(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 724 824**

(51) Int. CI.:

C12Q 1/6869 (2008.01)
C12N 15/10 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **13.03.2013 PCT/US2013/031023**

(87) Fecha y número de publicación internacional: **18.09.2014 WO14142850**

(96) Fecha de presentación y número de la solicitud europea: **13.03.2013 E 13711276 (9)**

(97) Fecha y número de publicación de la concesión europea: **20.02.2019 EP 2970951**

(54) Título: **Métodos para la secuenciación de ácidos nucleicos**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
16.09.2019

(73) Titular/es:

ILLUMINA, INC. (100.0%)
5200 Illumina Way
San Diego, CA 92122, US

(72) Inventor/es:

STEEMERS, FRANK;
AMINI, SASAN;
GUNDERSON, KEVIN;
PIGNATELLI, NATASHA y
GORYSHIN, IGOR

(74) Agente/Representante:

CONTRERAS PÉREZ, Yahel

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

DESCRIPCIÓN

Métodos para la secuenciación de ácidos nucleicos

Campo de la invención

Las realizaciones de la presente divulgación se refieren a la secuenciación de ácidos nucleicos. En particular, las realizaciones de los métodos y composiciones proporcionadas en el presente documento se refieren a la preparación de moldes de ácido nucleico y a la obtención de datos de secuencia de los mismos.

Antecedentes de la invención

La detección de secuencias específicas de ácidos nucleicos presentes en una muestra biológica se ha utilizado, por ejemplo, como un método de identificación y clasificación de microorganismos, de diagnóstico de enfermedades infecciosas, de detección y caracterización de anomalías genéticas, de identificación de cambios genéticos asociados con el cáncer, para el estudio de la susceptibilidad genética a las enfermedades y la medición de la respuesta a diversos tipos de tratamiento. Una técnica común para detectar secuencias específicas de ácidos nucleicos en una muestra biológica es la secuenciación de ácidos nucleicos.

La metodología de secuenciación de ácidos nucleicos ha evolucionado significativamente a partir de los métodos de degradación química utilizados por Maxam y Gilbert, y los métodos de elongación de cadena utilizados por Sanger. En la actualidad, se utilizan varias metodologías de secuenciación que permiten el procesamiento paralelo de ácidos nucleicos en una única ejecución de secuenciación. Por tanto, la información generada a partir de una única secuenciación puede ser enorme.

Sumario de la invención

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen un método para obtener información de secuencia de un ácido nucleico diana, comprendiendo dicho método: (a) obtener un ácido nucleico molde que comprende una pluralidad de transposomas insertados en dicho ácido nucleico diana, en donde al menos algunos de los transposomas insertados comprenden cada uno una primera secuencia de transposón, una segunda secuencia de transposón no contigua a dicha primera secuencia de transposón, y una transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón; (b) compartimentar el ácido nucleico molde que comprende dicha pluralidad de transposomas insertados en cada recipiente de una pluralidad de recipientes; (c) eliminar la transposasa del ácido nucleico molde; y (d) obtener información de secuencia del ácido nucleico molde de cada recipiente.

En algunas realizaciones, la etapa (b) comprende proporcionar a cada recipiente una cantidad de ácido nucleico molde igual a aproximadamente menos de un equivalente haploide, aproximadamente igual a un equivalente haploide o más de un equivalente haploide del ácido nucleico diana.

En algunas realizaciones, la etapa (b) comprende proporcionar a cada recipiente una cantidad de ácido nucleico molde de menos de aproximadamente un equivalente haploide del ácido nucleico diana.

En algunas realizaciones, la etapa (c) comprende un método seleccionado del grupo que consiste en añadir un detergente, cambiar la temperatura, cambiar el pH, añadir una proteasa, añadir una chaperona y añadir una polimerasa.

En algunas realizaciones, la primera secuencia de transposón comprende un primer sitio de cebador y las segundas secuencias de transposón comprenden un segundo sitio de cebador.

En algunas realizaciones, el primer sitio de cebador comprende adicionalmente un primer código de barras y el segundo sitio de cebador comprende adicionalmente un segundo código de barras.

En algunas realizaciones, el primer código de barras y el segundo código de barras son distintos.

En algunas realizaciones, el ácido nucleico diana comprende un ácido nucleico amplificado.

En algunas realizaciones, el ácido nucleico diana se obtiene enriqueciendo una pluralidad de ácidos nucleicos para una secuencia de interés antes o después de la transposición.

En algunas realizaciones, la etapa (a) comprende adicionalmente enriquecer el ácido nucleico molde para una secuencia de interés.

En algunas realizaciones, el ácido nucleico diana comprende ADN genómico.

En algunas realizaciones, la etapa (d) comprende adicionalmente ensamblar a partir de los datos de secuencia una representación de al menos una porción de dicho ácido nucleico molde de cada recipiente.

En algunas realizaciones, la información de secuencia comprende información de secuencia de haplotipos.

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen un método para preparar una biblioteca de ácidos nucleicos molde para obtener información de secuencia de un ácido nucleico diana, comprendiendo dicho método: (a) preparar un ácido nucleico molde que comprende una pluralidad de transposomas insertados en dicho ácido nucleico diana, en donde al menos alguno de los transposomas insertados comprenden cada uno una primera secuencia de transposón, una segunda secuencia de transposón no contigua a dicha primera secuencia de transposón, y una transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón; y (b) compartimentar el ácido nucleico molde que comprende dicha pluralidad de transposomas insertados en cada recipiente de una pluralidad de recipientes; y (c) eliminar la transposasa del ácido nucleico molde.

En algunas realizaciones, la etapa (b) comprende proporcionar a cada recipiente una cantidad de ácido nucleico molde igual a menos de un equivalente haploide, aproximadamente un equivalente o más de un equivalente del ácido nucleico diana.

En algunas realizaciones, la etapa (b) comprende proporcionar a cada recipiente una cantidad de ácido nucleico molde de menos de aproximadamente un equivalente haploide del ácido nucleico diana.

En algunas realizaciones, la etapa (c) comprende un método seleccionado del grupo que consiste en añadir un detergente, cambiar la temperatura, cambiar el pH, añadir una proteasa, añadir una chaperona y añadir una polimerasa.

En algunas realizaciones, la primera secuencia de transposón comprende un primer sitio de cebador y las segundas secuencias de transposón comprenden un segundo sitio de cebador.

En algunas realizaciones, el primer sitio de cebador comprende adicionalmente un primer código de barras y el segundo sitio de cebador comprende adicionalmente un segundo código de barras.

En algunas realizaciones, el primer código de barras y el segundo código de barras son distintos.

En algunas realizaciones, el ácido nucleico diana comprende un ácido nucleico amplificado.

En algunas realizaciones, el ácido nucleico diana se obtiene enriqueciendo una pluralidad de ácidos nucleicos para una secuencia de interés.

En algunas realizaciones, la etapa (a) comprende adicionalmente enriquecer el ácido nucleico molde para una secuencia de interés.

En algunas realizaciones, el ácido nucleico diana comprende ADN genómico.

En algunas realizaciones, la información de secuencia comprende información de secuencia de haplotipos.

Algunas realizaciones de los métodos y composiciones proporcionadas en el presente documento incluyen una biblioteca de ácidos nucleicos molde preparada de acuerdo con uno cualquiera de los métodos anteriores.

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen un método para obtener información de secuencia de un ácido nucleico diana, comprendiendo dicho método: (a) compartimentar el ácido nucleico diana en una pluralidad de primeros recipientes; (b) proporcionar un primer índice al ácido nucleico diana de cada primer recipiente, obteniendo de este modo un primer ácido nucleico indexado; (c) combinar los primeros ácidos nucleicos indexados; (d) compartimentar los primeros ácidos nucleicos molde indexados en una pluralidad de segundos recipientes; (e) proporcionar un segundo índice al primer nucleico molde indexado de cada segundo recipiente, obteniendo de este modo un segundo ácido nucleico indexado; y (f) obtener información de secuencia del segundo ácido nucleico indexado de cada segundo recipiente.

En algunas realizaciones, la etapa (b) comprende poner en contacto el ácido nucleico diana con una pluralidad de transposomas, comprendiendo cada uno una transposasa y una secuencia de transposón que comprende el primer índice en condiciones tales que al menos algunas de las secuencias de transposón se insertan en el ácido nucleico diana.

En algunas realizaciones, la etapa (b) comprende poner en contacto el ácido nucleico diana con una pluralidad de transposomas, comprendiendo cada transposón una primera secuencia de transposón que comprende un primer índice, una segunda secuencia de transposón no contigua a dicha primera secuencia de transposón, y una

transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón.

En algunas realizaciones, la etapa (d) comprende eliminar la transposasa de los primeros ácidos nucleicos molde indexados compartimentados.

En algunas realizaciones, la transposasa se elimina posteriormente a la etapa (b).

En algunas realizaciones, la transposasa se elimina antes de la etapa (f).

En algunas realizaciones, eliminar la transposasa comprende un método seleccionado del grupo que consiste en añadir un detergente, cambiar la temperatura, cambiar el pH, añadir una proteasa, añadir una chaperona y añadir una polimerasa de desplazamiento de cadena.

En algunas realizaciones, la primera de las secuencias de transposón comprende un primer sitio de cebador y la segunda de las secuencias de transposón comprende un segundo sitio de cebador.

En algunas realizaciones, el primer sitio de cebador comprende adicionalmente un primer código de barras y el segundo sitio de cebador comprende adicionalmente un segundo código de barras.

En algunas realizaciones, el primer código de barras y el segundo código de barras son distintos.

En algunas realizaciones, la etapa (b) comprende amplificar el ácido nucleico diana con al menos un cebador que comprende el primer índice.

En algunas realizaciones, la etapa (b) comprende ligar el ácido nucleico diana con al menos un cebador que comprende el primer índice.

En algunas realizaciones, el primer índice proporcionado al ácido nucleico diana de cada primer recipiente es distinto.

En algunas realizaciones, la etapa (a) comprende proporcionar a cada primer recipiente una cantidad de ácido nucleico diana mayor de aproximadamente uno o más equivalentes haploides del ácido nucleico diana.

En algunas realizaciones, la etapa (d) comprende proporcionar a cada recipiente una cantidad de los primeros ácidos nucleicos molde indexados mayor de aproximadamente uno o más equivalentes haploides del ácido nucleico diana.

En algunas realizaciones, la etapa (e) comprende amplificar el primer nucleico molde indexado con al menos un cebador que comprende el segundo índice.

En algunas realizaciones, la etapa (e) comprende ligar el primer nucleico molde indexado con al menos un cebador que comprende el segundo índice.

En algunas realizaciones, el segundo índice proporcionado al primer nucleico molde indexado de cada segundo recipiente es distinto.

En algunas realizaciones, el ácido nucleico diana comprende un ácido nucleico amplificado.

En algunas realizaciones, el ácido nucleico diana se obtiene enriqueciendo una pluralidad de ácidos nucleicos para una secuencia de interés.

En algunas realizaciones, el ácido nucleico diana comprende ADN genómico.

En algunas realizaciones, la etapa (f) comprende adicionalmente ensamblar a partir de los datos de secuencia una representación de al menos una porción de dicho ácido nucleico molde de cada recipiente.

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen un método de preparación de una biblioteca de ácidos nucleicos molde para obtener información de secuencia de un ácido nucleico diana, comprendiendo dicho método: (a) compartimentar el ácido nucleico diana en una pluralidad de primeros recipientes; (b) proporcionar un primer índice al ácido nucleico diana de cada primer recipiente, obteniendo de este modo un primer ácido nucleico indexado; (c) combinar los primeros ácidos nucleicos indexados; (d) compartimentar los primeros ácidos nucleicos molde indexados en una pluralidad de segundos recipientes; y (e) proporcionar un segundo índice al primer nucleico molde indexado de cada segundo recipiente, obteniendo de este modo un segundo ácido nucleico indexado.

En algunas realizaciones, la etapa (b) comprende poner en contacto el ácido nucleico diana con una pluralidad de

transposomas, comprendiendo cada uno una transposasa y una secuencia de transposón que comprende el primer índice en condiciones tales que al menos algunas de las secuencias de transposón se insertan en el ácido nucleico diana.

En algunas realizaciones, la etapa (b) comprende poner en contacto el ácido nucleico diana con una pluralidad de transposomas, comprendiendo cada transposón una primera secuencia de transposón que comprende un primer índice, una segunda secuencia de transposón no contigua a dicha primera secuencia de transposón, y una transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón.

En algunas realizaciones, la etapa (d) comprende eliminar la transposasa de los primeros ácidos nucleicos molde indexados compartimentados.

En algunas realizaciones, eliminar la transposasa comprende un método seleccionado del grupo que consiste en añadir un detergente, cambiar la temperatura, cambiar el pH, añadir una proteasa, añadir una chaperona y añadir una polimerasa.

En algunas realizaciones, la primera de las secuencias de transposón comprende un primer sitio de cebador y la segunda de las secuencias de transposón comprende un segundo sitio de cebador.

En algunas realizaciones, el primer sitio de cebador comprende adicionalmente un primer código de barras y el segundo sitio de cebador comprende adicionalmente un segundo código de barras.

En algunas realizaciones, el primer código de barras y el segundo código de barras son distintos.

En algunas realizaciones, la etapa (b) comprende amplificar el ácido nucleico diana con al menos un cebador que comprende el primer índice.

En algunas realizaciones, la etapa (b) comprende ligar el ácido nucleico diana con al menos un cebador que comprende el primer índice.

En algunas realizaciones, el primer índice proporcionado al ácido nucleico diana de cada primer recipiente es distinto.

En algunas realizaciones, la etapa (a) comprende proporcionar a cada primer recipiente una cantidad de ácido nucleico diana mayor de aproximadamente uno o más equivalentes haploides del ácido nucleico diana.

En algunas realizaciones, la etapa (d) comprende proporcionar a cada recipiente una cantidad de los primeros ácidos nucleicos molde indexados mayor de aproximadamente uno o más equivalentes haploides del ácido nucleico diana.

En algunas realizaciones, la etapa (e) comprende amplificar el primer nucleico molde indexado con al menos un cebador que comprende el segundo índice.

En algunas realizaciones, la etapa (e) comprende ligar el primer nucleico molde indexado con al menos un cebador que comprende el segundo índice.

En algunas realizaciones, el segundo índice proporcionado al primer nucleico molde indexado de cada segundo recipiente es distinto.

En algunas realizaciones, el ácido nucleico diana comprende un ácido nucleico amplificado.

En algunas realizaciones, el ácido nucleico diana se obtiene enriqueciendo una pluralidad de ácidos nucleicos para una secuencia de interés ya sea antes o después de la transposición.

En algunas realizaciones, el ácido nucleico diana comprende ADN genómico.

Algunas realizaciones de los métodos y composiciones proporcionadas en el presente documento incluyen una biblioteca de ácidos nucleicos molde preparada de acuerdo con uno cualquiera de los métodos anteriores.

Breve descripción de los dibujos

La **FIG. 1** representa un esquema de un transposoma que comprende una transposasa dimérica y dos secuencias de transposón no contiguas, y un transposoma que comprende una transposasa dimérica y una secuencia de transposón contigua.

La **FIG. 2** representa un método para preparar un transposoma con un enlazador que comprende una secuencia complementaria bicatenaria.

La **FIG. 3** representa una realización de la fabricación de una biblioteca de moldes usando transposomas que comprenden secuencias de transposón que comprenden un enlazador monocatenario que acopla las dos secuencias de transposón en cada transposoma en una orientación 5'-5'. Las secuencias se extienden utilizando cebadores.

La **FIG. 4** representa un esquema para preparar ácidos nucleicos molde para obtener información de secuencia, en el que un ácido nucleico diana se compartimenta en 96 tubos, se indexa por la inserción de transposones derivados de Tn5, los ácidos nucleicos indexados se combinan y se compartimentan adicionalmente en 96 tubos, se indexan adicionalmente por amplificación, después los ácidos nucleicos indexados dos veces se pueden combinar.

La **FIG. 5** representa una realización esquemática para obtener información de secuencia de haplotipos, en la que un ácido nucleico molde se indexa con el código de barras de un transposón y con un cebador. Se prepara un ácido nucleico molde mediante la inserción de un transposón en bucle en un ácido nucleico diana. El ácido nucleico molde se diluye en compartimentos. El ácido nucleico molde de cada compartimento se indexa por amplificación con un cebador. Se secuencian los ácidos nucleicos molde indexados, se alinean y se obtiene una representación de la secuencia.

La **FIG. 6** representa un esquema que incluye la preparación de un ácido nucleico diana utilizando parejas de compañeros (*matepair*) y amplificación por círculo rodante, seguido de la inserción de transposomas en el ácido nucleico diana, dilución del ácido nucleico diana para obtener información de haplotipos, eliminación de la transposasa por adición de SDS, generación de bibliotecas aleatorias (*shotgun*), indexación y secuenciación.

La **FIG. 7** representa un esquema que incluye la preparación de un ácido nucleico diana utilizando transposición por horquilla y amplificación por círculo rodante, seguido de la inserción de transposomas en el ácido nucleico diana, dilución del ácido nucleico diana, eliminación de la transposasa por adición de SDS, generación de bibliotecas aleatorias (*shotgun*), indexación y secuenciación para obtener información de haplotipos.

La **FIG. 8** representa un esquema de ejemplo para la generación de bibliotecas por parejas de compañeros.

La **FIG. 9** representa un esquema de ejemplo para la generación de bibliotecas por parejas de compañeros.

La **FIG. 10** es un gráfico que representa un modelo de tasa de error en la información de secuencia para el número de veces que se secuencia una secuencia particular asociada con un código de barras.

La **FIG. 11** representa imágenes de geles de agarosa que muestran oligonucleótidos unidos con un acoplamiento de bisoxiamina 5'-5', en que los transposones precursores en bucle están indicados por la banda del dímero.

La **FIG. 12** es una imagen de un gel de agarosa que muestra el peso molecular aparente de un ácido nucleico diana transpuesto asociado con la transposasa (carril izquierdo) y sin transposasa (+ SDS al 0,1 %, carril medio).

La **FIG. 13** resume que se observaron bloques de haplotipos de hasta 100 kb para muestras en las que se eliminó la transposasa mediante SDS después de la dilución.

La **FIG. 14** representa un gráfico que muestra las frecuencias de las lecturas de secuenciación para distancias particulares entre lecturas alineadas vecinas, para ácidos nucleicos molde preparados añadiendo SDS para eliminar la transposasa antes de la dilución, para obtener información de haplotipos, o después de la dilución para obtener información de haplotipos.

La **FIG. 15** muestra un gráfico de índices de códigos de barras y la proporción de lecturas, y demuestra que se observaron los 9216 distintos compartimentos en un esquema de indexación de 96 x 96.

La **FIG. 16** representa un análisis de apilamiento de información de haplotipos obtenido utilizando transposomas que comprenden Mu.

Descripción detallada

Las realizaciones de la presente invención se refieren a la secuenciación de ácidos nucleicos. En particular, las realizaciones de los métodos y composiciones proporcionadas en el presente documento se refieren a la preparación de moldes de ácido nucleico y a la obtención de datos de secuencia de los mismos. Los métodos y composiciones proporcionados en el presente documento están relacionados con los métodos y composiciones proporcionados en la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832. Algunas realizaciones de la presente invención se refieren a la preparación de moldes para obtener información de secuencia de haplotipos a partir de un ácido nucleico diana y a la obtención de información de secuencia de haplotipos a partir de tales moldes. Más realizaciones se refieren a la preparación de moldes para obtener información de secuencia a partir de una cadena de un ácido nucleico diana de bicatenario y a la obtención de información de secuencia a partir de dichos moldes. Las realizaciones particulares proporcionadas en el presente documento se refieren al uso de integrasas, por ejemplo transposasas, para mantener la proximidad física de los extremos asociados de ácidos nucleicos fragmentados; y al uso de compartimentos virtuales para permitir el uso de altas concentraciones de ácidos nucleicos.

La obtención de información de haplotipos de un ácido nucleico diana incluye la distinción entre distintos alelos (por ejemplo, los SNP, anomalías genéticas, etc.) en un ácido nucleico diana. Dichos métodos son útiles para caracterizar distintos alelos en un ácido nucleico diana y para reducir la tasa de error en la información de secuencia. Generalmente, los métodos para obtener información de secuencia de haplotipos incluyen obtener información de secuencia para una porción de un ácido nucleico molde. En una realización, puede diluirse un ácido nucleico molde y obtenerse la información de secuencia a partir de una cantidad de ácido nucleico molde equivalente a aproximadamente un haplotipo del ácido nucleico diana.

En realizaciones adicionales, puede compartimentarse un ácido nucleico molde de manera tal que pueden estar presentes en el mismo compartimento múltiples copias de un cromosoma, como resultado de la indexación doble o múltiple proporcionada en el presente documento, también se puede determinar un haplotipo. En otras palabras, se puede preparar un ácido nucleico molde utilizando compartimentos virtuales. En tales realizaciones, un ácido nucleico se puede distribuir entre varios primeros compartimentos, proporcionar un primer índice al ácido nucleico de cada compartimento, combinar los ácidos nucleicos, distribuir el ácido nucleico entre varios segundos compartimentos y proporcionar un segundo índice al ácido nucleico de cada compartimento. Ventajosamente, tal indexación permite obtener información de haplotipos en concentraciones más altas de ácido nucleico, en comparación con la mera dilución de un ácido nucleico en un único compartimento hasta una cantidad equivalente a un haplotipo del ácido nucleico.

En algunas realizaciones proporcionadas en el presente documento, las bibliotecas de moldes se preparan utilizando transposomas. En algunas de tales bibliotecas, el ácido nucleico diana puede estar fragmentado. Por consiguiente, algunas realizaciones proporcionadas en el presente documento se refieren a métodos para mantener la información de secuencia para la contigüidad física de fragmentos adyacentes. Dichos métodos incluyen el uso de integrasas para mantener la asociación de fragmentos de ácido nucleico molde adyacentes en el ácido nucleico diana. Ventajosamente, tal uso de integrasas para mantener la proximidad física de los ácidos nucleicos fragmentados aumenta la probabilidad de que los ácidos nucleicos fragmentados de la misma molécula original, por ejemplo, un cromosoma, se produzca en el mismo compartimento.

Otras realizaciones proporcionadas en el presente documento se refieren a la obtención de información de secuencia a partir de cada cadena de un ácido nucleico, la cual pueda ser útil para reducir la tasa de error en la información de secuenciación. Los métodos para preparar bibliotecas de ácidos nucleicos molde para la obtención de información de secuencia a partir de cada cadena de un ácido nucleico se pueden preparar de manera que cada cadena se pueda distinguir y que los productos de cada cadena también se puedan distinguir.

Algunos de los métodos proporcionados en el presente documento incluyen métodos de análisis de ácidos nucleicos. Dichos métodos incluyen la preparación de una biblioteca de ácidos nucleicos molde de un ácido nucleico diana, obtener datos de secuencia de la biblioteca de ácidos nucleicos molde y ensamblar una representación de secuencia del ácido nucleico diana a partir de tales datos de secuencia.

Generalmente, los métodos y composiciones proporcionados en el presente documento están relacionados con los métodos y composiciones proporcionados en la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832. Los métodos proporcionados en el presente documento se refieren al uso de transposomas útiles para insertar características en un ácido nucleico diana. Dichas características incluyen sitios de fragmentación, sitios de cebadores, códigos de barras, etiquetas de afinidad, fracciones indicadoras, etc.

En un método útil con las realizaciones proporcionadas en el presente documento, se prepara una biblioteca de ácidos nucleicos molde a partir de un ácido nucleico diana. La biblioteca se prepara insertando una pluralidad de códigos de barras distintivos en todo el ácido nucleico diana. En algunas realizaciones, cada código de barras incluye una primera secuencia de código de barras y una segunda secuencia de código de barras, que tiene un sitio de fragmentación dispuesto entre ellos. La primera secuencia de código de barras y la segunda secuencia de código de barras pueden identificarse o designarse para emparejarse entre sí. El emparejamiento puede ser informativo, de modo que un primer código de barras se asocie con un segundo código de barras. Ventajosamente, las secuencias de códigos de barras emparejadas se pueden usar para ensamblar datos de secuenciación procedente de la biblioteca de ácidos nucleicos molde. Por ejemplo, identificar un primer ácido nucleico molde que comprende una primera secuencia de código de barras y un segundo ácido nucleico molde que comprende una segunda secuencia de código de barras que está emparejado con el primero indica que los primero y segundo ácidos nucleicos molde representan secuencias adyacentes entre sí en una representación de secuencia del ácido nucleico diana. Dichos métodos pueden usarse para ensamblar una representación de secuencia de un ácido nucleico diana *de novo*, sin el requisito de un genoma de referencia.

Definiciones

Como se usa en el presente documento, la expresión "ácido nucleico" y/o "oligonucleótido", y/o sus equivalentes gramaticales pueden referirse a al menos dos monómeros nucleotídicos unidos entre sí. Un ácido nucleico generalmente puede contener enlaces fosfodiéster; sin embargo, en algunas realizaciones, los análogos de ácido nucleico pueden tener otros tipos de esqueletos, que comprenden, por ejemplo, fosforamida (Beaucage, *et al.*, Tetrahedron, 49:1925 (1993); Letsinger, J. Org. Chem., 35:3800 (1970); Sprinzl, *et al.*, Eur. J. Biochem., 81:579 (1977); Letsinger, *et al.*, Nucl. Acids Res., 14:3487 (1986); Sawai, *et al.*, Chem. Lett., 805 (1984), Letsinger, *et al.*, J. Am. Chem. Soc., 110:4470 (1988); y Pauwels, *et al.*, Chemica Scripta, 26:141 (1986)), fosforotioato (Mag, *et al.*, Nucleic Acids Res., 19:1437 (1991); y la patente de Estados Unidos N.º 5.644.048), fosforoditioato (Briu, *et al.*, J. Am. Chem. Soc., 111:2321 (1989), enlaces O-metilfosforamidita (véase Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press) y esqueletos y enlaces de ácido nucleico peptídico (véase Egholm, J.

Am. Chem. Soc., 114:1895 (1992); Meier, *et al.*, Chem. Int. Ed. Engl., 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson, *et al.*, Nature, 380:207 (1996)).

