas de ácidos nucleicos (PS2), en donde la secuencia de doble cadena de PS1 es diferente de la secuencia de doble cadena de PS2. En determinadas realizaciones, cada uno de los pares de códigos X^a-X^b puede estar asociado con y dispuesto entre un sitio de cebado de moléculas de ácidos nucleicos (PS1) aguas arriba y aguas abajo (*véase, p. ej.*, la Fig. 2).

En cualquiera de las realizaciones descritas en esta memoria, un primer sitio de cebado de moléculas de ácidos nucleicos PS1 estará situado aguas arriba (5') del primer código al azar X^a y el primer sitio de cebado de moléculas de ácidos nucleicos PS1 también se encontrará aguas abajo (3') del segundo código al azar X^b. En determinadas realizaciones, un cebador de oligonucleótido complementario a la cadena sentido de PS1 puede ser utilizado para cebar una reacción de secuenciación para obtener la secuencia de la cadena sentido del primer código al azar X^a o para cebar una reacción de secuenciación para obtener la secuencia de la cadena anti-sentido del segundo código al azar X^b, mientras que un cebador de oligonucleótidos complementario a la cadena anti-sentido de PS1 se puede utilizar para cebar una reacción de secuenciación para obtener la secuencia de la cadena anti-sentido del primer código al azar X^a o para cebar una reacción de secuenciación para obtener la secuencia de la cadena sentido del segundo código al azar X^b.

En cualquiera de las realizaciones descritas en esta memoria, el segundo sitio de cebado de moléculas de ácidos nucleicos PS2 estará situado aguas abajo (3') del primer código al azar X^a y el segundo sitio de cebado de moléculas de ácidos nucleicos PS2 también se encontrará aguas arriba (5') del segundo código al azar X^b. En

determinadas realizaciones, un cebador de oligonucleótido complementario a la cadena sentido de PS2 puede ser utilizado para cebar una reacción de secuenciación para obtener la secuencia de la cadena sentido desde el extremo 5' de la molécula de ácido nucleico diana de doble cadena asociada o para cebar una reacción de secuenciación para obtener la secuencia de la cadena anti-sentido desde el extremo 3' de la molécula de ácido nucleico diana de doble cadena asociada, mientras que un cebador de oligonucleótidos complementario a la cadena anti-sentido de PS2 se puede utilizar para cebar una reacción de secuenciación para obtener la secuencia de la cadena anti-sentido desde el extremo 5' de la molécula de ácido nucleico diana de doble cadena asociada o para cebar una reacción de secuenciación para obtener la secuencia de la cadena sentido desde el extremo 3' de la molécula de ácido nucleico diana de doble cadena asociada.

Dependiendo de la longitud de la molécula de ácido nucleico diana se puede obtener la secuencia completa de la molécula de ácido nucleico si es lo suficientemente corta, o solo se puede obtener una porción de la secuencia de molécula de ácido nucleico completa si es más larga que aproximadamente 100 nucleótidos a aproximadamente 250 nucleótidos. Una ventaja de las composiciones y los métodos de la presente divulgación es que, a pesar de que una molécula de ácido nucleico diana es demasiado larga para obtener datos de secuencia para la molécula o fragmento completo, los datos de secuencia obtenidos de un extremo de una molécula diana de doble cadena pueden ser específicamente vinculados a los datos de secuencia obtenidos a partir del extremo opuesto o de la segunda cadena de la misma molécula diana de doble cadena, ya que cada una de las moléculas diana en un banco de esta divulgación tendrá códigos de doble cadena, o un par único de códigos X^a-X^b. La vinculación de los datos de secuencia de las dos cadenas permite la identificación sensible de mutaciones "verdaderas", en las que la secuenciación más profunda aumenta en realidad la sensibilidad de la detección, y estos métodos pueden proporcionar datos suficientes para cuantificar el número de mutaciones de artefactos.

En cualquiera de las realizaciones descritas en esta memoria, una pluralidad de códigos al azar puede comprender, además, una primera secuencia de reconocimiento de endonucleasas de restricción (RE1) y una segunda secuencia de reconocimiento de endonucleasas de restricción (RE2), en donde la primera secuencia de reconocimiento de endonucleasas de restricción RE1 está localizada aguas arriba (5') del primer código al azar X^a y la segunda secuencia de reconocimiento de endonucleasas de restricción RE2 está localizada aguas abajo (3') del segundo código al azar X^b. En determinadas realizaciones, una primera secuencia de reconocimiento de endonucleasas de restricción RE1 y una segunda secuencia de reconocimiento de endonucleasas de restricción RE2 son iguales o diferentes. En determinadas realizaciones, RE1, RE2, o tanto RE1 como RE2 son endonucleasas de restricción "de corte raro" que tienen una secuencia de reconocimiento que ocurre solo raramente dentro de un genoma o dentro de una secuencia de moléculas de ácidos nucleicos diana o son "de corte romo" que generan moléculas de ácidos nucleicos con extremos romos después de la digestión (*p. ej.*, *Sma*I). Dichas enzimas de corte raro tienen generalmente sitios de reconocimiento más largos con siete u ocho nucleótidos o secuencias de reconocimiento más largas, tales como *Aar*I, *Abe*I, *Asc*I, *Asi*SI, *Bbv*CI, *Bst*RZ2461, *Bst*SWI, *Cci*NI, *Csi*BI, *Csp*BI, *Fse*I, *Not*I, *Mch*AI, *Msp*SWI, *Mss*I, *Pac*I, *Pme*I, *Sbf*I, *Sda*I, *Sgf*I, *Smi*I, *Srf*I, *Sse*232I, *Sse*8387I, *Swa*I, *Taq*II, *Vpa*K32I o similares,

En determinadas realizaciones, un banco de moléculas de ácidos nucleicos comprende moléculas de ácidos nucleicos que tienen una fórmula de 5'-RE1-PS1-X^a-PS2-Y-PS2-X^b-PS1-RE2-3', en donde RE1 es una primera secuencia de reconocimiento de endonucleasas de restricción, PS1 es un primer sitio de cebado de moléculas de ácidos nucleicos, PS2 es un segundo sitio de cebado de moléculas de ácidos nucleicos, RE2 es una segunda secuencia de reconocimiento de endonucleasas de restricción, Y comprende una molécula de ácido nucleico diana, y X^a y X^b son códigos que comprenden una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos o entre aproximadamente 6 nucleótidos y aproximadamente 15 nucleótidos o entre aproximadamente 7 nucleótidos y aproximadamente 9 nucleótidos. En realizaciones adicionales, RE1 y RE2 son secuencias reconocidas por la misma endonucleasa de restricción o un isoesquizómero o neoesquizómero de la misma, o RE1 y RE2 tienen diferentes secuencias reconocidas por diferentes endonucleasas de restricción. En realizaciones adicionales, PS1 y PS2 tienen diferentes secuencias. En realizaciones adicionales, las moléculas de ácidos nucleicos diana del banco de moléculas de ácidos nucleicos tendrán cada una códigos únicos duales X^a y X^b, en donde ninguno de los códigos X^a tiene la misma secuencia que cualquier otro código X^a, ninguno de los códigos X^b tiene la misma secuencia que cualquier otro código X^b y ninguno de los códigos X^a tiene la misma secuencia que cualquier código X^b. En aún otras realizaciones, las moléculas de ácidos nucleicos diana del banco de moléculas de ácidos nucleicos tendrán cada una un único código o par de códigos X^a-X^b , en donde ninguno de los códigos X^a o X^b tiene la misma secuencia.

También se contempla en la presente divulgación utilizar un banco de moléculas de ácidos nucleicos diana de doble cadena con código de barras o de doble cadena con doble código de barras para reacciones de amplificación y secuenciación para detectar mutaciones verdaderas. Con el fin de facilitar determinados métodos de amplificación o secuenciación, se pueden incluir otras características en las composiciones de la presente divulgación. Por ejemplo, la amplificación de puente puede implicar ligar secuencias de adaptador a cada uno de los extremos de una población de moléculas de ácidos nucleicos diana. Cebadores de oligonucleotídicos de cadena sencilla complementarios a los adaptadores se inmovilizan sobre un sustrato sólido, las moléculas diana que contienen las secuencias de adaptador se desnaturalizan en cadenas sencillas y se hibridan a cebadores complementarios sobre el sustrato sólido. Se utiliza una reacción de extensión para copiar la molécula diana hibridada y el producto de doble cadena se desnaturaliza nuevamente en cadenas sencillas. Las cadenas sencillas copiadas forman entonces un

bucle (forman un "puente") y se hibridan con un cebador complementario sobre el sustrato sólido, sobre el que se ejecuta nuevamente la reacción de extensión. De esta manera, muchas moléculas diana pueden amplificarse al mismo tiempo y el producto resultante está sujeto a secuenciación paralela masiva.

