(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 528 253**

(51) Int. CI.:

C12Q 1/68 (2006.01)
G01N 37/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **06.11.2009 E 09824428 (8)**

(97) Fecha y número de publicación de la concesión europea: **17.12.2014 EP 2245187**

(54) Título: **Métodos para la determinación precisa de datos de secuencia y de la posición de bases modificadas**

(30) Prioridad:

07.11.2008 US 112548 P
07.04.2009 US 167313 P
05.11.2009 US 613291

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.02.2015

(73) Titular/es:

INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE (100.0%)
No. 195, Sec. 4 Chung Hsing Road
Chutung Hsinchu 31040, Taiwan, TW

(72) Inventor/es:

PAN, CHAO-CHI;
FANN, JENN-YEH;
CHIOU, CHUNG-FAN;
CHIEN, HUNG-CHI y
CHEN, HUI-LING

(74) Agente/Representante:

PONTI SALES, Adelaida

DESCRIPCIÓN

Métodos para la determinación precisa de datos de secuencia y de la posición de bases modificadas

Campo de la invención

[0001] La presente invención se refiere a métodos de determinación de la secuencia de ácidos nucleicos y de identificación de las posiciones de bases modificadas en ácidos nucleicos.

Antecedentes de la invención

[0002] Los recientes desarrollos en la tecnología de secuenciación del ADN han planteado la posibilidad de un tipo de medicina preventiva muy personalizada a nivel genómico. Además, la posibilidad de obtener rápidamente grandes cantidades de datos de secuencia a partir de muchos individuos en una o más poblaciones podría marcar el comienzo de una nueva fase de la revolución genómica en la ciencia biomédica.

[0003] Las diferencias de bases individuales entre genotipos pueden tener efectos fenotípicos importantes. Por ejemplo, se han identificado más de 300 mutaciones en el gen que codifica la fenilalanina hidroxilasa (PAH), la enzima que convierte la fenilalanina a tirosina en el catabolismo de la fenilalanina y en la biosíntesis de proteínas y neurotransmisores que produce una actividad enzimática deficiente y da lugar a los trastornos hiperfenilalaninemia y fenilcetonuria. Véase, por ejemplo, Jennings y col., Eur J Hum Genet 8, 683-696 (2000).

[0004] Los datos de secuencia se pueden obtener usando el método de secuenciación de Sanger, en el que análogos de nucleótidos terminadores de cadena didesoxi marcados se incorporan a una reacción de extensión del cebador en bruto y se resuelven y analizan productos de longitudes diferentes para determinar la identidad del terminador incorporado. Véase, por ejemplo, Sanger y col., Proc Natl Acad Sci USA 74, 5463-5467 (1997). De hecho, muchas secuencias genómicas se han determinado usando esta tecnología. No obstante, el coste y la velocidad para obtener los datos de secuencia mediante la secuenciación de Sanger pueden ser una limitación.

[0005] Nuevas tecnologías de secuenciación pueden producir datos de secuencia a una velocidad sorprendente de cientos de megabases al día, con un coste por base inferior al de la secuenciación de Sanger. Véase, por ejemplo, Kato, Int J Clin Exp Med 2, 193-202 (2009). No obstante, los datos en bruto obtenidos usando estas tecnologías de secuenciación pueden ser más propensos a errores que la secuenciación de Sanger tradicional. Esto puede ser el resultado de obtener información a partir de moléculas de ADN individuales en lugar de una población mayor.

[0006] Por ejemplo, en la secuenciación por síntesis de moléculas sencillas, se podría saltar una base debido a que el dispositivo ignore una señal débil, o debido a la falta de señal como resultado del blanqueamiento del colorante fluorescente, o debido a que la polimerasa actúe demasiado deprisa para su detección por el dispositivo. Todos los casos anteriores producen un error de deleción de la secuencia en bruto. De forma similar, también se pueden producir errores de mutación y errores de inserción a una mayor frecuencia por razones tan simples como señales potencialmente más débiles y reacciones más rápidas que en los métodos convencionales.

[0007] Los datos de secuencia de baja precisión son más difíciles de ensamblar. En la secuenciación a gran escala, tal como la secuenciación de un genoma eucariota completo, las moléculas de ADN están fragmentadas en trozos más pequeños. Estos trozos se secuencian en paralelo, y a continuación las lecturas resultantes se ensamblan para reconstruir la secuencia completa de las moléculas de ADN de la muestra original. La fragmentación se puede conseguir, por ejemplo, mediante cizallamiento mecánico o escisión enzimática.

[0008] El ensamblaje de lecturas de secuencia pequeñas en un gran genoma requiere que las lecturas fragmentadas sean suficientemente precisas para que se puedan agrupar juntas de forma correcta. Esto generalmente es cierto para los datos de secuenciación en bruto generados en el método de Sanger, que puede tener una precisión de los datos en bruto superior al 95 %. La tecnología de secuenciación precisa de moléculas sencillas se podría aplicar para detectar modificaciones de bases individuales o mutaciones de muestras de ácidos nucleicos. No obstante, la precisión de los datos en bruto para las tecnologías de secuenciación de moléculas sencillas pueden ser inferiores debido a las limitaciones descritas anteriormente. La precisión de las lecturas individuales de datos de secuencia en bruto puede ser sólo del 60 al 80 %. Véase, por ejemplo, Harris y col., Science 320:106-109 (2008). Así, sería útil proporcionar métodos precisos de secuenciación de moléculas sencillas.

[0009] Además, la metilación del ADN desempeña un papel crítico en la regulación de la expresión génica; por ejemplo, la metilación de los promotores con frecuencia da lugar al silenciamiento transcripcional. También se sabe que la metilación es un mecanismo esencial en la impronta genómica y en la inactivación del cromosoma X. No obstante, el progreso para descifrar los perfiles de metilación genómicos completos y complejos ha sido limitado. Por tanto, podrían ser útiles métodos de determinación de perfiles de metilación del ADN de alto rendimiento, más incluso si los métodos también proporcionan una determinación de secuencia precisa.

[0010] El documento de Estados Unidos 2004/248161 A1 desvela un método para la secuenciación de un ácido nucleico. El método incluye el suministro de un complejo circular molde con cebador de anclaje imprimado y la combinación del complejo con polimerasa, y nucleótidos para generar copias complementarias lineales y concatenadas del molde circular.

[0011] El documento de Estados Unidos 2006/061754 A1 desvela un método para la secuenciación de moléculas de ADN diana. La reacción de la secuencia incluye un complejo individual de una enzima de polimerización por desplazamiento de la cadena y un ADN diana circular, que se inmoviliza por confinamiento óptico.

Sumario de la invención

[0012] La invención está limitada por las reivindicaciones 1-24.

[0013] En algunas realizaciones, se proporciona un método de determinación de la secuencia de una muestra de ácidos nucleicos que comprende (a) el suministro de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que el inserto tiene una secuencia conocida; (b) la obtención de los datos de secuencia que comprende la secuencia de al menos dos unidades de inserto-muestra, en la que se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra; (c) el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (b) al comparar las secuencias con la secuencia conocida del inserto; (d) aceptar o rechazar al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (b) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos; (e) la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (d); y (f) la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas. En el presente documento se desvela un sistema que comprende un aparato de secuenciación unido de manera operable a un aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario, el almacenamiento que está codificado mediante programación que comprende un sistema operativo, un *software* de interfaz de usuario, e instrucciones que, cuando las ejecuta el procesador, opcionalmente mediante la introducción por parte del usuario, realiza un método que comprende: (a) la obtención de los datos de secuencia de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que: (i) el inserto tiene una secuencia conocida, (ii) los datos de secuencia comprenden la secuencia de al menos dos unidades de inserto-muestra, y (iii) se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra; (b) el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (a) al comparar las secuencias con la secuencia conocida del inserto; (c) aceptar o rechazar al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (a) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos; (d) la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (c); y (e) la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas, en la que se usa un resultado del sistema para producir al menos una de (i) una secuencia de una muestra de ácidos nucleicos o (ii) una indicación de que existe una base modificada en al menos una posición en una muestra de ácidos nucleicos.

[0014] La invención está limitada por las reivindicaciones 1-24.

[0015] En algunas realizaciones, se proporciona un almacenamiento codificado mediante programación que comprende un sistema operativo, un *software* de interfaz de usuario, e instrucciones que, cuando las ejecuta el procesador sobre un sistema que comprende un aparato de secuenciación unido de manera operable a un aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario, opcionalmente con la introducción por parte del usuario, realiza un método que comprende: (a) la obtención de los datos de secuencia de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que: (i) el inserto tiene una secuencia conocida, (ii) los datos de secuencia comprenden la secuencia de al menos dos unidades de inserto-muestra, y (iii) se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra; (b) el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (a) al comparar las secuencias con la secuencia conocida del inserto; (c) aceptar o rechazar al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (a) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos; (d) la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (c); y (e) la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas, en la que el método da lugar a un resultado usado para producir al menos una de (i) una secuencia de una muestra de ácidos nucleicos o (ii) una indicación de que existe una base modificada en al

menos una posición en una muestra de ácidos nucleicos.

[0016] En algunas realizaciones, se describe un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena y una posición de al menos una base modificada en la secuencia, que comprende: (a) el cierre de las cadenas directa e inversa juntas para formar una molécula de par cerrado circular; (b) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; (c) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; (d) la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases en la molécula de par cerrado circular para producir una molécula de par cerrado circular alterada; (e) la obtención de los datos de secuencia de la molécula de par cerrado circular alterada en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa alteradas; y (f) la determinación de las posiciones de bases modificadas en la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa alteradas.

[0017] En algunas realizaciones, la divulgación proporciona un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena, que comprende: (a) el cierre de las cadenas directa e inversa de la muestra de ácidos nucleicos juntas para formar una molécula de par cerrado circular; (b) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; y (c) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular.

[0018] En algunas realizaciones, la divulgación proporciona un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena y una posición de al menos una base modificada en la secuencia, que comprende: (a) el cierre de las cadenas directa e inversa de la muestra de ácidos nucleicos juntas para formar una molécula de par cerrado circular; (b) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; y (c) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular.

[0019] En algunas realizaciones, la divulgación proporciona un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena y una posición de al menos una base modificada en la secuencia, que comprende: (a) el cierre de las cadenas directa e inversa de la muestra de ácidos nucleicos juntas para formar una molécula de par cerrado circular; (b) la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases en la molécula de par cerrado circular para producir una molécula de par cerrado circular alterada; (c) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; y (d) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena y la posición de al menos una base modificada en la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular.

[0020] En algunas realizaciones, la divulgación proporciona un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena y una posición de al menos una base modificada en la secuencia, que comprende: (a) el cierre de las cadenas directa e inversa de la muestra de ácidos nucleicos juntas para formar una molécula de par cerrado circular; (b) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que los datos de secuencia comprenden secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; (c) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular; (d) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que se usa al menos un análogo de nucleótido que discrimina entre una base y su forma modificada para obtener los datos de secuencia que comprenden al menos una posición en la que se incorporó el al menos un análogo de nucleótido marcado de forma distintiva; y (e) la determinación de las posiciones de bases modificadas en la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa.

[0021] En algunas realizaciones, la divulgación proporciona un método de determinación de una secuencia de una muestra de ácidos nucleicos de doble cadena y una posición de al menos una base modificada en la secuencia, que comprende: (a) el cierre de las cadenas directa e inversa de la muestra de ácidos nucleicos juntas para formar una molécula de par cerrado circular; (b) la obtención de los datos de secuencia de la molécula de par cerrado circular mediante la secuenciación de una sola molécula, en la que se usa al menos un análogo de nucleótido que discrimina entre una base y su forma modificada para obtener los datos de secuencia que comprenden al menos una posición en la que se incorporó el al menos un análogo de nucleótido marcado de forma distintiva; y (c) la determinación de la secuencia de la muestra de ácidos nucleicos de doble cadena y la posición de la al menos una

base modificada en la secuencia de la muestra de ácidos nucleicos de doble cadena al comparar las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular.

[0022] Objetos y ventajas adicionales de la invención se expondrán en parte en la descripción siguiente, y en parte serán obvias a partir de la descripción. Los objetos y ventajas de la invención se concretarán y conseguirán por medio de los elementos y combinaciones apuntados en las reivindicaciones anexas.

[0023] Se debe entender que tanto la descripción general anterior como la descripción detallada siguiente son únicamente ejemplares y explicativas y no son restrictivas de la invención, como se reivindica.