Otros ácidos nucleicos análogos incluyen aquellos con esqueletos positivos (Denpcy, *et al.*, Proc. Natl. Acad. Sci. USA, 92:6097 (1995)); esqueletos no iónicos (Patentes de Estados Unidos N.º 5.386.023; 5.637.684; 5.602.240; 5.216.141; y 4.469.863; Kiedrowshi, *et al.*, Angew. Chem. Intl. Ed. English, 30:423 (1991); Letsinger, *et al.*, J. Am. Chem. Soc., 110:4470 (1988); Letsinger, *et al.*, Nucleosides & Nucleotides, 13:1597 (1994); Capítulos 2 y 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui y P. Dan Cook; Mesmaeker, *et al.*, Bioorganic & Medicinal Chem. Lett., 4:395 (1994); Jeffs, *et al.*, J. Biomolecular NMR, 34:17 (1994); Tetrahedron Lett., 37:743 (1996)) y sin ribosa (patente de Estados Unidos N.º 5.235.033 y N.º 5.034.506, y los Capítulos 6 y 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui y P. Dan Coo). Los ácidos nucleicos también pueden contener uno o más azúcares carbocíclicos (véase Jenkins, *et al.*, Chem. Soc. Rev., (1995) pág. 169 176).

Se pueden efectuar modificaciones del esqueleto de ribosa-fosfato para facilitar la adición de fracciones adicionales, tales como marcadores, o para aumentar la estabilidad de tales moléculas en determinadas condiciones. Además, se pueden preparar mezclas de ácidos nucleicos de origen natural y sus análogos. Como alternativa, se pueden preparar mezclas de distintos análogos de ácidos nucleicos, y mezclas de ácidos nucleicos de origen natural y análogos. Los ácidos nucleicos pueden ser monocatenarios o bicatenarios, como se especifica, o contener porciones de ambas secuencias, bicatenaria o monocatenaria. El ácido nucleico puede ser ADN, por ejemplo, genómico o ADNc, ARN o un híbrido, de células individuales, múltiples células o de múltiples especies, como con las muestras para metagenómicas, tales como de muestras ambientales, además de muestras mixtas, por ejemplo, muestras de tejidos mixtas o muestras mixtas para distintos individuos de la misma especie, muestras de enfermedades tales como ácidos nucleicos relacionados con el cáncer, y similares. Un ácido nucleico puede contener cualquier combinación de desoxirribo- y ribonucleótidos, y cualquier combinación de bases, incluyendo uracilo, adenina, timina, citosina, guanina, inosina, xantina, hipoxantanina, isocitosina, isoguanina y análogos de bases tales como nitropirrol (incluyendo 3-nitropirrol) y el nitroindol (incluyendo 5-nitroindol), etc.

En algunas realizaciones, un ácido nucleico puede incluir al menos una base promiscua. Las bases promiscuas pueden emparejarse con más de un tipo distinto de base. En algunas realizaciones, una base promiscua puede emparejarse con al menos dos tipos distintos de bases y no más de tres tipos distintos de bases. Un ejemplo de una base promiscua incluye inosina, que puede emparejarse con adenina, timina o citosina. Otros ejemplos incluyen hipoxantina, 5-nitroindol, 5-nitroindol no cíclico, 4-nitropirazol, 4-nitroimidazol y 3-nitropirrol (Loakes *et al.*, Nucleic Acid Res. 22:4039 (1994); Van Aerschot *et al.*, Nucleic Acid Res. 23:4363 (1995); Nichols *et al.*, Nature 369:492 (1994); Bergstrom *et al.*, Nucleic Acid Res. 25:1935 (1997); Loakes *et al.*, Nucleic Acid Res. 23:2361 (1995); Loakes *et al.*, J. Mol. Biol. 270:426 (1997); y Fotin *et al.*, Nucleic Acid Res. 26:1515 (1998)). También pueden usarse bases promiscuas que pueden emparejarse con al menos tres, cuatro o más tipos de bases.

Como se usa en el presente documento, la expresión "análogo de nucleótido" y/o sus equivalentes gramaticales pueden referirse a análogos sintéticos que tienen porciones con bases nucleotídicas modificadas, porciones con pentosas modificadas y/o porciones con fosfatos modificados y, en el caso de polinucleótidos, enlaces internucleotídicos modificados, como se describe generalmente en otros lugares (por ejemplo, Scheit, Nucleotide Analogs, John Wiley, Nueva York, 1980; Englisch, Angew. Chem. Int. Ed. Engl. 30:613-29, 1991; Agarwal, Protocols for Polynucleotides and Analogs, Humana Press, 1994; y S. Verma y F. Eckstein, Ann. Rev. Biochem. 67:99-134, 1998). Generalmente, las porciones con fosfatos modificados comprenden análogos de fosfato en donde el átomo de fósforo está en el estado de oxidación +5 y uno o más de los átomos de oxígeno se reemplazan por una fracción que no es oxígeno, por ejemplo, azufre. Los ejemplos de análogos de fosfato incluyen, pero sin limitación, fosforotioato, fosforoditioato, fosforoselenoato, fosforodiselenoato, fosforoanilotioato, fosforanilidato, fosforamidato, boronofosfatos, incluyendo contraiones asociados, por ejemplo, H+, NH4+, Na+, si tales contraiones están presentes. Los ejemplos de porciones con bases nucleotídicas modificadas incluyen, pero sin limitación, 5-metilcitosina (5mC); análogos de C-5-propinilo, incluyendo, pero sin limitación, C-5 propinil-C y C-5 propinil-U; 2,6-diaminopurina, también conocida como 2-amino adenina o 2-amino-dA); hipoxantina, seudouridina, 2-tiopirimidina, isocitosina (isoC), 5-metil isoC e isoguanina (isoG; véase, por ejemplo, la Patente de Estados Unidos n.º 5.432.272). Las porciones con pentosas modificadas ejemplares incluyen, pero sin limitación, análogos de ácido nucleico bloqueado (LNA, forma siglada del inglés *locked nucleic acid*) que incluyen pero sin limitación Bz-A-LNA, 5-Me-Bz-C-LNA, dmf-G-LNA y T-LNA (véase, por ejemplo, The Glen Report, 16(2):5, 2003; Koshkin *et al.*, Tetrahedron 54:3607-30, 1998) y modificaciones 2' o 3' donde la posición 2' o 3' es hidrógeno, hidroxi, alcoxi (por ejemplo, metoxi, etoxi, aliloxi, isopropoxi, butoxi, isobutoxi y fenoxi), azido, amino, alquilamino, flúor, cloro o bromo. Los enlaces internucleotídicos modificados incluyen análogos de fosfato, análogos que tienen enlaces intersubunidades aquirales y no cargados (por ejemplo, Sterchak, E. P. *et al.*, Organic Chem., 52:4202, 1987) y polímeros a base de morfolino no cargados que tienen enlaces intersubunidades aquirales (véase, por ejemplo, la Patente de Estados Unidos n.º 5.034.506). Algunos análogos de enlaces internucleotídicos incluyen morfolidato, acetal y heterociclos unidos por poliamida. En una clase de análogos de nucleótido, conocidos como ácidos nucleicos peptídicos, que incluyen ácidos nucleicos peptídicos seudocomplementarios ("PNA", forma siglada del inglés *pseudocomplementary peptide nucleic acids*), un enlace convencional de azúcar e internucleótidos ha sido reemplazado por un polímero con esqueleto de 2-aminoetil glicinamida (véase, por ejemplo, Nielsen *et al.*, Science, 254:1497-1500, 1991; Egholm *et*

al., J. Am. Chem. Soc., 114: 1895-1897 1992; Demidov *et al*., Proc. Natl. Acad. Sci. 99:5953-58, 2002; Peptide Nucleic Acids: Protocols and Applications, Nielsen, ed., Horizon Bioscience, 2004).

Como se usa en el presente documento, la expresión "lectura de secuenciación" y/o sus equivalentes gramaticales puede referirse a un procedimiento repetitivo de etapas físicas o químicas que se lleva a cabo para obtener señales indicativas del orden de los monómeros en un polímero. Las señales pueden ser indicativas de un orden de monómeros con una resolución de monómeros individuales o una resolución más baja. En realizaciones particulares, las etapas pueden iniciarse en una diana de ácido nucleico y llevarse a cabo para obtener señales indicativas del orden de las bases en la diana de ácido nucleico. La expresión puede llevarse a cabo hasta su finalización típica, que generalmente se define por el punto en el que las señales procedentes del procedimiento ya no pueden distinguir las bases de la diana con un nivel de certeza razonable. Si se desea, la finalización puede producirse antes, por ejemplo, una vez que se ha obtenido una cantidad deseada de información de secuencia. Una lectura de secuenciación puede llevarse a cabo en una única molécula de ácido nucleico diana o simultáneamente sobre una población de moléculas de ácidos nucleicos diana que tienen la misma secuencia, o simultáneamente sobre una población de ácidos nucleicos diana que tienen distintas secuencias. En algunas realizaciones, una lectura de secuenciación termina cuando ya no se obtienen señales de una o más moléculas de ácido nucleico diana, a partir de las cuales se haya iniciado la adquisición de la señal. Por ejemplo, puede iniciarse una lectura de secuenciación para una o más moléculas de ácido nucleico diana que están presentes en un sustrato en fase sólida y terminarse tras la eliminación de una o más moléculas de ácido nucleico diana del sustrato. La secuenciación se puede terminar al detener la detección de los ácidos nucleicos diana que estaban presentes en el sustrato cuando se inició la secuenciación.

Como se usa en el presente documento, la expresión "representación de secuenciación" y/o los equivalentes gramaticales de la misma puede referirse a información que significa el orden y tipo de unidades monoméricas en el polímero. Por ejemplo, la información puede indicar el orden y el tipo de nucleótidos en un ácido nucleico. La información puede estar en cualquiera de una diversidad de formatos, incluyendo, por ejemplo, una representación, imagen, medio electrónico, serie de símbolos, serie de números, serie de letras, serie de colores, etc. La información puede estar con una resolución de monómeros individuales o una resolución más baja. Un polímero ejemplar es un ácido nucleico, tal como ADN o ARN, que tiene unidades nucleotídicas. Una serie de letras "A", "T", "G", y "C" es una representación de secuencia conocida para el ADN que se puede correlacionar, con una resolución de nucleótidos individuales, con la secuencia real de una molécula de ADN. Otros polímeros ejemplares son las proteínas, que tienen unidades aminoacídicas y los polisacáridos que tienen unidades de sacárido.

Como se usa en el presente documento, la expresión "al menos una porción" y/o los equivalentes gramaticales de la misma puede referirse a cualquier fracción de una cantidad total. Por ejemplo, "al menos una porción" puede referirse a al menos aproximadamente el 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, 50 %, 55 %, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 99 %, 99,9 % o 100 % de una cantidad total.

Transposomas

Un "transposoma" comprende una enzima de integración tal como una integrasa o transposasa, y un ácido nucleico que comprende un sitio de reconocimiento de integración, tal como un sitio de reconocimiento de transposasa. En realizaciones proporcionadas en el presente documento, la transposasa puede formar un complejo funcional con un sitio de reconocimiento de transposasa que tiene la capacidad de catalizar una reacción de transposición. La transposasa puede unirse al sitio de reconocimiento de transposasa e insertar el sitio de reconocimiento de transposasa en un ácido nucleico diana en un proceso denominado en ocasiones "etiquetación". En algunos de tales eventos de inserción, una cadena del sitio de reconocimiento de transposasa puede transferirse al ácido nucleico diana. La FIG. 1 representa dos ejemplos de transposomas. En un ejemplo, un transposoma (10) comprende una transposasa dimérica que comprende dos subunidades (20) y dos secuencias de transposón no contiguas (30). En otro ejemplo, un transposoma (50) comprende una transposasa, comprende una transposasa dimérica que comprende dos subunidades (60) y una secuencia de transposón contigua (70).

Algunas realizaciones pueden incluir el uso de una transposasa Tn5 hiperactiva y un sitio de reconocimiento de transposasa de tipo Tn5 (Goryshin y Reznikoff, J. Biol. Chem., 273:7367 (1998)), o una transposasa MuA y un sitio de reconocimiento de transposasa Mu que comprende secuencias finales para R1 y R2 (Mizuuchi, K., Cell, 35: 785, 1983; Savilahti, H, *et al*., EMBO J., 14: 4893, 1995). También se pueden usar secuencias de ME (forma siglada del inglés *mosaic elements*, elementos en mosaico) optimizadas por un experto en la materia.

Más ejemplos de sistemas de transposición que pueden usarse con determinadas realizaciones de las composiciones y métodos proporcionados en el presente documento incluyen Tn552 de *Staphylococcus aureus* (Colegio *et al*., J. Bacteriol., 183: 2384-8, 2001; Kirby C *et al*., Mol. Microbiol., 43: 173-86, 2002), Ty1 (Devine y Boeke, Nucleic Acids Res., 22: 3765-72, 1994 y la publicación internacional WO 95/23875), Transposón Tn7 (Craig, N L, Science. 271: 1512, 1996; Craig, N L, revisión en: Curr Top Microbiol Immunol., 204:27-48, 1996), Tn/O e IS10 (Kleckner N, *et al*., Curr Top Microbiol Immunol., 204:49-82, 1996), transposasa Mariner (Lampe DJ, *et al*., EMBO J., 15: 5470-9, 1996), Tc1 (Plasterk R H, Curr. Topics Microbiol. Immunol., 204: 125-43, 1996), Elemento P (Gloor, G B,

Methods Mol. Biol., 260: 97-114, 2004), Tn3 (Ichikawa y Ohtsubo, J Biol. Chem. 265:18829-32, 1990), secuencias de inserción bacterianas (Ohtsubo y Sekine, Curr. Top. Microbiol. Immunol. 204: 1-26, 1996), retrovirus (Brown, *et al*., Proc Natl Acad Sci USA, 86:2525-9, 1989) y retrotransposón de levadura (Boeke & Corces, Annu Rev Microbiol. 43:403-34, 1989). Más ejemplos incluyen IS5, Tn10, Tn903, IS911 y versiones técnicamente diseñadas de enzimas de la familia de las transposasas (Zhang *et al*., (2009) PLoS Genet. 5:e1000689. Epub 16 de oct. de 2009; Wilson C. *et al* (2007) J. Microbiol. Methods 71:332-5).

Más ejemplos de integrasas que se pueden usar con los métodos y composiciones proporcionadas en el presente documento incluyen integrasas retrovíricas y secuencias de reconocimiento de integrasa para tales integrasas retrovíricas, tales como las integrasas del VIH-1, VIH-2, VIS, VEP-1, VRS.

Secuencias de transposón

Algunas realizaciones de las composiciones y métodos proporcionados en el presente documento incluyen secuencias de transposón. En algunas realizaciones, una secuencia de transposón incluye al menos un sitio de reconocimiento de transposasa. En algunas realizaciones, una secuencia de transposón incluye al menos un sitio de reconocimiento de transposasa y al menos un código de barras. Las secuencias de transposón útiles con los métodos y composiciones proporcionadas en el presente documento se proporcionan en la de solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832. En algunas realizaciones, una secuencia de transposón incluye un primer sitio de reconocimiento de transposasa, un segundo sitio de reconocimiento de transposasa y un código de barras o códigos de barras dispuestos entre ellos.

Transposomas con secuencias de transposón no contiguas

Algunos transposomas proporcionados en el presente documento incluyen una transposasa que comprende dos secuencias de transposón. En algunas de tales realizaciones, las dos secuencias de transposón no están unidas entre sí, en otras palabras, las secuencias de transposón no están contiguas entre sí. Los ejemplos de tales transposones son bien conocidos en la técnica, véase, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2010/0120098. La **FIG. 1** representa un ejemplo de transposoma (10) que comprende una transposasa dimérica (20) y dos secuencias de transposón (30).

Estructuras en bucle

En algunas realizaciones, un transposoma comprende una ácido nucleico secuencia de transposón que se une a dos subunidades de transposasa para formar un "complejo en bucle" o un "transposoma en bucle". Esencialmente, un complejo de transposasa con transposones contiguos. La FIG. 1 representa un ejemplo de transposoma (50) que comprende una transposasa dimérica (60) y una secuencia de transposón (70). Los complejos en bucle pueden garantizar que los transposones se inserten en el ADN diana al tiempo que se mantiene la información de ordenamiento del ADN diana original y sin fragmentar el ADN diana. Como se apreciará, las estructuras en bucle pueden insertar cebadores, códigos de barras, índices y similares en un ácido nucleico diana, manteniendo la conectividad física del ácido nucleico diana. En algunas realizaciones, la secuencia de transposón de un transposoma en bucle puede incluir un sitio de fragmentación de forma que la secuencia de transposón puede fragmentarse para crear un transposoma que comprende dos secuencias de transposón. Dichos transposomas son útiles para garantizar que los fragmentos de ADN diana vecinos, en los que se insertan los transposones, reciban combinaciones de códigos que se pueden ensamblar de forma inequívoca en una fase posterior del ensayo.

Códigos de barras

Generalmente, un código de barras puede incluir una o más secuencias nucleotídicas que pueden usarse para identificar uno o más ácidos nucleicos particulares. El código de barras puede ser una secuencia artificial o puede ser una secuencia de origen natural generada durante la transposición, tal como secuencias de ADN genómico flanqueantes idénticas (códigos g) en el extremo de fragmentos de ADN anteriormente yuxtapuestos. Un código de barras puede comprender al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 o más nucleótidos consecutivos. En algunas realizaciones, un código de barras puede comprender al menos aproximadamente 10, 20, 30, 40, 50, 60, 70 80, 90, 100 o más nucleótidos consecutivos. En algunas realizaciones, al menos una porción de los códigos de barras en una población de ácidos nucleicos que comprende códigos de barras es distinta. En algunas realizaciones, al menos aproximadamente el 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 99 % de los códigos de barras son distintos. En más de tales realizaciones, todos los códigos de barras son distintos. La diversidad de distintos códigos de barras en una población de ácidos nucleicos que comprende códigos de barras puede generarse aleatoriamente o no aleatoriamente.

En algunas realizaciones, una secuencia de transposón comprende al menos un código de barras. En algunas realizaciones, tales como transposomas que comprenden dos secuencias de transposón no contiguas, la primera secuencia de transposón comprende un primer código de barras y la segunda secuencia de transposón comprende un segundo código de barras. En algunas realizaciones, tal como en transposomas en bucle, una secuencia de código de barras comprende una primera secuencia de código de barras y una segunda secuencia de código de

barras. En algunas de las realizaciones anteriores, la primera secuencia de código de barras puede identificarse o designarse para que se empareje con el segundo código de barras. Por ejemplo, puede saberse que una primera secuencia de código de barras conocida está emparejada con una segunda secuencia de código de barras conocida usando una tabla de referencia que comprende una pluralidad de primeras y segundas secuencias de código de barras que se sabe que se emparejan entre sí.

En otro ejemplo, la primera secuencia de código de barras puede comprender la misma secuencia que la segunda secuencia de código de barras. En otro ejemplo, la primera secuencia de código de barras puede comprender el complemento inverso la segunda secuencia de código de barras. En algunas realizaciones, la primera secuencia de código de barras y la segunda secuencia de código de barras son distintas. La primera y segunda secuencias de código de barras pueden comprender un bicódigo.

En algunas realizaciones de composiciones y métodos descritos en el presente documento, los códigos de barras se utilizan en la preparación de ácidos nucleicos molde. Como se entenderá, la gran cantidad de códigos de barras disponibles permite que cada molécula de ácido nucleico molde comprenda una identificación distintiva. La identificación distintiva de cada molécula en una mezcla de ácidos nucleicos molde se puede utilizar en varias aplicaciones. Por ejemplo, se pueden aplicar moléculas identificadas de manera distintiva para identificar moléculas de ácido nucleico individuales, en muestras que tiene múltiples cromosomas, en genomas, en células, en tipos celulares, en patologías celulares y en especies, por ejemplo, en secuenciación de haplotipos, en discriminación de alelos parentales, en secuenciación metagenómica y en la secuenciación de muestras de un genoma.

Enlazadores

Algunas realizaciones que comprenden transposomas en bucle en que una transposasa está formando complejo con transposones contiguos, incluyen secuencias de transposón que comprenden una primera secuencia de código de barras y una segunda secuencia de código de barras que tiene un enlazador dispuesto entre ellas. En otras realizaciones, el enlazador puede estar ausente o puede ser el esqueleto de azúcar-fosfato el que conecta un nucleótido con otro. El enlazador puede comprender, por ejemplo, uno o más de un nucleótido, un ácido nucleico, una fracción química no nucleotídica, un análogo de nucleótido, aminoácido, péptido, polipéptido o proteína. En realizaciones preferentes, el enlazador comprende un ácido nucleico. El enlazador puede comprender al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 o más nucleótidos. En algunas realizaciones, un enlazador puede comprender al menos aproximadamente 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 o más nucleótidos.

En algunas realizaciones, un enlazador puede ser amplificable, por ejemplo, por PCR, amplificación por círculo rodante, amplificación por desplazamiento de cadena y similares. En otras realizaciones, un enlazador puede comprender fracciones no amplificables. Los ejemplos de enlazadores no amplificables incluyen enlazadores químicos orgánicos tales como alquilo, propilo, PEG; bases no naturales tales como IsoC, isoG; o cualquier grupo que no se amplifique en esquemas de amplificación basados en ADN. Por ejemplo, los transposones que contienen parejas de isoC, isoG pueden amplificarse con mezclas de dNTP que carecen de isoG e isoC complementarios, lo que garantiza que no se produzca amplificación a través de los transposones insertados.

En algunas realizaciones, el enlazador comprende un ácido nucleico monocatenario. En algunas realizaciones, el enlazador acopla secuencias de transposón en una orientación 5'-3', una orientación 5'-5' o una orientación 3'-3'.

Sitios de fragmentación

En algunas realizaciones que comprenden transposomas en bucle, el enlazador puede comprender un sitio de fragmentación. Se puede usar un sitio de fragmentación para escindir la asociación física, pero no la informativa entre una primera secuencia de código de barras y una segunda secuencia de código de barras. La escisión puede ser por medios bioquímicos, químicos u otros. En algunas realizaciones, un sitio de fragmentación puede incluir un nucleótido o una secuencia nucleotídica que puede fragmentarse por diversos medios. Por ejemplo, un sitio de fragmentación puede comprender un sitio de endonucleasa de restricción; al menos un ribonucleótido escindible con una ARNasa; análogos de nucleótidos escindibles en presencia de determinado agente químico; un enlace diol escindible por tratamiento con peryodato; un grupo disulfuro escindible con un agente químico reductor; una fracción escindible que puede someter a escisión fotoquímica; y un péptido escindible mediante una enzima peptidasa u otro medio adecuado. Véase, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

Sitios de cebadores

En algunas realizaciones, una secuencia de transposón puede incluir un "adaptador de secuencia" o "sitio de adaptador de secuencia", es decir, una región que comprende uno o más sitios que pueden hibridar con un cebador. En algunas realizaciones, una secuencia de transposón puede incluir al menos un primer sitio de cebador útil para amplificación, secuenciación y similares. En algunas realizaciones que comprenden transposomas en bucle, un enlazador puede incluir un adaptador de secuencia. En más realizaciones que comprenden transposomas en bucle,

un enlazador comprende al menos un primer sitio de cebador y un segundo sitio de cebador. La orientación de los sitios de cebador en tales realizaciones puede ser tal que un cebador que hibride con el primer sitio de cebador y un cebador que hibride con el segundo sitio de cebador estén en la misma orientación, o en orientaciones distintas.

En algunas realizaciones, un enlazador puede incluir un primer sitio de cebador, un segundo sitio de cebador que tenga un sitio no amplificable dispuesto entre ellos. El sitio no amplificable es útil para bloquear la extensión de una cadena polinucleotídica entre el primer y el segundo sitios de cebador, en donde la cadena polinucleotídica hibrida con uno de los sitios de cebador. El sitio no amplificable también puede ser útil para prevenir concatámeros. Los ejemplos de sitios no amplificables incluyen un análogo de nucleótido, fracción química no nucleotídica, aminoácido, péptido y polipéptido. En algunas realizaciones, un sitio no amplificable comprende un análogo de nucleótido que no se empareja significativamente con A, C, G o T. Algunas realizaciones incluyen un enlazador que comprende un primer sitio de cebador, un segundo sitio de cebador que tenga un sitio de fragmentación dispuesto entre ellos. Otras realizaciones pueden usar un diseño de adaptador bifurcado o en forma de Y útil para la secuenciación direccional, como se describe en la patente de Estados Unidos N.º 7.741.463.

Etiquetas de afinidad

En algunas realizaciones, una secuencia de transposón o la transposasa puede incluir una etiqueta de afinidad. En algunas realizaciones que comprenden transposomas en bucle un enlazador puede comprender una etiqueta de afinidad. Las etiquetas de afinidad pueden ser útiles para una diversidad de aplicaciones, por ejemplo, la separación masiva de ácidos nucleicos diana hibridados con etiquetas de hibridación. La aplicación adicional incluye, pero sin limitación, utilizar por ejemplo etiquetas de afinidad para purificar complejos de transposasa/transposón y el ADN diana insertado por transposón. Como se usa en el presente documento, la expresión "etiqueta de afinidad" y los equivalentes gramaticales pueden referirse a un componente de un complejo de múltiples componentes, en donde los componentes del complejo de múltiples componentes interactúan específicamente o se unen entre sí. Por ejemplo, una etiqueta de afinidad puede incluir biotina o poli-His que puede unirse a estreptavidina o níquel, respectivamente. Se enumeran otros ejemplos de complejos de etiquetas de afinidad de componentes múltiples en, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

Fracciones indicadoras

En algunas realizaciones de las composiciones y métodos descritos en el presente documento, una secuencia de transposón o la transposasa puede incluir una fracción indicadora. En algunas realizaciones que comprenden transposomas en bucle un enlazador puede comprender una fracción indicadora. Como se usa en el presente documento, la expresión "fracción indicadora" y los equivalentes gramaticales pueden referirse a cualquier etiqueta, marcador o grupo identificable. El experto en la materia apreciará que pueden usarse muchas especies distintas de fracciones indicadoras con los métodos y composiciones descritos en el presente documento, ya sea individualmente o en combinación con una o más fracciones indicadoras distintas. En determinadas realizaciones, una fracción indicadora puede emitir una señal. Los ejemplos de una señal incluyen, pero sin limitación, una señal fluorescente, una quimioluminiscente, una bioluminiscente, una fosforescente, una radioactiva, una calorimétrica, una de actividad iónica, una electrónica o una electroquimioluminiscente. Se enumeran ejemplos de fracciones indicadoras en, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

Determinados métodos de fabricación de secuencias de transposón

Las secuencias de transposón proporcionadas en el presente documento pueden prepararse mediante una diversidad de métodos. Los métodos ejemplares incluyen síntesis directa, métodos de extensión de horquilla y PCR. En algunas realizaciones, las secuencias de transposón pueden prepararse por síntesis directa. Por ejemplo, una secuencia de transposón que comprende un ácido nucleico puede prepararse por métodos que comprenden síntesis química. Dichos métodos son bien conocidos en la materia, por ejemplo, la síntesis en fase sólida utilizando precursores de fosforamidita tales como los obtenidos de 2'-desoxinucleósidos protegidos, ribonucleósidos o análogos de nucleósidos. Los ejemplos de métodos para preparar la secuenciación por transposones se pueden encontrar en, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

En algunas realizaciones que comprenden transposomas en bucle, pueden prepararse secuencias de transposón que comprenden un enlazador monocatenario. En algunas realizaciones, el enlazador acopla las secuencias de transposón de un transposoma de forma que una secuencia de transposón que comprende una primera secuencia de reconocimiento de transposasa se acopla a una segunda secuencia de transposón que comprende una segunda secuencia de reconocimiento de transposasa en una orientación 5' a 3'. En algunas realizaciones, el enlazador acopla una secuencia de transposón que comprende una primera secuencia de reconocimiento de transposasa a una segunda secuencia de transposón que comprende una segunda secuencia de reconocimiento de transposasa en una orientación 5' a 5' o en una orientación 3' a 3'. El acoplamiento de secuencias de transposón de un transposoma en una orientación 5' a 5' o en una orientación 3' a 3' puede ser ventajoso para evitar que elementos de

reconocimiento de transposasa, en particular elementos en mosaico (ME o M), interactúen entre sí. Por ejemplo, las secuencias de transposón acopladas pueden prepararse preparando secuencias de transposón que comprendan un grupo aldehído o un grupo oxiamina. Los grupos aldehído y oxiamina pueden interactuar para formar un enlace covalente acoplando así las secuencias de transposón.