En determinadas realizaciones, un banco de moléculas de ácidos nucleicos comprende moléculas de ácidos nucleicos que tienen una fórmula de 5'-RE1-AS-PS1-X^a-PS2-Y-PS2-X^b-PS1-AS-RE2-3', en donde RE1 y RE2 son primera y segunda secuencias de reconocimiento de endonucleasas de restricción, PS1 y PS2 son primer y segundo sitios de cebado de moléculas de ácidos nucleicos, AS es una secuencia de adaptador que comprende una longitud que oscila entre aproximadamente 20 nucleótidos y aproximadamente 100 nucleótidos, Y comprende una molécula de ácidos nucleicos diana, y X^a y X^b son códigos que comprenden una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos o entre aproximadamente 6 nucleótidos y aproximadamente 15 nucleótidos o entre aproximadamente 7 nucleótidos y aproximadamente 9 nucleótidos.

En realizaciones adicionales, un banco de moléculas de ácido nucleico comprende moléculas de ácidos nucleicos que tienen una fórmula de 5'-RE1-AS-PS1-X^a-Y-X^b-PS1-AS-RE2-3', en donde RE1 y RE2 son primera y segunda secuencias de reconocimiento de endonucleasas de restricción, PS1 es un primer sitio de cebado de moléculas de ácido nucleico, AS es una secuencia de adaptador que comprende una longitud que oscila entre aproximadamente 20 nucleótidos y aproximadamente 100 nucleótidos, Y comprende una molécula de ácido nucleico diana, y X^a y X^b son códigos que comprenden una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos o entre aproximadamente 6 nucleótidos y aproximadamente 15 nucleótidos o entre aproximadamente 7 nucleótidos y aproximadamente 9 nucleótidos. En realizaciones relacionadas, la secuencia de adaptador AS del vector mencionado anteriormente puede comprender, además, un PS2 que es un segundo sitio de cebado de moléculas de ácidos nucleicos o el PS2 puede ser una parte de la secuencia AS original. En aún otras realizaciones, el banco de moléculas de ácidos nucleicos puede comprender, además, una secuencia de índice (que comprende una longitud que oscila entre aproximadamente 4 nucleótidos y aproximadamente 25 nucleótidos) localizada entre cada uno de la primera y segunda AS y el PS1, de modo que el banco puede agruparse con otros bancos que tienen diferentes secuencias de índice para facilitar la secuenciación multiplex (también denominada multiplexación) antes o después de la amplificación.

Cada una de las moléculas de ácidos nucleicos diana duales de código de barras anteriormente mencionadas se puede ensamblar en un banco de soportes en forma de, por ejemplo, un vector autorreplicante, tal como un plásmido, cósmido, YAC, vector viral u otros vectores conocidos en la técnica. En determinadas realizaciones, cualquiera de las moléculas de ácidos nucleicos de doble cadena antes mencionadas que comprenden una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar, están contenidas en un vector. En aún otras realizaciones, un banco de vectores de este tipo es portado en una célula huésped tal como bacterias, levaduras o células de mamíferos.

La presente divulgación también proporciona vectores útiles para generar un banco de moléculas de ácidos nucleicos diana duales de código de barras de acuerdo con esta divulgación. Vectores ilustrativos que comprenden códigos y otros elementos de esta divulgación se ilustran en las Figuras 1 y 2.

En determinadas realizaciones, se proporciona una pluralidad de vectores de ácidos nucleicos que comprenden una pluralidad de códigos al azar, en donde cada uno de los vectores comprende una región que tiene una fórmula de 5'-RE1-PS1-X^a-PS2-RE3-PS2-X^b-PS1-RE2-3', en donde (a) RE1 es una primera secuencia de reconocimiento de endonucleasas de restricción, (b) PS1 es un primer sitio de cebado de moléculas de ácidos nucleicos, (c) X^a comprende un primer código al azar, (d) RE3 es un tercera secuencia de reconocimiento de endonucleasas de restricción, en donde RE3 es un sitio en el que se puede insertar una molécula de ácido nucleico diana, (e) X^b comprende un segundo código al azar, (f) PS2 es un segundo sitio de cebado de moléculas de ácidos nucleicos y (g) RE2 es una segunda secuencia de reconocimiento de endonucleasas de restricción; y en donde cada uno de la pluralidad de códigos al azar comprende una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos, preferiblemente entre aproximadamente 7 nucleótidos y aproximadamente 9 nucleótidos; y en donde la pluralidad de vectores de ácidos nucleicos es útil para preparar un banco de moléculas de ácidos nucleicos de doble cadena en el que cada uno de los vectores tiene una inserción de molécula de ácido nucleico diana diferente. En determinadas realizaciones, la secuencia del código X^a es diferente de la secuencia del código X^b en cada uno de los vectores (es decir, cada uno de los vectores tiene un par único). En realizaciones adicionales, la pluralidad de vectores de ácidos nucleicos puede comprender, además, al menos una secuencia de adaptador (AS) entre RE1 y PS1 y al menos una AS entre PS1 y RE2, o puede comprender al menos una AS entre RE1 y el código X^a y al menos una AS entre el código X^b y RE2, en donde la AS tiene opcionalmente un sitio de cebado.

En realizaciones de vectores adicionales, la pluralidad de códigos al azar puede tener cada uno el mismo o diferente número de nucleótidos, y puede comprender de aproximadamente 6 nucleótidos a aproximadamente 8 nucleótidos a aproximadamente 10 nucleótidos a aproximadamente 12 nucleótidos a aproximadamente 15 nucleótidos. En aún otras realizaciones, una pluralidad de moléculas de ácidos nucleicos diana que comprenden de aproximadamente 10 nucleótidos a aproximadamente 10.000 nucleótidos o que comprenden de aproximadamente 100 nucleótidos a aproximadamente 750 nucleótidos o a aproximadamente 1.000 nucleótidos, pueden insertarse en el vector en RE3.

En determinadas realizaciones, RE3 escindirá ADN en extremos romos y la pluralidad de moléculas de ácidos nucleicos diana ligadas en este sitio también tendrá extremos romos.

En determinadas realizaciones, la pluralidad de vectores de ácidos nucleicos en donde cada uno de los vectores comprende una región que tiene una fórmula de 5'-RE1-PS1-X^a-PS2-RE3-PS2-X^b-PS1-RE2-3', los códigos X^a y los códigos X^b en cada uno de los vectores se secuencia antes de que una molécula de ácido nucleico diana se inserte en cada uno de los vectores. En realizaciones adicionales, la pluralidad de vectores de ácidos nucleicos en donde cada uno de los vectores comprende una región que tiene una fórmula de 5'-RE1-PS1-X^a-PS2-RE3-PS2-X^b-PS1-RE2-3', los códigos X^a y los códigos X^b de cada uno de los vectores se secuencian después de que una molécula de ácido nucleico diana se inserte en cada uno de los vectores o se secuencia al mismo tiempo que se secuencia una inserción de molécula de ácido nucleico diana.

Las moléculas de ácidos nucleicos diana duales de código de barras y los vectores que contienen este tipo de moléculas de esta divulgación se pueden utilizar, además, en reacciones de secuenciación para determinar la secuencia y la frecuencia de mutación de las moléculas en el banco. En determinadas realizaciones, esta divulgación proporciona un método para obtener una secuencia de ácido nucleico preparando un banco de ácidos nucleicos dual de código de barras de doble cadena tal como se describe en esta memoria y luego secuenciando cada una de las cadenas de la pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos al azar. En determinadas realizaciones, las moléculas de ácidos nucleicos diana y los códigos asociados se escinden para la secuenciación directa del vector utilizando enzimas endonucleasas de restricción antes de la amplificación. En determinadas realizaciones, los métodos de secuenciación de nueva generación se utilizan para determinar la secuencia de moléculas del banco, tal como secuenciación por síntesis, pirosecuenciación, secuenciación de colorante-terminador reversible o secuenciación de polonias.