[0024] Los dibujos acompañantes, que se incorporan y constituyen una parte de esta memoria descriptiva, ilustran diversas realizaciones de la divulgación y, junto con la descripción, sirven para explicar los principios de la invención, como se definen las reivindicaciones 1-24.

Breve descripción de los dibujos

[0025] Los aspectos y ventajas anteriores de esta divulgación podrán ser evidentes a partir de la siguiente descripción detallada con referencia a los dibujos acompañantes en los que:

Figura 1. Preparación de una molécula de ADN circular de acuerdo con algunas realizaciones descritas en el presente documento. Se fragmentó una muestra de ADN 1; un fragmento 2 se ligó en su extremo 5' (diamante) a un enlazador 3 y en su extremo 3' (flecha) a otro enlazador 4. Los enlazadores 3 y 4 son complementarios a segmentos contiguos de un oligonucleótido 5. La hibridación de 5 a 3 y 4 proporciona un sustrato para la circularización por ligación, reacción que produce una molécula circular 6 que comprende un inserto de ácidos nucleicos (de la secuencia de los enlazadores 3 y 4) y una muestra de ácidos nucleicos (de la secuencia de fragmento 2).

Figura 2. Amplificación por círculo rodante. Un oligonucleótido 5, hibridado a una molécula circular 6 producida como en la Figura 1, se unió mediante una polimerasa 7 anclada a una superficie 8. La extensión del oligonucleótido proporciona una copia lineal complementaria 9 de la molécula circular. La extensión continua produce un desplazamiento de la cadena y la síntesis de una molécula 10 que contiene múltiples copias de la molécula circular.

Figura 3. Molécula de par cerrado circular. (A) Una molécula de doble cadena que contiene una cadena directa 11 y una cadena inversa 12 se pueden combinar con insertos que forman las horquillas 13 y 14, que pueden ser idénticas o diferentes, para formar una molécula de par cerrado circular. En algunas realizaciones, los enlazadores tienen extremos escalonados y rebajados (37 y 38). Estos se pueden rellenar usando una polimerasa o pueden ser complementarios a extremos escalonados de la molécula de doble cadena (no mostrado). En una molécula de par cerrado circular completa, 37 y 38 se rellenan y se sellan de forma que la molécula tenga un esqueleto circular continuo y de cadena sencilla. (B) Después de rellenar los huecos y unir los extremos según sea adecuado, se forma un ADN circular que contiene la cadena directa 11, el enlazador 14, la cadena inversa 12, y el enlazador 13, presentados en este documento en forma fundida. La molécula se puede convertir a la forma de doble cadena, por ejemplo, mediante la hibridación de un cebador a uno de los enlazadores y su extensión usando una polimerasa sin actividad de desplazamiento de la cadena, por ejemplo, ADN polimerasa I de *E. coli*, seguido de ligación.

Figura 4. Esquemas para la determinación de la secuencia y determinación del perfil de secuencia y de metilación usando moléculas de par cerrado circulares. (Izquierda) Una molécula de par cerrado circular se puede secuenciar durante al menos una longitud completa de la molécula para proporcionar lecturas de la secuencia complementaria; se puede usar la secuenciación continua para proporcionar mayor redundancia. Los datos de secuencia se pueden alinear y evaluar en base a las secuencias de los ácidos nucleicos de inserto para así obtener una secuencia precisa de los ácidos nucleicos de muestra. (Derecha) Se puede usar la conversión de un tipo específico de nucleótido, tal como mediante conversión con bisulfito o transición fotoquímica, seguido de secuenciación, alineamiento, y comparación de la secuencia modificada y su complemento sin modificar para obtener datos de secuencia y perfiles de metilación precisos. Se pueden usar lecturas de secuencia extendida que contienen múltiples repeticiones de la secuencia de ácidos nucleicos de muestra para incrementar la precisión.

Figura 5. Conversión de nucleótidos. (A) Una molécula de par cerrado circular que contienen los insertos 13 y 14, una cadena directa 15 que contiene al menos un resto de 5-metilcitosina (^{m}C), y una cadena inversa 16 se someten a tratamiento, tal como transición fotoquímica, para convertir la ^{m}C en T, produciendo una cadena directa 17 convertida. Los nucleótidos complementarios en la cadena inversa están inalterados, produciendo un par inestable G-T. (Los restos ^{m}C en la cadena inversa, si están presentes, se convertirían mediante el tratamiento). (B) Una molécula de par cerrado circular que contiene los insertos 13 y 14, una cadena directa 15 que contiene al menos un resto de 5-metilcitosina (^{m}C), y una cadena inversa 16 se someten a tratamiento, tal como conversión con bisulfito, para convertir C (pero no ^{m}C) en U, produciendo una cadena directa 39 convertida y una cadena inversa 40 convertida. Los nucleótidos complementarios a los nucleótidos convertidos están inalterados, dando lugar a pares inestables G-U.

Figura 6. Obtención de los datos de secuencia y de un perfil de metilación de una molécula de par cerrado circular.

(A) Un cebador 18 se hibridó a la molécula de par cerrado circular convertida de la Figura 5A y se extendió con una polimerasa, produciendo la síntesis de una cadena con los segmentos 19, 20, y 21, complementarios a las secuencias de 16, 14, y 17, respectivamente. (B) Se obtiene la secuencia que comprende al menos dos repeticiones: al menos una de una repetición de la muestra 17 y una repetición del complemento recién sintetizado de la cadena directa 21; y al menos una de una repetición del complemento recién sintetizado de la cadena inversa 19 y una repetición de la cadena inversa 16. Estas repeticiones se alinean; una posición 41 en la que hay discrepancia entre las repeticiones significa que en esa posición se ha modificado una base. Dependiendo del tipo de modificación usada, se pueden determinar las bases presentes originalmente en la posición correspondiente de la muestra de ácidos nucleicos. En este ejemplo, en el que la molécula de par cerrado circular se ha modificado por conversión de ^{m}C en T (véase Figura 5A), la discrepancia indica que había presente una ^{m}C en la muestra de ácidos nucleicos en la cadena directa en posición 41; lo lógico es que en una posición en la que las secuencias discrepan, la base que es el producto de la reacción de conversión, T, haya sustituido el sustrato de la reacción de conversión, ^{m}C, que estaba presente en la muestra de ácidos nucleicos.

Figura 7. Datos de secuencia en bruto y procesados obtenidos de un molde de una molécula de ácidos nucleicos circular.

(A) El contenido de secuencia que se puede obtener a partir de un molde circular está representado esquemáticamente. La secuencia de la muestra de ácidos nucleicos está representada por líneas y la secuencia del inserto de ácidos nucleicos está representada por círculos. La secuencia ilustrada comienza con una secuencia parcial 22 de una muestra de ácidos nucleicos, seguida de la secuencia de un inserto de ácidos nucleicos 23; éstas van seguidas de una secuencia 24 de la muestra de ácidos nucleicos, una secuencia 25 de un inserto de ácidos nucleicos, una secuencia 26 de la muestra de ácidos nucleicos, y una secuencia 27 de un inserto de ácidos nucleicos. 28 representa una secuencia adicional no mostrada en esta figura, que va seguida de una secuencia 29 de un inserto de ácidos nucleicos, una secuencia 30 de la muestra de ácidos nucleicos, una secuencia 31 de un inserto de ácidos nucleicos, y una secuencia parcial 32 de una muestra de ácidos nucleicos.

Si el molde circular comprende una muestra sencilla de ácidos nucleicos y un inserto sencillo de ácidos nucleicos, entonces tanto 22 como 24, junto con las subsiguientes secuencias de muestra de ácidos nucleicos 26, 30, y 32, son secuencias de la misma muestra sencilla de ácidos nucleicos; asimismo, en este caso 23, 25, 27, 29, y 31 son secuencias del mismo inserto sencillo de ácidos nucleicos. Si el molde circular comprende repeticiones directa e inversa de la secuencia de la muestra de ácidos nucleicos y los insertos de ácidos nucleicos que tienen secuencias conocidas, que pueden ser idénticas o diferentes, como en el caso de una molécula de par cerrado circular, entonces las secuencias de muestra de ácidos nucleicos tienen orientaciones alternadas y corresponden a las dos repeticiones de la muestra de ácidos nucleicos de forma alterna (por ejemplo, 22 podría estar en orientación directa, es decir, es una secuencia de la repetición inversa, y 24 podría estar en orientación inversa, es decir, es una secuencia de la repetición directa, o viceversa). Asimismo, las secuencias del inserto de ácidos nucleicos 23, 25, etc., también corresponderían a los dos insertos de ácidos nucleicos, que pueden ser idénticos o diferentes, del molde circular de forma alternada.

(B) La secuencia mostrada en la Figura 7A se puede descomponer en segmentos, cada uno que contiene una repetición de la secuencia de muestra de ácidos nucleicos, por ejemplo, 24; los elementos también comprenden al menos una repetición del inserto de ácidos nucleicos, por ejemplo, dos repeticiones del inserto de ácidos nucleicos, por ejemplo, 23 y 25. Algunos segmentos pueden contener únicamente una secuencia parcial, por ejemplo, 33, o una secuencia excepcionalmente larga, por ejemplo, 34. Dichos segmentos pueden ser el resultado de errores durante la secuenciación. En algunas realizaciones, dichos segmentos se excluyen de consideraciones adicionales.

Figura 8. Diagrama de las etapas del procesamiento de secuencia. En algunas realizaciones, se examinan los datos de secuencia en bruto, se procesan y se aceptan o se rechazan como se muestra.

Fig. 9. Productos de amplificación por círculo rodante. Los productos de las reacciones descritas en el Ejemplo 1 se sometieron a electroforesis y el gel se visualizó como se describe. Partiendo de la izquierda, C1 y C2 son los carriles de los controles negativos. El carril Mr más a la izquierda contiene la escalera patrón de 1 kb FERMENTAS GENERULER, Cat. No. SM0311, con unos tamaños de banda que oscilan entre 250 y 10.000 pb. Los siguientes 10 carriles contienen productos de las reacciones de amplificación por círculo rodante como se ha indicado, generados usando dos cebadores o un cebador (control de la amplificación) y productos de las reacciones L0 (control de ligación negativo) o L3, reacciones de ligación tomadas en los tiempos indicados; véase Ejemplo 1. El siguiente carril Mr contiene la escalera Plus de 100 pb FERMENTAS GENERULER, Cat. No. SM0321, con unos tamaños de banda que oscilan entre 100 y 3000 pb. Los siguientes 10 carriles contienen los mismos productos que en los 10 carriles de producto anteriores excepto porque estos productos se mezclaron con colorante de carga que contiene SDS al 1 %.

Figura 10. Alineamientos que muestran secuencias de repetición y la secuencia original deducida de una muestra de ácidos nucleicos simulada. Las posiciones en las que todas las secuencias alineadas concuerdan están marcadas por asteriscos. (A) Las lecturas a y b del Ejemplo 2 se muestran junto con la secuencia original deducida, marcada "o", de la cadena directa de la muestra de ácidos nucleicos. La secuencia original se dedujo usando las reglas mostradas en la Tabla 5. Las posiciones en las que las tres secuencias mostradas tienen C son posiciones en las

que la muestra de ácidos nucleicos simulada contenía una citosina metilada en la cadena directa. Las posiciones en las que las tres secuencias mostradas tienen G son posiciones en las que la muestra de ácidos nucleicos simulada contenía una citosina metilada en la cadena inversa. (B) Las lecturas a y b del Ejemplo 3 se muestran junto con la secuencia original deducida de la cadena directa, marcada "r_a". La secuencia original se dedujo usando las reglas de la Tabla 6. Las posiciones en las que la secuencia original deducida tiene una C que discrepan con la lectura a son posiciones en las que la muestra de ácidos nucleicos simulada contenía una citosina metilada en la cadena directa. Las posiciones en las que la secuencia original deducida tiene una G que discrepan con la lectura b son posiciones en las que la muestra de ácidos nucleicos simulada contenía una citosina metilada en la cadena inversa.

Figura 11. Aparato de computación y almacenamiento. (A) En algunas realizaciones, la divulgación se refiere a un aparato de secuenciación 51 unido de manera operable a un aparato de computación 52 que comprende al menos un elemento de interfaz de usuario seleccionado entre una pantalla 57, un teclado 58, y un ratón 59, y al menos un ordenador 53 que comprende un almacenamiento 54 (véase panel B), un sistema de bus 55, y un procesador 56. (B) En algunas realizaciones, se describe un almacenamiento 54 que comprende un sistema operativo 60, un *software* de interfaz de usuario 61, y un *software* de procesamiento 62. El almacenamiento además puede comprender datos de secuencia 63 obtenidos del aparato de secuenciación (51 en la Figura 11A).