En algunas realizaciones, se pueden preparar transposomas que comprenden secuencias complementarias. La FIG. 2 ilustra una realización en la que se carga una transposasa con secuencias de transposón que comprenden colas complementarias. Las colas hibridan para formar una secuencia de transposón enlazada. La hibridación puede producirse en condiciones diluidas para disminuir la probabilidad de hibridación entre transposomas.

Ácidos nucleicos diana

Un ácido nucleico diana puede incluir cualquier ácido nucleico de interés. Los ácidos nucleicos diana pueden incluir ADN, ARN, ácido nucleico peptídico, ácido nucleico morfolino, ácido nucleico bloqueado, ácido glicol nucleico, ácido treosa nucleico, muestras mixtas de ácidos nucleicos, ADN poliploide (es decir, ADN vegetal), mezclas de los mismos e híbridos de los mismos. En una realización preferente, el ADN genómico o copias amplificadas del mismo se utilizan como el ácido nucleico diana. En otra realización preferente, se utiliza ADNc, ADN mitocondrial o ADN de cloroplasto.

Un ácido nucleico diana puede comprender cualquier secuencia nucleotídica. En algunas realizaciones, el ácido nucleico diana comprende ADN secuencias homopoliméricas. Un ácido nucleico diana puede incluir también secuencias de repetición. Las secuencias de repetición pueden ser de cualquiera de una diversidad de longitudes, incluyendo, por ejemplo, 2, 5, 10, 20, 30, 40, 50, 100, 250, 500 o 1000 nucleótidos o más. Las secuencias de repetición pueden estar repetidas, ya sea de forma contigua o no contigua, cualquiera de una diversidad de veces incluyendo, por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 o 20 nucleótidos o más.

Algunas realizaciones descritas en el presente documento pueden utilizar un único ácido nucleico diana. Otras realizaciones pueden utilizar una pluralidad de ácidos nucleicos diana. En tales realizaciones, una pluralidad de ácidos nucleicos diana puede incluir una pluralidad de los mismos ácidos nucleicos diana, una pluralidad de ácidos nucleicos diana distintos donde algunos ácidos nucleicos diana son iguales, o una pluralidad de ácidos nucleicos diana donde todos los ácidos nucleicos diana son distintos. Las realizaciones que utilizan una pluralidad de ácidos nucleicos diana pueden llevarse a cabo en formatos múltiples de forma que los reactivos se suministren simultáneamente a los ácidos nucleicos diana, por ejemplo, en una o más cámaras o en una superficie de matriz. En algunas realizaciones, la pluralidad de ácidos nucleicos diana puede incluir substancialmente todo el genoma de un organismo particular. La pluralidad de ácidos nucleicos diana puede incluir al menos una porción del genoma de un organismo particular, que incluya, por ejemplo, al menos aproximadamente el 1 %, 5 %, 10 %, 25 %, 50 %, 75 %, 80 %, 85 %, 90 %, 95% o 99% del genoma. En realizaciones particulares, la porción puede tener un límite superior que sea como máximo de alrededor del 1 %, 5 %, 10 %, 25 %, 50 %, 75 %, 80 %, 85 %, 90 %, 95% o 99% del genoma.

Los ácidos nucleicos diana pueden obtenerse de cualquier fuente. Por ejemplo, los ácidos nucleicos diana pueden prepararse a partir de moléculas de ácido nucleico obtenidas de un único organismo o de poblaciones de moléculas de ácido nucleico obtenidas de fuentes naturales que incluyen uno o más organismos. Las fuentes de moléculas de ácido nucleico incluyen, pero sin limitación, orgánulos, células, tejidos, órganos u organismos. Las células que pueden usarse como fuentes de moléculas de ácido nucleico diana pueden ser procariotas (células bacterianas, por ejemplo, *Escherichia*, *Bacillus*, *Serratia*, *Salmonella*, *Staphylococcus*, *Streptococcus*, *Clostridium*, *Chlamydia*, *Neisseria*, *Treponema*, *Mycoplasma*, *Borrelia*, *Legionella*, *Pseudomonas*, *Mycobacterium*, *Helicobacter*, *Erwinia*, *Agrobacterium*, *Rhizobium* y géneros de *Streptomyces*); arquea, tal como la *crenarchaeota*, *nanoarchaeota* o *euryarchaeotia*; o eucariotas tales como hongos, (por ejemplo, levaduras), plantas, protozoos y otros parásitos, y animales (incluyendo los insectos (por ejemplo, *Drosophila* spp.), nematodos (por ejemplo, *Caenorhabditis elegans*) y mamíferos (por ejemplo rata, ratón, mono, primate no humano y humano). Los ácidos nucleicos diana y los ácidos nucleicos molde pueden enriquecerse en determinadas secuencias de interés usando diversos métodos bien conocidos en la técnica. Los ejemplos de tales métodos se proporcionan en la Pub. Int. N.º WO/2012/108864. En algunas realizaciones, los ácidos nucleicos pueden enriquecerse adicionalmente durante los métodos de preparación de bibliotecas de moldes. Por ejemplo, los ácidos nucleicos pueden enriquecerse en determinadas secuencias, antes de la inserción de transposomas, después de la inserción de transposomas y/o después de la amplificación de los ácidos nucleicos.

Además, en algunas realizaciones, los ácidos nucleicos diana y/o ácidos nucleicos molde pueden estar altamente purificados, por ejemplo, los ácidos nucleicos pueden estar al menos aproximadamente el 70%, 80 %, 90 %, 95 %, 96 %, 97 %, 98 %, 99% o 100% exentos de contaminantes antes de su uso con los métodos proporcionados en el presente documento. En algunas realizaciones, es beneficioso utilizar métodos conocidos en la técnica que mantengan la calidad y el tamaño del ácido nucleico diana, por ejemplo, el aislamiento y/o la transposición directa del ADN diana se puede realizar utilizando tacos de agarosa. La transposición también se puede realizar directamente en las células, con una población de células, con lisados y con ADN no purificado.

Determinados métodos de preparación de ácidos nucleicos molde

Algunas realizaciones incluyen métodos de preparación de ácidos nucleicos molde. Como se usa en el presente documento, "ácido nucleico molde" puede referirse a un sustrato para obtener información de secuencia. En algunas realizaciones, un ácido nucleico molde puede incluir un ácido nucleico diana, un fragmento del mismo o cualquier copia del mismo que comprenda al menos una secuencia de transposón, un fragmento del mismo o cualquier copia del mismo. En algunas realizaciones, un ácido nucleico molde puede incluir un ácido nucleico diana que comprenda un adaptador de secuenciación, tal como un sitio de cebador de secuenciación.

Algunos métodos para preparar ácidos nucleicos molde incluyen insertar una secuencia de transposón en un ácido nucleico diana, preparando de este modo un ácido nucleico molde. Algunos métodos de inserción incluyen poner en contacto una secuencia de transposón proporcionada en el presente documento con un ácido nucleico diana en presencia de una enzima, tal como una transposasa o integrasa, en condiciones suficientes para la integración de la secuencia o secuencias de transposón en el ácido nucleico diana.

En algunas realizaciones, la inserción de secuencias de transposón en un ácido nucleico diana puede ser no aleatoria. En algunas realizaciones, las secuencias de transposón pueden ponerse en contacto con ácidos nucleicos diana que comprenden proteínas que inhiben la integración en determinados sitios. Por ejemplo, se puede inhibir la integración de secuencias de transposón en el ADN genómico que comprende proteínas, ADN genómico que comprende cromatina, ADN genómico que comprende nucleosomas o ADN genómico que comprende histonas. En algunas realizaciones, las secuencias de transposón pueden asociarse con etiquetas de afinidad para integrar la secuencia de transposón en una secuencia particular en un ácido nucleico diana. Por ejemplo, una secuencia de transposón puede estar asociada con una proteína que se dirige a secuencias específicas de ácido nucleico, por ejemplo, histonas, proteínas de unión a cromatina, factores de transcripción, factores de iniciación, etc., y anticuerpos o fragmentos de anticuerpos que se unen a proteínas de unión a ácidos nucleicos específicas de secuencia particulares. En una realización ejemplar, una secuencia de transposón se asocia con una etiqueta de afinidad, tal como biotina; la etiqueta de afinidad se puede asociar con una proteína de unión a ácido nucleico.

Se entenderá que durante la integración de algunas secuencias de transposón en un ácido nucleico diana, varios nucleótidos consecutivos del ácido nucleico diana en el sitio de integración están duplicadas en el producto integrado. Por lo tanto, el producto integrado puede incluir una secuencia duplicada en cada extremo de la secuencia integrada en el ácido nucleico diana. Como se usa en el presente documento, la expresión "etiqueta de hospedador" o "etiqueta-g" puede referirse a una secuencia de ácido nucleico diana que está duplicada en cada extremo de una secuencia de transposón integrada. Las porciones monocatenarias de ácidos nucleicos que pueden generarse mediante la inserción de secuencias de transposón pueden repararse mediante una diversidad de métodos bien conocidos en la técnica, por ejemplo, mediante el uso de ligasas, oligonucleótidos y/o polimerasas.

En algunas realizaciones, se inserta una pluralidad de las secuencias de transposón proporcionadas en el presente documento en un ácido nucleico diana. Algunas realizaciones incluyen condiciones de selección suficientes para lograr la integración de una pluralidad de secuencias de transposón en un ácido nucleico diana, de forma que la distancia promedio entre cada secuencia de transposón integrada comprende un determinado número de nucleótidos consecutivos en el ácido nucleico diana.

Algunas realizaciones incluyen condiciones de selección suficientes para lograr la inserción de una secuencia o secuencias de transposón en un ácido nucleico diana, pero no en otra secuencia o secuencias de transposón. Se puede usar una diversidad de métodos para reducir la probabilidad de que una secuencia de transposón se inserte en otra secuencia de transposón. Los ejemplos de tales métodos útiles con las realizaciones proporcionadas en el presente documento pueden encontrarse en, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

En algunas realizaciones, las condiciones pueden seleccionarse de modo que la distancia promedio en un ácido nucleico diana entre secuencias de transposón integradas sea al menos aproximadamente de 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 o más nucleótidos consecutivos. En algunas realizaciones, la distancia promedio en un ácido nucleico diana entre secuencias de transposón integradas sea al menos aproximadamente de 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 o más nucleótidos consecutivos. En algunas realizaciones, la distancia promedio en un ácido nucleico diana entre secuencias de transposón integradas es de al menos aproximadamente 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 90 kb, 100 kb o más nucleótidos consecutivos. En algunas realizaciones, la distancia promedio en un ácido nucleico diana entre secuencias de transposón integradas es de al menos aproximadamente 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1000 kb o más nucleótidos consecutivos. Como se entenderá, algunas condiciones que pueden seleccionarse incluyen poner en contacto un ácido nucleico diana con un determinado número de secuencias de transposón.

Algunas realizaciones de los métodos descritos en el presente documento incluyen condiciones de selección suficientes para lograr que al menos una parte de las secuencias de transposón integradas en un ácido nucleico diana sean distintas. En realizaciones preferentes de los métodos y composiciones descritos en el presente

documento, cada secuencia de transposón integrada en un ácido nucleico diana es distinta. Algunas condiciones que pueden seleccionarse para lograr una determinada porción de secuencias de transposón integradas en secuencias diana que sean distintas incluyen la selección del grado de diversidad de la población de secuencias de transposón. Como se entenderá, la diversidad de secuencias de transposón surge en parte debido a la diversidad de los códigos de barras de tales secuencias de transposón. Por consiguiente, algunas realizaciones incluyen proporcionar una población de secuencias de transposón en la que al menos una porción de los códigos de barras son distintos. En algunas realizaciones, al menos aproximadamente el 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 98 %, 99 % o 100 % de los códigos de barras en una población de secuencias de transposón son distintos. En algunas realizaciones, al menos una porción de las secuencias de transposón integradas en un ácido nucleico diana son iguales.

Algunas realizaciones para la preparación de ácidos nucleicos molde pueden incluir copiar las secuencias que comprenden el ácido nucleico diana. Por ejemplo, algunas realizaciones incluyen la hibridación de un cebador con un sitio de cebador de una secuencia de transposón integrada en el ácido nucleico diana. En algunas de tales realizaciones, el cebador puede hibridar con el sitio de cebador y se puede extender. Las secuencias copiadas pueden incluir al menos una secuencia de código de barras y al menos una porción del ácido nucleico diana. En algunas realizaciones, las secuencias copiadas pueden incluir una primera secuencia de código de barras, una segunda secuencia de código de barras y al menos una porción de un ácido nucleico diana dispuesto entre ellas. En algunas realizaciones, al menos un ácido nucleico copiado puede incluir al menos una primera secuencia de código de barras de un primer ácido nucleico copiado que puede identificarse o designarse para emparejarse con una segunda secuencia de código de barras de un segundo ácido nucleico copiado. En algunas realizaciones, el cebador puede incluir un cebador de secuenciación. En algunas realizaciones, los datos de secuenciación se obtienen usando el cebador de secuenciación. En más realizaciones, los adaptadores que comprenden sitios de cebador pueden ligarse a cada extremo de un ácido nucleico, y el ácido nucleico amplificarse a partir de tales sitios de cebador.

Algunas realizaciones para la preparación de un ácido nucleico molde pueden incluir secuencias de amplificación que comprenden al menos una porción de una o más secuencias de transposón y al menos una porción de un ácido nucleico diana. En algunas realizaciones, al menos una porción de un ácido nucleico diana puede amplificarse usando cebadores que hibridan con sitios de cebador de secuencias de transposón integradas, integradas en un ácido nucleico diana. En algunas de tales realizaciones, un ácido nucleico amplificado puede incluir una primera secuencia de código de barras y una segunda secuencia de código de barras que tiene al menos una porción del ácido nucleico diana dispuesta entre ellos. En algunas realizaciones, al menos un ácido nucleico amplificado puede incluir al menos una primera secuencia de código de barras de un primer ácido nucleico amplificado que puede identificarse como emparejado con una segunda secuencia de código de barras de una segunda secuencia amplificada.

Algunos métodos para preparar ácidos nucleicos molde incluyen insertar secuencias de transposón que comprenden enlazadores monocatenarios. La FIG. 3 ilustra un ejemplo en el que las secuencias de transposón (ME-P1-enlazador-P2-ME; extremo de mosaico-sitio de cebador 1-enlazador-sitio de cebador 2-extremo de mosaico) se insertan en un ácido nucleico diana. El ácido nucleico diana que tiene las secuencias de transposón/enlazador insertadas puede extenderse y amplificarse.

En una realización de las composiciones y métodos descritos en el presente documento, se utilizan transposomas que tienen secuencias finales transponibles simétricas para producir un fragmento de ácido nucleico diana etiquetado en el extremo (fragmento sometido a etiquetación o etiquetado). Cada fragmento sometido a etiquetación, por lo tanto, contiene extremos idénticos, que carecen de direccionalidad. Después puede emplearse una única PCR con cebadores, utilizando las secuencias finales de transposón, para amplificar el número de copias de molde de 2n a 2n*2x, donde x corresponde al número de ciclos de PCR. En una etapa posterior, una PCR con cebadores puede añadir secuencias adicionales, tales como secuencias adaptador de secuenciación.

En algunas realizaciones, puede ser ventajoso que cada ácido nucleico molde incorpore al menos un sitio de cebador universal. Por ejemplo, un ácido nucleico molde puede incluir primeras secuencias finales que comprenden un primer sitio de cebador universal, y segundas secuencias finales que comprenden un segundo sitio de cebador universal. Los sitios de cebador universal pueden tener diversas aplicaciones, tales como el uso en amplificación, secuenciación y/o identificación de uno o más ácidos nucleicos molde. El primero y segundo sitios de cebador universal pueden ser los mismos, sustancialmente similares, similares o distintos. Los sitios de cebador universal pueden introducirse en ácidos nucleicos mediante diversos métodos bien conocidos en la técnica, por ejemplo, ligamiento de sitios de cebador con ácidos nucleicos, amplificación de ácidos nucleicos utilizando cebadores con cola e inserción de una secuencia de transposón que comprende un sitio de cebador universal.

Inserción dirigida

En algunas realizaciones de los métodos y composiciones proporcionados en el presente documento, las secuencias de transposón pueden insertarse en secuencias dirigidas particulares de un ácido nucleico diana. La transposición en ADNbc puede ser más eficaz que en dianas de ADNmc. En algunas realizaciones, el ADNbc se

desnaturaliza en ADNmc y se aparea con sondas oligonucleotídicas (20-200 bases). Estas sondas crean sitios de ADNbc que se pueden usar de manera eficaz como sitios de integración con los transposomas que se proporcionan en el presente documento. En algunas realizaciones, el ADNbc puede tenerse como objetivo utilizando una formación de bucle D con sondas oligonucleotídicas recubiertas con recA y la posterior formación de tríplex. En algunas de tales realizaciones, la estructura de horquilla de replicación es el sustrato preferente para los transposomas que comprenden la transposasa Tn4430. En más realizaciones, las regiones de interés en el ADNbc pueden tenerse como objetivo utilizando proteínas de unión a ADN específicas de secuencia, tales como los complejos de dedos de zinc y otros ligandos de afinidad por regiones de ADN específicas.

En algunas realizaciones, se pueden usar para dirigir la inserción en el ácido nucleico diana transposomas que comprenden una transposasa que tiene un sustrato preferente de posiciones desapareadas en un ácido nucleico diana. Por ejemplo, algunas transposasas MuA, tales como HYPERMU (Epicenter), tienen una preferencia por dianas desapareadas. En algunas de tales realizaciones, las sondas oligonucleotídicas que comprenden un desapareamiento se aparean con un ácido nucleico diana monocatenario. Se pueden usar transposomas que comprenden transposasas MuA, tales como HYPERMU, para dirigirse a las secuencias desapareadas del ácido nucleico diana.

Ácidos nucleicos molde de fragmentación

Algunas realizaciones para la preparación de un ácido nucleico molde pueden incluir la fragmentación de un ácido nucleico diana. En algunas realizaciones, la inserción de transposomas que comprenden secuencias de transposón no contiguas puede dar como resultado la fragmentación de un ácido nucleico diana. En algunas realizaciones que comprenden transposomas en bucle, un ácido nucleico diana que comprende secuencias de transposón puede fragmentarse en los sitios de fragmentación de las secuencias de transposón. Ejemplos adicionales de métodos útiles para fragmentar ácidos nucleicos diana útiles con las realizaciones proporcionadas en el presente documento se pueden encontrar en, por ejemplo, la solicitud de patente de Estados Unidos Pub. n.º 2012/0208705, la solicitud de patente de Estados Unidos Pub. No. 2012/0208724 y solicitud de patente Int. Pub. n.º WO 2012/061832.

Etiquetado de moléculas individuales

La presente divulgación proporciona métodos para etiquetar moléculas de forma que las moléculas individuales puedan rastrearse e identificarse. Después, los datos masivos se pueden desconvolucionar y convertir de nuevo a la molécula individual. La capacidad para distinguir moléculas individuales y de relacionar la información con la molécula de origen es especialmente importante cuando los procedimientos desde la molécula original hasta el producto final cambian la representación (estequiométrica) de la población original. Por ejemplo, la amplificación conduce a la duplicación (por ejemplo, duplicados de PCR o amplificación sesgada) que puede desviar la representación original. Esto puede alterar la identificación del estado de metilación, el número de copias, la proporción alélica debido a la amplificación no uniforme y/o el sesgo de amplificación. Al identificar moléculas individuales, el etiquetado con códigos distingue entre moléculas idénticas después del procesamiento. Por tanto, las duplicaciones y el sesgo de amplificación se pueden filtrar, permitiendo una determinación precisa de la representación original de una molécula o población de moléculas.

Una ventaja de etiquetar de forma distintiva moléculas individuales es que las moléculas idénticas en el grupo original se identifican de forma distintiva en virtud de su etiquetado. En análisis adicionales posteriores, estas moléculas etiquetadas de forma distintiva ahora se pueden distinguir. Esta técnica se puede explotar en esquemas de ensayo en los que se emplea amplificación. Por ejemplo, se sabe que la amplificación distorsiona la representación original de una población mixta de moléculas. Si no se empleara el etiquetado distintivo, la representación original (tal como el número de copias o la proporción alélica) debería tener en cuenta los sesgos (conocidos o desconocidos) para cada molécula en la representación. Con el etiquetado distintivo, la representación se puede determinar con precisión eliminando duplicados y contando la representación original de las moléculas, cada una teniendo una etiqueta distintiva. Por lo tanto, los ADNc pueden amplificarse y secuenciarse, sin temor al sesgo, debido a que los datos se pueden filtrar de forma que solo se seleccionen las secuencias auténticas o las secuencias de interés para un análisis adicional. Las lecturas precisas se pueden construir tomando el consenso entre muchas lecturas con el mismo código de barras.

En algunas realizaciones de las composiciones y métodos descritos en el presente documento, es preferente etiquetar la población original en las fases tempranas del ensayo, aunque el etiquetado puede producirse en fases posteriores si las etapas tempranas no introducen sesgos o no son importantes. En cualquiera de estas aplicaciones, la complejidad de las secuencias de códigos de barras debe ser mayor que la cantidad de moléculas individuales que se deben etiquetar. Esto asegura que distintas moléculas diana reciban etiquetas distintas y distintivas. Por tanto, es conveniente un conjunto de oligonucleótidos aleatorios de una determinada longitud (por ejemplo, 5, 10, 20, 30, 40, 50, 100 o 200 nucleótidos de longitud). Un conjunto aleatorio de etiquetas representa una gran complejidad de etiquetas con espacio de código de 4" donde n es el número de nucleótidos. Se pueden incorporar códigos adicionales (ya sean diseñados o aleatorios) en distintas fases, para que sirvan como un control adicional, tal como un control de paridad para la corrección de errores.

En una realización de las composiciones y métodos descritos en el presente documento, las moléculas individuales (tal como el ADN diana) están unidas a etiquetas distintivas, tales como secuencias oligonucleotídicas y/o códigos de barras distintivos. La unión de las etiquetas puede producirse a través de ligamiento, química de acoplamiento, adsorción, inserción de secuencias de transposón, etc. Otros medios incluyen amplificación (tal como por PCR, RCA (forma siglada de *rolling circle amplification*) o LCR), copia (tal como la adición por una polimerasa) e interacciones no covalentes.

Los métodos específicos comprenden la inclusión de códigos de barras (por ejemplo, secuencias diseñadas o aleatorias) a los cebadores de PCR, de forma que cada molde recibirá un código individual dentro del espacio de código, obteniendo de este modo amplicones únicos que pueden discriminarse de otros amplicones. Este concepto se puede aplicar a cualquier método que use amplificación por polimerasa, tal como los ensayos GoldenGate como se divulga en las patentes de Estados Unidos N.º 7.582.420, N.º 7.955.794 y N.º 8.003.354. Las secuencias diana etiquetadas con código se pueden circularizar y amplificar por métodos tales como la amplificación por círculo rodante, para producir amplicones etiquetados con códigos. De forma similar, el código también se puede añadir ARN

Métodos de análisis de ácidos nucleicos molde

Algunas realizaciones de la tecnología descrita en el presente documento incluyen métodos de análisis de ácidos nucleicos molde. En tales realizaciones, la información de secuenciación se puede obtener a partir de ácidos nucleicos molde y esta información se puede usar para generar una representación de secuencia de uno o más ácidos nucleicos diana.

En algunas realizaciones de los métodos de secuenciación descritos en el presente documento, se puede usar una estrategia de lectura vinculada. Una estrategia de lectura vinculada puede incluir la identificación de datos de secuenciación que vincula al menos dos lecturas de secuenciación. Por ejemplo, una primera lectura de secuenciación puede contener un primer marcador y una segunda lectura de secuenciación puede contener un segundo marcador. El primero y segundo marcadores pueden identificar los datos de secuenciación procedentes de cada lectura de secuenciación como adyacentes en una representación de secuencia del ácido nucleico diana. En algunas realizaciones de las composiciones y métodos descritos en el presente documento, los marcadores pueden comprender una primera secuencia de código de barras y una segunda secuencia de código de barras, en las que la primera secuencia de código de barras puede emparejarse con la segunda secuencia de código de barras. En otras realizaciones, los marcadores pueden comprender una primera etiqueta de hospedador y una segunda etiqueta de hospedador. En más realizaciones, los marcadores pueden comprender una primera secuencia de código de barras con una primera etiqueta de hospedador y una segunda secuencia de código de barras con una segunda etiqueta de hospedador.

Una realización ejemplar de un método para secuenciar un ácido nucleico molde puede comprender las siguientes etapas: (a) secuenciar la primera secuencia de código de barras utilizando un cebador de secuenciación que hibrida con el primer sitio de cebador; y (b) secuenciar la segunda secuencia de código de barras utilizando un cebador de secuenciación que hibrida con el segundo cebador. El resultado son dos lecturas de secuencia que ayudan a vincular el ácido nucleico molde con sus vecinos genómicos. Dadas las lecturas suficientemente largas y los fragmentos de biblioteca lo suficientemente cortos, estas dos lecturas se pueden combinar mediante programas informáticos para hacer una lectura larga que cubra todo el fragmento. Usando las lecturas de secuencias de códigos de barras y la secuencia duplicada de 9 nucleótidos presente desde la inserción, las lecturas ahora se pueden vincular a sus vecinos genómicos para formar "lecturas vinculadas" mucho más largas *in silico*.

Como se entenderá, una biblioteca que comprende ácidos nucleicos molde puede incluir fragmentos de ácido nucleico duplicados. La secuenciación de fragmentos de ácido nucleico duplicados es ventajosa en los métodos que incluyen la creación de una secuencia de consenso para fragmentos duplicados. Dichos métodos pueden aumentar la precisión para proporcionar una secuencia de consenso para un ácido nucleico molde y/o biblioteca de ácidos nucleicos molde.