En aún realizaciones adicionales se proporcionan métodos para determinar la tasa de errores debida a la amplificación y secuenciación determinando la secuencia de una cadena de una molécula de ácido nucleico diana asociada con el primer código al azar y alineando con la secuencia de la cadena complementaria asociada con el segundo código al azar para distinguir entre una mutación pre-existente y una mutación de artefacto por amplificación o secuenciación, en donde la tasa de errores de secuenciación medida variará de aproximadamente 10^{-6} a aproximadamente 5 x 10^{-6} a aproximadamente 10^{-7} a aproximadamente 5 x 10^{-7} a aproximadamente 10^{-8} a aproximadamente 10^{-9}. En otras palabras, utilizando los métodos de esta divulgación, una persona de experiencia ordinaria en la técnica puede asociar cada una de las secuencias de ADN leída a un ADN molde original. Dado que ambas cadenas del ADN de doble cadena original son códigos de barras asociados con códigos de barras, esto aumenta la sensibilidad de la llamada base de secuenciación al identificar más fácilmente los cambios de secuencia de "mutaciones" de artefactos introducidos durante el proceso de secuenciación.

En determinadas realizaciones, las composiciones y los métodos de la presente divulgación serán útiles para detectar mutantes raros frente a una gran señal de fondo, tal como cuando se monitorizan células tumorales circulantes; detectar ADN mutante circulante en sangre, monitorizar o detectar enfermedades y mutaciones raras mediante secuenciación directa, monitorizar o detectar mutaciones asociadas a la enfermedad o a la respuesta a fármacos. Se pueden utilizar realizaciones adicionales para cuantificar el daño del ADN, cuantificar o detectar mutaciones en genomas virales (*p. ej.*, VIH y otras infecciones virales) u otros agentes infecciosos que pueden ser indicativos de respuesta a la terapia o pueden ser útiles para controlar la progresión o recurrencia de la enfermedad. En aún otras realizaciones, estas composiciones y métodos pueden ser útiles para detectar daño al ADN de la quimioterapia, o en la detección y cuantificación de la metilación específica de secuencias de ADN.

EJEMPLOS

EJEMPLO 1

SECUENCIACIÓN DE UN CÓDIGO DUAL DE UNA GENOTECA DE TUMORES

Las células cancerosas contienen numerosas mutaciones clonales, es decir, mutaciones que están presentes en la mayoría o en todas las células malignas de un tumor y que han sido seleccionadas, presumiblemente, porque confieren una ventaja proliferativa. Una cuestión importante es si las células cancerosas también contienen un gran número de mutaciones al azar, es decir, mutaciones no seleccionadas distribuidas aleatoriamente que se producen en solo una o unas pocas células de un tumor. Mutaciones al azar de este tipo podrían contribuir a la heterogeneidad morfológica y funcional de los cánceres e incluir mutaciones que confieren resistencia a la terapia. La presente divulgación proporciona composiciones y métodos para distinguir mutaciones clonales de mutaciones al azar.

Para examinar si células malignas exhiben un fenotipo mutador que da como resultado la generación de mutaciones al azar en todo el genoma, la secuenciación dual del código de la presente divulgación se realizará en genotecas normales y tumorales. En síntesis, ADN genómico del tejido normal y tumoral compatible con el paciente se prepara utilizando kits Qiagen® (Valencia, CA) y se cuantifica por absorbancia óptica y PCR cuantitativa (qPCR). El ADN genómico aislado se fragmenta a un tamaño de aproximadamente 150-250 pares de bases (banco de inserción corta) o a un tamaño de aproximadamente 300-700 pares de bases (banco de inserción larga) por cizallamiento. Los

fragmentos de ADN que tienen extremos colgantes se reparan (es decir, se hacen romos) utilizando T4 ADN polimerasa (que tiene actividad de exonucleasa tanto 3' a 5' como actividad polimerasa 5' a 3') y los extremos 5' del ADN romo se fosforilan con T4 polinucleótido quinasa (Quick Blunting Kit I, New England Biolabs), y luego se purifica. Los fragmentos de ADN reparados en el extremo se ligan al sitio *Sma*I del banco de vectores de código dual mostrados en la Figura 2 para generar una genoteca diana.

El banco de vectores de código ligado se purifica y los fragmentos de la genoteca diana se amplifican utilizando, por ejemplo, el siguiente protocolo de PCR: 30 segundos a 98°C; cinco a treinta ciclos de 10 segundos a 98°C, 30 segundos a 65°C, 30 segundos a 72°C; 5 minutos a 72°C; y luego almacenar a 4°C. La amplificación se realiza utilizando cebadores de cadena sentido y anti-sentido que se reasocian con una secuencia localizada dentro de la región de adaptador (en determinadas realizaciones, el cebador se reasociará a una secuencia aguas arriba de la AS), y está aguas arriba del código único y la inserción genómica diana (y, si está presente, aguas arriba de una secuencia de índice si se desea una secuenciación múltiple, *véase, p. ej.*, la Fig. 2) para la secuenciación del puente de Illumina. La secuencia del banco descrita anteriormente se realizará utilizando, por ejemplo, un instrumento de secuenciación Genome Analyzer II de Illumina® tal como se especifica por el fabricante.

Las etiquetas de código único se utilizan para descontextuar computacionalmente los datos de secuenciación y asignar todas las lecturas de secuencia a moléculas individuales (es decir, distinguir los errores de la PCR y secuenciación de las mutaciones reales). La llamada de bases y el alineamiento de secuencias se realizarán utilizando, por ejemplo, la secuencia Eland (Illumina, San Diego, CA). Los datos generados permitirán la identificación de la heterogeneidad del tumor a nivel de un solo nucleótido y revelarán tumores que tienen un fenotipo mutador.

EJEMPLO 2

SECUENCIACIÓN DEL CÓDIGO DUAL DE UN BANCO DE ADNMT

Las mutaciones en el ADN mitocondrial (ADNmt) conducen a una colección diversa de enfermedades que son difíciles de diagnosticar y tratar. Cada una de las células humanas tiene cientos a miles de genomas mitocondriales y mutaciones de ADNmt asociadas a la enfermedad son de naturaleza homoplásmica, es decir, la mutación idéntica está presente en una preponderancia de mitocondrias dentro de un tejido (Taylor y Turnbull, *Nat. Rev. Genet. 6*:389, 2005; Chatterjee *et al.*, *Oncogene 25*:4663, 2006). Aunque los mecanismos precisos de la acumulación de mutación de ADN mitocondrial en la patogénesis de la enfermedad siguen siendo esquivos, se han documentado múltiples mutaciones homoplásmicas en cánceres colorrectales, de mama, cervicales, ováricos, de próstata, hígado y pulmón (Copeland *et al.*, *Cancer Invest. 20*:557, 2002; Brandon *et al.*, *Oncogene 25*:4647, 2006). Por lo tanto, el genoma mitocondrial proporciona un excelente potencial como un biomarcador específico de la enfermedad, que puede permitir mejores resultados de tratamiento y una mayor supervivencia general.

La secuenciación de código dual de la presente divulgación puede aprovecharse para cuantificar las células tumorales circulantes (CTC) y el ADNmt tumoral circulante (ADNctmt) podría utilizarse para diagnosticar y clasificar el cáncer, evaluar la respuesta al tratamiento y evaluar la progresión y la recurrencia después de la cirugía. En primer lugar, ADNmt aislado para cáncer de próstata y células de la sangre periférica del mismo paciente se secuenciarán para identificar mutaciones de ADNmt homoplásmicas somáticas. Estos biomarcadores de ADNmt se evaluarán estadísticamente por su potencial importancia clínica y fundamental con respecto a la puntuación de Gleason, la fase clínica, la recurrencia, la respuesta terapéutica y la progresión.

Una vez identificadas las mutaciones homoplásmicas específicas de los tumores individuales, se examinarán muestras de sangre del paciente para detectar la presencia de mutaciones idénticas en el plasma y la capa leucocitaria para determinar las frecuencias de ADNctmt y CTC respetuosamente. Esto se logrará utilizando la tecnología de secuenciación del código dual de esta divulgación, y como se describe en el Ejemplo 1, para monitorizar con sensibilidad múltiples mutaciones de ADNmt concurrentes. Se determinará la distribución de CTCs en la sangre periférica de pacientes con niveles séricos de PSA y puntuaciones de Gleason variables.