Figura 12. Esquema general de determinación de la secuencia y la posición de la 5-metilcitosina usando conversión con bisulfito con una molécula de par cerrado lineal. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende 5-metilcitosina (en la parte superior). Se construye una molécula de par cerrado lineal ligando un inserto de horquilla a un extremo de doble cadena de la molécula (debajo de la primera flecha, a la derecha), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Además, hay unidas solapas (*flaps*, en inglés) lineales al otro extremo de doble cadena (a la izquierda). Se realiza la conversión con bisulfito, convirtiendo las citosinas en uracilos pero dejando las 5-metilcitosinas inalteradas. La molécula se copia al proporcionar un cebador que se une a la solapa lineal unida en el extremo 3' de la molécula de par cerrado lineal y al extender el cebador con una polimerasa. Los extremos se pueden procesar, por ejemplo, mediante digestión por restricción, para preparar la molécula para su posterior clonación y/o secuenciación.

Figura 13. Esquema general de determinación de la secuencia y la posición de la 5-metilcitosina usando transición fotoquímica con una molécula de par cerrado lineal. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende 5-metilcitosina (en la parte superior). Se construye una molécula de par cerrado lineal ligando un inserto de horquilla a un extremo de doble cadena de la molécula (debajo de la primera flecha, a la derecha), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Además, hay unidas solapas lineales al otro extremo de doble cadena (a la izquierda). Se realiza la transición fotoquímica, convirtiendo las 5-metilcitosinas en timinas pero dejando las citosinas inalteradas. La molécula se copia al proporcionar un cebador que se une a la solapa lineal unida en el extremo 3' de la molécula de par cerrado lineal y al extender el cebador con una polimerasa. Los extremos se pueden procesar, por ejemplo, mediante digestión por restricción, para preparar la molécula para su posterior clonación y/o secuenciación.

Figura 14. Esquema general de determinación de la secuencia usando una molécula de par cerrado lineal. Se proporciona una muestra de ácidos nucleicos de doble cadena (en la parte superior). Se construye una molécula de par cerrado lineal ligando un inserto de horquilla a ambos extremos de doble cadena de la molécula (debajo de la primera flecha, a la derecha y a la izquierda), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Se realiza la secuenciación y los datos de secuencia se analizan para determinar la secuencia de la muestra; véase, por ejemplo, Ejemplo 5.

Figura 15. Esquema general de determinación de la secuencia y la posición de la 5-metilcitosina usando conversión con bisulfito y una molécula de par cerrado circular. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende 5-metilcitosina (en la parte superior). Se construye una molécula de par cerrado circular ligando un inserto de horquilla a ambos extremos de doble cadena de la molécula (debajo de la primera flecha, a la derecha y a la izquierda), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Se realiza la conversión con bisulfito, convirtiendo las citosinas en uracilos pero dejando las 5-metilcitosinas inalteradas. Se realiza la secuenciación y los datos de secuencia se analizan para determinar la secuencia de la muestra y las posiciones de la 5-metilcitosina; véase, por ejemplo, Ejemplo 6.

Figura 16. Esquema general de determinación de la secuencia y la posición de la 5-metilcitosina usando transición fotoquímica y una molécula de par cerrado circular. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende 5-metilcitosina (en la parte superior). Se construye una molécula de par cerrado circular ligando un inserto de horquilla a ambos extremos de doble cadena de la molécula (debajo de la primera flecha, a la derecha y a la izquierda), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Se realiza la transición fotoquímica, convirtiendo las 5-metilcitosinas en timinas pero dejando las citosinas inalteradas. Se realiza la secuenciación y los datos de secuencia se analizan para determinar la secuencia de la muestra y las posiciones de la 5-metilcitosina; véase, por ejemplo, Ejemplo 7.

Figura 17. Esquema general de determinación de la secuencia y la posición del 5-bromouracilo usando una molécula de par cerrado circular. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende 5-

bromouracilo (en la parte superior). Se construye una molécula de par cerrado circular ligando un inserto de horquilla a ambos extremos de doble cadena de la molécula (debajo de la primera flecha, a la derecha y a la izquierda), cerrando juntas de esta forma las cadenas directa e inversa de la muestra de doble cadena. Se realiza la secuenciación y los datos de secuencia se analizan para determinar la secuencia de la muestra y las posiciones de la 5-metilcitosina; véase, por ejemplo, Ejemplo 7.

Descripción detallada de la invención

Definiciones

[0026] Para facilitar la comprensión de esta invención, a continuación se definen una serie de términos. Los términos no definidos en el presente documento tienen los significados que entiende habitualmente el experto en la materia en las áreas pertinentes para la presente invención. Términos tales como "un", "una" y "el/la" no está previsto que se refieran únicamente a una entidad singular, sino que incluyan la clase general, de la cual se puede usar un ejemplo específico para su ilustración. La terminología del presente documento se usa para describir realizaciones específicas de la invención, pero su uso no delimita la invención, excepto por lo que se indica en las reivindicaciones 1-24.

[0027] El término ácido nucleico incluye oligonucleótidos y polinucleótidos.

[0028] Condiciones muy rigurosas para la hibridación se refiere a condiciones en las cuales dos ácidos nucleicos deben poseer un alto grado de homología entre sí para hibridarse. Ejemplos de condiciones muy rigurosas para la hibridación incluyen hibridación en cloruro sódico/citrato sódico 4X (SSC), a 65 o 70 ºC, o hibridación en SSC 4X más 50 % de formaldehído a 42 o 50 ºC aproximadamente, seguido de al menos uno, al menos dos o al menos tres lavados en SSC 1X, a 65 o 70 ºC.

[0029] La temperatura de fusión se refiere a la temperatura a la cual la mitad de un ácido nucleico en solución se encuentra en estado fundido y la otra mitad se encuentra en estado no fundido, asumiendo la presencia de ácidos nucleicos complementarios suficientes. En el caso de un oligonucleótido presente en exceso sobre la secuencia complementaria, la temperatura de fusión es la temperatura a la cual la mitad de la secuencia complementaria se hibrida con el oligonucleótido. En el caso de un inserto de ácidos nucleicos capaz de formar una horquilla, la temperatura de fusión es la temperatura a la cual la mitad del inserto está en forma de "horquilla" parcialmente auto-hibridada. Puesto que la temperatura de fusión depende de las condiciones, las temperaturas de fusión de los oligonucleótidos descritas en el presente documento se refieren a la temperatura de fusión en solución acuosa de cloruro sódico 50 mM, con el oligonucleótido a 0,5 µM. Las temperaturas de fusión se pueden estimar por diversos métodos conocidos en la técnica, por ejemplo, usando los parámetros termodinámicos vecinos más próximos encontrados en Allawi y col., Biochemistry, 36, 10581-10594 (1997) junto con ecuaciones termodinámicas convencionales.

[0030] Un sitio en una molécula de ácidos nucleicos es adecuado para la unión de un cebador si tiene una secuencia única en la molécula de ácidos nucleicos y es de una longitud y composición tales que el oligonucleótido complementario tiene una temperatura de fusión aceptable, por ejemplo, una temperatura de fusión que oscila entre 45 ºC y 70 ºC, entre 50 ºC y 70 ºC, entre 45 ºC y 65 ºC, entre 50 ºC y 65 ºC, entre 55 ºC y 70 ºC, entre 60 ºC y 70 ºC, entre 55 ºC y 60 ºC, entre 60 ºC y 65 ºC, o entre 50 ºC y 55 ºC.

[0031] La extensión de un cebador, un oligonucleótido, o un ácido nucleico se refiere a la adición de al menos un nucleótido al cebador, oligonucleótido, o ácido nucleico. Esto incluye reacciones catalizadas por la actividad polimerasa o ligasa.

[0032] Un cebador de secuenciación es un oligonucleótido que se puede unir a un sitio en una molécula de ácidos nucleicos que es adecuado para la unión del cebador y que se extiende en la reacción de secuenciación para así producir los datos de secuencia.

[0033] Un inserto de ácidos nucleicos es capaz de formar una horquilla si se puede auto-hibridar parcialmente, y si la forma auto-hibridada tiene una temperatura de fusión de al menos 15 ºC.

[0034] Un extremo escalonado es un segmento de cadena sencilla en el extremo de una molécula u horquilla de ácidos nucleicos de doble cadena.

[0035] Una repetición o secuencia de repetición es una secuencia que se encuentra más de una vez en un ácido nucleico. Cuando las repeticiones están presentes en una molécula de ácidos nucleicos, todas las unidades de la secuencia, incluyendo la primera unidad, se consideran repeticiones. Las repeticiones incluyen secuencias que son complementos inversos entre sí, tales como los que se encuentran en una molécula de par cerrado circular. Las repeticiones también incluyen secuencias que no son exactamente idénticas sino que proceden de la misma secuencia, por ejemplo, secuencias que difieren debido a eventos de incorporación errónea u otros errores de la polimerasa durante la síntesis, o secuencias que inicialmente eran idénticas o complementos inversos perfectos pero

que difieren debido a su modificación por un procedimiento tal como transición fotoquímica o tratamiento con bisulfito.

[0036] Un inserto de ácidos nucleicos y una muestra de ácidos nucleicos se encuentran inmediatamente aguas arriba o aguas abajo entre sí si no hay otras repeticiones del inserto o de la muestra que intervengan entre el inserto y la muestra. En una molécula de cadena sencilla, aguas arriba se refiere a la dirección 5' y aguas abajo se refiere a la dirección 3'. En una molécula de doble cadena, la polaridad se puede determinar arbitrariamente o se puede determinar según la polaridad de los elementos direccionales tales como promotores, secuencias codificantes, etc., si la mayoría de dichos elementos están orientados de la misma forma. La polaridad de un promotor es aquella en la que la dirección de una síntesis de iniciación de ARN polimerasa es aguas abajo. La polaridad de una secuencia codificante es aquella en la que la dirección desde el codón de iniciación al de detención es aguas abajo.

[0037] Dos repeticiones están en orientaciones directa e inversa una con respecto a la otra, y tienen orientaciones opuestas, si son complementos inversos entre sí o uno o los dos proceden del complemento inverso entre sí. Cuál de las dos repeticiones se considera directa puede ser arbitrario o se puede determinar según la polaridad de los elementos en la repetición, como se ha descrito en el párrafo anterior.

[0038] Una base modificada es una base distinta a la adenina, timina, guanina, citosina o uracilo que se puede incluir en lugar de una o más de las bases anteriormente mencionadas en un ácido nucleico o nucleótido.

[0039] Los códigos de ambigüedad son códigos que representan una combinación de bases en una secuencia, en el sentido de que podría estar presente cualquiera de las bases representadas, por ejemplo: Y = pirimidina (C, U o T); R = purina (A o G); W = débil (A, T, o U); S = fuerte (G o C); K = ceto (T, U, o G); M = amino (C o A); D = no C (A, G, T, o U); V = no T o U (A, C o G); H = no G (A, C, T, o U); B = no A (C, G, T o U).

[0040] Una matriz de peso por posición es una matriz en la que las filas corresponden a posiciones en la secuencia de ácidos nucleicos y las columnas corresponden a bases, o viceversa, y cada elemento en la matriz es un peso para una base particular en una posición particular. Una secuencia se puede puntuar contra una matriz de peso por posición al sumar los pesos correspondientes a cada base de la secuencia; por ejemplo, si la secuencia es ACG, la puntuación sería la suma del peso para A en la primera columna de la matriz, el peso para C en la segunda columna, y el peso para G en la tercera columna, asumiendo que las columnas se corresponden con las posiciones. Una matriz de peso por posición se puede ejecutar sobre una secuencia con una longitud mayor al número de posiciones de la matriz mediante la puntuación iterativa de la secuencia contra la matriz, en la que la posición de partida se incrementa en una posición en cada iteración. De esta forma, se puede identificar una secuencia que produzca una puntuación máxima o mínima contra la matriz.

[0041] El almacenamiento se refiere a un repositorio de información digital accesible por medio de un ordenador. Incluye RAM, ROM, discos duros, memoria de estado sólido no volátil, discos ópticos, discos magnéticos, y sus equivalentes.

[0042] Una estructura de datos es un objeto o variable en un almacenamiento que contiene datos. Una estructura de datos puede contener datos escalares (por ejemplo, un carácter, número o cadena individuales), un conjunto de datos escalares (por ejemplo, una matriz o colección de escalares), o un conjunto recursivo (por ejemplo, una lista, que puede ser multidimensional, que comprende sub-listas, matrices, colecciones, y/o escalares como elementos, con las sub-listas que pueden contener ellas mismas sub-listas, matrices, colecciones, y/o escalares como elementos).