En algunas realizaciones de la tecnología de secuenciación descrita en el presente documento, el análisis de secuencia se realiza en tiempo real. Por ejemplo, la secuenciación en tiempo real se puede realizar mediante la adquisición y el análisis simultáneos de los datos de secuenciación. En algunas realizaciones, un procedimiento de secuenciación para obtener datos de secuenciación puede terminarse en diversos puntos, incluso después de obtener al menos una parte de los datos de la secuencia de un ácido nucleico diana o antes de secuenciar la lectura completa del ácido nucleico. Se proporcionan métodos ejemplares, sistemas y realizaciones adicionales en la publicación de patente internacional N.º WO 2010/062913.

En una realización ejemplar de un método para ensamblar lecturas de secuenciación cortas usando una estrategia de lectura vinculada, las secuencias de transposón que comprenden códigos de barras se insertan en ADN genómico, se prepara una biblioteca y se obtienen datos de secuenciación para la biblioteca de ácidos nucleicos molde. Se pueden ensamblar bloques de moldes identificando códigos de barras emparejados y después se ensamblan cóntigos más grandes. En una realización, las lecturas ensambladas se pueden ensamblar

adicionalmente en cóntigos más grandes a través del emparejamiento de códigos, utilizando lecturas solapantes.

Algunas realizaciones de la tecnología de secuenciación descritas el presente documento incluyen características de detección y corrección de errores. Los ejemplos de errores pueden incluir errores en la identificación de bases durante un procedimiento de secuenciación y errores en el ensamblaje de fragmentos en cóntigos más grandes. Como se entendería, la detección de errores puede incluir detectar la presencia o la probabilidad de errores en un conjunto de datos, y por tanto, detectar la ubicación de un error o la cantidad de errores puede no ser necesario. Para la corrección de errores, es útil la información respecto a la ubicación de un error y/o la cantidad de errores en un conjunto de datos. Los métodos para la corrección de errores son bien conocidos en la técnica. Los ejemplos incluyen el uso de distancias de Hamming y el uso de un algoritmo de suma de verificación (véase, por ejemplo, la solicitud de patente de Estados Unidos publicación N.º 2010/0323348; la patente de Estados Unidos N.º 7.574.305; y la Patente de Estados Unidos N.º 6.654.696).

Bibliotecas anidadas

Un método alternativo implica los métodos de etiquetado de unión anteriores y la preparación de bibliotecas de secuenciación anidadas. Las sub-bibliotecas anidadas se crean a partir de fragmentos de ADN etiquetados con códigos. Esto puede permitir el etiquetado de transposones menos frecuente en el genoma. Además, puede crear una mayor diversidad de lecturas de secuenciación (anidadas). Estos factores pueden conducir a una cobertura y precisión mejoradas.

El submuestreo y la amplificación del genoma completo pueden crear muchas copias de una determinada población de moléculas de partida. Después, se generan fragmentos de ADN por fragmentación específica por transposón, donde cada fragmento recibe un código que permite vincular el fragmento de nuevo al vecino original, que tiene un código coincidente (ya sea idéntico, complementario o vinculado de otra manera mediante programas informáticos). Los fragmentos etiquetados se fragmentan al menos una segunda vez por métodos aleatorios o métodos específicos de secuencia, tales como digestión enzimática, corte al azar, corte basado en transposones u otros métodos, creando de este modo sub-bibliotecas de los fragmentos de ADN etiquetados con códigos. En una variación útil del método descrito anteriormente, los fragmentos etiquetados con códigos se pueden aislar preferentemente utilizando transposones que contienen funcionalidad de biotina u otra de afinidad para fines de enriquecimiento cadena abajo. La preparación posterior de la biblioteca convierte los fragmentos de ADN anidados en moldes de secuenciación. La secuenciación de extremos emparejados da como resultado la determinación de la secuencia de la etiqueta de código de los fragmentos de ADN y del ADN diana. Dado que se crean bibliotecas anidadas para la misma etiqueta de código, los fragmentos largos de ADN se pueden secuenciar con lecturas cortas.

Métodos de secuenciación

Los métodos y la composición descritos en el presente documento se pueden usar junto con una diversidad de técnicas de secuenciación. En algunas realizaciones, el procedimiento para determinar la secuencia de nucleótidos de un ácido nucleico diana puede ser un proceso automatizado.

Algunas realizaciones de los métodos de secuenciación descritos el presente documento incluyen tecnologías de secuenciación por síntesis (SBS, forma siglada del inglés *sequencing by synthesis*), por ejemplo, técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPi) a medida que se incorporan nucleótidos particulares a la cadena naciente (Ronaghi *et al*., Analytical Biochemistry 242(1): 84-9 (1996); Ronaghi, M. Genome Res. 11(1):3-11 (2001); Ronaghi *et al*., Science 281(5375):363 (1998); la patente de Estados Unidos N.º 6.210.891; la patente de Estados Unidos N.º 6.258.568 y la patente de Estados Unidos N.º 6.274.320).

En otro tipo ejemplar de SBS, la secuenciación de ciclo se lleva a cabo mediante la adición escalonada de nucleótidos terminadores reversible que contienen, por ejemplo, un marcador colorante escindible o fotoblanqueable como se describe, por ejemplo, en la patente de Estados Unidos N.º 7.427.67, patente de Estados Unidos N.º 7.414.1163 y patente de Estados Unidos N.º 7.057.026. Este enfoque, que está siendo comercializado por Illumina Inc., también se describe en las solicitudes de patente internacional n.º de publicación WO 91/06678 y WO 07/123744. La disponibilidad de terminadores marcados con fluorescencia, en los que la terminación se puede invertir y el marcador fluorescente escindir, facilita una secuenciación de terminación reversible cíclica eficaz (CRT, forma siglada del inglés *cyclic reversible termination*). Las polimerasas también se pueden codiseñar técnicamente para incorporar y extenderse desde estos nucleótidos modificados.

Los sistemas SBS ejemplares adicionales y métodos que se pueden utilizar con los métodos y composiciones descritos en el presente documento se describen en la solicitud de patente de Estados Unidos publicación N.º 2007/0166705, la solicitud de patente de Estados Unidos publicación N.º 2006/0188901, la patente de Estados Unidos N.º 7057026, la solicitud de patente de Estados Unidos publicación N.º 2006/0240439, la solicitud de patente de Estados Unidos publicación N.º 2006/0281109, la publicación PCT N.º WO 05/065814, la solicitud de patente de Estados Unidos publicación N.º 2005/0100900, la publicación PCT N.º WO 06/064199 y la publicación PCT N.º WO 07/010251.

Algunas realizaciones de la tecnología de secuenciación descritas el presente documento pueden utilizar técnicas de secuenciación por ligamiento. Dichas técnicas utilizan la ADN ligasa para incorporar nucleótidos e identificar la incorporación de tales nucleótidos. Los sistemas y métodos de SBS ilustrativos que se pueden utilizar con los métodos y composiciones descritos en el presente documento se describen en la patente de Estados Unidos N.º 6.969.488, la patente de Estados Unidos N.º 6.172.218 y la patente de Estados Unidos N.º 6.306.597.

Algunas realizaciones de la tecnología de secuenciación descritas en el presente documento pueden incluir técnicas tales como las tecnologías de *next-next*. Un ejemplo puede incluir técnicas de secuenciación por nanoporos (Deamer, D.W. y Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing". Trends Biotechnol. 18, 147-151 (2000); Deamer, D. y D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li *et al.*, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). En tales realizaciones, el ácido nucleico diana pasa a través de un nanoporo. El nanoporo puede ser un poro sintético o una proteína de membrana biológica, tal como α-hemolisina. A medida que el ácido nucleico atraviesa el nanoporo, cada pareja de bases se puede identificar midiendo las fluctuaciones en la conductancia eléctrica del poro. (Patente de Estados Unidos n.º 7.001.792; Soni y Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores". Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis". Nanomed. 2:459-481 (2007); Cockroft *et al.*, "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution". J. Am. Chem. Soc. 130:818-820 (2008)). En algunas de tales realizaciones, las técnicas de secuenciación por nanoporos pueden ser útiles para confirmar la información de secuencia generada por los métodos descritos en el presente documento.

Algunas realizaciones de la tecnología de secuenciación descritas en el presente documento pueden utilizar métodos que implican el control en tiempo real de la actividad de la ADN polimerasa. Las incorporaciones de nucleótidos se pueden detectar mediante interacciones transferencia de energía de resonancia de fluorescencia (FRET, forma siglada del inglés *fluorescence resonance energy transfer*) entre una polimerasa que porta un fluoróforo y nucleótidos marcados con γ-fosfato, como se describe, por ejemplo, en la patente de Estados Unidos N.º 7.329.492 y la patente de Estados Unidos N.º 7.211.414, o las incorporaciones de nucleótidos pueden detectarse con ondas guía en modo cero como se describe, por ejemplo, en la patente de Estados Unidos N.º 7.315.019 y usando análogos de nucleótidos fluorescentes y polimerasas técnicamente diseñadas como se describe, por ejemplo, en la patente de Estados Unidos N.º 7.405.281 y la solicitud de patente publicación N.º 2008/0108082. La iluminación se puede restringir a un volumen en la escala de zeptolitros alrededor de una polimerasa anclada a la superficie de forma que la incorporación de los nucleótidos marcados de forma fluorescente se puede observar con un fondo bajo (Levene, M.J. *et al.* "Zeromode waveguides for single-molecule analysis at high concentrations". Science 299, 682-686 (2003); Lundquist, P.M. *et al.* "Parallel confocal detection of single molecules in real time". Opt. Lett. 33, 1026-1028 (2008); Korlach, J. *et al.* "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures". Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). En un ejemplo, la tecnología de secuenciación de ADN de molécula única en tiempo real (SMRT, forma siglada del inglés *single molecule, real-time*) proporcionada por Pacific Biosciences Inc. se puede utilizar con los métodos descritos en el presente documento. En algunas realizaciones, se puede utilizar un chip de SMRT o similar (por ejemplo, la patente de Estados Unidos N.º 7.181.122, patente de Estados Unidos N.º 7.302.146 y la patente de Estados Unidos N.º 7.313.308). Un chip de SMRT comprende una pluralidad de ondas guía en modo cero (ZMW, forma siglada del inglés *zero-mode waveguides*). Cada ZMW comprende un orificio cilíndrico de decenas de nanómetros de diámetro que perfora una película delgada de metal sustentado por un sustrato transparente. Cuando la ZMW se ilumina a través del sustrato transparente, la luz atenuada puede penetrar en los 20-30 nm inferiores de cada ZMW creando un volumen de detección de aproximadamente 1 x 10-21 l. Los volúmenes de detección más pequeños aumentan la sensibilidad de la detección de señales fluorescentes al reducir la cantidad de fondo que se puede observar.

Los chips de SMRT y tecnologías similares se pueden usar en asociación con monómeros nucleotídicos marcados con fluorescencia en el fosfato terminal del nucleótido (Korlach J. *et al.*, "Long, processive enzymatic DNA synthesis using 100 % dye-labeled terminal phosphate-linked nucleotides." Nucleosides, Nucleotides and Nucleic Acids, 27:1072-1083, 2008). El marcador se escinde del monómero nucleotídico en la incorporación del nucleótido al polinucleótido. Por consiguiente, el marcador no se incorpora al polinucleótido, aumentando la proporción señal: fondo. Además, se reduce la necesidad de condiciones para escindir un marcador de los monómeros nucleotídicos marcados.

Un ejemplo adicional de una plataforma de secuenciación que puede usarse en asociación con algunas de las realizaciones descritas en el presente documento lo proporciona Helicos Biosciences Corp. En algunas realizaciones, se puede utilizar la secuenciación de molécula única verdadera (Harris T.D. *et al.*, "Single Molecule DNA Sequencing of a viral Genome" Science 320:106-109 (2008)). En una realización, se puede preparar una biblioteca de ácidos nucleicos diana mediante la adición de una cola de poli(A) 3' a cada ácido nucleico diana. La cola de poli(A) hibrida con los oligonucleótidos de poli(T) anclados en un cubreobjetos. El oligonucleótido de poli(T) se puede usar como cebador para la extensión de un polinucleótido complementario al ácido nucleico diana. En una realización, se suministran monómeros nucleotídicos marcados con fluorescencia, concretamente, A, C, G o T, uno a la vez, al ácido nucleico diana en presencia de ADN polimerasa. Se detecta la incorporación de un nucleótido marcado en el polinucleótido complementario al ácido nucleico diana y la posición de la señal fluorescente en el

cubreobjetos indica la molécula que se ha extendido. El marcador fluorescente se elimina antes de añadir el siguiente nucleótido para continuar el ciclo de secuenciación. El seguimiento de la incorporación de nucleótidos en cada cadena de polinucleótidos puede proporcionar información de secuencia para cada ácido nucleico diana individual.

Un ejemplo adicional de una plataforma de secuenciación que puede usarse en asociación con los métodos descritos en el presente documento lo proporciona Complete Genomics Inc. Se pueden preparar bibliotecas de ácidos nucleicos diana donde las secuencias de ácido nucleico diana se intercalan aproximadamente cada 20 pb con secuencias adaptadoras. Los ácidos nucleicos diana pueden amplificarse utilizando la replicación por círculo rodante y los ácidos nucleicos diana amplificados pueden usarse para preparar una matriz de ácidos nucleicos diana. Los métodos de secuenciación de tales matrices incluyen la secuenciación por ligamiento, en particular, secuenciación por ligamiento sonda-anclaje combinatorio (cPAL, forma sigada de *combinatorial probe-anchor ligation*).

En algunas realizaciones usando cPAL, se pueden determinar aproximadamente 10 bases contiguas adyacentes a un adaptador. Se usa un conjunto de sondas que incluye cuatro marcadores distintos para cada base (A, C, T, G) para leer las posiciones adyacentes a cada adaptador. Un grupo separado se utiliza para leer cada posición. Se suministra al ácido nucleico diana un conjunto de sondas y un anclaje específico para un adaptador particular en presencia de ligasa. El anclaje hibrida con el adaptador y una sonda hibrida con el ácido nucleico diana adyacente al adaptador. El anclaje y la sonda están ligados entre sí. Se detecta la hibridación y se elimina el complejo anclaje-sonda. Después, se suministra un anclaje y un conjunto de sondas distintos al ácido nucleico diana en presencia de ligasa.

Los métodos de secuenciación descritos en el presente documento se pueden llevar a cabo ventajosamente en formatos de múltiples de forma que múltiples ácidos nucleicos diana distintos se manipulan simultáneamente. En realizaciones particulares, se pueden tratar distintos ácidos nucleicos en un recipiente de reacción común o sobre una superficie de un sustrato particular. Esto permite un suministro conveniente de los reactivos de secuenciación, la eliminación de reactivos que no han reaccionado y la detección de eventos de incorporación de una forma múltiple. En realizaciones que utilizan ácidos nucleicos diana unidos a superficies, los ácidos nucleicos diana pueden estar en un formato de matriz. En un formato de matriz, los ácidos nucleicos diana pueden estar acoplados normalmente a una superficie en una manera distinguible desde el punto de vista espacial. Por ejemplo, los ácidos nucleicos diana pueden estar unidos mediante una unión covalente directa, unión a una perla u otra partícula, o asociados a una polimerasa u otra molécula que está unida a la superficie. La matriz puede incluir una copia única de un ácido nucleico diana en cada sitio (también denominado como característica) o múltiples copias que tienen la misma secuencia pueden estar presente en cada sitio o característica. Se pueden producir múltiples copias mediante métodos de amplificación tales como, amplificación puente o PCR de emulsión, como se describe con más detalle en el presente documento.

Los métodos expuestos en el presente documento pueden usar matrices que tienen características en cualquiera de una diversidad de densidades que incluyen, por ejemplo, al menos aproximadamente 10 características/cm^2, 100 características/cm^2, 500 características/cm^2, 1.000 características/cm^2, 5.000 características/cm^2, 10.000 características/cm^2, 50.000 características/cm^2, 100.000 características/cm^2, 1.000.000 características/cm^2, 5.000.000 características/cm^2, 10 características/cm^2, $5x10^7$ características/cm^2, 10^8 características/cm^2, $5x\ 10^8$ características/cm^2, 10^9 características/cm^2, $5x\ 10^9$ características/cm^2 o más.

Superficies

En algunas realizaciones, el molde de ácido nucleico proporcionado en el presente documento se puede unir a un soporte sólido ("sustrato"). Los sustratos pueden ser de bi- o tridimensionales y pueden comprender una superficie plana (por ejemplo, un portaobjetos) o se pueden conformar. Un sustrato puede incluir vidrio (por ejemplo, vidrio de poro controlado (CPG, forma siglada del inglés *controlled pore glass*)), cuarzo, plástico (tal como poliestireno (poliestireno de baja reticulación y alta reticulación), policarbonato, polipropileno y poli(metilmetacrilato)), copolímero acrílico, poliamida, silicio, metal (por ejemplo, oro derivatizado con alcanotiolato), celulosa, nailon, látex, dextrano, matriz de gel (por ejemplo, gel de sílice), poliacroleína o materiales compuestos.

Los sustratos tridimensionales adecuados incluyen, por ejemplo, esferas, micropartículas, perlas, membranas, portaobjetos, placas, chips micromaquinados, tubos (por ejemplo, tubos capilares), micropocillos, dispositivos microfluídicos, canales, filtros o cualquier otra estructura adecuada para anclar un ácido nucleico. Los sustratos pueden incluir matrices planas o matrices capaces de tener regiones que incluyen poblaciones de ácidos nucleicos molde o cebadores. Los ejemplos incluyen CPG derivatizado con nucleósidos y portaobjetos de poliestireno; portaobjetos magnéticas derivatizados; poliestireno injertado con polietilenglicol y similares. Pueden usarse diversos métodos bien conocidos en la técnica para unir, anclar o inmovilizar ácidos nucleicos a la superficie del sustrato.

Métodos para reducir las tasas de error en los datos de secuenciación

Algunas realizaciones de los métodos y composiciones proporcionadas en el presente documento incluyen reducir

las tasas de error en los datos de secuenciación. En algunas de tales realizaciones, las cadenas sentido y antisentido de un ácido nucleico diana bicatenario están asociadas cada una con un código de barras distinto. Cada cadena se amplifica, la información de secuencia se obtiene a partir de múltiples copias de las cadenas amplificadas y se genera una representación de la secuencia consenso del ácido nucleico diana a partir de la información de secuencia redundante. Por lo tanto, la información de secuencia puede originarse e identificarse a partir de cada cadena. Por consiguiente, los errores de secuencia pueden identificarse y reducirse cuando la información de secuencia que se origina en una cadena es inconsistente con la información de secuencia de la otra cadena.

En algunas realizaciones, las cadenas sentido y antisentido de un ácido nucleico diana están asociadas con un código de barras distinto. Los códigos de barras pueden asociarse con el ácido nucleico diana mediante una diversidad de métodos que incluyen ligamiento de adaptadores e inserción de secuencias de transposón. En algunas de tales realizaciones, un adaptador en Y puede ligarse a al menos un extremo de un ácido nucleico diana. El adaptador en Y puede incluir una secuencia bicatenaria y cadenas no complementarias, comprendiendo cada cadena un código de barras distinto. El ácido nucleico diana con el adaptador en Y ligado se puede amplificar y secuenciar de forma tal que cada código de barras pueda usarse para identificar las cadenas originales sentido o antisentido. Un método similar se describe en Kinde I. *et al.*, (2011) PNAS 108:9530-9535. En algunas realizaciones, las cadenas sentido y antisentido de un ácido nucleico diana se asocian con un código de barras distinto, insertando las secuencias de transposón proporcionadas en el presente documento. En algunas de tales realizaciones, las secuencias de transposón pueden comprender códigos de barras no complementarios.

Algunas realizaciones de tales métodos incluyen la obtención de información de secuencia de una cadena de un ácido nucleico bicatenario diana, que comprende (a) obtener datos de secuencia de un ácido nucleico molde que comprende un primer adaptador de secuenciación y un segundo adaptador de secuenciación que tiene al menos una porción del ácido nucleico diana bicatenario dispuesta entre ellos, en donde: (i) el primer adaptador de secuenciación comprende un primer código de barras bicatenario, un primer sitio de cebador monocatenario y un segundo sitio de cebador monocatenario, en donde el primer y segundo sitios de cebador no son complementarios y (ii) el segundo adaptador de secuenciación comprende un segundo código de barras bicatenario, un tercer sitio de cebador monocatenario y un cuarto sitio de cebador monocatenario, en donde los tercero y cuarto sitios de cebador no son complementarios. En algunas realizaciones, el primer sitio de cebador de la cadena sentido del ácido nucleico molde y el tercer sitio de cebador de la cadena antisentido del ácido nucleico molde comprenden la misma secuencia. En algunas realizaciones, cada código de barras es distinto. En algunas realizaciones, el primer adaptador de secuenciación comprende una horquilla monocatenaria que acopla el primer sitio de cebador y el segundo sitio del cebador.

En otra realización, cada extremo de un ácido nucleico diana está asociado con un adaptador que comprende un código de barras distinto, de forma que los productos de extensión de la cadena sentido y antisentido de un ácido nucleico pueden distinguirse uno del otro. En algunas realizaciones, las secuencias del sitio de cebador y los códigos de barras se seleccionan de forma que la extensión a partir de un cebador apareado con la cadena sentido produzca productos que puedan distinguirse de los productos de extensión a partir de un cebador apareado con la cadena antisentido. En un ejemplo, el sitio de cebador sentido 3' es el mismo que el sitio de cebador antisentido 3', pero distinto de los sitios de cebador sentido 5' y antisentido 5'. La extensión de los cebadores apareados con el sitio de cebador sentido 3' y el sitio de cebador antisentido 3' produciría los siguientes productos a partir de cada cadena:

Cadena codificante: (5') código de barras 2 - [secuencia diana] - código de barras 1 (3')
Cadena antisentido: (5') código de barras 1 - [secuencia diana] - código de barras 2 (3')

Por lo tanto, se pueden distinguir entre sí los productos de extensión a partir de la cadena sentido y antisentido de un ácido nucleico. Un método ejemplar se ilustra en Schmitt M.W., *et al.*, PNAS (2012) 109:14508-13. En algunos de tales métodos, los códigos de barras y los sitios de cebador pueden asociarse con el ácido nucleico diana mediante una diversidad de métodos que incluyen ligamiento de adaptadores e inserción de secuencias de transposón. En algunas realizaciones, las secuencias de transposón pueden diseñarse para proporcionar adaptadores con horquillas. Las horquillas proporcionan la capacidad de mantener la contigüidad física de las cadenas sentido y antisentido de un ácido nucleico diana. Puede prepararse un ácido nucleico molde que comprenda horquillas utilizando secuencias de transposón que comprenden enlazadores descritos en el presente documento. Los ejemplos de enlazadores incluyen ácidos nucleicos monocatenarios.

Algunas realizaciones para la preparación de una biblioteca de ácidos nucleicos molde para obtener información de secuencia de cada cadena de un ácido nucleico diana bicatenario incluyen (a) proporcionar una población de transposomas que comprende una transposasa y una primera secuencia de transposón que comprende: (i) un primer sitio de reconocimiento de transposasa, un primer sitio de cebador y un primer código de barras, y (ii) una segunda secuencia de transposón que comprende un segundo sitio de reconocimiento de transposasa, un segundo sitio de cebador y un segundo código de barras, en donde la primera secuencia de transposón no es contigua con la segunda secuencia de transposón; y (b) poner en contacto los transposomas con un ácido nucleico bicatenario en condiciones tales que dichas primera y segundas secuencias de transposón se inserten en el ácido nucleico diana bicatenario, preparando de este modo una biblioteca de ácidos nucleicos molde para obtener información de secuencia de cada cadena del ácido nucleico diana bicatenario. En algunas realizaciones, la población de

transposomas comprende adicionalmente transposomas que comprenden una transposasa y una secuencia de transposón que comprende un tercer sitio de reconocimiento de transposasa y un cuarto sitio de reconocimiento de transposasa que tienen una secuencia de código de barras dispuesta entre ellos, comprendiendo dicha secuencia de código de barras un tercer código de barras y un cuarto código de barras que tienen un adaptador de secuencia dispuesto entre ellos, comprendiendo dicho adaptador de secuenciación un tercer sitio de cebador y un cuarto sitio de cebador que tiene un enlazador dispuesto entre ellos. En algunas realizaciones, el primer sitio de cebador de la cadena sentido del ácido nucleico molde y el tercer sitio de cebador de la cadena antisentido del ácido nucleico molde comprenden la misma secuencia. Algunas realizaciones también incluyen una etapa (c) que selecciona ácidos nucleicos molde que comprenden secuencias de transposón en donde la primera secuencia de transposón no es contigua a la segunda secuencia de transposón, y secuencias de transposón que comprenden un enlazador. En algunas realizaciones, el enlazador comprende una etiqueta de afinidad adaptada para unirse con una sonda de captura. En algunas realizaciones, la etiqueta de afinidad se selecciona del grupo que consiste en His, biotina y estreptavidina. En algunas realizaciones, cada código de barras es distinto. En algunas realizaciones, el enlazador comprende un ácido nucleico monocatenario. En algunas realizaciones, el ácido nucleico diana comprende ADN genómico.

Métodos de obtención de información de haplotipos

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen métodos para obtener información de haplotipos a partir de un ácido nucleico diana. La información de haplotipos puede incluir la determinación de la presencia o ausencia de distintas secuencias en locus especificados en un ácido nucleico diana, tal como un genoma. Por ejemplo, se puede obtener información de secuencia para copias maternas y paternas de un alelo. En un organismo poliploide, se puede obtener información de secuencia para al menos un haplotipo. Dichos métodos también son útiles para reducir la tasa de error en la obtención de información de secuencia a partir del ácido nucleico diana.

Generalmente, los métodos para obtener información de haplotipos incluyen la distribución de un ácido nucleico en uno o más compartimentos, de forma que cada compartimento comprenda una cantidad de ácido nucleico equivalente a aproximadamente un equivalente haploide del ácido nucleico, o equivalente a menos de aproximadamente un equivalente haploide del ácido nucleico. Después, se puede obtener la información de secuencia de cada compartimento, obteniendo de este modo información de haplotipos. La distribución del ácido nucleico molde en una pluralidad de recipientes aumenta la probabilidad de que un único recipiente incluya una única copia de un alelo o SNP, o que la información de la secuencia consenso obtenida de un único recipiente refleje la información de secuencia de un alelo o SNP. Como se entenderá, en algunas de tales realizaciones, un ácido nucleico molde se puede diluir antes de compartimentar el ácido nucleico molde en una pluralidad de recipientes. Por ejemplo, cada recipiente puede contener una cantidad de ácidos nucleicos diana igual a aproximadamente un equivalente haploide del ácido nucleico diana. En algunas realizaciones, un recipiente puede incluir menos de aproximadamente un equivalente haploide de un ácido nucleico diana.