EJEMPLO 3

DETECCIÓN DE ALTA RESOLUCIÓN DE MUTACIONES DE TP53

Un estudio genómico reciente determinó que TP53 está mutado en el 96% del carcinoma ovárico seroso de alto grado (HGSC), responsable de dos tercios de todas las muertes por cáncer de ovario (Cancer Genome Atlas Research Network, *Nature 474*:609, 2011), y los modelos actuales indican que la pérdida de TP53 es un evento temprano en la patogénesis de HGSC (Bowtell, *Nat. Rev. Cancer 10*:803, 2010). Por lo tanto, la casi universalidad y la aparición temprana de mutaciones de TP53 en HGSC hacen de TP53 un candidato biomarcador prometedor para la detección temprana y la monitorización de la enfermedad de HGSC. La secuenciación de código dual de la presente divulgación se utilizó para detectar mutaciones somáticas de TP53 que surgieron durante la replicación en *E. coli*.

Construcción de Vector de Código Dual

Se preparó un oligonucleótido que contiene sitios de enzima de restricción *EcoR*I y *BamH*I, secuencias de adaptador, índices y códigos de barras al azar de 7 nucleótidos que flanquean un sitio de enzima de restricción *Sma*I con la siguiente secuencia (Integrated DNA Technologies):

GATACAGGATCCAATGATACGGCGACCACCGAGATCTACACTAGATCGCGCCTCCCTCGCGCCATCAGAGATGT GTATAAGAGACAGNNNNNNNCCCGGGNNNNNNNCTGTCTCTTATACACATCTCTGAGCGGGCTGGCAAGGCAG ACCGTAAGGCGAATCTCGTATGCCGTCTTCTGCTTGGAATTCGATACA (SEQ ID NO:2). Para amplificar y crear un producto de doble cadena a partir de este oligonucleótido de ADN de cadena sencilla se realizaron 30 ciclos de PCR utilizando ADN polimerasa de alta fidelidad PfuUltra (Agilent Technologies) según las instrucciones del fabricante (secuencia de cebador directo: GATACAGGATCCAATGATACGG, SEQ ID NO: 3; secuencia de cebador inverso: TGTATCGAATTCCAAGCAGAAG, SEQ ID NO: 4). Se utilizaron las siguientes condiciones de ciclado: 95°C durante 2 minutos, seguido de 30 ciclos de 95°C durante 1 minuto y 64°C durante 1 minuto. La naturaleza de doble cadena del producto se verificó utilizando una digestión por restricción *Sma*I (New England BioLabs). El producto se purificó después (Zymo Research DNA Clean & Concentrator-5) y se sometió a digestión por restricción de *EcoR*I/*BamH*I (New England BioLabs) y *EcoR*I-HF (New England BioLabs) para preparar la construcción para el ligamiento en una cadena principal de pUC19 digerida con *EcoR*I/*BamH*I. El vector y la construcción digeridos se realizaron en un gel de agarosa de bajo punto de fusión UltraPure en un gel de electroforesis de agarosa de bajo punto de fusión UltraPure (Invitrogen) al 1,5% con IX SybrSafe (Invitrogen) y las bandas apropiadas fueron eliminadas. El ADN en los fragmentos de gel se purificó utilizando un kit de recuperación de ADN en gel Zymo-Clean (Zymo Research) y se cuantificó utilizando un espectrofotómetro (Nanofotómetro, Implen). Se llevaron a cabo reacciones de ligamiento utilizando T4 ADN ligasa HC (Invitrogen) y un vector 1:3 para insertar la relación molar a temperatura ambiente durante 2 horas, luego se precipitó con etanol y se resuspendió en agua. El ADN purificado (2 µl) se sometió a electroporación en células resistentes a fagos ElectroMAX DH10B T1 (Invitrogen). Las células transformadas se sembraron en una dilución 1:100 en medio agar LB que contenía 100 µg/mL de carbenicilina y se incubaron durante la noche a 37ºC para determinar los recuentos de colonias, y el resto de la transformación se esparció en cultivos LB para crecimiento durante la noche a 37°C. El ADN de los cultivos durante la noche se purificó utilizando el Kit QIAquick Spin Minipreo (Qiagen).

Una única operación de secuenciación de nueva generación en MiSeq® demostró una cobertura y diversidad óptimas en el código de siete pares de bases del banco de vectores. La Figura 3A muestra que cada uno de los nucleótidos se detectó aproximadamente a la misma velocidad en cada posición aleatoria del código (en este caso, se secuenciaron los códigos 5').

Construcción del Banco del Exón 4 de TP53

En síntesis, células SKOV-3 (línea celular de carcinoma de ovario humano) se cultivaron en medio 5a de McCoy complementado con Suero Bovino Fetal al 10%, L-glutamina 1,5 mM, 2200 mg/L de bicarbonato de sodio y Penicilina/Estreptomicina. Se recogieron células SKOV-3 y el ADN se extrajo utilizando un kit DNeasy de sangre y tejidos (Qiagen). Los cebadores de la PCR se diseñaron para amplificar el exón 4 de TP53 humano; secuencia del cebador directo: TCTGTCTCCTTCCTCTTCCTACA (SEQ ID NO: 5) y secuencia del cebador inverso: AACCAGCCCTGTCGTCTCT (SEC ID NO: 6). Se realizaron treinta ciclos de PCR en ADN de SKOV-3 utilizando cebadores de 0,5 µM y mezcla maestra GoTaq Hot Start Colorless (Promega) en las siguientes condiciones de ciclado: 95ºC durante 2 minutos; 30 ciclos de 95°C durante 30 segundos, 63°C durante 30 segundos, 72°C durante 1 minuto; seguido de 72°C durante 5 minutos. Cada uno de los productos de la PCR se clonó en vectores TOPO (Invitrogen), se transformó en células de *E. coli* One Shot TOP 10 químicamente competentes (Invitrogen), se sembró en medio agar LB que contenía 100 µg/mL de carbenicilina y se incubó durante la noche a 37°C.

Se recogieron diez colonias y se cultivaron durante la noche. El ADN de los cultivos de LB durante la noche se purificó utilizando el kit QIAquick Spin Miniprep (Qiagen). La secuenciación de los clones TOPO se realizó utilizando la secuenciación basada en electroforesis capilar en un Analizador de ADN 3730x1 de Applied Biosystems. Se seleccionó un clon de TOPO que contenía la secuencia del exón 4 de TP53 de tipo salvaje apropiada. El ADN se sometió a digestión con *EcoR*I para escindir la inserción del exón 4 de TP53 y se realizó en un gel de agarosa de bajo punto de fusión UltraPure al 1,5%. La banda de ADN del exón 4 de TP53 se escindió manualmente y se purificó utilizando el kit de recuperación de ADN en gel Zymo-Clean, seguido de extracción con fenol/cloroformo/alcohol isoamílico y precipitación con etanol. El ADN digerido se hizo luego romo y se fosforiló utilizando el kit Quick Blunting (New England BioLabs) y se purificó con una extracción con fenol/cloroformo/alcohol isoamílico y precipitación con etanol.

El banco de vectores Cypher Seq se digirió con *Sma*I, se trató con fosfatasa antártica (New England BioLabs) y se realizó en un gel de agarosa de bajo punto de fusión UltraPure al 1,5%. La banda apropiada se escindió y se purificó utilizando el kit de recuperación de ADN en gel Zymo-Clean, seguido de extracción con fenol/cloroformo/alcohol isoamílico y precipitación con etanol. Los ligamientos de extremos romos del vector y el ADN del exón 4 de TP53 se llevaron a cabo luego en reacciones de 20 µl utilizando ADN de T4 Ligasa HC (Invitrogen) y un vector 1:10 para insertar la relación molar. Los ligamientos se incubaron a 16°C durante la noche, se precipitaron con etanol y se transformaron en células resistentes al fago T1 de ElectroMAX DH10b. Las bacterias se cultivaron durante la noche a 37°C en LB que contenía 100 µg/mL de carbenicilina y el ADN se purificó utilizando el kit QIAquick Spin Miniprep. La presencia de la inserción apropiada se verificó por digestión de restricción diagnóstica y electroforesis en gel.