Muestra de ácidos nucleicos

[0043] Los métodos comprenden la determinación de la secuencia de una muestra de ácidos nucleicos y/o la determinación de las posiciones de bases modificadas en una muestra de ácidos nucleicos. El término "muestra de ácidos nucleicos" se refiere al ácido nucleico cuya secuencia y/o posiciones de bases modificadas se deben determinar de acuerdo con los métodos descritos en el presente documento.

[0044] La muestra de ácidos nucleicos se puede obtener de una fuente que incluye, sin limitación, ADN (incluyendo sin limitación ADN genómico, ADNc, ADN mitocondrial, ADN de cloroplastos, y ADN extracromosómico o extracelular) o ARN (incluyendo sin limitación ARNm, ARN transcrito primario, ARNt, ARNr, miARN, ARNip, y ARNnop). La muestra de ácidos nucleicos puede proceder de un individuo, paciente, espécimen, cultivo celular, biopelícula, órgano, tejido, célula, espora, animal, planta, hongo, protista, bacteria, archaea, virus, o virión. En algunas realizaciones, la muestra de ácidos nucleicos se obtiene en forma de muestra del entorno, por ejemplo, del suelo o de una masa de agua; la muestra de ácidos nucleicos se puede obtener en forma de muestra del entorno sin conocimiento específico de si el ácido nucleico es de origen celular, extracelular, o vírico. Además, el ácido nucleico se puede obtener a partir de una reacción química o enzimática, incluyendo reacciones en las que el ácido nucleico sintético, recombinante o de origen natural se modifica con una enzima, por ejemplo, una metiltransferasa

[0045] En algunas realizaciones, la muestra de ácidos nucleicos es una muestra procesada procedente de una fuente tal como una de las enumeradas anteriormente. Por ejemplo, el ácido nucleico aislado se puede fragmentar por cizalladura, tal como por sonicación o pipeteado a través de una apertura estrecha, o digestión enzimática, tal como con una endonucleasa, que puede ser un endonucleasa de restricción. En algunas realizaciones, la muestra de ácidos nucleicos tiene al menos un extremo escalonado. El ácido nucleico aislado en primer lugar se puede clonar y propagar en una célula hospedadora y/o vector, por ejemplo, como cromosoma artificial de bacteria o levadura, mini-cromosoma, plásmido, cósmido, elemento extracromosómico, o constructo integrado en el cromosoma.

Suministro de una molécula de ácidos nucleicos circular

[0046] En algunas realizaciones, los métodos comprenden el suministro de una molécula de ácidos nucleicos circular que comprende una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que el inserto tiene una secuencia conocida. La molécula de ácidos nucleicos circular puede ser de cadena sencilla o de doble cadena.

[0047] En algunas realizaciones, la molécula de ácidos nucleicos circular se suministra aislándola en forma circular a partir de su fuente, si parte de su secuencia es conocida y de esta manera puede servir como inserto de ácidos nucleicos (por ejemplo, se puede conocer un motivo conservado dentro de la secuencia de un gen contenido en la molécula circular, o se puede saber que la molécula contiene una secuencia basada en su capacidad para hibridarse en condiciones muy rigurosas a otro ácido nucleico de secuencia conocida). En algunas realizaciones, la secuencia del inserto de ácidos nucleico sólo se conoce de forma inexacta, como en el caso en el que el conocimiento de la secuencia procede de las propiedades de hibridación rigurosas. En algunas realizaciones, la secuencia del inserto de ácidos nucleicos se conoce con exactitud, como en el caso en el que la molécula de ácidos nucleicos circular tiene una secuencia estructural conocida o se ha manipulado genéticamente para que contenga una secuencia conocida.

[0048] En algunas realizaciones, la molécula de ácidos nucleicos circular se suministra al llevar a cabo una reacción o reacciones *in vitro* para incorporar la muestra de ácidos nucleicos a una molécula circular junto con un inserto de ácidos nucleicos. La reacción o reacciones *in vitro* en algunos casos pueden comprender la ligación mediante una ligasa y/u otras reacciones de unión de las cadenas de manera que se puede catalizar mediante diversas enzimas, incluyendo recombinasas y topoisomerasas. Se puede usar ADN ligasa o ARN ligasa para unir enzimáticamente los dos extremos de un molde lineal, con o sin molécula adaptadora o enlazadores, para formar un círculo. Por ejemplo, también se pueden usar pares de ARN ligasa de T4 de ADN o ARN de cadena sencilla, como se describe en Tessier y col., Anal Biochem, 158: 171-78 (1986). También se puede usar CIRCLIGASE(TM) (Epicentre, Madison, Wis.) Para catalizar la ligación de un ácido nucleico de cadena sencilla. De manera alternativa, se puede usar una ligasa de doble cadena, tal como ligasa de ADN de *E. coli* o de T4, para llevar a cabo la reacción de circularización.

[0049] En algunas realizaciones, el suministro de la molécula de ácidos nucleicos circular comprende la amplificación de un molde de ácidos nucleicos con cebadores (que pueden ser cebadores aleatorios con solapas 5' de secuencia conocida que pueden servir como inserto de ácidos nucleicos) que comprenden regiones complementarias y la circularización del ácido nucleico amplificado, de manera que puede estar catalizado por una ligasa o una recombinasa; el ácido nucleico amplificado en algunas realizaciones se puede procesar en sus extremos, por ejemplo, mediante restricción o fosforilación, antes de su circularización.

[0050] En algunas realizaciones, la molécula de ácidos nucleicos circular se suministra al realizar una circularización química. Los métodos químicos emplean agentes de acoplamiento conocidos tales como BrCN más imidazol y un metal divalente, N-cianoimidazol con $ZnCl_2$, clorhidrato de 1-(3-dimetilaminopropil)-3-etilcarbodiimida, y otras carbodiimidas y carbonil diimidazoles. Los extremos de un molde lineal también se pueden unir al condensar un 5'-fosfato y un 3'-hidroxilo, o un 5'-hidroxilo y un 3'-fosfato.

[0051] En algunas realizaciones, la molécula de ácidos nucleicos circular es una molécula de par cerrado circular (MPCc). Este tipo de molécula se describe con detalle a continuación.

Suministro de repeticiones directa e inversa de la muestra de ácidos nucleicos; moléculas de par cerrado circular

[0052] En algunas realizaciones, los métodos comprenden el suministro de repeticiones directa e inversa de una muestra de ácidos nucleicos y el cierre conjunto de las cadenas directa e inversa para formar una MPCc. La estructura general de una MPCc se muestra en la Figura 3A. Una MPCc es una molécula de ácidos nucleicos circular de cadena sencilla que comprende repeticiones directa e inversa de una muestra de ácidos nucleicos; las repeticiones están flanqueadas por insertos de ácidos nucleicos, como se muestra en la Figura 3A. Los insertos de ácidos nucleicos pueden ser idénticos o diferentes. En algunas realizaciones, los insertos tienen una longitud de al menos 50 nucleótidos o al menos 100 nucleótidos. En algunas realizaciones, los insertos tienen una longitud que oscila entre 50 o 100 nucleótidos y 10.000 o 50.000 nucleótidos.

[0053] Las cadenas de una muestra de ácidos nucleicos de doble cadena se pueden cerrar juntas para formar una MPCc, por ejemplo, al ligar insertos de ácidos nucleicos que forman horquillas a cada extremo de la molécula. En algunas realizaciones, los insertos de ácidos nucleicos que forman horquillas tienen temperaturas de fusión de al menos 20 ºC, 25 ºC, 30 ºC, 35 ºC, 40 ºC, 45 ºC, 50 ºC, 55 ºC, 60 ºC, 65 ºC, o 70 ºC. La ligación puede ser una ligación por extremos romos o por extremos cohesivos. Las estructuras en horquilla tienen regiones de tronco de bases emparejadas y regiones en bucle desemparejadas. En algunas realizaciones, el ácido nucleico de inserto comprende una región en bucle de un tamaño de al menos 20, 22, 25, 30, o 35 nucleótidos. En algunas realizaciones, esta región en bucle es adecuada para la unión del cebador. En algunas realizaciones, la región en bucle se une a un cebador con una temperatura de fusión de al menos 45 ºC, 50 ºC, 55 ºC, 60 ºC, 65 ºC, o 70 ºC.

[0054] En algunas realizaciones, la muestra de ácidos nucleicos comprende extremos cohesivos diferentes, tales como los que se pueden generar por digestión usando enzimas de restricción con sitios de restricción diferentes, y estos extremos cohesivos diferentes favorecen la ligación de diferentes insertos de ácidos nucleicos. En algunas realizaciones, el ácido nucleico de doble cadena a convertir de esta forma se puede obtener mediante la extensión de un cebador aleatorio que comprende una solapa 5' de secuencia conocida junto con un molde que comprende la secuencia de muestra deseada.

[0055] Las cadenas de un ácido nucleico de doble cadena también se pueden cerrar juntas para formar una MPCc mediante tratamiento con una enzima que convierte los extremos de doble cadena en horquillas, por ejemplo, recombinasas que forman un enlace fosfotirosina con una cadena de una molécula de doble cadena seguido por la formación de una horquilla mediante ataque nucleófilo sobre el enlace fosfotirosina por la otra cadena. Ejemplos de dichas recombinasas son miembros de la familia tal como la λ integrasa y la Flp recombinasa. Véase, por ejemplo, Chen y col., Cell 69, 647-658 (1992); Roth y col., Proc Natl Acad Sci USA 90, 10788-10792 (1993). En algunas realizaciones, la muestra de ácidos nucleicos comprende secuencias de reconocimiento para la enzima que convierte los extremos de doble cadena en horquillas. En algunas realizaciones, las secuencias de reconocimiento para la enzima que convierte los extremos de doble cadena en horquillas están unidas a la muestra de ácidos nucleicos, por ejemplo, mediante ligación.

[0056] En algunas realizaciones, el ácido nucleico de muestra inicialmente se obtiene en forma de cadena sencilla y se convierte a la forma de doble cadena antes de la formación de la MPCc. Esto se puede conseguir, por ejemplo, ligando una horquilla con un extremo escalonado en el extremo 3' del ácido nucleico de muestra, y a continuación la extensión desde el extremo 3' de la horquilla ligada para sintetizar una cadena complementaria. A continuación se puede unir una segunda horquilla a la molécula para formar una MPCc.

Inserto de ácidos nucleicos

[0057] Los métodos descritos en el presente documento comprenden el suministro y/o uso de moléculas de ácidos nucleicos circulares, que incluyen MPCc, que comprenden al menos un inserto de ácidos nucleicos. En algunas realizaciones, el al menos un inserto de ácidos nucleicos tiene una secuencia parcialmente conocida, conocida sin precisión, o completamente conocida, como se ha descrito anteriormente. En algunas realizaciones, la secuencia del al menos un inserto de ácidos nucleicos es completamente conocida. En algunas realizaciones, el al menos un inserto de ácidos nucleicos comprende un sitio de unión adecuado para un oligonucleótido, que incluye un cebador de secuenciación. En algunas realizaciones, el al menos un ácido nucleico de inserto forma una horquilla.

[0058] En algunas realizaciones, el al menos un inserto de ácidos nucleicos tienen una longitud que oscila entre 10 y 300, 15 y 250, 30 y 200, o 30 y 100 restos de nucleótidos. En algunas realizaciones, el al menos un inserto de ácidos nucleicos tiene una temperatura de fusión que oscila entre 45 ºC y 70 ºC o entre 50 ºC y 65 ºC.

[0059] En algunas realizaciones, el al menos un inserto de ácidos nucleicos comprende un promotor, por ejemplo, el promotor de la ARN polimerasa de T7. Véase, por ejemplo, Guo y col., J Biol Chem 280, 14956-14961 (2005). El promotor es reconocido por una ARN polimerasa como sitio para iniciar la síntesis de ARN. También se conocen promotores adicionales en la técnica.

Unidad de inserto-muestra

[0060] Las moléculas de ácidos nucleicos circulares usadas en los métodos descritos en el presente documento comprenden al menos una muestra de ácidos nucleicos y al menos un inserto de ácidos nucleicos agrupados en forma de al menos una unidad de inserto-muestra. Una unidad de inserto-muestra es un segmento de ácidos nucleicos en el que un inserto de ácidos nucleicos está inmediatamente aguas arriba o aguas abajo de una muestra de ácidos nucleicos.