Método de haplotipado con compartimentos virtuales

Algunos métodos para obtener información de haplotipos proporcionados en el presente documento incluyen el uso de compartimentos virtuales. Ventajosamente, algunos de tales métodos permiten que los compartimentos incluyan cantidades de ácidos nucleicos equivalentes a al menos uno o más equivalentes haploides. En otras palabras, tales métodos permiten el uso de mayores concentraciones de ácidos nucleicos en los compartimentos en comparación con otros métodos de haplotipado, aumentando de este modo la eficacia y los rendimientos de diversas manipulaciones.

En algunos métodos para obtener información de haplotipos con compartimentos virtuales, se compartimenta un ácido nucleico en una pluralidad de primeros recipientes y los ácidos nucleicos de cada compartimento están provistos de un primer índice; los ácidos nucleicos con el primer índice se combinan y después se compartimentan en una pluralidad de segundos recipientes, y los ácidos nucleicos de cada compartimento están provistos de un segundo índice. Se puede preparar un ácido nucleico molde realizando al menos 2, 3, 4, 5, 6, 7, 8, 9, 10 o más rondas de compartimentación, indexación y agrupamiento. De tal manera, se proporciona una pluralidad de índices distintos a un ácido nucleico molde en un método escalonado. Después de la indexación, los ácidos nucleicos molde indexados se pueden agrupar y distribuir en una pluralidad de compartimentos, de forma que cada compartimento incluya probablemente una cantidad de un ácido nucleico molde particular que tenga una combinación particular de índices que sea equivalente a aproximadamente un equivalente haploide del ácido nucleico diana, o equivalente a menos de aproximadamente un equivalente haploide del ácido nucleico diana, o equivalente a más de aproximadamente un equivalente de haploide. En otras palabras, cada recipiente puede recibir una cantidad de ácido nucleico molde que comprende más del equivalente de un equivalente haploide, sin embargo, es probable que cada copia de un alelo o SNP esté asociada con una combinación distinta de índices. Por consiguiente, puede reducirse el número de recipientes para compartimentar un ácido nucleico molde de forma que cada recipiente incluya aproximadamente una cantidad de ácido nucleico molde equivalente a un haploide o menos de un ácido nucleico diana. Además, la cantidad de ácido nucleico en cada recipiente puede ser mayor que la cantidad de aproximadamente un equivalente haploide, aumentando de este modo la eficacia y los rendimientos de diversas

manipulaciones.

Existen diversos métodos para indexar ácidos nucleicos. Por ejemplo, en algunas realizaciones, los índices pueden insertarse en los ácidos nucleicos utilizando transposomas proporcionados en el presente documento; los índices pueden ligarse a los ácidos nucleicos; y los índices pueden añadirse a los ácidos nucleicos durante el copiado, por ejemplo, la amplificación de un ácido nucleico. En algunas realizaciones, se puede preparar un ácido nucleico molde que comprenda un índice utilizando transposomas que comprenden una secuencia de transposón contigua. Véase, por ejemplo, el transposoma (50) en la **FIG. 1.** La inserción de secuencias de transposón contiguas puede tener como resultado la conservación de la información de la posición para una molécula de ácido nucleico particular después de la distribución de ácidos nucleicos molde entre varios compartimentos. En algunas realizaciones, se puede preparar un ácido nucleico molde que comprenda un índice utilizando transposomas que comprenden secuencias de transposón no contiguas. Véase, por ejemplo, el transposoma (10) en la **FIG. 1.** Los ejemplos de tales secuencias de transposón se exponen en la solicitud de patente de Estados Unidos número de publicación 2010/0120098. La inserción de secuencias de transposón no contiguas puede dar como resultado la fragmentación de una molécula de ácido nucleico particular. Por lo tanto, en algunas realizaciones, la inserción de secuencias de transposón no contiguas en un ácido nucleico molde puede reducir la información de la posición para una molécula de ácido nucleico particular después de la distribución de ácidos nucleicos molde entre varios compartimentos. En otras palabras, distintos fragmentos de una molécula de ácido nucleico particular pueden distribuirse en distintos recipientes.

En un ejemplo con un genoma diploide, después de agrupar y diluir en compartimentos, se puede añadir una mayor cantidad de ácidos nucleicos a cada compartimento, dado que la probabilidad de tener una copia del padre y una copia de la madre de la misma región con los mismos índices es menor. Por ejemplo, pueden estar presentes en el mismo compartimento una copia del padre y una copia de la madre para la misma región, siempre que cada una contenga un índice distinto, por ejemplo, uno proviene de una reacción de transposición con un primer índice (índice-1) y el otro proviene de una reacción de transposición con un primer índice distinto (índice-2). En otras palabras, pueden estar presentes en el mismo compartimento copias de la misma región/cromosoma, dado que se pueden distinguir por su índice único incorporado en la primera reacción de transposición. Esto permite que se distribuya más ADN en cada compartimento en comparación con los métodos de dilución alternativos. El esquema de indexación doble crea un número total de compartimentos virtuales de, número de reacciones de transposición indexadas iniciales multiplicados por el número de reacciones de PCR indexadas.

La **FIG. 4** representa una realización de ejemplo para obtener información de haplotipos utilizando compartimentos virtuales. Un ácido nucleico diana que comprende ADN genómico se distribuye en un primer conjunto de 96 recipientes y a los ácidos nucleicos de cada recipiente se les proporciona un primer índice distinto utilizando un transposón derivado de Tn5. Por lo tanto, se obtiene una pluralidad de ácidos nucleicos molde con el primer índice (por ejemplo, Tn5-1, Tn5-2...y Tn5-96). La pluralidad de ácidos nucleicos molde con el primer índice se combinan y después se redistribuyen en un segundo conjunto de 96 recipientes, y a los ácidos nucleicos de cada recipiente se les proporciona un segundo índice distinto mediante la amplificación de los ácidos nucleicos utilizando cebadores que comprenden los segundos índices. Por lo tanto, se obtiene una pluralidad de ácidos nucleicos molde con el segundo índice (por ejemplo, PCR1, PCR2...y PCR96). La pluralidad de ácidos nucleicos molde con el segundo índice se puede combinar y se puede obtener información de secuencia. El uso de recipientes físicos de 96 x 96 es equivalente a 9216 compartimentos virtuales.

Métodos para obtener información de haplotipos extendida

Como se describe anteriormente, la inserción de secuencias de transposón no contiguas en el ácido nucleico molde puede reducir la información de la posición para una molécula de ácido nucleico particular, por ejemplo, después de la distribución de los ácidos nucleicos molde entre varios compartimentos. Sin embargo, el solicitante ha descubierto métodos para conservar tal información de la posición para una molécula de ácido nucleico particular. Sin quedar ligado a teoría alguna, se ha observado que, después de la transposición, los dos fragmentos adyacentes resultantes de una molécula de ácido nucleico particular tenderán a estar distribuidos en el mismo recipiente en condiciones que mantienen la transposasa en el sitio de inserción de una secuencia de transposón. En otras palabras, la transposasa puede mantener juntos los dos fragmentos adyacentes resultantes de una molécula de ácido nucleico particular.

En algunas realizaciones, puede eliminarse una transposasa de un ácido nucleico molde después de distribuir el ácido nucleico molde en varios recipientes. Puede eliminarse una transposasa del sitio de una inserción mediante diversos métodos bien conocidos en la técnica, incluyendo la adición de un detergente, tal como SDS, cambiar la temperatura, por digestión con proteinasa, captura por chaperona y cambiando el pH. Las ADN polimerasas, con o sin propiedades de desplazamiento de cadena incluyendo, pero sin limitación, la ADN polimerasa phi29, la ADN polimerasa Bst, etc. también se pueden usar para retirar la transposasa del ADN.

La **FIG. 5** representa un esquema de ejemplo en que un ácido nucleico diana se distribuye en un conjunto de primeros recipientes y se indexa por inserción de transposomas, tales como transposomas que comprenden secuencias de transposón no contiguas. Los primeros ácidos nucleicos molde indexados se agrupan y distribuyen en

un conjunto de segundos recipientes y se indexan mediante amplificación por PCR. La información de secuencia se puede obtener a partir de los segundos ácidos nucleicos molde indexados.

Algunos métodos de obtención de información de haplotipos extendida a partir de un ácido nucleico diana incluyen (a) obtener un ácido nucleico molde que comprende una pluralidad de transposomas insertados en el ácido nucleico diana, en donde al menos alguno de los transposomas insertados comprenden cada uno una primera secuencia de transposón, una segunda secuencia de transposón no contigua a la primera secuencia de transposón, y una transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón; (b) compartimentar el ácido nucleico molde que comprende la pluralidad de transposomas insertados en cada recipiente de una pluralidad de recipientes; (c) eliminar la transposasa del ácido nucleico molde; y (d) obtener información de secuencia del ácido nucleico molde de cada recipiente, obteniendo de este modo información de haplotipos a partir del ácido nucleico diana. En algunas realizaciones, compartimentar el ácido nucleico molde incluye proporcionar a cada recipiente una cantidad de ácido nucleico molde equivalente a mayor de aproximadamente un equivalente haploide del ácido nucleico diana, una cantidad de ácido nucleico molde equivalente a aproximadamente un equivalente haploide del ácido nucleico diana, o una cantidad de ácido nucleico molde equivalente a menos de aproximadamente un equivalente haploide del ácido nucleico diana.

Una realización adicional para el mantenimiento de la contigüidad de los ácidos nucleicos diana para las aplicaciones de secuenciación comprende utilizar eventos transposicionales unilaterales (es decir, un extremo de transposón) en lugar de eventos transposicionales bilaterales (es decir, dos extremos de transposón), como se describe en el presente documento. Por ejemplo, se ha demostrado que las transposasas, incluyendo, pero sin limitación, Mu, mutante de MuE392Q, Tn5 presentan transposición unilateral de una secuencia de transposón en un ácido nucleico diana (Haapa *et al*., 1999, Nucl. Acids Res. 27(3): 2777-2784). El mecanismo transposicional unilateral de estas transposasas se puede utilizar en los métodos descritos en el presente documento para mantener la contigüidad de una muestra para secuenciación, por ejemplo, para haplotipar o para ensamblar un ácido nucleico diana.
En un ejemplo de transposición unilateral en un ADN diana, el transposoma, una transposasa dimérica Tn5, se asocia con un solo extremo de secuencia de transposón. En realizaciones preferentes, el extremo de transposón podría comprender adicionalmente secuencias adicionales tales como secuencias de índice, códigos de barras y/o secuencias de cebador y similares, que podrían usarse, por ejemplo, para identificar una muestra, amplificar o extender el ácido nucleico diana y alinear las secuencias de fragmentos. El complejo de transposoma se asocia con el ácido nucleico diana, en ese caso ADNbc. En el sitio de asociación del transposoma, la transposasa escinde esa cadena del ADN diana e inserta el transposón, y cualquier otra secuencia adicional en el punto de escisión. La transposasa permanece asociada con el ADN diana hasta que se elimina, por ejemplo, después de la partición de la muestra como se describe en el presente documento, la transposasa puede eliminarse por degradación (por ejemplo, con el uso de SDS u otros métodos como se describe en el presente documento). El ácido nucleico diana, en este caso ADNbc, no se fragmenta después de la eliminación de la transposasa, por tanto, el transposón y cualquier secuencia adicional pueden incorporarse en el ADN diana sin fragmentar el ADN. Una vez eliminada la transposasa, para crear bibliotecas para secuenciación se puede realizar una amplificación de la diana por cualquier medio conocido en la técnica, en este ejemplo, amplificación de un cebador única o múltiple (debido a la incorporación de secuenciación múltiple de cebadores distintos, incluidos en uno o más transposones) ya sea exponencial o lineal, tal como PCR dirigida o la amplificación del genoma completo (por ejemplo, por desplazamiento de múltiples cadenas). Como se describe en el presente documento con respecto a las secuencias de transposón bilaterales, podrían incluirse como parte de la secuencia de transposón, dependiendo de las necesidades del usuario, una diversidad de distintas combinaciones de índice, código de barras, sitio (o sitios) de endonucleasas de restricción y/o de secuencia de cebador. Por tanto, para los métodos divulgados en el presente documento para determinar el haplotipo de un ácido nucleico diana, también podrían utilizarse para mantener la contigüidad de un ácido nucleico diana los complejos de transposoma unilateral.

Los transposomas unilaterales también pueden crearse a partir de los dos complejos de transposón/transposasa o los complejos de transposón/transposasa en bucle descritos en el presente documento. Por ejemplo, una de las secuencias de transposón de un complejo de dos transposones o un extremo del transposón en bucle podrían, por ejemplo, modificarse químicamente o bloquearse de forma que la transposición no se produzca o que se produzca, mínimamente, en ese extremo. Por ejemplo, podría incorporarse al final de uno de los extremos del transposón un didesoxinucleótido, un hapteno tal como una biotina, lo que inhibiría la transposición en ese extremo, permitiendo de este modo que solo un transposón, o un extremo de un transposón en bucle, se inserte en el ácido nucleico diana.

En una realización, un método para obtener información de secuencia de un ácido nucleico diana comprende obtener un ácido nucleico molde que comprende una pluralidad de transposones insertados en dicho ácido nucleico diana, de forma que se retiene la contigüidad del molde, compartimentar los ácidos nucleicos que comprenden la pluralidad de transposones insertados en una pluralidad de recipientes, generar bibliotecas indexadas específicas de compartimentos a partir de las dianas de ácidos nucleicos transpuestas y obtener información de secuencia de los ácidos nucleicos molde en cada recipiente de la pluralidad de recipientes.

Determinados métodos de preparación de ácidos nucleicos diana para haplotipado

Algunas realizaciones de los métodos y composiciones proporcionados en el presente documento incluyen la preparación de ácidos nucleicos diana para haplotipado utilizando los métodos proporcionados en el presente documento. Utilizando un método de pre-amplificación, el número de lecturas singulares se aumenta al generar múltiples copias idénticas del mismo fragmento de ácido nucleico que un producto contiguo. En algunas de tales realizaciones, una biblioteca se amplifica mediante métodos tales como la amplificación por círculo rodante (RCA). En algunas realizaciones, se preparan bibliotecas circulares de un ácido nucleico diana y la biblioteca se amplifica por RCA. Dichos métodos generan ácidos nucleicos largos extendidos.

Se muestra un esquema de ejemplo en la **FIG. 6.** La **FIG. 6** describe un método que incluye preparar ácidos nucleicos diana para haplotipado mediante la generación de una biblioteca que comprende moléculas circulares por parejas de compañeros y la selección de ácidos nucleicos de tamaños específicos o en el intervalo de 1-10 kb o 10-20 kb, o 20 kb-50 kb, 50-200 kb; amplificar la biblioteca mediante RCA para generar ácidos nucleicos solitarios extendidos; insertar índices en la biblioteca amplificada con transposones; compartimentar la biblioteca insertada; eliminar la transposasa con SDS; indexar adicionalmente la biblioteca; y obtener información de secuencia de la biblioteca.

Otro esquema de ejemplo se muestra en la **FIG. 7.** La **FIG. 7** representa un método que incluye preparar ácidos nucleicos diana para haplotipado generando una biblioteca que comprende moléculas circulares mediante transposición en horquilla, el relleno de huecos y la selección de ácidos nucleicos de un tamaño específico o en el intervalo de 1-10 kb o 10-20 kb, o 20 kb-50 kb, 50-200 kb5; amplificar la biblioteca mediante RCA para generar ácidos nucleicos solitarios extendidos; insertar índices en la biblioteca amplificada con transposones; compartimentar la biblioteca insertada; eliminar la transposasa con SDS; indexar adicionalmente la biblioteca; y obtener información de secuencia de la biblioteca.

Métodos para generar bibliotecas de compañeros emparejados

Los métodos para la generación de bibliotecas de parejas de compañeros incluyen: fragmentar el ADN genómico en grandes fragmentos normalmente mayores de (aunque no limitado a) 1000 pb; circularizar fragmentos individuales mediante un método que etiqueta la unión ligada; fragmentar el ADN adicionalmente; enriquecer las secuencias de uniones etiquetadas y ligar adaptadores a las secuencias de unión enriquecidas, de forma que puedan secuenciarse, produciendo información acerca de la pareja de secuencias en los extremos del fragmento de ADN largo original. Estos procedimientos implican al menos 2 etapas, en que el ADN se fragmenta, ya sea física o enzimáticamente. En al menos una o más etapas distintas, los adaptadores se ligan a los extremos de los fragmentos. Las preparaciones de parejas de compañeros normalmente llevan 2-3 días en realizarse y comprenden múltiples etapas de manipulaciones de ADN. La diversidad de la biblioteca resultante se correlaciona directamente con el número de etapas necesarias para preparar la biblioteca.

El método proporcionado en el presente documento simplifica el número de etapas en el protocolo de generación de bibliotecas al emplear una reacción mediada por transposasa que, de forma simultánea, fragmenta y añade secuencias de adaptador a los extremos de los fragmentos. Al menos una o ambas etapas de fragmentación (fragmentación inicial del ADN genómico y fragmentación de fragmentos circularizados) se pueden realizar con un transposoma, reemplazando así la necesidad de etapas separadas de fragmentación y ligamiento de adaptadores. Obviar la depuración, la preparación y el ligamiento de los extremos de los fragmentos reduce el número de etapas del procedimiento y, así, se aumenta el rendimiento de los datos utilizables en la preparación y se hace que el procedimiento sea más robusto. En una realización, el protocolo se puede realizar sin recurrir a métodos que purifican una selección de tamaños basados en electroforesis. Este método produce una gama de tamaños de fragmentos más amplia que la que se puede lograr con los métodos electroforéticos en gel, pero aun así produce datos utilizables. La ventaja es que se evita una etapa de trabajo intensivo.

La **FIG. 8** proporciona un esquema de ejemplo donde solo la fragmentación inicial se reemplaza por una etapa de etiquetación por transposoma. El ADN circularizado se fragmenta por métodos físicos o por métodos químicos/enzimáticos, y los fragmentos se convierten en una biblioteca a través de la aplicación de protocolos convencionales de preparación de muestras (por ejemplo, TRUSEQ). La FIG. 9 ilustra un esquema de ejemplo en que tanto la fragmentación inicial como la fragmentación del ADN circularizado se realizan con una etapa de etiquetación por transposoma. Para el adaptador utilizado para la etiquetación inicial y la etiquetación posterior del círculo, las secuencias de adaptador para el transposoma (incluyendo las secuencias de ME) pueden o no ser distintas.

La amplificación del ácido nucleico molde mediante la generación de múltiples copias de cada molécula antes de la transposición o introducción de índices moleculares crea una redundancia que puede ser útil para obtener una mayor cobertura de SNP en cada bloque de haplotipos y también para el ensamblaje del genoma *de novo*, de forma similar a un enfoque de aleatorio (*shotgun*). El ácido nucleico molde se puede convertir en una biblioteca de tamaños definidos mediante transposición de baja frecuencia, corte físico o digestión enzimática y, después, amplificado durante un número finito de ciclos por PCR o un esquema de amplificación del genoma completo (por ejemplo, utilizando phi29). La biblioteca amplificada que ya contiene la redundancia incorporada se puede utilizar como material de entrada para el procedimiento de haplotipado. De esta forma, cada región del genoma está representada

múltiples veces por múltiples copias generadas por adelantado, contribuyendo cada copia con una cobertura parcial de esa región; sin embargo, la cobertura por consenso estará más cerca de completarse.

Ejemplos

Ejemplo 1-Reducción de las tasas de error

Se preparó una biblioteca de ácidos nucleicos molde con cada fragmento que comprendía un código de barras distinto. Se amplificó cada fragmento y se obtuvo la información de secuencia de al menos un producto amplificado a partir de cada fragmento. Se determinó una secuencia de consenso a partir de la información de secuencia de los productos amplificados a partir de cada fragmento. En particular, se secuenció una biblioteca de secuenciación preparada por NEXTERA durante 500 ciclos en un instrumento MISEQ. La biblioteca consistía en una distribución de tamaños, con longitudes de lectura máximas que se extendieron hasta -300 nt. Las tasas de error en el ciclo 250 fueron aproximadamente del 15 %. Si un molde estaba representado solo tres veces, la tasa de error disminuía a ~ 1 % en el ciclo 250. La **FIG. 10** ilustra un modelo de tasas de error en función del número de productos amplificados secuenciados (cobertura de cada código de barras). Las tasas de error disminuyen a medida que aumenta el número de productos amplificados secuenciados a partir de un fragmento.

Ejemplo 2-Acoplamiento de secuencias de transposón

Este ejemplo ilustra métodos para acoplar dos secuencias de transposón en diversas orientaciones, incluyendo una orientación 5'-5' y una orientación 3'-3'. En un método ejemplar, se utiliza aldehído y oxiamina para formar oligonucleótidos unidos a través de la formación de éter de oxima. Se combina un oligonucleótido modificado con aldehído (ya sea en el extremo 5' o en el 3') con un oligonucleótido modificado con oxiamina en el extremo 5', en el tampón de reacción, y se deja incubar durante 2 horas. El producto final se puede aislar, por ejemplo, mediante purificación de PAGE.

En otro método ejemplar, se realizó acoplamiento por bisoxiamina. Oligonucleótidos modificados con aldehído se dimerizaron con un enlazador de bis-oxiamina (por ejemplo, dioxiamino butano) usando concentraciones localmente altas para forzar la bisustitución. Se sintetizaron oligonucleótidos de 100 meros con un aldehído en el extremo 5' y se purificaron. Se sintetizó el oligonucleótido bisoxiamina. Se preparó una solución 1 mM del oligonucleótido bisoxiamina en un tampón de reacción de pH bajo que contenía un catalizador (urea 5 M, anilina 100 mM, citrato 10 mM, NaCl 150 mM, pH 5,6) y se añadió a una solución 665 µM de oligonucleótido aldehído en agua. El volumen completo de la solución se diluyó 1:1 con tampón de reacción y se dejó incubar a temperatura ambiente durante 2 horas. Una titulación de diversas proporciones de aldehído:bisoxiamina mostró dimerización a altas proporción de bisoxiamina. Las condiciones más satisfactorias se repitieron con oligonucleótidos con aldehído en 3'. La **FIG. 11** muestra los resultados de las reacciones de acoplamiento de bisoxiamina 5'-5' en las que se observaron transposones precursores en bucle en la banda de dímero indicada. Se observaron resultados similares para las reacciones de acoplamiento de bisoxiamina 3'-3'.

Ejemplo 3-Control de la estabilidad del transposón

Se prepararon transposomas utilizando secuencias de transposón largas y cortas cargadas en la transposasa. Los productos de transposoma incluían: A (2 secuencias cortas); B (secuencias largas y cortas); y C (2 secuencias largas). Las cantidades relativas de cada especie de transposoma se midieron en diversas condiciones, tales como la temperatura, tampones, proporciones de secuencias de transposón con respecto a la transposasa. Generalmente, la sal NaCl o KCl alta aumentó el intercambio de secuencias de transposón entre transposomas. Los tampones glutamato y acetato eliminaron o redujeron el intercambio, con concentraciones preferentes entre 100-600 mM. Se determinaron las condiciones óptimas de almacenamiento.

Ejemplo 4-Mantenimiento de la contigüidad del molde

Este ejemplo ilustra un método para el mantenimiento de la información de contigüidad de un ácido nucleico molde preparado utilizando transposomas que comprenden secuencias de transposón no contiguas, en que la transposasa Tn5 permanece unida al ADN molde después de la transposición. El ácido nucleico diana se puso en contacto con transposomas que comprendían transposasa Tn5 y secuencias de transposón no contiguas. La **FIG. 12** muestra que las muestras tratadas adicionalmente con SDS aparecieron como una mancha de diversos fragmentos de ácido nucleico molde; las muestras no tratadas con SDS mostraron retención de ácido nucleico molde de supuesto alto peso molecular. Por lo tanto, a pesar de que el ácido nucleico diana puede estar fragmentado, las secuencias adyacentes aún pueden estar asociadas entre sí por la transposasa (como lo demuestra el ADN unido a Tn5 que queda en los pocillos).

En otro método ejemplar más, se preparó una biblioteca de ácidos nucleicos molde utilizando transposomas que comprendían secuencias de transposón no contiguas con ácido nucleico diana que comprendía al cromosoma 22 humano. La **FIG. 13** resume que se observaron bloques de haplotipos de hasta 100 kb para muestras en las que se eliminó la transposasa mediante SDS después de la dilución. Por lo tanto, practicando los métodos como se

describe en el presente documento, los ácidos nucleicos diana pueden mantener la integridad de la diana cuando se transponen, diluyen y transformarse en bibliotecas de secuenciación.

Ejemplo 5-Mantenimiento de la contigüidad del molde

Los ácidos nucleicos diana se sometieron a etiquetación con transposomas que comprendían secuencias de transposón no contiguas (NEXTERA), se diluyeron a la concentración deseada y después se trataron con SDS para eliminar la enzima transposasa antes de la PCR. Como control, se sometió a etiquetación la misma cantidad de ADN de entrada, primero se trató con SDS y después se diluyó a la concentración deseada. El tratamiento con SDS antes de la dilución elimina la información de proximidad, dado que la enzima transposasa se disocia del ADN diana con SDS, fragmentando de este modo el ADN diana. Se configuraron dos reacciones de etiquetación en 50 ng de un ADNg de Coriell, se detuvo una reacción con SDS al 0,1 % y se diluyó a 6 pg. A continuación, la otra reacción se diluyó en primer lugar a 6 pg y después se detuvo con SDS al 0,1 %. Las reacciones completas se utilizaron para configurar una reacción de PCR de 30 ciclos y se secuenciaron en una plataforma Gene Analyzer (Illumina), de acuerdo con las instrucciones del fabricante. Las lecturas se mapearon en un genoma de ser humano de referencia y se calculó y representó la distribución de distancias.

Como se muestra en la **FIG. 14**, en la muestra de SDS posdilución, la distancia mediana cambió a tamaños más pequeños y se hace evidente una gran subpoblación de lecturas ubicadas de forma proximal. Si hubiere algún haplotipado, se esperaba una acumulación de lecturas proximales. La distribución bimodal de la muestra posdilución demostró que hay un enriquecimiento de lecturas proximales.