La construcción de secuenciación que contenía los adaptadores Illumina, códigos de barras y ADN de TP53 se amplificó utilizando 10 ciclos de PCR y cebadores diseñados contra los extremos del adaptador (cebador directo: AATGATACGGCGACCACCGA, SEQ ID NO: 7 y cebador inverso: CAAGCAGAAGACGGCATACGA, SEQ ID NO: 8). Las condiciones del ciclo de PCR eran las siguientes: 95°C durante 2 minutos; 10 ciclos de 95°C durante 30 segundos, 63°C durante 30 segundos, 72°C durante 1 minuto; seguido de 72°C durante 5 minutos. La construcción de secuenciación se purificó en gel (kit de recuperación de ADN en gel Zymo-Clean), se extrajo con fenol/cloroformo/alcohol isoamílico y se precipitó con etanol. El banco se cuantificó utilizando el ensayo Quant-iT PicoGreen (Invitrogen) antes de la carga en la celda de flujo Illumina MiSeq®. Finalmente, el banco fue secuenciado. La secuenciación se realizó según las instrucciones del protocolo del fabricante con MiSeq® al nivel de calidad Q30 (Illumina). Una puntuación Q se define como una propiedad que está relacionada logarítmicamente con las probabilidades de error de llamada de bases ($Q = -10 \log_{10}P$). En el caso de una puntuación Q asignada de 30 (Q30) a una base, esto significa que la probabilidad de una llamada de bases incorrecta es 1 en 1.000 veces - es decir, la precisión de la llamada de bases (es decir, la probabilidad de una llamada de bases correcta) es 99.9% - considerado el patrón de oro para la secuenciación de nueva generación. Los códigos de barras se utilizaron para descontextuar los datos de secuenciación.

Resultados

El ADN del exón 4 de TP53 de un banco de vectores de código dual producido en *E. coli* se secuenció con una profundidad de más de un millón, y todas las lecturas de la secuenciación con pares de códigos idénticos y sus complementos inversos se agruparon en familias para crear una secuencia consenso. Tal como se ilustra en la Figura 3B, los errores introducidos durante la preparación del banco (círculo en blanco) y durante la secuenciación (círculo gris y triángulo) se eliminaron computacionalmente de la secuencia consenso y solo las mutaciones presentes en todas las lecturas (diamantes negros, Figura 3B) de una familia de códigos se contaron como mutaciones verdaderas (véase la parte inferior de la Figura 3B).

Se comparó la secuencia del exón 4 de TP53 de tipo salvaje con los resultados de la secuencia reales y se representaron gráficamente las sustituciones antes (Figura 4A) y después de la corrección con Cypher Seq (Figura 4B). Antes de la corrección, la frecuencia de errores detectada era 3.9×10^{-4}/pb (Figura 4A). En resumen, la frecuencia de errores inicial refleja errores relacionados con el ensayo (*p. ej.*, PCR, secuenciación y otros errores introducidos después de la codificación de barras). Esto significa que la detección de una mutación rara es difícil debido a que la relación de ruido a señal es muy alta. Sin embargo, después de la corrección Cypher Seq, la frecuencia de errores cayó a 8.8×10^{-7}/pb (Figura 4B). En otras palabras, las sustituciones restantes son muy probablemente de naturaleza biológica y lo más probable es que reflejen los errores introducidos durante la replicación en *E. coli* antes del ligamiento en los vectores con códigos de barras. Por lo tanto, las mutaciones verdaderas (es decir, las que surgen de forma natural en una célula durante la replicación) son fácilmente detectables utilizando el sistema de codificación de la presente divulgación.

LISTA DE SECUENCIAS

<110> Fred Hutchinson Cancer Research Center

<120> COMPOSICIONES Y MÉTODOS PARA IDENTIFICAR MUTACIONES DE MANERA PRECISA

<130> 360056.409WO

<150> US 61/600,535
<151>17-02-2012

<160> 8

<170> PatentIn versión 3.5

<210> 1
<211> 7
<212> ADN
<213> Secuencia Artificial

<220>
<223> secuencia de códigos al azar

<220>
<221> característica miscelánea
<222> (1)..(7)
<223> n = A,T,C o G

<400> 1
nnnnnnn 7

<210> 2
<211> 195
<212> ADN
<213> Secuencia Artificial

<220>
<223> Un oligonucleótio que contiene sitios de enzima de restricción EcoRI y BamHI, secuencias de adaptador, índices y códigos de barras al azar de 7 nucleótidos que flanquean un sitio de enzima de restricción *Sma*I

<220>
<221> característica miscelánea
<222> (1)..(195)
<223> n = A,T,C o G

<400> 2
gatacaggat ccaatgatac ggcgaccacc gagatctaca ctagatcgcg cctccctcgc 60

gccatcagag atgtgtataa gagacagnnn nnnncccggg nnnnnnnctg tctcttatac 120

acatctctga gcgggctggc aaggcagacc gtaaggcgaa tctcgtatgc cgtcttctgc 180

ttggaattcg ataca 195

<210> 3
<211> 22
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador directo

<400> 3
gatacaggat ccaatgatac gg 22

<210> 4
<211> 22
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador inverso

<400> 4
tgtatcgaat tccaagcaga ag 22

<210> 5
<211> 23
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador directo

<400> 5
tctgtctcct tcctcttcct aca 23

<210> 6
<211> 19
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador inverso

<400> 6
aaccagccct gtcgtctct 19

<210> 7
<211> 20
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador directo

<400> 7
aatgatacgg cgaccaccga 20

<210> 8
<211> 21
<212> ADN
<213> Secuencia Artificial

<220>
<223> Secuencia de cebador inverso

<400> 8
caagcagaag acggcatacg a 21

REIVINDICACIONES

1. Un método para detectar una mutación verdadera en una molécula de ácido nucleico, que comprende:

amplificar un banco de ácidos nucleicos de doble cadena, en donde el banco de ácidos nucleicos de doble cadena comprende una pluralidad de moléculas de ácidos nucleicos diana y una pluralidad de códigos de doble cadena, en donde el banco de ácidos nucleicos comprende moléculas que tienen una fórmula de X^a-Y-X^b (en orden 5' a 3'), en donde:

(a) X^a comprende un primer código;

(b) Y comprende una molécula de ácido nucleico diana, y

(c) X^b comprende un segundo código,

en donde cada una de la pluralidad de moléculas de ácidos nucleicos diana está asociada con un par único de primero y segundo códigos de doble cadena, en donde cada uno de la pluralidad de códigos comprende una longitud que oscila entre aproximadamente 5 nucleótidos y aproximadamente 50 nucleótidos, en donde se amplifican cada una de las cadenas de la pluralidad de moléculas de ácidos nucleicos diana y de la pluralidad de códigos de doble cadena;

secuenciar cada una de las cadenas amplificadas de la pluralidad de moléculas de ácidos nucleicos diana y de la pluralidad de códigos para obtener lecturas de secuenciación para la pluralidad de moléculas de ácidos nucleicos diana y la pluralidad de códigos, y de sus complementos inversos;

agrupar las lecturas de secuenciación de moléculas de ácidos nucleicos que comprenden pares de códigos idénticos en familias de lecturas de secuenciación, y

detectar la mutación verdadera a lo largo de una tasa de fondo de mutaciones de artefactos, comprendiendo dicha detección identificar como mutación verdadera una mutación presente sustancialmente en todas las lecturas en una familia de lecturas de secuenciación.

2. El método de la reivindicación 1, en el que las familias de lecturas de secuenciación comprenden los complementos inversos de las lecturas de secuenciación de moléculas de ácidos nucleicos diana que comprenden pares de códigos idénticos.

3. El método de la reivindicación 1 o 2, que comprende generar secuencias consenso para las familias de lecturas de secuenciación.

4. El método de la reivindicación 3, en el que la generación de la secuencia consenso comprende eliminar por computación mutaciones que surgen durante la preparación del banco o durante la secuenciación.

5. El método de la reivindicación 1, en el que la secuencia de doble cadena del código X^a para cada una de las moléculas de ácidos nucleicos diana es diferente de la secuencia de doble cadena del código X^b.

6. El método de la reivindicación 5, en el que ninguna de las secuencias de doble cadena del código X^a es la misma que la secuencia de doble cadena de cualquier otro código X^a, en el que ninguna de las secuencias de doble cadena del código X^b es la misma que la secuencia de doble cadena de cualquier otro código X^b y en el que ninguna de las secuencias de doble cadena del código X^a y del código X^b son las mismas.

7. El método de la reivindicación 1, en el que la secuencia de doble cadena del código X^a es idéntica a la secuencia de doble cadena del código X^b para una o más moléculas de ácidos nucleicos diana, con la condición de que el código de doble cadena para cada uno de los ácidos nucleicos diana sea diferente.