[0061] En algunas realizaciones, la molécula de ácidos nucleicos circular es una MPCc, que comprende dos unidades de inserto-muestra; las muestras de ácidos nucleicos en estas dos unidades están en orientaciones opuestas entre sí, es decir, una es una repetición directa de la muestra de ácidos nucleicos y la otra es una repetición inversa. Cabe señalar que la MPCc se puede considerar que comprende dos unidades de inserto-muestra

en las que los insertos están aguas arriba o aguas abajo de las muestras; es decir, una MPCc conforme a la estructura mostrada en la Figura 3B contiene, en este orden, los elementos 11 (repetición directa), 14 (inserto), 12 (repetición inversa), y 13 (inserto), con 13 que vuelve a conectar con 11 para cerrar el círculo. Independientemente de que las unidades de inserto-muestra se considere que son 11 con 14, y 12 con 13, o 13 con 11, y 14 con 12, la molécula contiene dos unidades de inserto-muestra. En realizaciones en las que la orientación del inserto y/o su posición relativa a la muestra es funcionalmente importante, por ejemplo, el inserto comprende un promotor y un sitio de unión al cebador, puede ser más eficiente agrupar las unidades de inserto-muestra para así agrupar el inserto con la muestra hacia la cual está orientado el sitio de unión al cebador o promotor, es decir, la muestra que se copiaría en primer lugar por una polimerasa que comienza desde el sitio de unión al cebador o promotor.

Obtención de los datos de secuencia

Método de secuenciación

[0062] Los métodos descritos en el presente documento comprenden la obtención de datos de secuencia. En algunas realizaciones, se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra durante la etapa de obtención de los datos de secuencia. En algunas realizaciones, la molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra se puede producir sintetizándola a partir de la molécula de ácidos nucleicos circular proporcionada. En algunas realizaciones, la molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra se pueden producir alterando la molécula de ácidos nucleicos circular proporcionada, por ejemplo, convirtiendo la molécula de ácidos nucleicos circular en una molécula de ácidos nucleicos lineal, que en algunas realizaciones puede ser de cadena sencilla. En algunas realizaciones, en una molécula de ácidos nucleicos, en la etapa de obtención de los datos de secuencia, se forma o se rompe al menos un enlace fosfodiéster, que puede ser la molécula de ácidos nucleicos circular proporcionada o uno de sus productos de síntesis del molde.

[0063] En algunas realizaciones, los datos de secuencia se obtienen usando la secuenciación mediante un método de síntesis. En algunas realizaciones, los datos de secuencia se obtienen usando un método de secuenciación de una sola molécula. En algunas realizaciones, el método de secuenciación de una sola molécula se selecciona entre pirosecuenciación, secuenciación con terminador reversible, secuenciación por ligación, secuenciación con nanoporos, y secuenciación de tercera generación.

[0064] En algunas realizaciones, los datos de secuencia se obtienen usando un método de secuenciación en bruto, por ejemplo, la secuenciación de Sanger o la secuenciación de Maxam-Gilbert.

[0065] Los métodos de secuenciación de una sola molécula se distinguen de los métodos de secuenciación en bruto dependiendo de si se aísla una sola molécula de ácidos nucleicos como parte del proceso de secuenciación. La molécula de ácidos nucleicos puede ser de cadena sencilla o de doble cadena; para este fin, las dos cadenas de ácidos nucleicos hibridadas se consideran una sola molécula. El aislamiento de la molécula sola se puede producir en un micropocillo, mediante el uso de un nanoporo, mediante unión directa o indirecta que se puede resolver ópticamente en un sustrato tal como un portaobjetos de microscopio, o de cualquier otra forma que permita la obtención de los datos de secuencia a partir de la molécula individual. En la unión indirecta, la molécula sola se une al sustrato mediante una estructura de enlace que se une a la molécula sola, por ejemplo, una proteína o un oligonucleótido. Cabe destacar que los métodos en los que se aísla una sola molécula, y a continuación se amplifica, y los datos de secuencia se obtienen directamente a partir del producto(s) de amplificación aún se consideran métodos de molécula sola debido a que se aísla una sola molécula y sirve como fuente definitiva de los datos de secuencia. (En contraste, en los métodos de secuenciación en bruto, se usa una muestra de ácidos nucleicos que contiene varias moléculas y se obtienen datos que contienen señales que proceden de múltiples moléculas). En algunas realizaciones, se realiza la secuenciación de una sola molécula en la que se obtiene una secuencia redundante a partir de la misma molécula. La secuencia redundante se puede obtener por secuenciación de al menos dos repeticiones directas o inversas dentro de una molécula, o por secuenciación del mismo segmento de la molécula más de una vez. La secuencia redundante puede ser completamente redundante o parcialmente redundante con alguna variación, por ejemplo, debido a diferencias introducidas por la alteración de la especificidad de emparejamiento de bases de un cierto tipo de bases, o debido a errores que se pudieran producir durante el proceso de secuenciación. En algunas realizaciones, la alteración de la especificidad de emparejamiento de bases se puede producir antes de la secuenciación. En algunas realizaciones, la misma molécula se secuencia varias veces, opcionalmente con un tratamiento de intervención que altera selectivamente la especificidad de emparejamiento de bases de un cierto tipo de bases que se produce entre las iteraciones de la secuenciación.

[0066] La secuenciación de Sanger, que supone el uso de terminadores de cadena didesoxi marcados, es muy conocida en la técnica; véase, por ejemplo, Sanger y col., Proc Natl Acad Sci USA 74, 5463-5467 (1997). La secuenciación de Maxam-Gilbert, que supone llevar a cabo múltiples reacciones de degradación química parcial sobre fracciones de la muestra de ácidos nucleicos seguido de detección y análisis de los fragmentos para inferir la secuencia, también es muy conocida en la técnica; véase, por ejemplo, Maxam y col., Proc Natl Acad Sci USA 74, 560-564 (1977). Otro método de secuenciación en bruto es la secuenciación por hibridación, en el que se deduce la secuencia de una muestra en base a sus propiedades de hibridación a una pluralidad de secuencias, por ejemplo,

sobre un *microarray* o chip de genes; véase, por ejemplo, Drmanac, y col., Nat Biotechnol 16, 54-58 (1998).

[0067] Los métodos de secuenciación de una sola molécula se describen en general, por ejemplo, en Kato, Int J Clin Exp Med 2, 193-202 (2009) y en las referencias allí citadas.

[0068] La pirosecuenciación, la secuenciación con terminador reversible, y la secuenciación por ligación se consideran métodos de secuenciación de segunda generación. En general, estos métodos usan los productos de amplificación generados a partir de una sola molécula, que están separados espacialmente de los productos de amplificación generados por otras moléculas. La separación espacial se puede poner en práctica usando una emulsión, un pocillo de picolitros, o mediante la unión a un portaobjetos de vidrio. La información de secuencia se obtiene mediante fluorescencia tras la incorporación de un nucleótido; después de obtener los datos, se elimina la fluorescencia del nucleótido recién incorporado y el proceso se repite para el siguiente nucleótido.

[0069] En la pirosecuenciación, el ion pirofosfato liberado mediante la reacción de polimerización se hace reaccionar con adenosina-5'-fosfosulfato mediante la ATP sulfurilasa para producir ATP; a continuación el ATP dirige la conversión de luciferina en oxiluciferina más luz mediante la luciferasa. Puesto que la fluorescencia es transitoria, en este método no es necesaria una etapa aparte para eliminar la fluorescencia. En un momento dado se añade un tipo de desoxirribonucleótido trifosfato (dNTP), y se elucida la información de secuencia según la cual el dNTP genera una señal importante en el sitio de reacción. El instrumento Roche GS FLX disponible en el mercado obtiene la secuencia usando este método. Esta técnica y sus aplicaciones se describen con detalle, por ejemplo, en Ronaghi y col., Anal Biochem 242, 84-89 (1996) y Margulies y col., Nature 437, 376-380 (2005) (corrección de errores en Nature 441, 120 (2006)).

[0070] En la secuenciación con terminador reversible, se incorpora un análogo de nucleótido marcado con colorante fluorescente que es un terminador de cadena reversible debido a la presencia de un grupo bloqueante en una reacción de extensión de una sola base. La identidad de la base se determina según el fluoróforo; en otras palabras, cada base se empareja con un fluoróforo diferente. Después de la obtención de la fluorescencia/datos de secuencia, el fluoróforo y el grupo bloqueante se eliminan químicamente, y se repite el ciclo para obtener la siguiente base de la información de secuencia. El instrumento Illumina GA funciona con este método. Esta técnica y sus aplicaciones se describen con detalle, por ejemplo, en Ruparel y col., Proc Natl Acad Sci USA 102, 5932-5937 (2005), y Harris y col., Science 320, 106-109 (2008).

[0071] En la secuenciación por ligación se usa una enzima ligasa para unir un oligonucleótido de doble cadena parcial con un extremo escalonado al ácido nucleico sometido a secuenciación, que tiene un extremo escalonado; para que se produzca la ligación, los extremos escalonados deben ser complementarios. Las bases en el extremo escalonado del oligonucleótido de doble cadena parcial se pueden identificar según el fluoróforo conjugado al oligonucleótido de doble cadena parcial y/o a un oligonucleótido secundario que se hibrida a otra parte del oligonucleótido de doble cadena parcial. Después de la obtención de los datos de fluorescencia, el complejo ligado se escinde aguas arriba del sitio de ligación, tal como mediante una enzima de restricción de tipo IIs, por ejemplo, Bbvl, que corta en un sitio a una distancia fija desde su sitio de reconocimiento (que estaba incluido en el oligonucleótido de doble cadena parcial). Esta reacción de escisión expone un nuevo extremo escalonado justo aguas arriba del extremo escalonado anterior, y el proceso se repite. Esta técnica y sus aplicaciones se describen con detalle, por ejemplo, en Brenner y col., Nat Biotechnol 18, 630-634 (2000). En algunas realizaciones, la secuenciación por ligación se adapta a los métodos descritos en el presente documento mediante la obtención de un producto de amplificación por círculo rodante de una molécula de ácidos nucleicos circular, y usando el producto de amplificación por círculo rodante como molde para la secuenciación por ligación.

[0072] En la secuenciación por nanoporos, una molécula de ácidos nucleicos de cadena sencilla se ensarta a través de un poro, por ejemplo, usando una fuerza impulsora electroforética, y se deduce la secuencia analizando los datos obtenidos a medida que la molécula de ácidos nucleicos de cadena sencilla pasa a través del poro. Los datos pueden ser datos de corriente iónica, en los que cada base altera la corriente, por ejemplo, bloqueando parcialmente el paso de corriente a través del poro en un grado distinguible y diferente.

[0073] En la secuenciación de tercera generación, se usa un portaobjetos con un revestimiento de aluminio con muchos orificios pequeños (~50 nm) como guía de ondas de modo cero (véase, por ejemplo, Levene y col., Science 299, 682-686 (2003)). La superficie de aluminio se protege frente a la unión de ADN polimerasa mediante la química de polifosfonato, por ejemplo, química de polivinilfosfonato (véase, por ejemplo, Korlach y col., Proc Natl Acad Sci USA 105, 1176-1181 (2008)). Esto da lugar a uniones preferidas de las moléculas de ADN polimerasa a la sílice expuesta en los orificios del revestimiento de aluminio. Esta configuración permite el uso del fenómeno de ondas evanescentes para reducir el ruido de fluorescencia de fondo, permitiendo la utilización de mayores concentraciones de dNTPs marcados por fluorescencia. El fluoróforo se une al fosfato terminal de los dNTPs, de manera que se libera la fluorescencia tras la incorporación del dNTP, pero el fluoróforo no permanece unido al nucleótido recién incorporado, es decir, el complejo está listo inmediatamente para otra ronda de incorporación. Mediante este método se puede detectar la incorporación de dNTPs a complejos cebador-molde individuales presentes en los orificios del revestimiento de aluminio. Véase, por ejemplo, Eid y col., Science 323, 133-138 (2009).

Molde de secuenciación; cantidad de datos de secuenciación obtenidos

[0074] En algunas realizaciones, los datos de secuencia se obtienen directamente a partir de una molécula de ácidos nucleicos circular, es decir, usando la molécula de ácidos nucleicos circular como molde. La molécula de ácidos nucleicos circular usada como molde puede ser una molécula de par cerrado circular. En algunas realizaciones, los datos de secuencia se obtienen a partir de una molécula de ácidos nucleicos producto que fue sintetizada usando una molécula de ácidos nucleicos circular como molde; es decir, un molde a partir del cual se obtienen los datos de secuencia puede ser una molécula de ácidos nucleicos producto sintetizada a partir de un molde de una molécula de ácidos nucleicos circular. En algunas realizaciones, los datos de secuencia se obtienen a partir tanto de un molde de una molécula de ácidos nucleicos circular como de una molécula de ácidos nucleicos producto sintetizada a partir del molde de la molécula de ácidos nucleicos circular.