La distribución de distancias fue una medida del tamaño de la muestra (es decir, cuanto más singulares son las lecturas, más corta es la distancia). El histograma de distancia para las muestras de SDS predilución y SDS posdilución se muestra en la **FIG. 14**. Para corregir la diferencia en el número de lecturas singulares, se redujo el muestreo de la predilución para proporcionar el mismo número de lecturas mapeadas singulares (664.741). Se observó un enriquecimiento significativo para las lecturas que son vecinos inmediatos (es decir, uniones). Esto se midió observando la distribución de la "distancia al siguiente alineamiento" y encontrando las lecturas cuya distancia a su próximo alineamiento es la longitud de la lectura de secuencia menos 9 (lo que corresponde a la duplicación de 9 pb provocada por Tn5 en el sitio de inserción). Dichas lecturas constituyen el 10 % de los datos (**FIG. 15**) y, con el empleo de un sistema de cebador único para la amplificación de las bibliotecas NEXTERA, se puede duplicar. Además, la implementación de una preparación de muestras más conservadora que permite una menor pérdida de muestras permite recuperar más datos de unión. El poder de resolución de haplotipos disminuyó cuando se aumentó el ADN de entrada. Por otro lado, la reducción de la entrada de ADN precisó más amplificación y, por lo tanto, más ciclos de PCR, generando muchos duplicados de PCR. El uso de complejos de Tn5 con códigos de barras de forma individual permitió que la etiquetación y las diluciones posteriores se llevaran a cabo en compartimentos separados. Se combinaron bajos niveles de entrada de material con código de barras y sometido a etiquetación individualmente para elevar las cantidades de ADN de entrada de la PCR al nivel que permitiera una amplificación más específica con menos desaprovechamiento de la capacidad de secuenciación en lecturas redundantes. Por consiguiente, el uso de complejos con código de barras suficientes permitió la organización en fases de la mayoría del genoma humano. Para aumentar el poder de resolución de haplotipos, la asignación de códigos de barras se implementó tanto a nivel del complejo como a nivel de cebador de PCR. Dicho esquema de indexación combinatorio permite el uso de ADN de entrada muy bajo de cada complejo con códigos de barras asignados individualmente en la reacción de PCR, lo que permitiría una potente resolución de haplotipos. Usando solo 40 oligonucleótidos de indexación (8+12=20 para los complejos de NEXTERA, lo que genera 8* 12=96 complejos individuales y 8+12=20 para los cebadores de PCR, lo que permitiría 8*12=96 índices adicionales), se generaron 96*96=9216 compartimentos virtuales para el procedimiento de haplotipado mencionado anteriormente. Usando un protocolo de secuenciación modificado, se secuenciaron todos los datos en un HiSeq-2000. En los resultados de secuenciación se observaron todas las combinaciones posibles de los 9216 códigos de barras.

Ejemplo 6-Obtención de información de haplotipos con Mu

Se utilizaron transposomas que comprendían Mu para obtener información de haplotipos. Se trató 1 ng de ADN genómico con Mu-TSM en un volumen de reacción de 50 µl con tampón TA 1 X y 1, 2, 4 u 8 µl de complejos Mu-TSM 25 µM. Las reacciones se incubaron a 37 ºC durante 2 horas. Las muestras se diluyeron a 1 pg/µl. Para la inactivación de mu, se prepararon 10 µl de cada muestra que contenían 1 pg o 5 pg de ADN genómico total. Se añadió SDS a una concentración final de 0,05 %. Las muestras se incubaron a 55 ºC durante 20 minutos. La muestra completa se usó para configurar una reacción de PCR de 50 µl utilizando NPM. La PCR se realizó durante 30 ciclos. Las muestras de PCR se limpiaron con SPRI 0,6 X y se resuspendieron en 20 µl de tampón de resuspensión. Se obtuvo información de secuenciación. La FIG. 16 muestra las lecturas proximales acumuladas observadas en el contenido subhaploide utilizando transposomas que comprendían Mu.

El término "que comprende" como se usa en el presente documento es sinónimo de "que incluye", "que contiene", o "caracterizado por", y es inclusivo o abierto y no excluye elementos o etapas de método adicionales y que se hayan mencionado.

REIVINDICACIONES

1. Un método de preparación de una biblioteca de ácidos nucleicos molde para obtener información de secuencia a partir de un ácido nucleico diana, comprendiendo dicho método:

(a) compartimentar el ácido nucleico diana en una pluralidad de primeros recipientes proporcionando a cada primer recipiente una cantidad de ácido nucleico diana mayor que aproximadamente uno o más equivalentes haploides del ácido nucleico diana;
(b) proporcionar un primer índice al ácido nucleico diana de cada primer recipiente, en donde el primer índice proporcionado al ácido nucleico diana de cada primer recipiente es distinto, en donde la etapa comprende poner en contacto el ácido nucleico diana con una pluralidad de transposomas, comprendiendo cada transposoma una primera secuencia de transposón que comprende un primer índice, una segunda secuencia de transposón no contigua a dicha primera secuencia de transposón, y una transposasa asociada con la primera secuencia de transposón y la segunda secuencia de transposón en condiciones tales que al menos algunas de las secuencias de transposón se inserten en el ácido nucleico diana, obteniendo de este modo una pluralidad de primeros ácidos nucleicos molde indexados;
(c) combinar los primeros ácidos nucleicos molde indexados;
(d) compartimentar los primeros ácidos nucleicos molde indexados en una pluralidad de segundos recipientes; y
(e) proporcionar un segundo índice a los primeros ácidos nucleicos molde indexados de cada segundo recipiente, en donde el segundo índice proporcionado a los primeros ácidos nucleicos molde indexados de cada segundo recipiente es distinto, obteniendo de este modo una pluralidad de segundos ácidos nucleicos molde indexados.

2. El método de la reivindicación 1, en donde la etapa (d) comprende eliminar la transposasa de los primeros ácidos nucleicos molde indexados compartimentados, en donde la eliminación de la transposasa comprende un método seleccionado del grupo que consiste en añadir un detergente, cambiar la temperatura, cambiar el pH, añadir una proteasa, añadir una chaperona y añadir una polimerasa.

3. El método de cualquiera de las reivindicaciones 1-2, en donde la primera secuencia de transposón comprende un primer sitio de cebador y la segunda secuencia de transposón comprende un segundo sitio de cebador.

4. El método de la reivindicación 3, en donde el primer sitio de cebador comprende adicionalmente un primer código de barras y el segundo sitio de cebador comprende adicionalmente un segundo código de barras, en donde el primer código de barras y el segundo código de barras son distintos.

5. El método de cualquiera de las reivindicaciones 1-4, en donde la etapa (d) comprende proporcionar a cada primer recipiente una cantidad del primer ácido nucleico molde indexado mayor que aproximadamente uno o más equivalentes haploides del ácido nucleico diana.

6. El método de cualquiera de las reivindicaciones 1-5, en donde la etapa (e) comprende amplificar los primeros ácidos nucleicos molde indexados con al menos un cebador que comprende el segundo índice.

7. El método de cualquiera de las reivindicaciones 1-5, en donde la etapa (e) comprende ligar los primeros ácidos nucleicos molde indexados con al menos un cebador que comprende el segundo índice.

8. El método de cualquiera de las reivindicaciones 1-7, en donde el ácido nucleico diana se obtiene enriqueciendo una pluralidad de ácidos nucleicos para una secuencia de interés ya sea antes o después de la transposición.

9. El método de la reivindicación 1, que comprende adicionalmente obtener información de secuencia a partir del ácido nucleico molde, en donde, opcionalmente, la información de secuencia comprende información de secuencia de haplotipos.

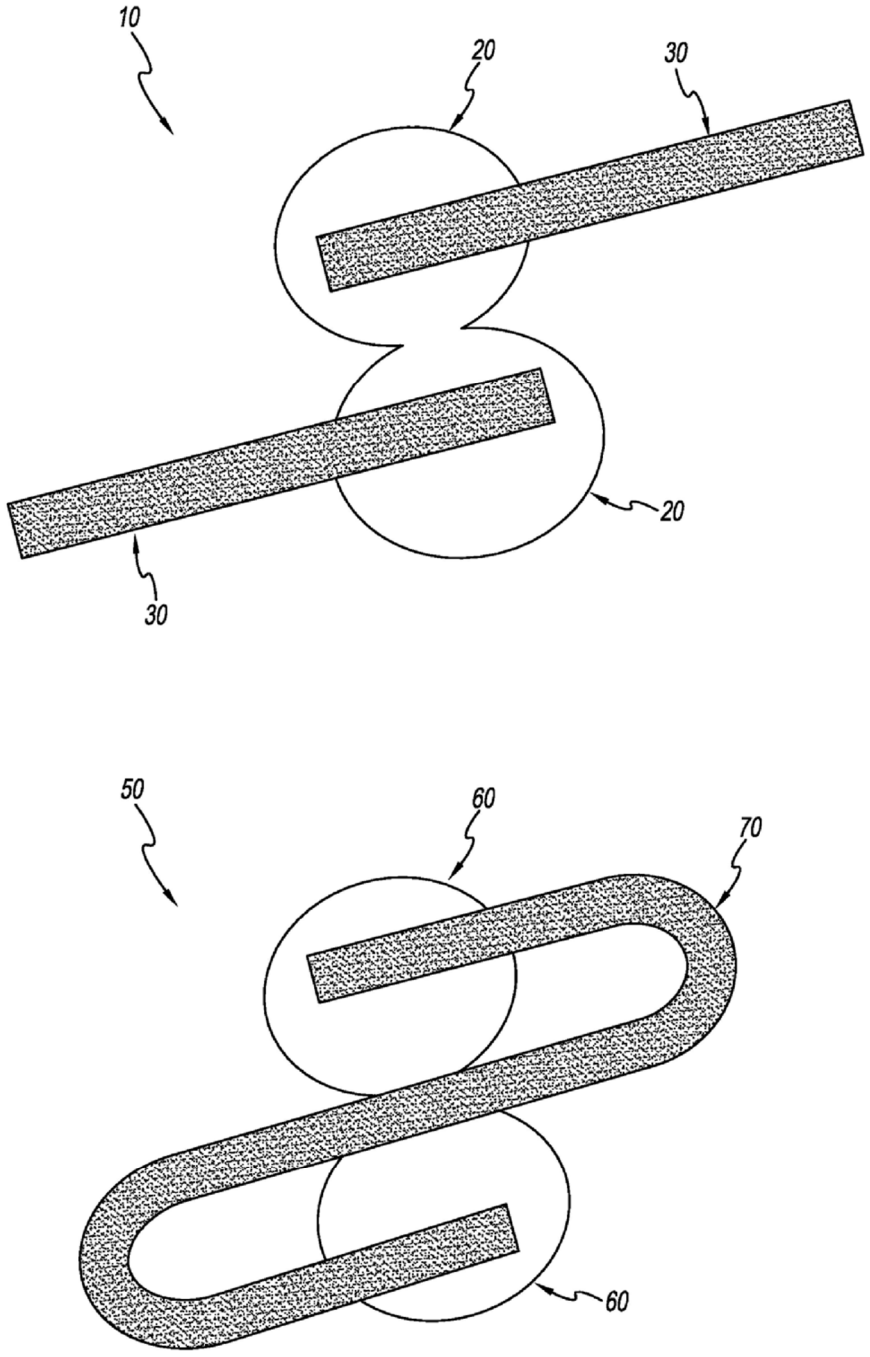

FIG. 1

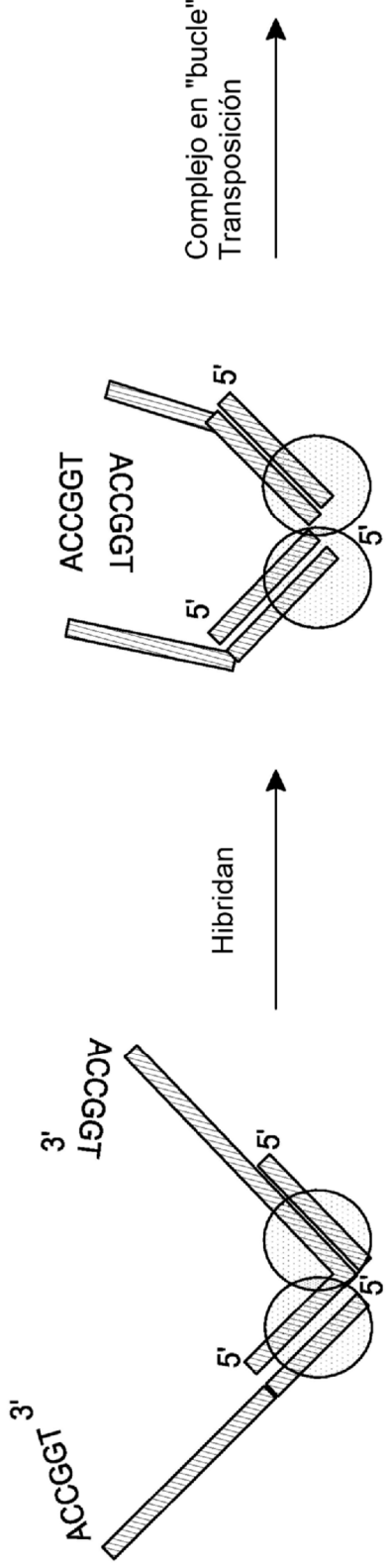
3'
ACCGGT
5'
5'
5'
ACCGGT
3'
Hibridan
ACCGGT
5'
5'
5'
Complejo en "bucle"
Transposición

FIG. 2

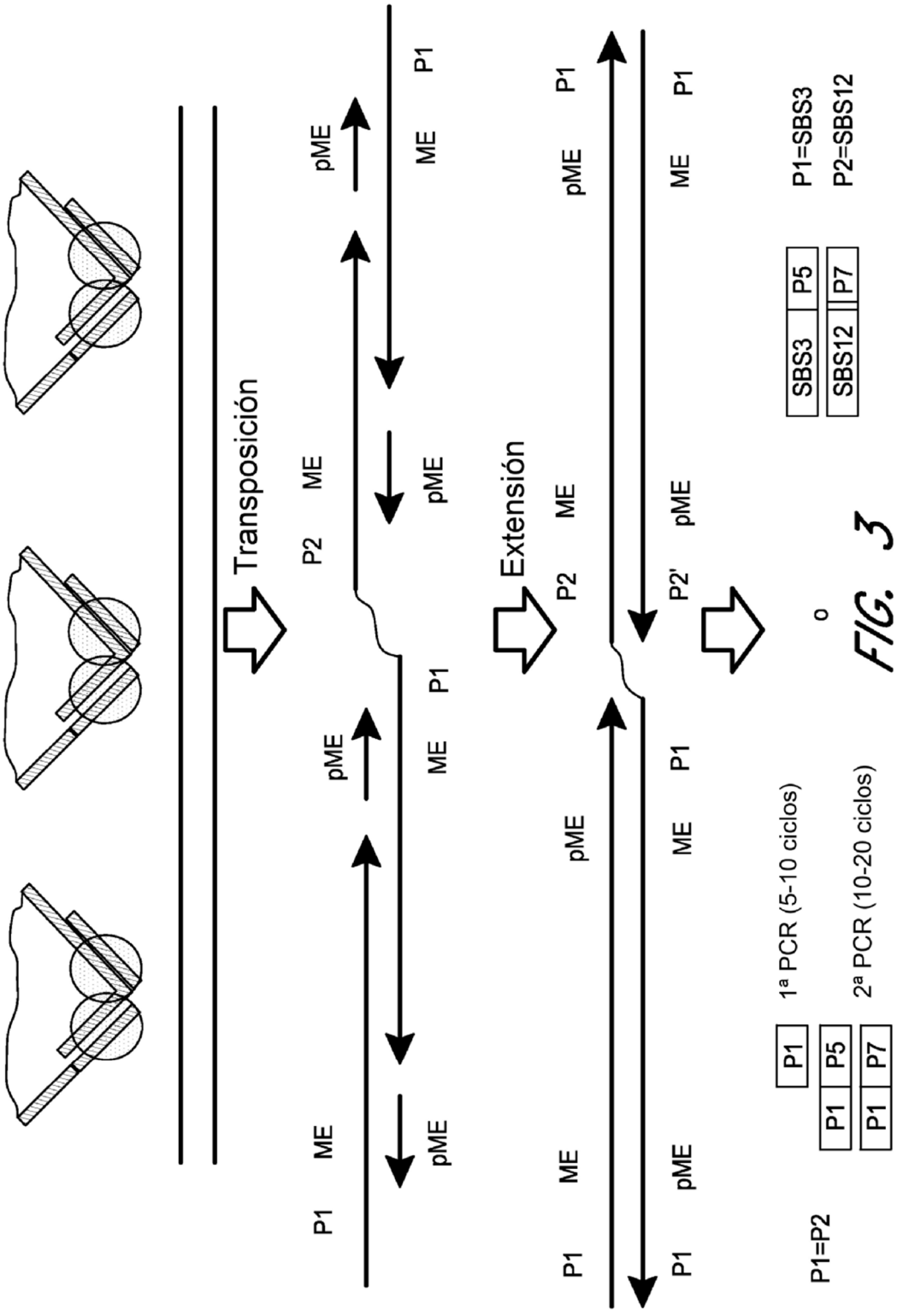
Transposición
Extensión
P1
ME
pME
P2
P2'
1ª PCR (5-10 ciclos)
2ª PCR (10-20 ciclos)
P1=P2
P1=SBS3
P2=SBS12
SBS3
SBS12
P5
P7

FIG. 3

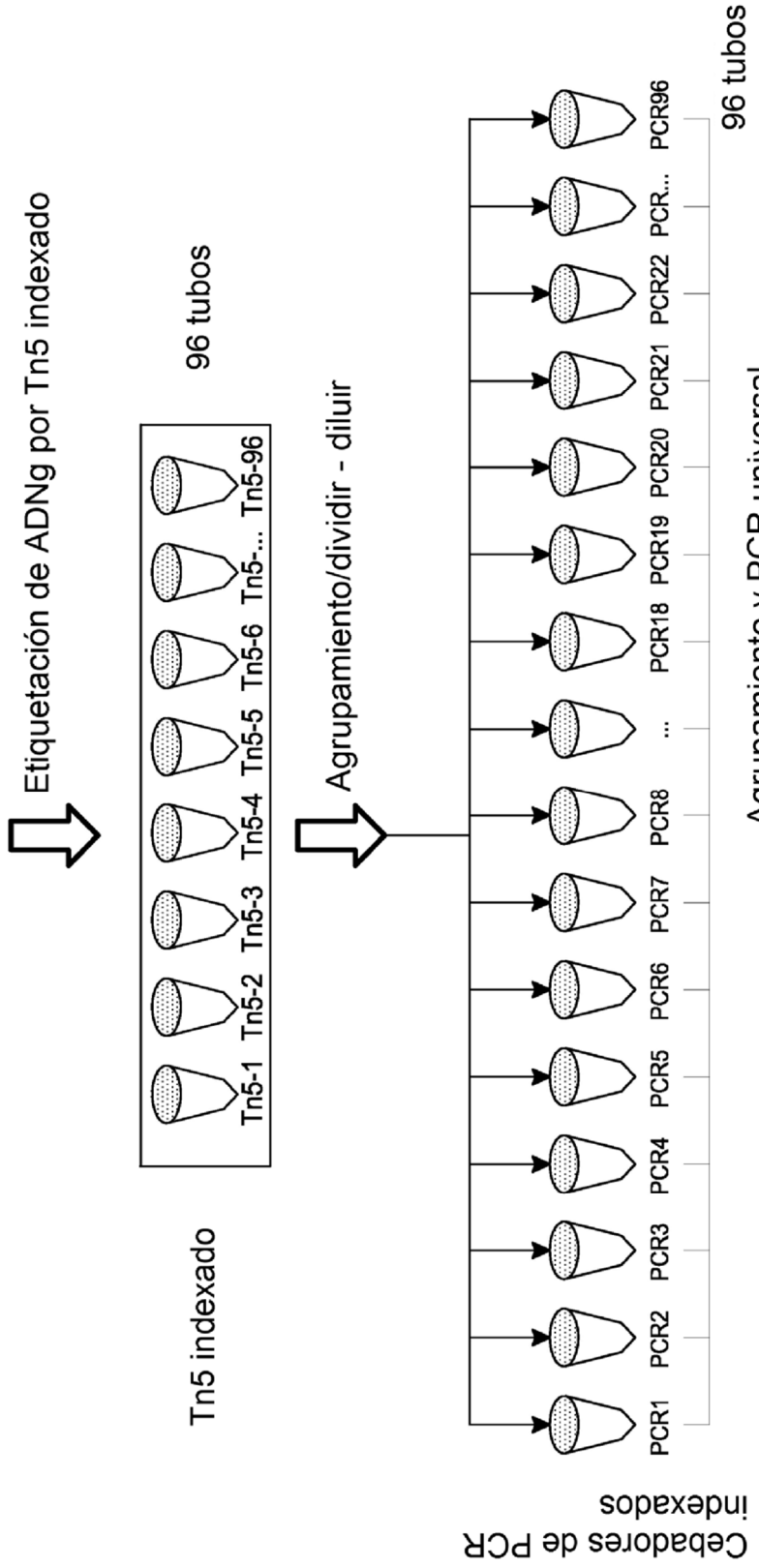
Etiquetación de ADNg por Tn5 indexado
96 tubos
Tn5-1
Tn5-2
Tn5-3
Tn5-4
Tn5-5
Tn5-6
Tn5-...
Tn5-96
Tn5 indexado
Agrupamiento/dividir - diluir
PCR1
PCR2
PCR3
PCR4
PCR5
PCR6
PCR7
PCR8
...
PCR18
PCR19
PCR20
PCR21
PCR22
PCR...
PCR96
96 tubos
Agrupamiento y PCR universal
Cebadores de PCR
indexados

FIG. 4

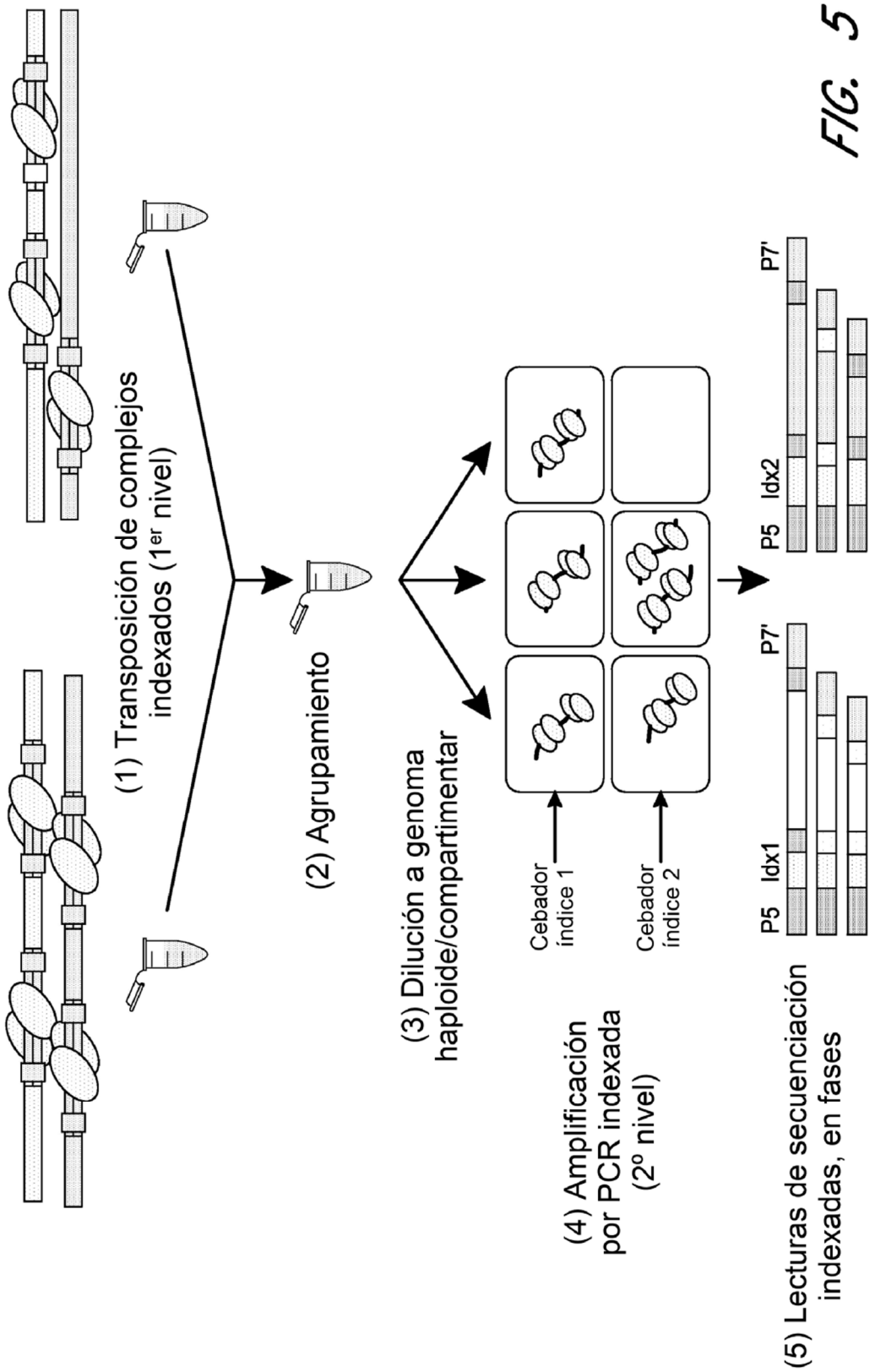
(1) Transposición de complejos indexados (1er nivel)
(2) Agrupamiento
(3) Dilución a genoma haploide/compartimentar
Cebador índice 1
Cebador índice 2
(4) Amplificación por PCR indexada (2º nivel)
P5
Idx1
P7'
P5
Idx2
P7'
(5) Lecturas de secuenciación indexadas, en fases