8. El método de la reivindicación 1, en el que la detección de la mutación verdadera comprende:

(i) secuenciar la pluralidad de moléculas de ácidos nucleicos con una tasa de errores que oscila entre aproximadamente 10^{-6} y aproximadamente 10^{-8};

(ii) secuenciar simultáneamente una pluralidad de moléculas de ácidos nucleicos diana diferentes con una tasa de errores de 5×10^{-6} o menos, 10^{-6} o menos, 5×10^{-7} o menos, 10^{-7} o menos, 5×10^{-8} o menos o 10^{-8} o menos;

(ii) secuenciar una molécula de ácido nucleico diana sencilla en profundidad con una tasa de errores de 5×10^{-7} o menos, 10^{-7} o menos, 5×10^{-8} o menos o 10^{-8} o menos.

9. El método de la reivindicación 1, en el que el primer y/o segundo códigos son:

(i) códigos al azar;

(ii) códigos catalogados; o

(iii) códigos al azar catalogados.

10. El método de la reivindicación 1, en el que las lecturas de secuenciación no cubren la secuencia completa de una molécula de ácido nucleico diana de doble cadena.

11. El método de la reivindicación 10, que comprende enlazar lecturas de secuenciación obtenidas de un extremo de la molécula diana de doble cadena con lecturas de secuenciación obtenidas del extremo opuesto o de la segunda cadena de la misma molécula diana de doble cadena.

12. El método de la reivindicación 1, en el que la pluralidad de moléculas de ácidos nucleicos diana comprende una molécula de ácido nucleico diana derivada de una célula tumoral circulante (CTC), un ADN mitocondrial de tumor circulante (ADNctmt) o un ADN viral.

13. El método de la reivindicación 1, en el que la pluralidad de códigos tiene cada uno el mismo número de nucleótidos y comprende una longitud de aproximadamente 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 nucleótidos.

14. El método de la reivindicación 1, en el que la molécula de ácido nucleico diana de Y comprende de aproximadamente 10 nucleótidos a aproximadamente 10.000 nucleótidos o de aproximadamente 100 nucleótidos a aproximadamente 1.000 nucleótidos.

15. El método de la reivindicación 1, en el que la amplificación es mediante amplificación de puente, amplificación en emulsión, amplificación por nanoesferas o amplificación por PCR.

16. El método de la reivindicación 1, en el que la secuenciación es secuenciación por síntesis, pirosecuenciación, secuenciación de colorante-terminador reversible o secuenciación de polonias.

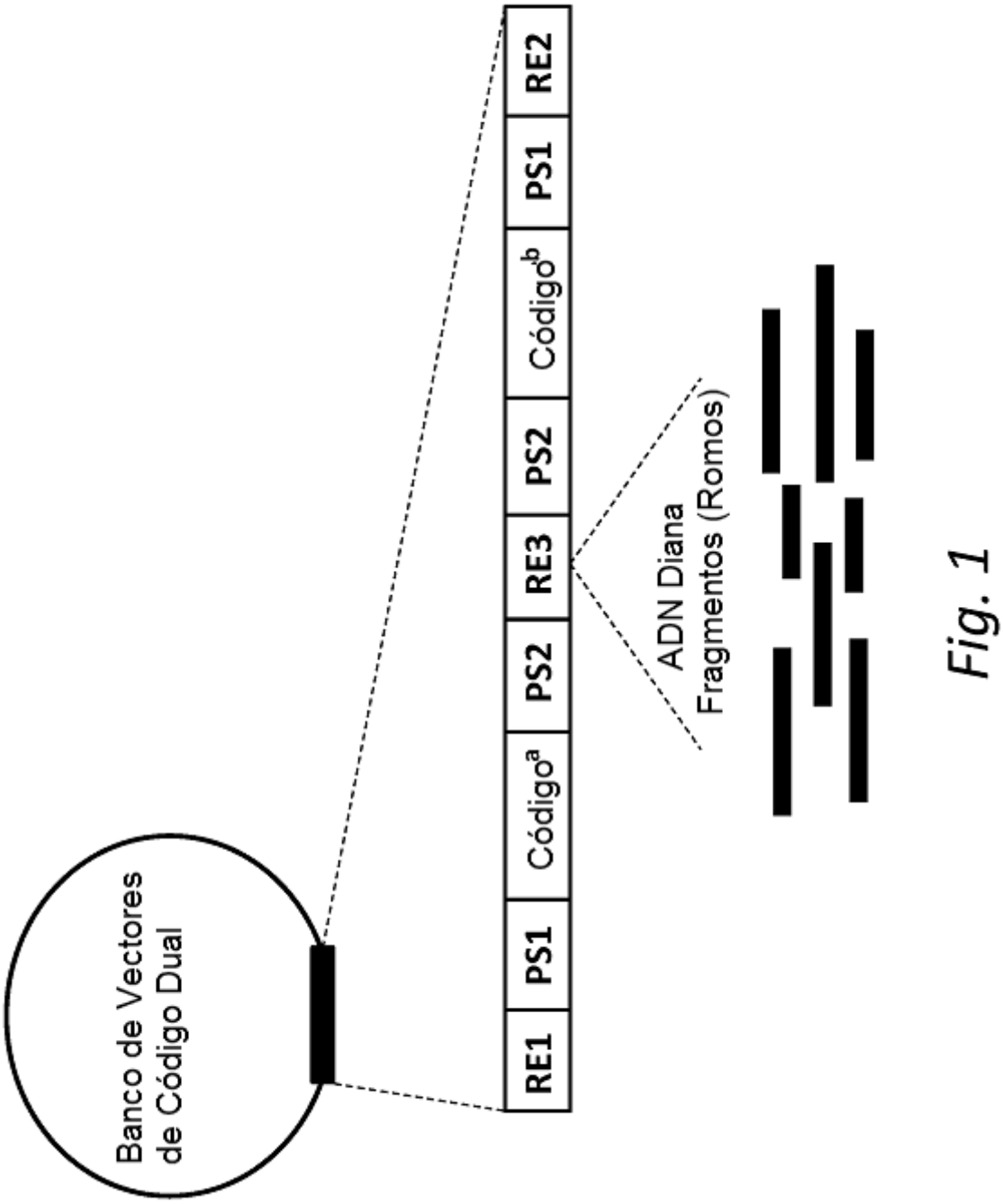
Banco de Vectores
de Código Dual
RE1
PS1
Código[a]
PS2
RE3
PS2
Código[b]
PS1
RE2
ADN Diana
Fragmentos (Romos)

Fig. 1

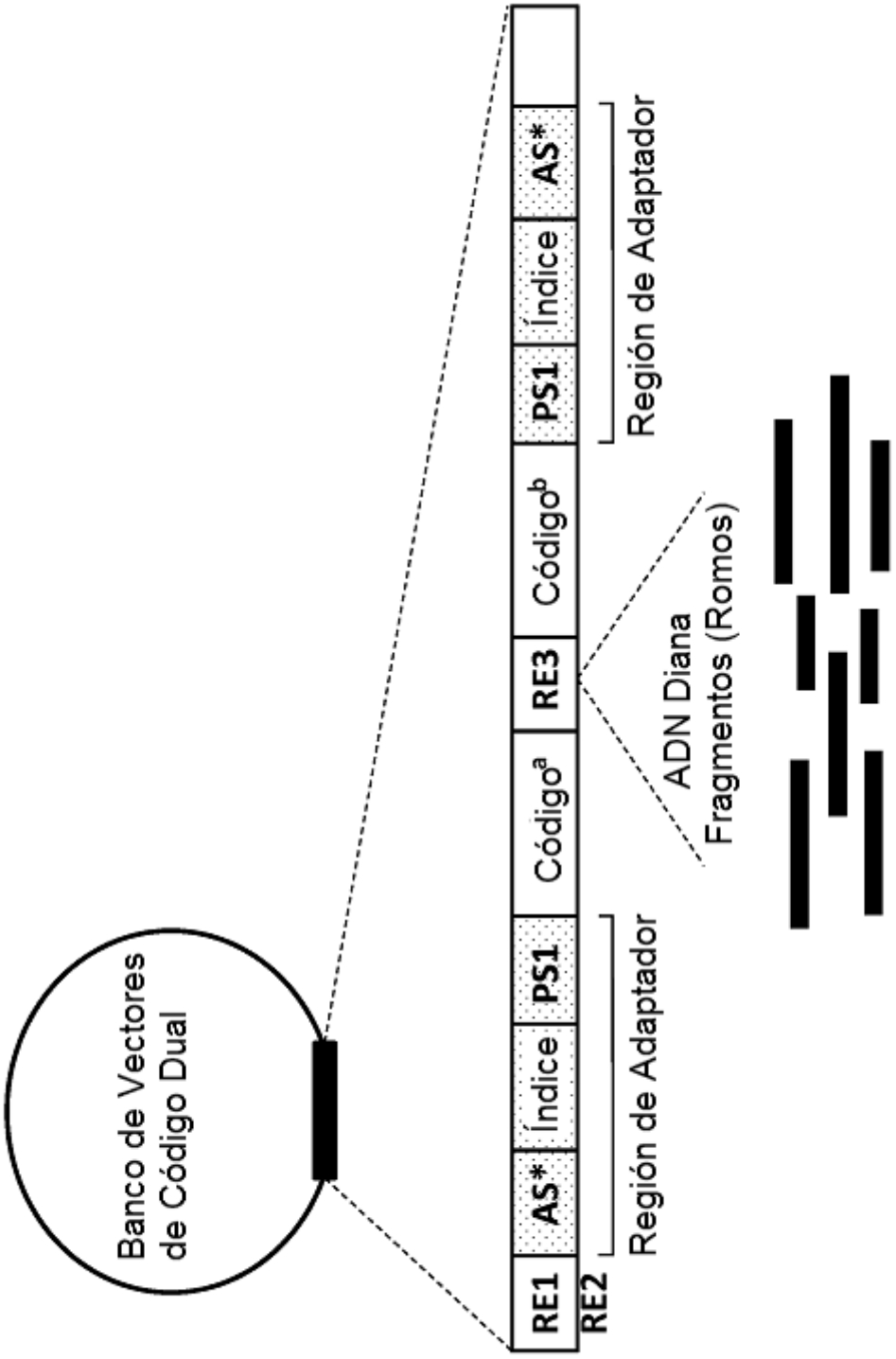
Banco de Vectores de Código Dual
RE1
AS*
Índice
PS1
Código[a]
RE3
Código[b]
PS1
Índice
AS*
RE2
Región de Adaptador
Región de Adaptador
ADN Diana
Fragmentos (Romos)