[0075] En algunas realizaciones, se realiza la amplificación por círculo rodante, que comprende la síntesis de una molécula de ácidos nucleicos producto que comprende al menos dos unidades de inserto-muestra usando la molécula de ácidos nucleicos circular como molde. En algunas realizaciones, la amplificación por círculo rodante comprende la síntesis de una molécula de ácidos nucleicos producto que comprende al menos 3, 4, 5, 10, 15, 20, 25, 50, o 100 unidades de inserto-muestra. El uso de amplificación por círculo rodante para producir una serie de copias de un molde es muy conocido en la técnica; véase, por ejemplo, Blanco y col., J Biol Chem 264, 8935-8940 (1989) y Banér y col., Nucleic Acids Res 26, 5073-5078 (1998). La amplificación por círculo rodante se puede realizar como parte de la secuenciación en la que una molécula de ácidos nucleicos circular es el molde de secuenciación, o para sinterizar una molécula de ácidos nucleicos producto que se ha de usar como molde de secuenciación.

[0076] Independientemente del molde, los datos de secuencia obtenidos de acuerdo con los métodos descritos en el presente documento comprenden al menos dos repeticiones de la secuencia de muestra de ácidos nucleicos; estas al menos dos repeticiones pueden incluir, en algunas realizaciones, al menos una repetición directa de la secuencia de muestra de ácidos nucleicos y al menos una repetición inversa de la secuencia de muestra de ácidos nucleicos. En algunas realizaciones, los datos de secuencia comprenden al menos 3, 4, 5, 10, 15, 20, 25, 50, o 100 repeticiones de la secuencia de muestra de ácidos nucleicos. En algunas realizaciones, los datos de secuencia comprenden al menos 2, 3, 4, 5, 10, 15, 20, 25, 50, o 100 repeticiones directas de la secuencia de muestra de ácidos nucleicos. En algunas realizaciones, los datos de secuencia comprenden al menos 2, 3, 4, 5, 10, 15, 20, 25, 50, o 100 repeticiones inversas de la secuencia de muestra de ácidos nucleicos. En algunas realizaciones, los datos de secuencia comprenden al menos 2, 3, 4, 5, 10, 15, 20, 25, 50, o 100 de cada una de las repeticiones directa e inversa de la secuencia de muestra de ácidos nucleicos.

Cálculo de puntuaciones

[0077] En algunas realizaciones, los métodos comprenden el cálculo de puntuaciones de las secuencias de al menos dos insertos en los datos de secuencia al comparar las secuencias con la secuencia conocida del inserto. En realizaciones en las que la secuencia del inserto sólo se conoce parcialmente o de forma inexacta, la secuencia conocida del inserto de ácidos nucleicos puede comprender posiciones ambiguas o desconocidas, por ejemplo, con el uso de códigos de ambigüedad o una matriz de peso por posición.

[0078] La comparación de las secuencias con la secuencia conocida del inserto incluye la identificación de las secuencias de al menos dos insertos en los datos de secuencia. La identificación de las secuencias se puede realizar en algunas realizaciones mediante inspección visual, es decir, una persona que explore visualmente los datos de secuencia e identifique las secuencias de ácidos nucleicos de inserto contenidas en ella, o mediante un método de alineamiento asistido por ordenador. Véase, por ejemplo, publicación de solicitud de patente internacional WO 2009/017678. En algunas realizaciones, la identificación de las secuencias se puede realizar mediante la exploración de los datos de secuencia usando un algoritmo que reconoce las secuencias, por ejemplo, mediante el cálculo de las puntuaciones de forma iterativa o heurística para múltiples posiciones dentro de los datos de secuencia a fin de identificar los extremos locales que corresponden más estrechamente con la secuencia conocida del inserto de ácidos nucleicos. En algunas realizaciones, la identificación de la secuencia de los al menos dos insertos de ácidos nucleicos se realiza simultáneamente con el cálculo de las puntuaciones, ya que ambos procesos pueden utilizar la misma puntuación.

[0079] En algunas realizaciones, el cálculo de las puntuaciones comprende la realización de un alineamiento usando un algoritmo de alineamiento adecuado, de los cuales se conocen muchos en la técnica y están fácilmente disponibles, por ejemplo, BLAST, MEGABLAST, alineamiento de Smith-Waterman, y alineamiento de Needleman-Wunsch. Véase, por ejemplo, Altschul y col., J Mol Biol 215, 403-410 (1990). Los algoritmos de alineamiento adecuados incluyen tanto algoritmos que permiten huecos como algoritmos que no permiten huecos. De manera alternativa, en algunas realizaciones, el cálculo de puntuaciones comprende el análisis de las secuencias usando un algoritmo tal como la ejecución de una matriz de peso por posición sobre las secuencias y el cálculo de la suma de los elementos de la matriz correspondientes a la secuencia. De esta forma, se puede calcular la puntuación en forma de máximo local encontrado al aplicar la matriz a una lectura de secuencia de manera escalonada.

[0080] En algunas realizaciones, las puntuaciones están correlacionadas positivamente con la cercanía de las al menos dos secuencias de inserto de ácidos nucleicos a la secuencia conocida (por ejemplo, la máxima puntuación posible es el resultado de una coincidencia exacta). Dichas puntuaciones correlacionadas positivamente incluyen, sin limitación, identidad en porcentaje, "bit-scores" y recuento de bases coincidentes.

[0081] En algunas realizaciones, las puntuaciones están correlacionadas negativamente con la cercanía de las al menos dos secuencias de inserto de ácidos nucleicos a la secuencia conocida (por ejemplo, la mínima puntuación posible es el resultado de una coincidencia exacta). Dichas puntuaciones correlacionadas negativamente incluyen, sin limitación, valor e, número de desemparejamientos, número de desemparejamientos y huecos, porcentaje de desemparejamientos, y porcentaje de desemparejamientos/huecos.

[0082] En algunas realizaciones, las puntuaciones se calculan en base a un porcentaje. El intervalo posible de puntuaciones calculado en base al porcentaje no varía en función de la longitud de las secuencias a comparar. Ejemplos de puntuaciones calculadas en base al porcentaje incluyen, sin limitación, identidad en porcentaje y porcentaje de desemparejamientos/huecos.

[0083] En algunas realizaciones, las puntuaciones se calculan en base a un recuento. El intervalo posible de puntuaciones calculadas en base a un recuento varía en función de la longitud de las secuencias a comparar. Ejemplos de puntuaciones calculadas en base a un recuento incluyen, sin limitación, “bit scores”, número de desemparejamientos, número de desemparejamientos y huecos, y recuento de bases coincidentes.

Aceptación o rechazo de repeticiones de la secuencia de la muestra de ácidos nucleicos; grupo de secuencias aceptadas

[0084] En algunas realizaciones, los métodos comprenden aceptar o rechazar repeticiones de la secuencia de la muestra de ácidos nucleicos en los datos de secuencia de acuerdo con las puntuaciones de una o de las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos. Así, en diversas realizaciones, las puntuaciones tanto de los insertos de ácidos nucleicos inmediatamente aguas arriba como los inmediatamente aguas abajo, la puntuación de cualquiera de ellos, o la puntuación de uno o del otro se usan específicamente para decidir si aceptar o rechazar una secuencia de muestra de ácidos nucleicos en los datos de secuencia.

[0085] En realizaciones en las que las puntuaciones están correlacionadas positivamente con la cercanía de las al menos dos secuencias del inserto de ácidos nucleicos a la secuencia conocida, es necesario que las puntuaciones sean mayores, o mayores o iguales, que un valor límite para aceptar una secuencia. La elección de un valor límite apropiado depende de varios factores, incluyendo el tipo de puntuación usada, la tasa de error del método de secuenciación, y consideraciones temporales y de redundancia.

[0086] Aceptar o rechazar repeticiones de la secuencia de la muestra de ácidos nucleicos se puede implementar de diversas formas de manera que para determinar la secuencia de la muestra de ácidos nucleicos se usa al menos una repetición aceptada, y no se usa ninguna repetición rechazada. La aceptación o rechazo de repeticiones se puede realizar, o no, de forma concertada con la recopilación de un grupo de secuencias aceptadas. Por ejemplo, las secuencias de repeticiones aceptadas se pueden copiar en una nueva estructura de datos a medida que son aceptadas, que se convierte en el grupo de secuencias aceptadas. O bien las secuencias de las repeticiones rechazadas se pueden borrar o sobrescribir (por ejemplo, con los caracteres "0" o "X" que representan datos nulos o excluidos) a medida que son rechazadas; en este caso, una vez que las secuencias rechazadas se han borrado o sobrescrito, la estructura de datos originales ha sido modificada para así convertirse en el grupo de secuencias aceptadas. En estos ejemplos, la aceptación o rechazo de repeticiones se considera que se realiza de forma concertada con la recopilación de un grupo de secuencias aceptadas.

[0087] En algunas realizaciones, las repeticiones de las secuencias de la muestra de ácidos nucleicos se puede rechazar por otro motivo, tal como que tenga una longitud que se desvía de la longitud de las otras repeticiones de la secuencia de la muestra de ácidos nucleicos (véase, por ejemplo, Figura 7B). Por ejemplo, una repetición de la secuencia de la muestra de ácidos nucleicos se puede rechazar si se desvía en un grado límite de la longitud media o mediana de las otras secuencias de muestra de ácidos nucleicos, o de una versión preliminar del grupo de secuencias aceptadas que comprende repeticiones de la secuencia de la muestra de ácidos nucleicos aceptada según las puntuaciones de una o de las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos como se ha descrito anteriormente, que puede o puede no tener en consideración la repetición de la secuencia de la muestra de ácidos nucleicos para su posible rechazo eliminado temporalmente del cálculo de la longitud mediana o media. El grado límite se puede expresar en términos de longitud absoluta, por ejemplo, 1, 2, 5, 10, 20, o 50 nucleótidos; longitud relativa, por ejemplo, 1 %, 2 %, 5 %, 10 %, 20 %, o 50 %; o en términos de medida estadística, tal como desviación estándar, por ejemplo, desviaciones estándar de 0,5, 1, 1,5, 2, 2,5, 3, 3,5, 4, o 5.

[0088] De manera alternativa, las secuencias se pueden marcar como aceptadas o rechazadas, y después de que se haya completado el proceso de marcado, las secuencias aceptadas se pueden copiar a una nueva estructura

de datos, o las secuencias rechazadas se pueden borrar o sobrescribir, para generar un grupo de secuencias aceptadas de una forma no concertada.

[0089] El grupo de secuencias aceptadas se puede seleccionar entre formas que incluyen una única cadena de datos, que comprende la al menos una repetición aceptada de la secuencia de la muestra de ácidos nucleicos y cualquier repetición aceptada adicional de forma concatenada, y una variable multi-elementos, en la que cada elemento representa una repetición aceptada de la secuencia de la muestra de ácidos nucleicos o una fracción de la misma. En algunas realizaciones, la variable multi-elementos se selecciona de una lista, conjunto, *hash*, y matriz. Es adecuada para su uso cualquier forma de estructura de datos que permita el almacenamiento de la al menos una repetición aceptada de la secuencia de la muestra de ácidos nucleicos y la posterior determinación de la secuencia de la muestra de ácidos nucleicos.

[0090] En realizaciones en las que la forma del grupo de secuencias aceptadas difiere de la forma de los datos de secuencia en bruto (por ejemplo, los datos de secuencia en bruto están en forma de cadena y el grupo de secuencias aceptadas está en forma de estructura de datos multi-elementos tal como un conjunto), los datos de secuencia en bruto se pueden analizar en elementos que contienen repeticiones, unidades de inserto-muestra, o repeticiones de muestra flanqueadas por los insertos inmediatamente aguas arriba y aguas abajo en un punto del método después de que se hayan obtenido los datos de secuencia en bruto y antes de que se haya generado el grupo final de secuencias aceptadas. La etapa de análisis se puede producir antes o después de la etapa de puntuación descrita anteriormente.

Determinación de la secuencia de la muestra de ácidos nucleicos; secuencias consenso; niveles de confianza

[0091] En algunas realizaciones, los métodos comprenden la determinación de la secuencia de la muestra de ácidos nucleicos.