FIG. 5

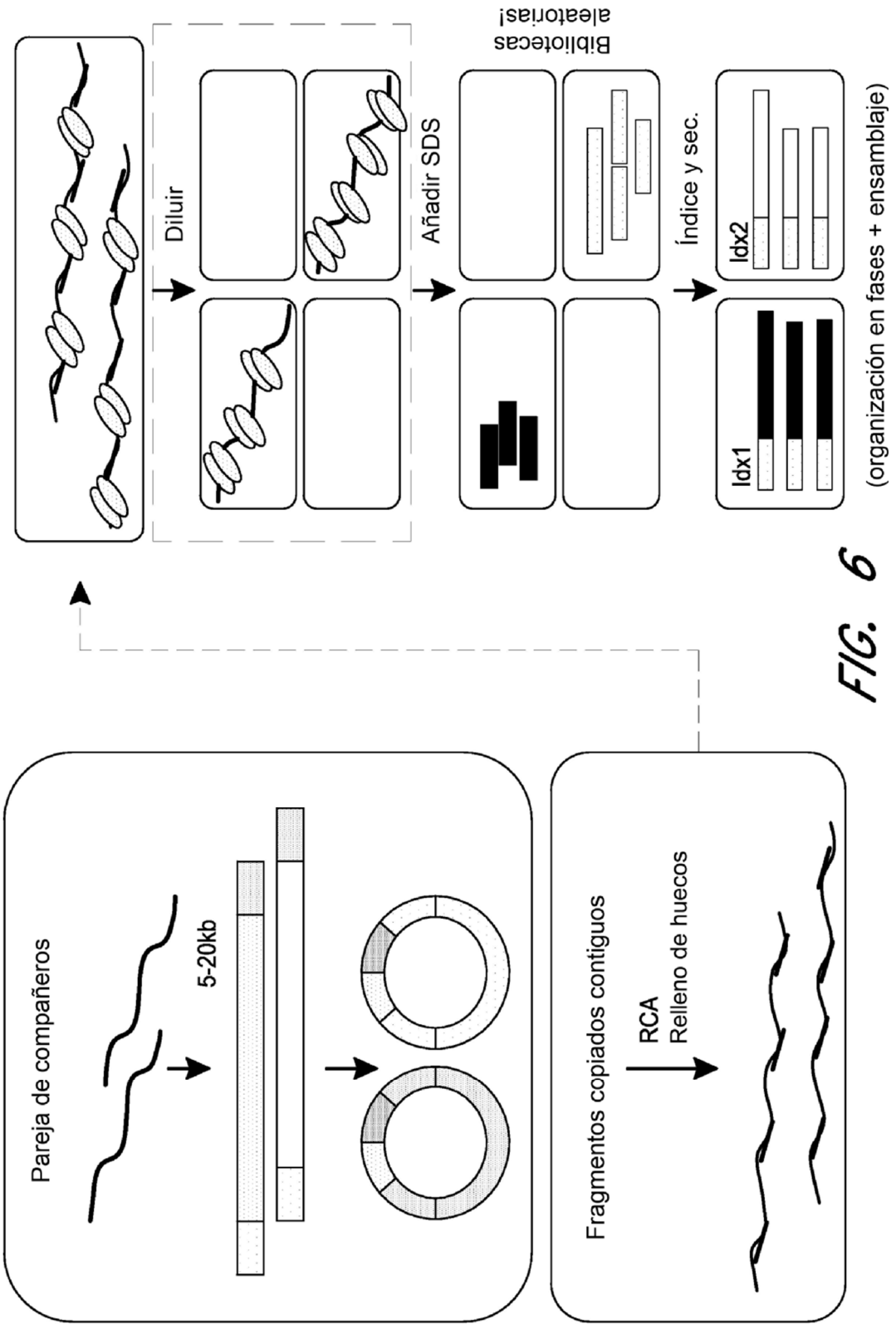
Pareja de compañeros
5-20kb
Fragmentos copiados contiguos
RCA
Relleno de huecos
Diluir
Añadir SDS
Bibliotecas aleatorias!
Índice y sec.
Idx1
Idx2
(organización en fases + ensamblaje)

FIG. 6

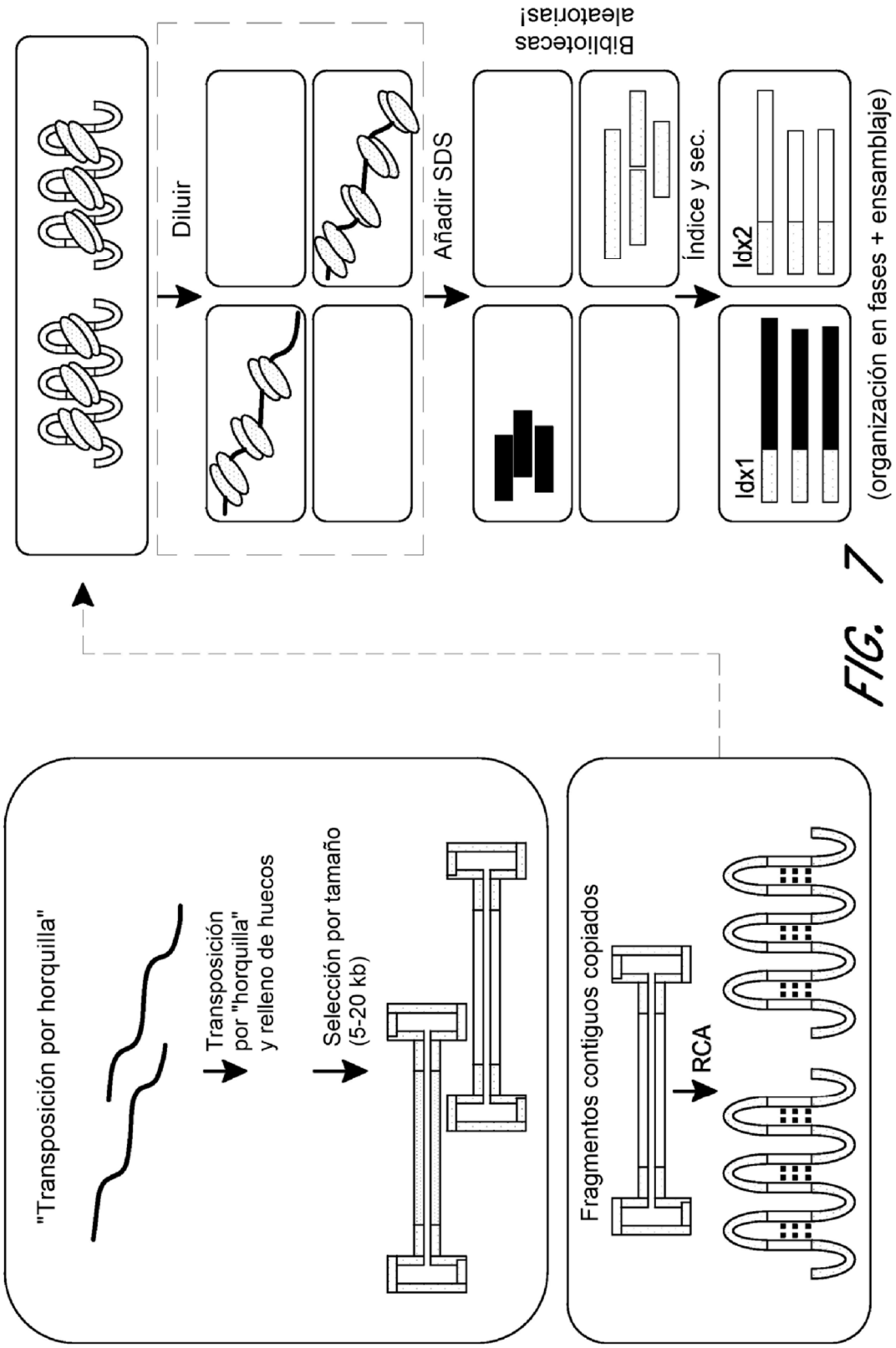
"Transposición por horquilla"
Transposición por "horquilla" y relleno de huecos
Selección por tamaño (5-20 kb)
Fragmentos contiguos copiados
RCA
Diluir
Añadir SDS
Bibliotecas aleatorias¡
Índice y sec.
Idx1
Idx2
(organización en fases + ensamblaje)

FIG. 7

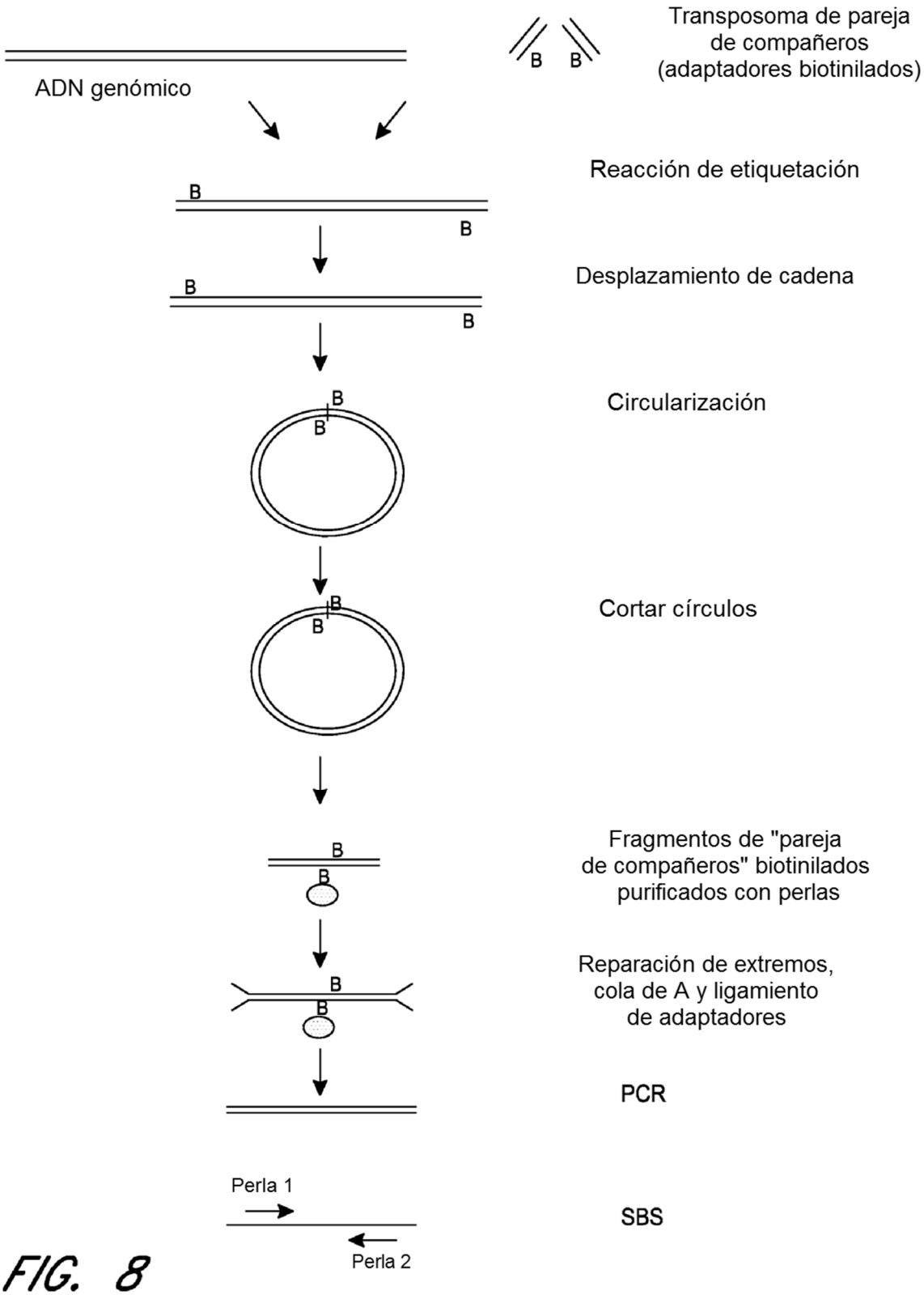
Transposoma de pareja de compañeros (adaptadores biotinilados)
B
B
ADN genómico
Reacción de etiquetación
Desplazamiento de cadena
Circularización
Cortar círculos
Fragmentos de "pareja de compañeros" biotinilados purificados con perlas
Reparación de extremos, cola de A y ligamiento de adaptadores
PCR
Perla 1
Perla 2
SBS

FIG. 8

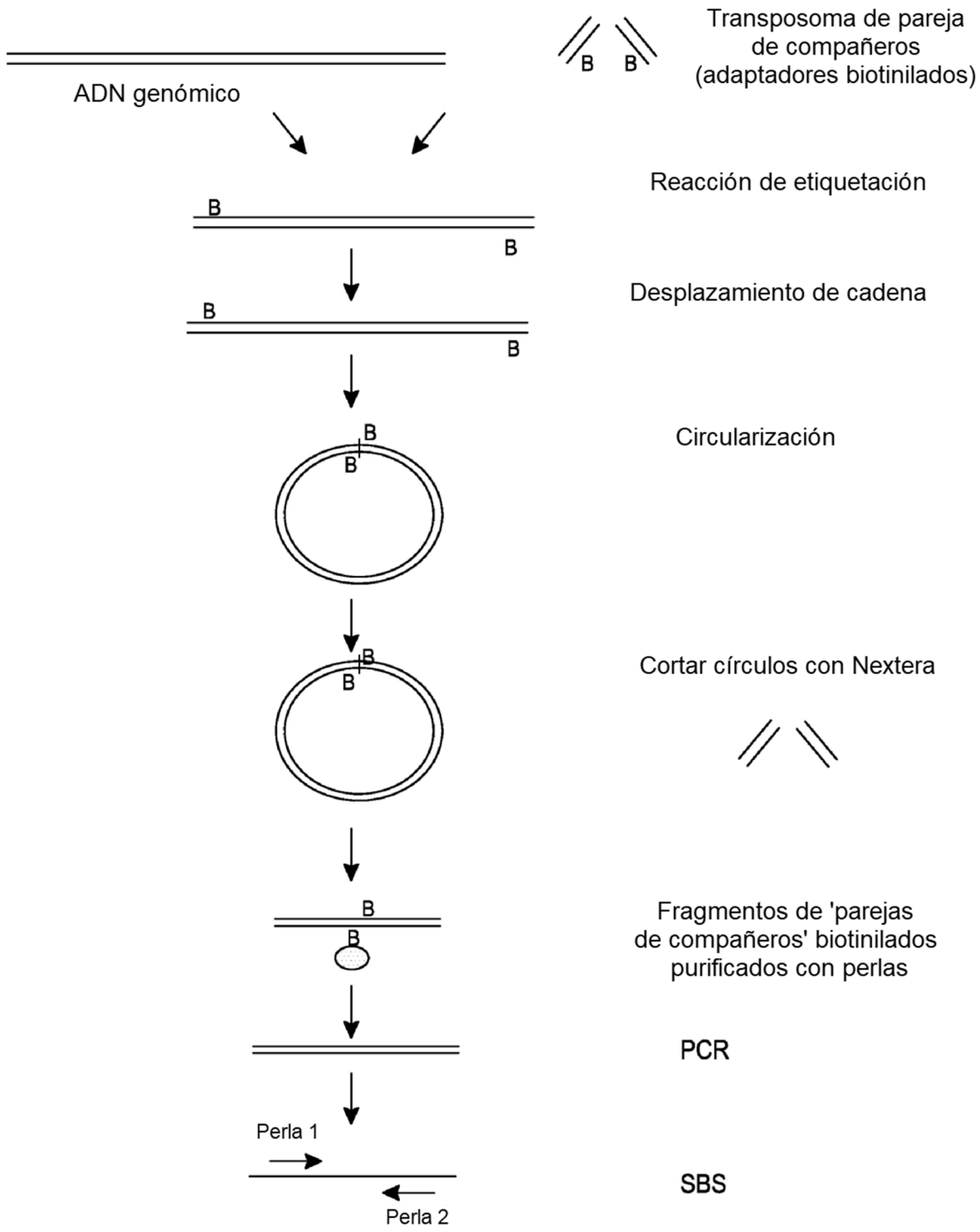
Transposoma de pareja de compañeros (adaptadores biotinilados)
B
B
ADN genómico
Reacción de etiquetación
Desplazamiento de cadena
Circularización
Cortar círculos con Nextera
Fragmentos de 'parejas de compañeros' biotinilados purificados con perlas
PCR
Perla 1
Perla 2
SBS

FIG. 9

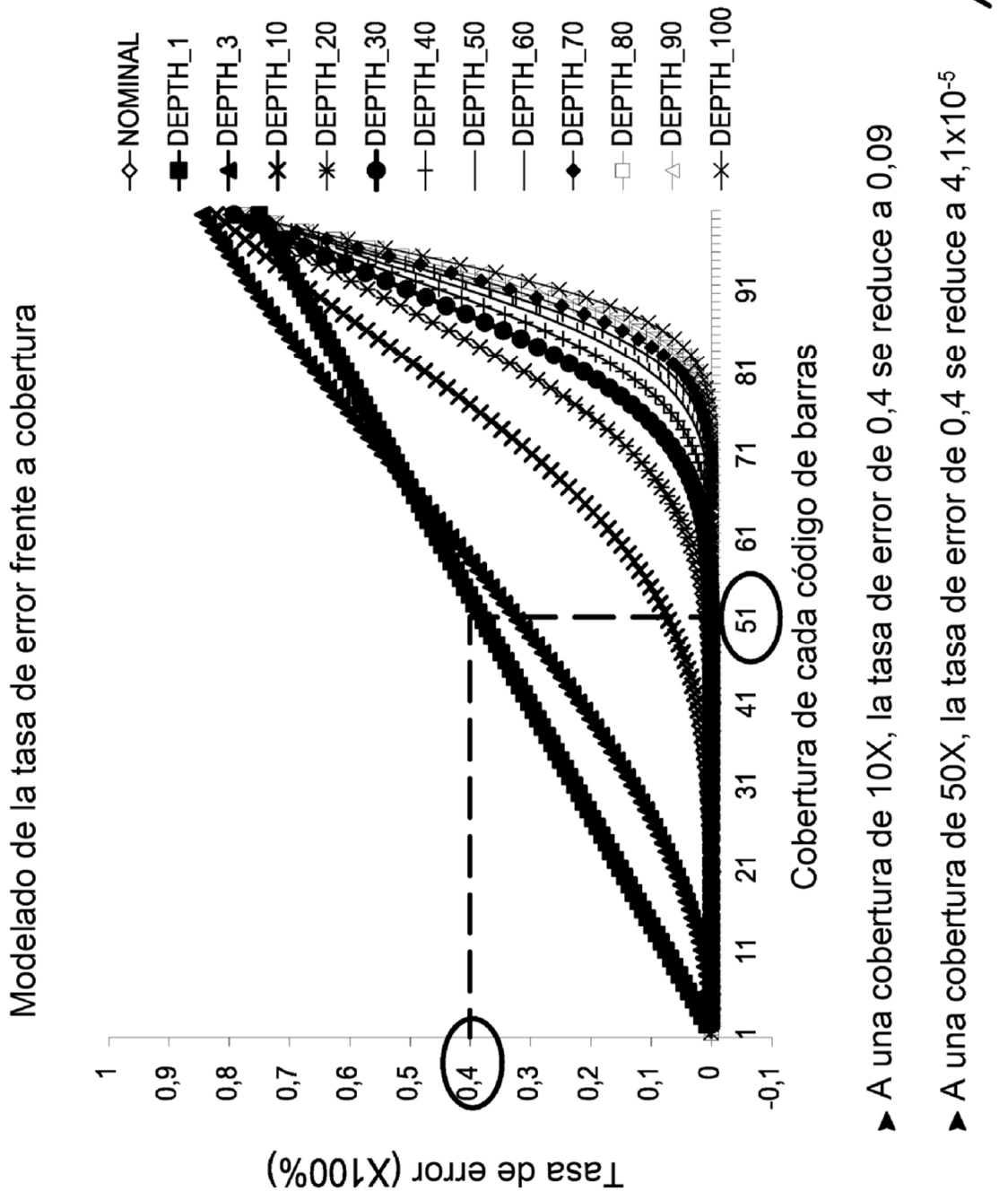
Modelado de la tasa de error frente a cobertura
NOMINAL
DEPTH_1
DEPTH_3
DEPTH_10
DEPTH_20
DEPTH_30
DEPTH_40
DEPTH_50
DEPTH_60
DEPTH_70
DEPTH_80
DEPTH_90
DEPTH_100
Tasa de error (X100%)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
1
11
21
31
41
51
61
71
81
91
Cobertura de cada código de barras
▲ A una cobertura de 10X, la tasa de error de 0,4 se reduce a 0,09
▲ A una cobertura de 50X, la tasa de error de 0,4 se reduce a 4,1x10-5

FIG. 10

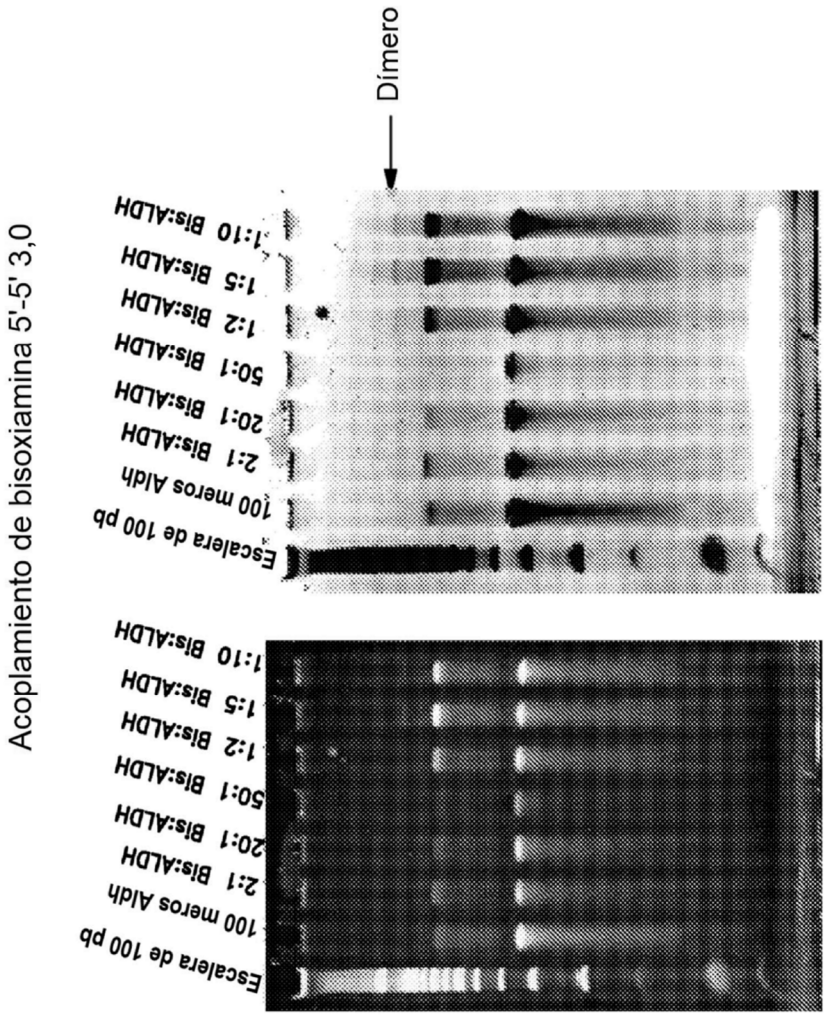
Acoplamiento de bisoxiamina 5'-5' 3,0
Dímero
Escalera de 100 pb
100 meros Aldh
2:1 Bis:ALDH
20:1 Bis:ALDH
50:1 Bis:ALDH
1:2 Bis:ALDH
1:5 Bis:ALDH
1:10 Bis:ALDH
Escalera de 100 pb
100 meros Aldh
2:1 Bis:ALDH
20:1 Bis:ALDH
50:1 Bis:ALDH
1:2 Bis:ALDH
1:5 Bis:ALDH
1:10 Bis:ALDH

FIG. 11

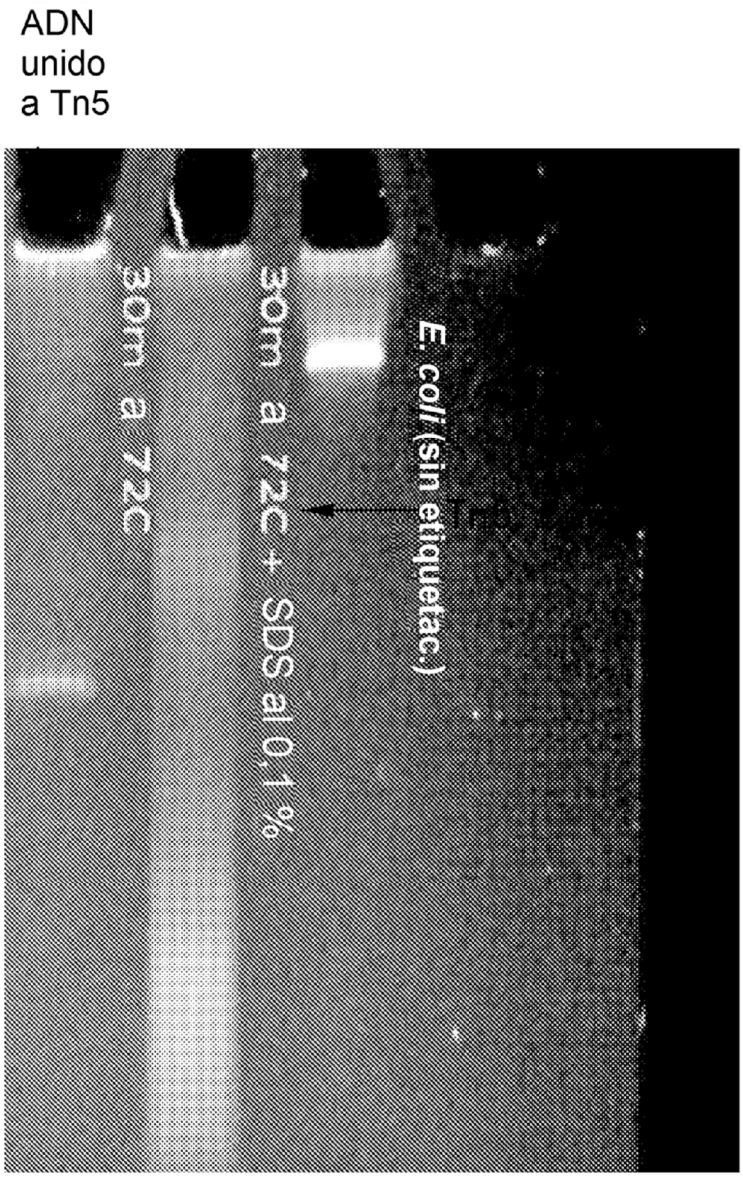
ADN
unido
a Tn5
E. coli (sin etiquetac.)
30m a 72C + SDS al 0,1 %
30m a 72C