Fig. 2

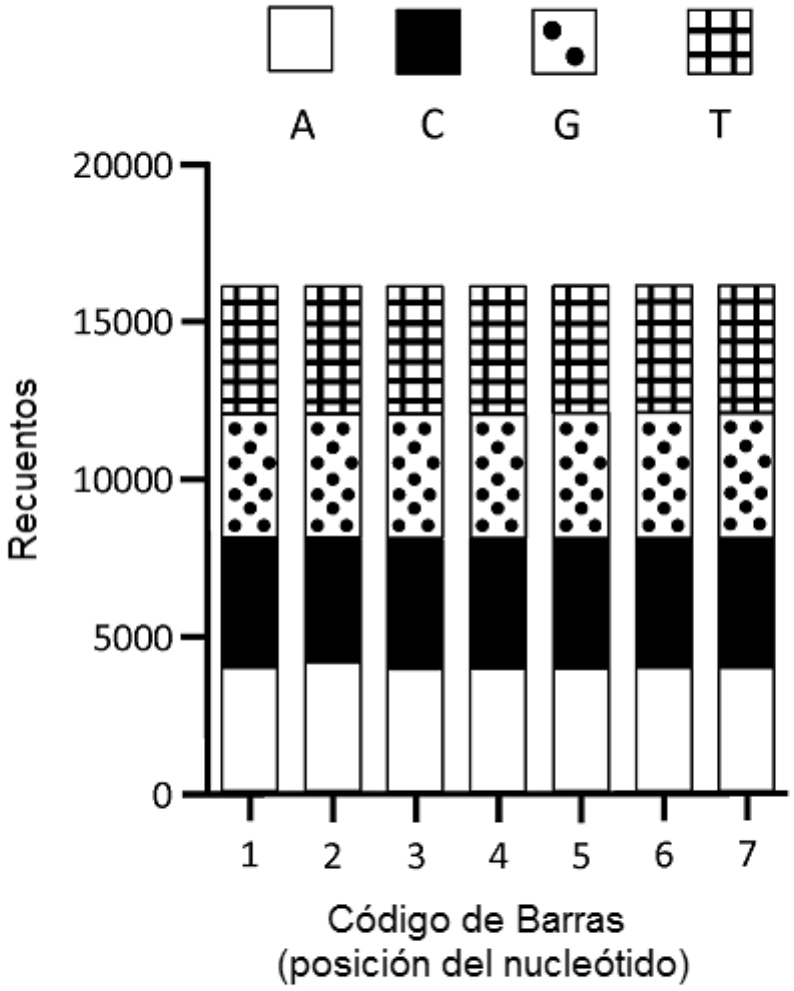
A
C
G
T
20000
15000
10000
5000
0
Recuentos
1
2
3
4
5
6
7
Código de Barras
(posición del nucleótido)

Fig. 3A

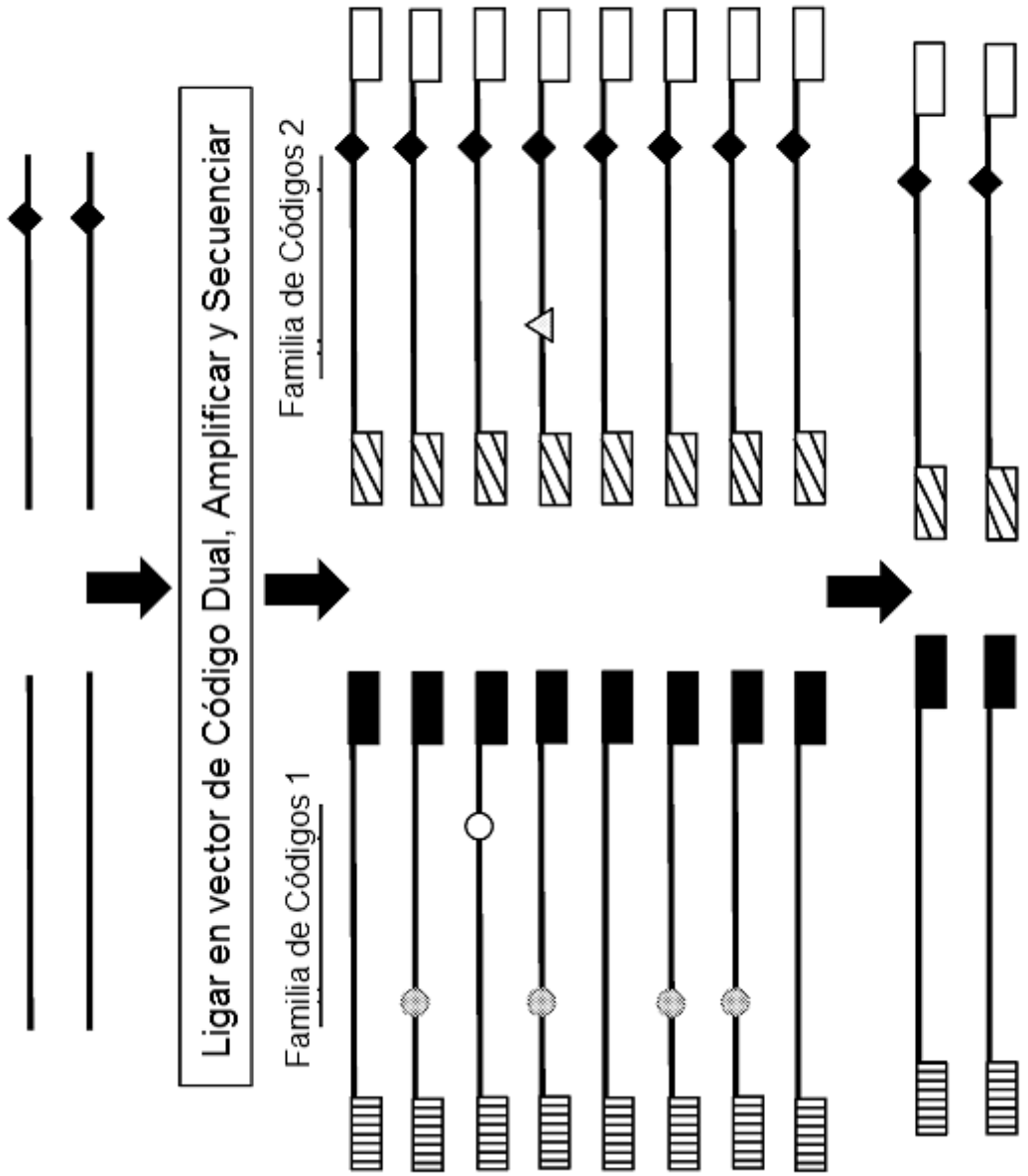
Ligar en vector de Código Dual, Amplificar y Secuenciar
Familia de Códigos 1
Familia de Códigos 2