[0092] La forma de determinar la secuencia de la muestra de ácidos nucleicos se puede seleccionar de forma condicionada en base al número de repeticiones de la muestra de ácidos nucleicos en el grupo de secuencias aceptadas. Por ejemplo, cuando el grupo de secuencias aceptadas contiene únicamente una repetición aceptada, se puede determinar que la secuencia de la muestra de ácidos nucleicos es la secuencia de la repetición aceptada. Cuando el grupo de secuencias aceptadas contiene únicamente dos, o al menos tres, repeticiones aceptadas, se puede determinar que la secuencia de la muestra de ácidos nucleicos es la secuencia consenso (véase a continuación) de las repeticiones aceptadas. Cuando el grupo de secuencias aceptadas contiene al menos tres repeticiones aceptadas hay disponibles más opciones en cuanto a cómo se determina la secuencia consenso.

Secuencia consenso

[0093] La secuencia consenso se determina a partir de un alineamiento (realizado como se ha descrito anteriormente, en sección "Cálculo de puntuaciones") de las repeticiones aceptadas; en las posiciones del alineamiento en las que las repeticiones aceptadas contienen la misma base, la secuencia consenso contiene esa base. En algunas realizaciones, en las posiciones del alineamiento en las que las repeticiones aceptadas no contienen la misma base, la secuencia consenso contiene el código de ambigüedad adecuado (por ejemplo, R cuando las repeticiones aceptadas contienen A y G en una posición). En algunas realizaciones, en las posiciones del alineamiento en las que las repeticiones aceptadas no contienen la misma base, la secuencia consenso contiene una N u otro símbolo indicativo de una base desconocida. En algunas realizaciones, en las posiciones del alineamiento en las que las repeticiones aceptadas no contienen la misma base, la secuencia consenso contiene la base de la repetición aceptada que proporcionó una señal más fuerte o más robusta durante la obtención de la secuencia (por ejemplo, si los datos en bruto estaban en forma de fluorescencia, en la secuencia consenso se pone la base determinada en base a la emisión de fluorescencia más brillante (en algunas realizaciones, después de una normalización y/o estandarización adecuada).

[0094] Cuando se determina una secuencia consenso a partir del grupo de secuencias aceptadas que contienen al menos tres repeticiones aceptadas, la base en cada posición de la secuencia consenso se puede determinar en algunas realizaciones por voto mayoritario; es decir, la base presente en una posición en más de la mitad de las repeticiones aceptadas se pone en esa posición en la secuencia consenso. Cuando las repeticiones aceptadas presentan una discrepancia en una posición tal que no hay voto mayoritario en esa posición, la base en esa posición en la secuencia consenso se determina mediante otro método, por ejemplo, se puede usar mayoría relativa (es decir, la base presente más frecuentemente en una posición de las repeticiones aceptadas se pone en esa posición en la secuencia consenso), o se puede usar uno de los procedimientos descritos en el párrafo anterior.

[0095] En algunas realizaciones, cuando se determina una secuencia consenso a partir de un grupo de secuencias aceptadas que contienen al menos tres repeticiones aceptadas, la base en cada posición de la secuencia consenso en algunas realizaciones se puede determinar según la frecuencia de cada base en esa posición en las repeticiones aceptadas. Así, la secuencia consenso puede ser una representación estadística de la probabilidad de que cada base esté presente en cada posición en la muestra de ácidos nucleicos. Dicha

representación puede adoptar la forma de matriz de peso por posición. En algunas realizaciones, los elementos de la matriz de peso por posición son las frecuencias con las que se observa cada base en cada posición en el alineamiento de las repeticiones aceptadas.

[0096] En algunas realizaciones, los elementos de la matriz de peso por posición se calculan a partir de las frecuencias con las que se observa cada base en cada posición en el alineamiento de las repeticiones aceptadas; para este cálculo también se pueden usar otros factores, por ejemplo, cuando se obtienen algunas secuencias de repeticiones aceptadas con señales más fuertes o más robustas que otras repeticiones durante la obtención de la secuencia, se puede proporcionar más peso a las secuencias de repeticiones aceptadas, y/o se puede proporcionar menos peso a las otras repeticiones. El grado en el que se modifican los pesos se puede determinar cuantitativamente, en base, por ejemplo, a la intensidad de la señal, o puede ser una modificación fija; por ejemplo, el peso de bases obtenidas con una señal relativamente fuerte se puede incrementar en un valor tal como el 50 % o el 100 %, y/o el peso de bases con una señal relativamente débil se puede reducir en un valor tal como el 33 % o el 50 %.

[0097] En algunas realizaciones, los elementos de la matriz de peso por posición son valores que se han obtenido a partir de frecuencias transformadas de cada base en cada posición (opcionalmente ponderadas como se ha descrito anteriormente). Las frecuencias se pueden transformar, por ejemplo, de forma logarítmica o exponencial; en algunas realizaciones, la transformación tiene el efecto de infra-ponderar bases observadas pocas veces en una posición y/o sobre-ponderar bases observadas habitualmente en una posición. Por ejemplo, si T está presente en una posición en un alineamiento de N secuencias de repeticiones aceptadas M veces, en la que $N > 2$ y $M < N/2$, y C está presente cada dos veces (es decir, N - M veces), en algunas realizaciones la transformación de estas frecuencias produciría que el peso de T en la matriz de peso por posición fuese inferior a M/N (o su porcentaje correspondiente) y/o el peso de C fuese superior a (N-M)/N (o su porcentaje correspondiente). En algunas realizaciones, la transformación se selecciona para así sobre ponderar únicamente la base (o bases, en el caso de igualdad en la frecuencia) observada más habitualmente.

Niveles de confianza

[0098] En algunas realizaciones, se determina el nivel de confianza para al menos una posición en la secuencia de la muestra de ácidos nucleicos. El nivel de confianza se puede expresar de diferentes formas, por ejemplo, como valor global de precisión en la determinación de la base, expresado como porcentaje o como puntuación Phred, o como tasa de error. En algunas realizaciones, el nivel de confianza se determina a partir de la frecuencia de la base o bases más habituales en una posición, o la frecuencia combinada de las bases que no son las más habituales. En algunas realizaciones, estas frecuencias se transforman, se sobre-ponderan, y/o se infra-ponderan como se ha descrito anteriormente.

Determinación del nivel de confianza de la secuencia en su conjunto; determinación de la secuencia de la muestra de ácidos nucleicos y niveles de confianza en tiempo real y/o hasta un nivel de confianza deseado

[0099] En algunas realizaciones, los métodos comprenden la determinación del nivel de confianza de la secuencia en su conjunto. El nivel de confianza de la secuencia en su conjunto se puede expresar de diferentes formas, por ejemplo, como valor global de precisión en la determinación de la base expresado como porcentaje o como puntuación Phred; como tasa de error; o como número esperado de errores en la secuencia.

[0100] Para calcular el nivel de confianza de la secuencia en su conjunto se pueden usar los niveles de confianza de las posiciones individuales, como se ha descrito en la sección anterior. Por ejemplo, se puede determinar un nivel de confianza global como media aritmética, media geométrica, mediana, o nivel de confianza modal de la población estadística de niveles de confianza en cada posición de la secuencia de la muestra de ácidos nucleicos. En algunas realizaciones, la población estadística de niveles de confianza en cada posición de la secuencia de la muestra de ácidos nucleicos se procesa antes del cálculo del nivel de confianza de la secuencia en su conjunto, por ejemplo, para rechazar valores atípicos.

[0101] En algunas realizaciones, los métodos comprenden la determinación de la secuencia de la muestra de ácidos nucleicos y los niveles de confianza en tiempo real. En estas realizaciones, los datos obtenidos en la etapa de secuenciación se procesan para determinar la secuencia y los niveles de confianza simultáneamente con la obtención de los datos de secuencia adicionales, por ejemplo, a partir de repeticiones adicionales de un producto de amplificación por círculo rodante. A medida que se obtienen los datos de secuencia adicionales, se actualizan tanto la secuencia determinada como los niveles de confianza. En algunas realizaciones, se prosigue con el proceso en tiempo real hasta que se alcanza un nivel de confianza preseleccionado. El nivel de confianza preseleccionado puede ser, por ejemplo, de una precisión de determinación de bases del 90 %, 95 %, 99 %, 99,5 %, 99,9 %, 99,95 %, o 99,99 %. El nivel de confianza preseleccionado puede ser para la secuencia en su conjunto o para una fracción de las posiciones en la secuencia, y se puede seleccionar a partir de valores tales como, por ejemplo, 50 %, 67 %, 75 %, 80 %, 85 %, 90 %, 95 %, 98 %, 99 %, 99,5 % y 99,9 %.

Muestras múltiples; ensamblaje de un cóntigo

[0102] En algunas realizaciones, el método comprende la repetición de las etapas del método usando al menos otra muestra de la misma fuente, especie, o cepa que la muestra de ácidos nucleicos que tiene la secuencia, que se solapa parcialmente con la secuencia de la muestra de ácidos nucleicos, determinando así al menos otra secuencia, y ensamblando la al menos otra secuencia con la secuencia de la muestra original para formar un cóntigo. En algunas realizaciones, el método comprende la repetición de las etapas del método con muchas muestras, para así generar cóntigos de tamaños superiores a 0,5, 1, 2, 5, 10, o 100 kb, o 1, 2, 5, 10, 100, o 1000 Mb. En algunas realizaciones, el cóntigo representa la secuencia completa, o la secuencia completa excepto las regiones heterocromáticas o refractarias, de una molécula de ácidos nucleicos, que puede ser, por ejemplo sin limitación, un cromosoma, un minicromosoma, un cromosoma artificial, un genoma vírico, o un elemento extracromosómico. El ensamblaje del cóntigo se puede llevar a cabo usando métodos conocidos en la técnica.

Bases modificadas

[0103] En algunas realizaciones, la muestra de ácidos nucleicos comprende al menos una base modificada, por ejemplo, 5-metilcitosina, 5-bromouracilo, uracilo, 5,6-dihidrouracilo, ribotimina, 7-metilguanina, hipoxantina, o xantina. El uracilo se puede considerar una base modificada en una cadena de ADN, y la ribotimina se puede considerar una base modificada en una cadena de ARN. En algunas realizaciones, al menos una base modificada en la muestra de ácidos nucleicos de doble cadena se empareja con una especificidad de emparejamiento de bases diferente de su base homóloga preferida. Esto se puede producir, por ejemplo, cuando una base en una molécula de doble cadena ha experimentado una reacción (por ejemplo, debido a oxidación esporádica, o exposición a un agente mutágeno tal como radiación o un mutágeno químico) que la transforma de una de las bases convencionales en una base modificada que no tiene las mismas bases homólogas preferidas.

[0104] Las bases homólogas preferidas se basan en las reglas de emparejamiento de bases de Watson y Crick. Por ejemplo, la base homóloga preferida de la adenina es la timina (o uracilo), y viceversa; la base homóloga preferida de la citosina es la guanina, y viceversa. Las bases homólogas preferidas de bases modificadas en general son conocidas por los expertos en la materia o se pueden predecir en base a la presencia de donadores y aceptores de enlaces de hidrógeno en posiciones análogas a las de las bases convencionales. Por ejemplo, la hipoxantina tiene un aceptor de enlaces de hidrógeno (un oxígeno con un doble enlace) en posición 6 del anillo purina, como guanina, y por tanto su base homóloga preferida es la citosina, que tiene un aceptor de enlaces hidrógeno (un grupo amina) en posición 6 del anillo pirimidina. Cabe destacar que la hipoxantina se puede formar por desaminación de la adenina. Puesto que la adenina normalmente se emparejaría con la timina en el ADN, esta reacción de desaminación puede dar lugar a un par hipoxantina-timina, en el que la base modificada hipoxantina no se empareja con su base homóloga preferida. La citosina también se puede desaminar para formar uracilo. En el contexto del ADN, el uracilo se puede considerar una base modificada, y si se empareja a la guanina (como puede ser el caso de la desaminación de la citosina en ADN normal de doble cadena), entonces esta también es una situación en la que la base modificada uracilo no se empareja con su base homóloga preferida.

Detección de bases modificadas; alteración de la especificidad de emparejamiento de bases de un tipo específico de bases

[0105] En algunas realizaciones, los métodos comprenden la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases. La alteración de la especificidad de emparejamiento de bases de un tipo específico de bases puede comprender la alteración específica de la especificidad de emparejamiento de bases de una versión sin modificar de una base, por ejemplo, citosina. En este caso, la especificidad de emparejamiento de bases de al menos una forma modificada de la base, por ejemplo, 5-metilcitosina, no se ve alterada.