FIG. 12

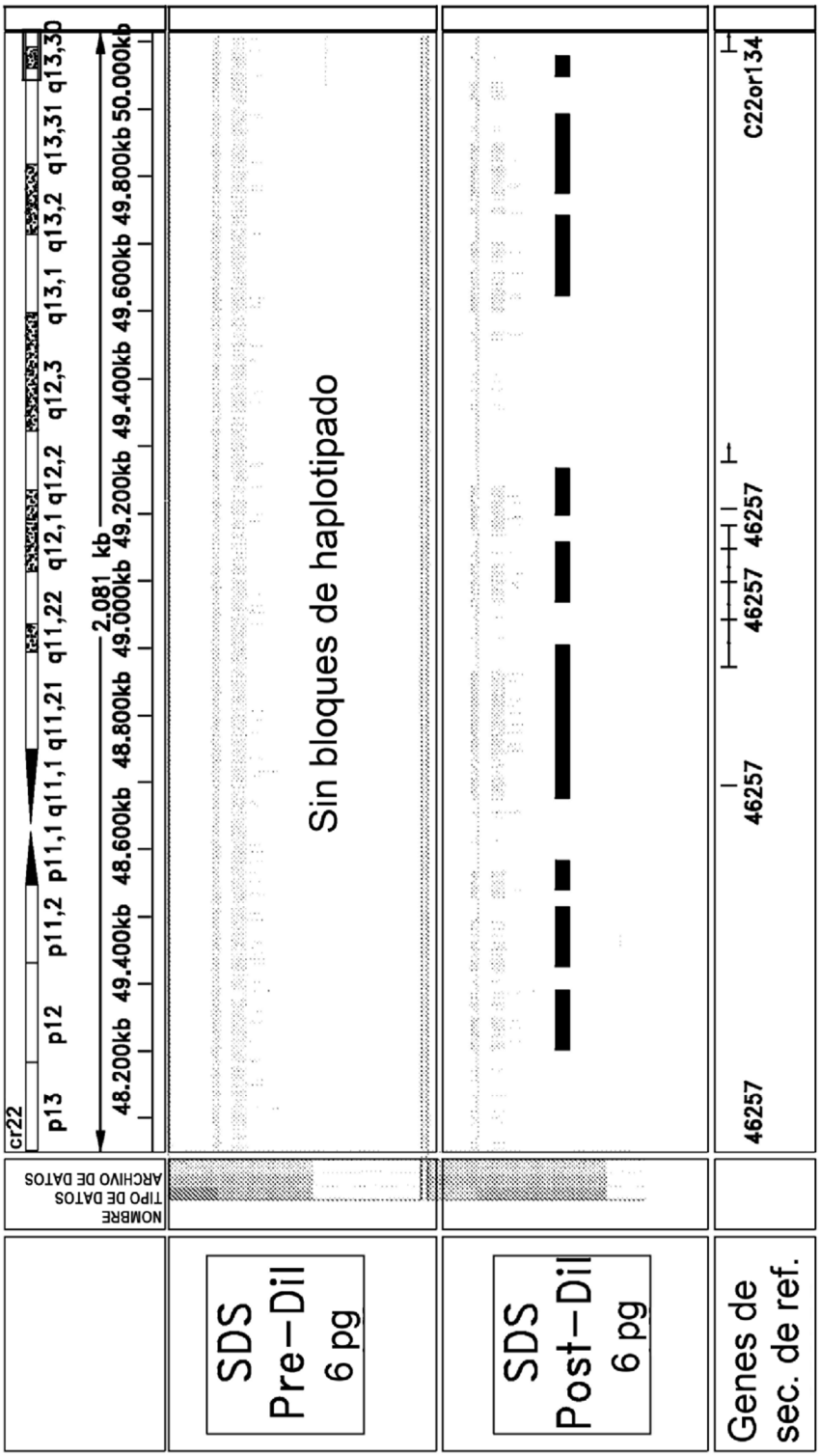
NOMBRE
TIPO DE DATOS
ARCHIVO DE DATOS
chr22
p13 p12 p11,2 p11,1 q11,1 q11,21 q11,22 q12,1 q12,2 q12,3 q13,1 q13,2 q13,31 q13,33
48.200kb 49.400kb 48.600kb 48.800kb 49.000kb 2,081 kb 49.200kb 49.400kb 49.600kb 49.800kb 50.000kb
Sin bloques de haplotipado
SDS Pre-DiI 6 pg
SDS Post-DiI 6 pg
Genes de sec. de ref.
46257
46257
46257
46257
C22or134

FIG. 13

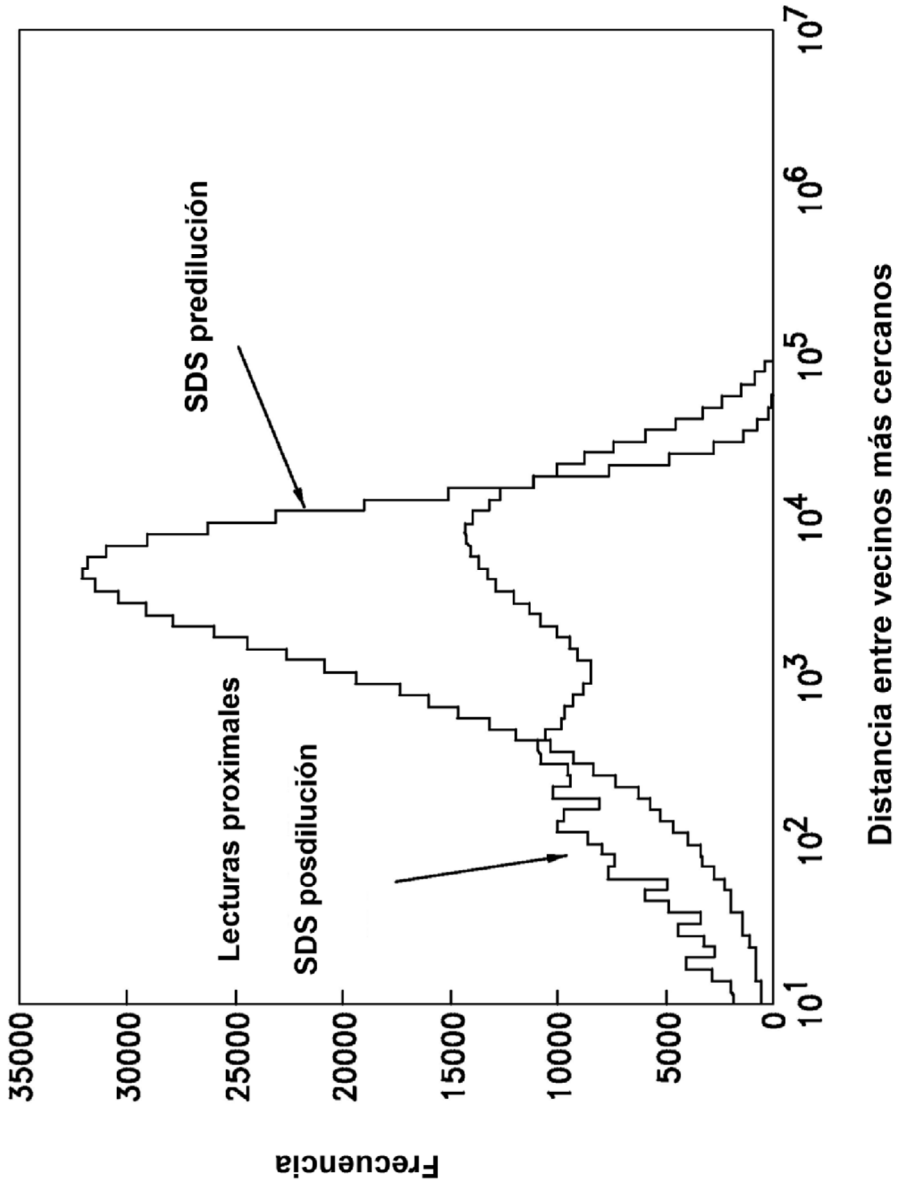
SDS predilución
Lecturas proximales
SDS posdilución
Frecuencia
Distancia entre vecinos más cercanos
0
5000
10000
15000
20000
25000
30000
35000
10^1
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 14

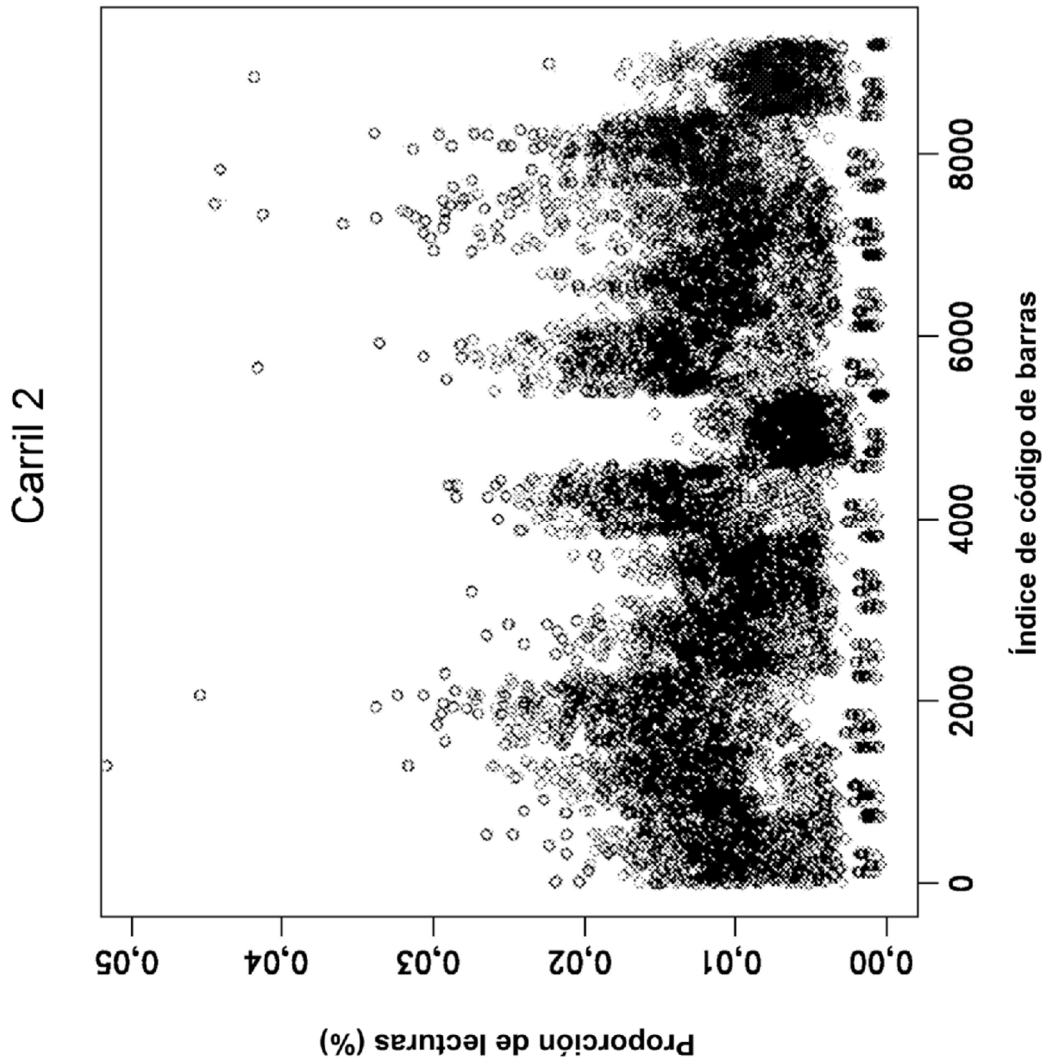
Carril 2
Proporción de lecturas (%)
0,00
0,01
0,02
0,03
0,04
0,05
0
2000
4000
6000
8000
Índice de código de barras

FIG. 15

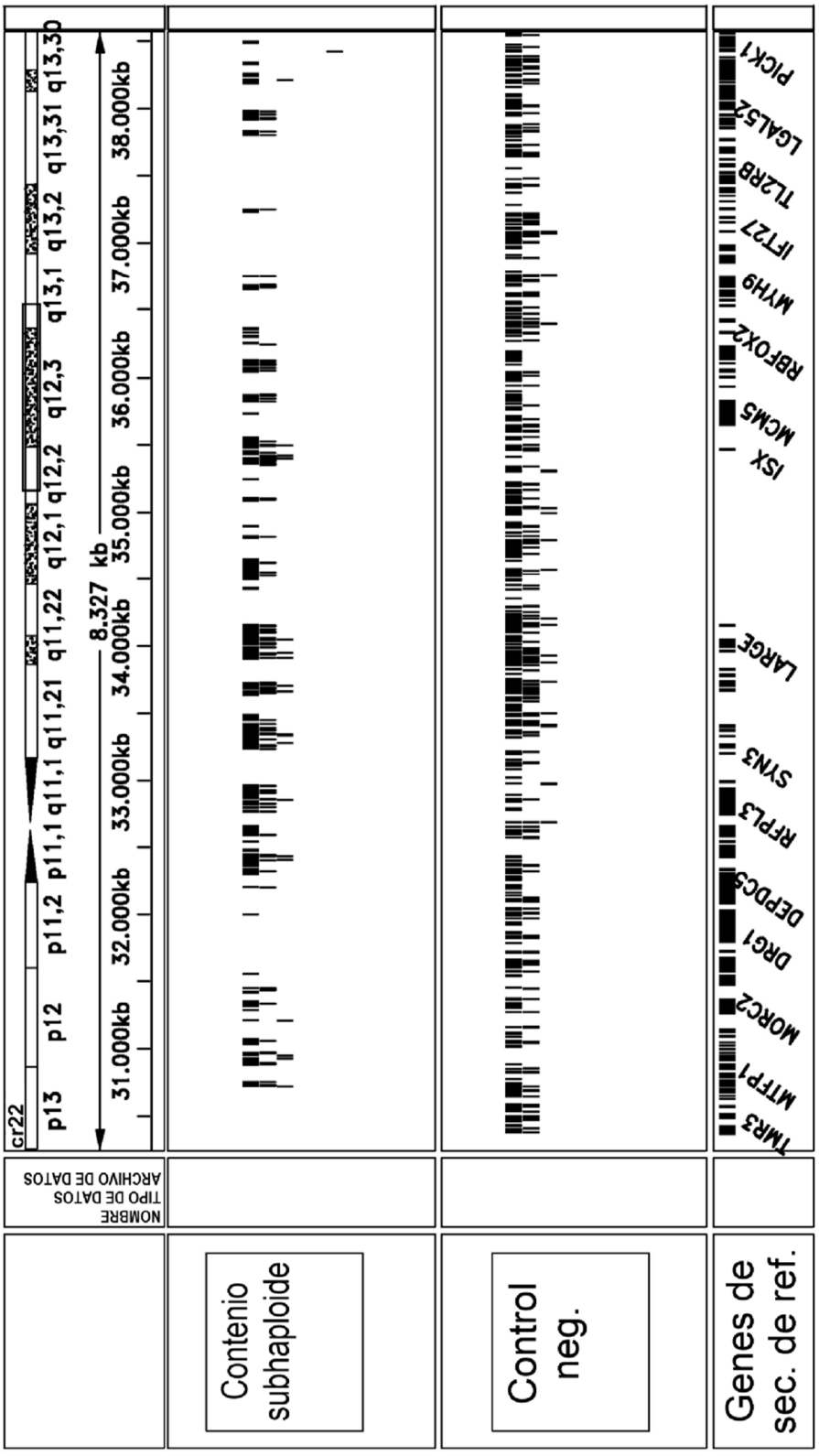
Contenio subhaploide
Control neg.
Genes de sec. de ref.
NOMBRE
TIPO DE DATOS
ARCHIVO DE DATOS
chr22
p13 p12 p11,2 p11,1 q11,1 q11,21 q11,22 q12,1 q12,2 q12,3 q13,1 q13,2 q13,31 q13,33
31.000kb 32.000kb 33.000kb 34.000kb 8.327 kb 35.000kb 36.000kb 37.000kb 38.000kb
TMR3
MTFP1
MORC2
DRG1
DEPDC5
RFPL3
SYN3
LARGE
ISX
MCM5
RBFOX2
MYH9
IFT27
TL2RB
LGALS2
PICK1

FIG. 16

REFERENCIAS CITADAS EN LA DESCRIPCIÓN

Esta lista de referencias citadas por el solicitante es únicamente para la comodidad del lector. No forma parte del documento de la patente europea. A pesar del cuidado tenido en la recopilación de las referencias, no se pueden excluir errores u omisiones y la EPO niega toda responsabilidad en este sentido.

Documentos de patentes citados en la descripción

- US 20120208705 A **[0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]**
- US 20120208724 A **[0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]**
- WO 2012061832 A **[0073] [0079] [0093] [0103] [0106] [0107] [0108] [0121] [0131]**
- US 5644048 A **[0081]**
- US 5386023 A **[0082]**
- US 5637684 A **[0082]**
- US 5602240 A **[0082]**
- US 5216141 A **[0082]**
- US 4469863 A **[0082]**
- US 5235033 A **[0082]**
- US 5034506 A **[0082] [0085]**
- US 5432272 A **[0085]**
- WO 9523875 A **[0091]**
- US 20100120098 A **[0094] [0171]**
- US 7741463 B **[0105]**
- WO 2012108864 A **[0114]**
- US 7582420 B **[0136]**
- US 7955794 B **[0136]**
- US 8003354 B **[0136]**
- WO 2010062913 A **[0141]**
- US 20100323348 A **[0143]**
- US 7574305 B **[0143]**
- US 6654696 B **[0143]**
- US 6210891 B **[0147]**
- US 6258568 B **[0147]**
- US 6274320 A **[0147]**
- US 742767 A **[0148]**
- US 74141163 B**[0148]**
- US 7057026 B **[0148] [0149]**
- WO 9106678 A **[0148]**
- WO 07123744 A **[0148]**
- US 20070166705 A**[0149]**
- US 20060188901 A**[0149]**
- US 20060240439 A**[0149]**
- US 20060281109 A**[0149]**
- WO 05065814 A **[0149]**
- US 20050100900 A**[0149]**
- WO 06064199 A **[0149]**
- WO 07010251 A **[0149]**
- US 6969488 B **[0150]**
- US 6172218 B **[0150]**
- US 6306597 B **[0150]**
- US 7001792 B **[0151]**
- US 7329492 B **[0152]**
- US 7211414 B **[0152]**
- US 7315019 B **[0152]**
- US 7405281 B **[0152]**
- US 20080108082 A**[0152]**
- US 7181122 B **[0152]**
- US 7302146 B **[0152]**
- US 7313308 B **[0152]**

Literatura diferente de patentes citada en la descripción

- **BEAUCAGE *et al.*** *Tetrahedron,* 1993, vol. 49, 1925 **[0081]**
- **LETSINGER, J.** *Org. Chem.,* 1970, vol. 35, 3800 **[0081]**
- **SPRINZL *et al.*** *Eur. J. Biochem.,* 1977, vol. 81,579 **[0081]**
- **LETSINGER *et al.*** *Nucl. Acids Res.,* 1986, vol. 14, 3487 **[0081]**
- **SAWAI *et al.*** *Chem. Lett.,* vol. 805, 1984 **[0081]**

- **LETSINGER *et al.*** *J. Am. Chem. Soc.,* 1988, vol. 110, 4470 **[0081] [0082]**
- **PAUWELS *et al.*** *Chemica Scripta,* 1986, vol. 26, 141 **[0081]**
- **MAG *et al.*** *Nucleic Acids Res.,* 1991, vol. 19, 1437 **[0081]**
- **BRIU *et al.*** *J. Am. Chem. Soc.,* 1989, vol. 111,2321 **[0081]**
- **ECKSTEIN.** Oligonucleotides and Analogues: A Practical Approach. Oxford University Press **[0081]**
- **EGHOLM, J.** *Am. Chem. Soc.,* 1992, vol. 114, 1895 **[0081]**
- **MEIER *et al.*** *Chem. Int. Ed. Engl.,* 1992, vol. 31,1008 **[0081]**
- **NIELSEN.** *Nature,* 1993, vol. 365, 566 **[0081]**
- **CARLSSON *et al.*** *Nature,* 1996, vol. 380, 207 **[0081]**
- **DENPCY *et al.*** *Proc. Natl. Acad. Sci. USA,* 1995, vol. 92, 6097 **[0082]**
- **KIEDROWSHI *et al.*** *Angew. Chem. Intl. Ed. English,* 1991, vol. 30, 423 **[0082]**
- **LETSINGER *et al.*** *Nucleosides & Nucleotides,* 1994, vol. 13, 1597 **[0082]**
- Carbohydrate Modifications in Antisense Research. ASC Symposium Series 580 **[0082]**
- **MESMAEKER *et al.*** *Bioorganic & Medicinal Chem. Lett.,* 1994, vol. 4, 395 **[0082]**
- **JEFFS *et al.*** *J. BiomolecularNMR,* 1994, vol. 34, 17 **[0082]**
- *Tetrahedron Lett.,* 1996, vol. 37, 743 **[0082]**
- **JENKINS *et al.*** *Chem. Soc. Rev.,* 1995, 169-176 **[0082]**
- **LOAKES *et al.*** *Nucleic Acid Res.,* 1994, vol. 22, 4039 **[0084]**
- **AERSCHOT *et al.*** *Nucleic Acid Res.,* 1995, vol. 23, 4363 **[0084]**
- **NICHOLS *et al.*** *Nature,* 1994, vol. 369, 492 **[0084]**
- **BERGSTROM *et al.*** *Nucleic Acid Res.,* 1997, vol. 25, 1935 **[0084]**
- **LOAKES *et al.*** *Nucleic Acid Res.,* 1995, vol. 23, 2361 **[0084]**
- **LOAKES *et al.*** *J. Mol. Biol.,* 1997, vol. 270, 426 **[0084]**
- **FOTIN *et al.*** *Nucleic Acid Res.,* 1998, vol. 26, 1515 **[0084]**
- **SCHEIT.** Nucleotide Analogs. John Wiley, 1980 **[0085]**
- **ENGLISCH.** *Angew. Chem. Int. Ed. Engl.,* 1991, vol. 30, 613-29 **[0085]**
- **AGARWAL.** Protocols for Polynucleotides and Analogs. Humana Press, 1994 **[0085]**
- **S. VERMA; F. ECKSTEIN.** *Ann. Rev. Biochem.,* 1998, vol. 67, 99-134 **[0085]**
- *The Glen Report,* 2003, vol. 16 (2), 5 **[0085]**
- **KOSHKIN *et al.*** *Tetrahedron,* 1998, vol. 54, 3607-30 **[0085]**
- **STERCHAK, E. P. *et al.*** *Organic Chem.,* 1987, vol. 52, 4202 **[0085]**
- **NIELSEN *et al.*** *Science,* 1991, vol. 254, 1497-1500 **[0085]**
- **EGHOLM *et al.*** *J. Am. Chem. Soc.,* 1992, vol. 114, 1895-1897 **[0085]**
- **DEMIDOV *et al.*** *Proc. Natl. Acad. Sci.,* 2002, vol. 99, 5953-58 **[0085]**
- Peptide Nucleic Acids: Protocols and Applications. Horizon Bioscience. 2004 **[0085]**
- **GORYSHIN ; REZNIKOFF.** *J. Biol. Chem.,* 1998, vol. 273, 7367 **[0090]**
- **MIZUUCHI, K.** *Cell,* 1983, vol. 35, 785 **[0090]**
- **SAVILAHTI, H *et al.*** *EMBO J.,* 1995, vol. 14, 4893 **[0090]**
- **COLEGIO *et al.*** *J. Bacteriol.,* 2001, vol. 183, 2384-8 **[0091]**
- **KIRBY C *et al.*** *Mol. Microbiol.,* 2002, vol. 43, 173-86 **[0091]**
- **DEVINE ; BOEKE.** *Nucleic Acids Res.,* 1994, vol. 22, 3765-72 **[0091]**
- **CRAIG, N L.** *Science,* 1996, vol. 271, 1512 **[0091]**
- **CRAIG, N L.** *Curr Top Microbiol Immunol.,* 1996, vol. 204, 27-48 **[0091]**
- **KLECKNER N *et al.*** *Curr Top Microbiol Immunol.,* 1996, vol. 204, 49-82 **[0091]**
- **LAMPE D J *et al.*** *EMBO J.,* 1996, vol. 15, 5470-9 **[0091]**
- **PLASTERK R H.** *Curr. Topics Microbiol. Immunol.,* 1996, vol. 204, 125-43 **[0091]**
- **GLOOR, G B.** *Methods Mol. Biol.,* 2004, vol. 260, 97-114 **[0091]**
- **ICHIKAWA ; OHTSUBO.** *J Biol. Chem.,* 1990, vol. 265, 18829-32 **[0091]**
- **OHTSUBO ; SEKINE.** *Curr. Top. Microbiol. Immunol.,* 1996, vol. 204, 1-26 **[0091]**
- **BROWN *et al.*** *Proc Natl Acad Sci USA,* 1989, vol. 86, 2525-9 **[0091]**
- **BOEKE ; CORCES.** *Annu Rev Microbiol.,* 1989, vol. 43, 403-34 **[0091]**
- **ZHANG *et al.*** *PLoS Genet.,* 2009, vol. 5, e1000689 **[0091]**
- **WILSON C. *et al.*** *J. Microbiol. Methods,* 2007, vol. 71,332-5 **[0091]**
- **RONAGHI *et al.*** *Analytical Biochemistry,* 1996, vol. 242 (1), 84-9 **[0147]**
- **RONAGHI, M.** *Genome Res.,* 2001, vol. 11 (1), 3-11 **[0147]**
- **RONAGHI *et al.*** *Science,* 1998, vol. 281 (5375), 363 **[0147]**
- **DEAMER, D.W.; AKESON, M.** Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.,* 2000, vol. 18, 147-151 **[0151]**
- **DEAMER, D.; D. BRANTON.** Characterization of nucleic acids by nanopore analysis. *Acc. Chem. Res.,* 2002, vol. 35, 817-825 **[0151]**
- **LI *et al.*** DNA molecules and configurations in a solid-state nanopore microscope. *Nat. Mater.,* 2003, vol. 2, 611-615 **[0151]**
- **SONI; MELLER.** A. Progress toward ultrafast DNA sequencing using solid-state nanopores. *Clin. Chem.,* 2007, vol. 53, 1996-2001 **[0151]**
- **HEALY, K.** Nanopore-based single-molecule DNA analysis. *Nanomed.,* 2007, vol. 2, 459-481 **[0151]**
- **COCKROFT *et al.*** A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide

resolution. *J. Am. Chem. Soc.,* 2008, vol. 130, 818-820 **[0151]**

- **LEVENE, M.J. *et al.*** Zero-mode waveguides for single-molecule analysis at high concentrations. *Science,* 2003, vol. 299, 682-686 **[0152]**
- **LUNDQUIST, P.M. *et al.*** Parallel confocal detection of single molecules in real time. *Opt. Lett.,* 2008, vol. 33, 1026-1028 **[0152]**
- **KORLACH *et al.*** Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. USA,* 2008, vol. 105, 1176-1181 **[0152]**
- **KORLACH J. *et al.*** Long, processive enzymatic DNA synthesis using 100 % dye-labeled terminal phosphate-linked nucleotides. *Nucleosides, Nucleotides and Nucleic Acids,* 2008, vol. 27, 1072-1083 **[0153]**
- **HARRIS T.D. *et al.*** Single Molecule DNA Sequencing of a viral Genome. *Science,* 2008, vol. 320, 106-109 **[0154]**
- **KINDE I. *et al.*** *PNAS,* 2011, vol. 108, 9530-9535 **[0162]**
- **SCHMITT M.W. *et al.*** *PNAS,* 2012, vol. 109, 14508-13 **[0165]**
- **HAAPA *et al.*** *Nucl. Acids Res.,* 1999, vol. 27 (3), 2777-2784 **[0178]**