Fig. 3B

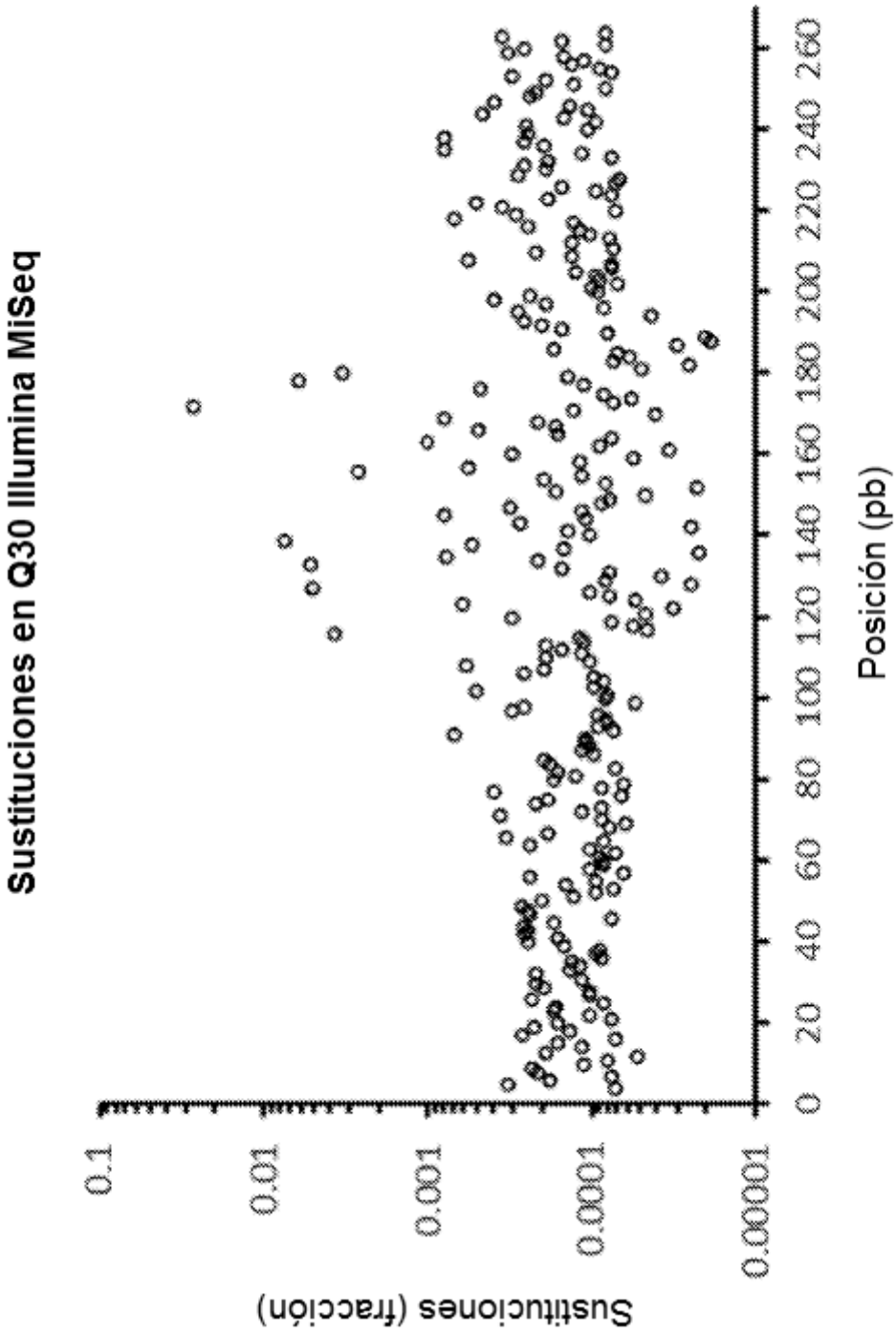
Sustituciones en Q30 Illumina MiSeq
Sustituciones (fracción)
0.1
0.01
0.001
0.0001
0.00001
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Posición (pb)

Fig. 4A

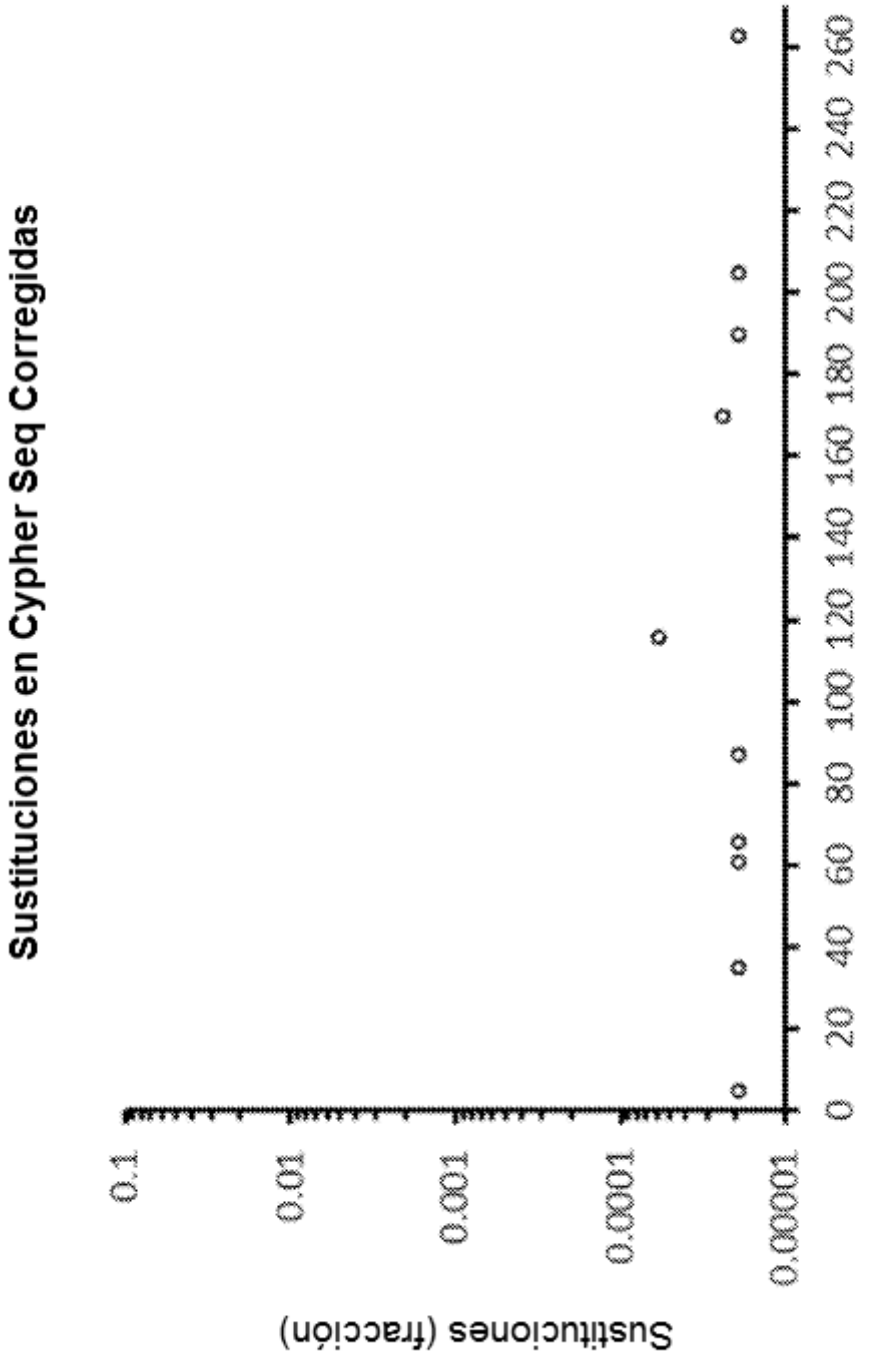
Sustituciones en Cypher Seq Corregidas
Sustituciones (fracción)
0.1
0.01
0.001
0.0001
0.00001
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Posición (pb)

Fig. 4B