[0106] De manera alternativa, la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases puede comprender la alteración específica de la especificidad de emparejamiento de bases de una versión modificada de una base (por ejemplo, 5-metilcitosina) pero no de la versión sin modificar de la base (citosina).

[0107] En algunas realizaciones, la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases comprende la transición fotoquímica, que convierte la 5-metilcitosina (pero no la citosina sin modificar) en timina. Véase, por ejemplo, Matsumura y col., Nucleic Acids Symp Ser No. 51, 233-234 (2007). Esta reacción altera la especificidad de emparejamiento de bases de las bases que experimentan transición fotoquímica de guanina a adenina (la guanina se empareja con la 5-metilcitosina mientras que la adenina se empareja con la timina).

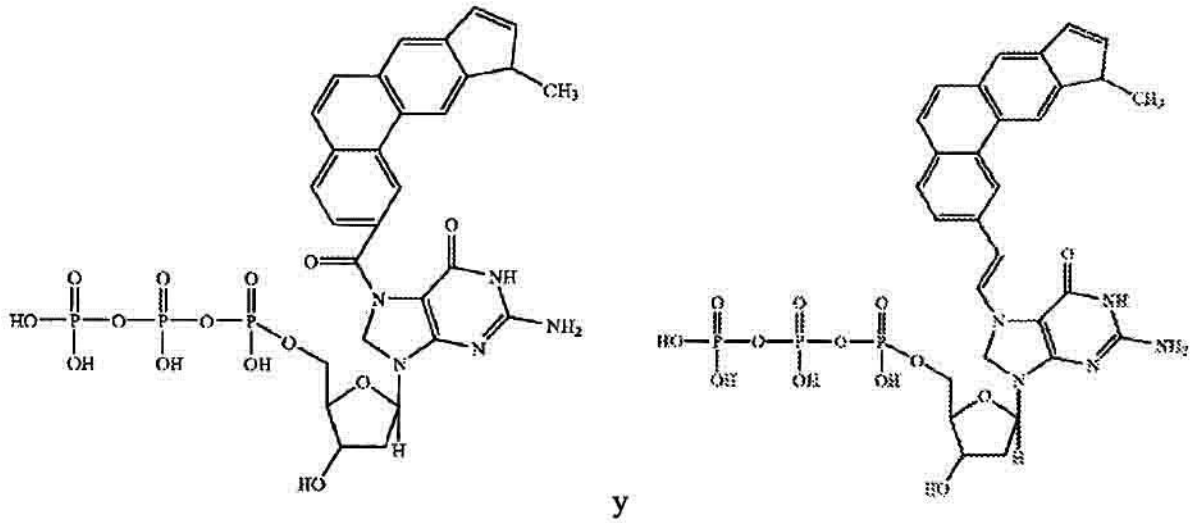
[0108] En otras realizaciones, la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases comprende la conversión con bisulfito, que convierte la citosina (pero no la 5-metilcitosina) en uracilo. Véase, por ejemplo, Laird y col., Proc Natl Acad Sci USA 101, 204-209 (2004), y Zilberman y col., Development 134, 3959-3965 (2007). Esta reacción altera la especificidad de emparejamiento de bases de bases que experimentan conversión con bisulfito de guanina a adenina (la guanina se empareja con la citosina mientras que la adenina se empareja con el uracilo).

[0109] En otras realizaciones adicionales, las bases modificadas se pueden detectar sin la etapa de alteración, tal como en los casos en los que la base modificada tiene una especificidad de emparejamiento de bases alterada con respecto a la versión sin modificar de la base. Ejemplos de dichas bases pueden incluir 5-bromouracilo, uracilo, 5,6-dihidrouracilo, ribotimina, 7-metilguanina, hipoxantina, y xantina. Véase, por ejemplo, Brown, Genomes, 2nd Ed., John Wiley & Sons, Inc., Nueva York, NY, 2002, capítulo 14, "Mutation, Repair, and Recombination", que describe la propensión del 5-bromouracilo a experimentar tautomerización cetoenólica que da lugar a un mayor emparejamiento a la guanina con respecto a la adenina, y la formación de hipoxantina (que se empareja preferentemente a la citosina frente a la timina) por desaminación de la adenina.

Análogos de nucleótidos que discriminan entre una base y su forma modificada

[0110] En algunas realizaciones, los datos de secuencia se obtienen usando al menos un análogo de nucleótido que discrimina entre una base y su forma modificada (un "análogo discriminador"; se empareja preferentemente con una, pero no con la otra, de la base y su forma modificada). El análogo de nucleótidos se puede usar y detectar como si fuera una quinta base, además de las cuatro bases convencionales, por ejemplo, mediante el uso de marcadores diferenciales en la secuenciación con terminador reversible o secuenciación por ligación, o cuando se incorpora en la pirosecuenciación, en la que los nucleótidos se pueden añadir de uno en uno y a continuación se pueden eliminar lavando. En algunas realizaciones, el análogo discriminador se añade antes que su nucleótido natural correspondiente (por ejemplo, en la pirosecuenciación) o se proporciona en una concentración que oscila entre 10 y 100 veces superior a la concentración de su nucleótido natural cognado (por ejemplo, en la secuenciación con terminador reversible). Por ejemplo, el análogo discriminador puede ser un análogo de desoxiguanosina trifosfato que discrimina entre la citosina y la 5-metilcitosina (por ejemplo, se emparejará con la citosina pero no con la 5-metilcitosina); el análogo se puede proporcionar a una concentración que oscila entre 10 y 100 veces superior a la concentración de desoxiguanosina trifosfato. De esta forma, el análogo en general se deberá incorporar frente a la versión de la base con la que preferentemente se empareja, pero la base natural en general se debe incorporar frente a la versión de la base con la que el análogo preferentemente no se empareja.

[0111] Ejemplos de análogos discriminadores se pueden encontrar en la patente de Estados Unidos 7.399.614, e incluyen, por ejemplo, las siguientes moléculas, que discriminan entre la citosina sin modificar y la 5-metilcitosina, y que preferentemente se emparejan con la primera.



y

Estas moléculas se denominan Análogo discriminador 1 y Análogo discriminador 2, respectivamente.

Determinación de las posiciones de bases modificadas en la muestra de ácidos nucleicos

[0112] En algunas realizaciones, los métodos comprenden la determinación de las posiciones de bases modificadas en la muestra de ácidos nucleicos. Estas realizaciones comprenden (i) el suministro de la muestra de ácidos nucleicos en forma de doble cadena; (ii) la conversión de la muestra de ácidos nucleicos en una molécula de par cerrado circular, en la que la molécula de par cerrado circular comprende repeticiones directa e inversa de la secuencia de la muestra de ácidos nucleicos y dos insertos de ácidos nucleicos que tienen secuencias conocidas, que pueden ser idénticas o diferentes; (iii) opcionalmente la alteración de la especificidad de emparejamiento de bases de un tipo específico de bases en la molécula de par cerrado circular; (iv) a continuación, la obtención de los datos de secuencia modelada por las repeticiones directa e inversa de la molécula de par cerrado circular o por una de sus secuencias complementarias; y (v) la determinación de las posiciones de las bases modificadas en una muestra de ácidos nucleicos usando los datos de secuencia de al menos las repeticiones directa e inversa o sus

copias. Cabe señalar que la secuencia modelada por una repetición directa tendrá el mismo sentido que la repetición inversa (y viceversa), pero puede o puede no ser completamente idéntica a la repetición inversa; pueden aparecer diferencias en la repetición directa que contiene bases que se pueden emparejar a una base distinta a la base correspondiente en la repetición inversa. Un ejemplo de dicha situación es si la repetición directa en una MPCc contiene 5-bromouracilo que se ha emparejado a adenina en la cadena inversa pero sirve de molde para la adición de una guanina en una reacción de secuenciación por síntesis.

[0113] Se obtienen datos de secuencia que comprenden al menos dos repeticiones: al menos una de una repetición de la muestra (por ejemplo, la repetición marcada como 17 en la Figura 5A) y una repetición del complemento recién sintetizado de la cadena directa (por ejemplo, la repetición marcada como 21 en la Figura 6A); y al menos una de una repetición del complemento recién sintetizado de la cadena inversa (por ejemplo, la repetición marcada como 19 en la Figura 6A) y una repetición de la cadena inversa (por ejemplo, la repetición marcada como 16 en la Figura 6A). Estas repeticiones se alinean. El alineamiento se puede realizar usando cualquier algoritmo adecuado, como se ha descrito anteriormente. Una posición en la que existe una discrepancia entre las repeticiones (por ejemplo, la posición marcada como 41 en la Figura 6B) significa que una base en la muestra de ácidos nucleicos en esa posición ha sufrido una alteración en su especificidad de emparejamiento de bases. Dependiendo del tipo de modificación, de la base modificada, y/o del análogo discriminador usado en el proceso o presente en la muestra, se pueden determinar las bases presentes originalmente en la posición correspondiente de la muestra de ácidos nucleicos.

[0114] Por ejemplo, cuando la molécula de par cerrado circular se ha alterado mediante la conversión de ^{m}C en T (véase Figura 5A), la discrepancia indica que había presente una ^{m}C en la muestra de ácidos nucleicos en la posición en la que hay una T o complementaria a uno A en una lectura, y es una C o complementaria a uno G en otra lectura; lo lógico es que en una posición en la que las secuencias discrepan, la base que es el producto de la reacción de conversión, T, haya sustituido al sustrato de la reacción de conversión, ^{m}C, que estaba presente en la muestra de ácidos nucleicos.

[0115] En otro ejemplo, en la que en la molécula de par cerrado circular ya se ha alterado por conversión de C a U, la discrepancia indica que había presente una C en la muestra de ácidos nucleicos en la posición ocupada por un U o T, o es complementario a una A en una lectura, y es una C o complementario a una G en otra lectura; lo lógico es que en una posición en la que las secuencias discrepan, la base que es el producto de la reacción de conversión, U (que el sistema de secuenciación puede leer como T), haya sustituido al sustrato de la reacción de conversión, C, que estaba presente en la muestra de ácidos nucleicos. Puesto que los restos ^{m}C no se modificarían por la conversión de C en U, las posiciones en las que las lecturas concuerdan al mostrar C en una posición y/o G como su complemento indican que ^{m}C estaba presente en esta posición en la muestra original.

[0116] En realizaciones en las que se usa un análogo discriminador como se ha descrito anteriormente, la presencia de la base a la que se une preferentemente se puede inferir en la secuencia original en la posición de la secuencia original correspondiente a la posición en la que aparece el análogo discriminador.

Sistema/soporte legible por ordenador

[0117] En algunas realizaciones, la divulgación se refiere a un sistema que comprende un aparato de secuenciación unido de manera operable a un aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario. El elemento de interfaz de usuario se puede seleccionar entre una pantalla, un teclado, y un ratón. En algunas realizaciones, el sistema comprende al menos un circuito integrado y/o al menos un semiconductor.

[0118] En algunas realizaciones, el aparato de secuenciación se selecciona entre aparatos de secuenciación configurados para llevar a cabo al menos uno de los métodos de secuenciación descritos anteriormente.

[0119] En algunas realizaciones, la pantalla puede ser una pantalla táctil, que sirve como único elemento de interfaz de usuario. El almacenamiento está codificado mediante programación que comprende un sistema operativo, un *software* de interfaz de usuario, e instrucciones que, cuando las ejecuta el procesador sobre el sistema que comprende un aparato de secuenciación unido de manera operable a aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario, opcionalmente mediante la introducción por parte del usuario, realiza un método de la invención como se ha descrito anteriormente. En algunas realizaciones, el almacenamiento además comprende datos de secuencia, que pueden estar en cualquiera de las formas descritas anteriormente, por ejemplo, datos de secuencia en bruto, un grupo de secuencias aceptadas, una secuencia consenso, etc.

[0120] En algunas realizaciones, el almacenamiento y todo su contenido se encuentran dentro de un único ordenador. En otras realizaciones, el almacenamiento está dividido entre al menos dos ordenadores, por ejemplo, ordenadores conectados mediante una conexión de red. En algunas realizaciones, la interfaz de usuario es parte de un ordenador que está conectado con al menos otro ordenador que comprende al menos un componente del sistema, por ejemplo, el *software* de procesamiento.

[0121] En algunas realizaciones, el resultado del sistema o el método ejecutado por el procesador da lugar a un indicio de que existe una base modificada en al menos una posición en una muestra de ácidos nucleicos. El indicio puede aparecer en diferentes formas, por ejemplo, una lista de las posiciones modificadas en la secuencia, una representación de texto o gráfica de la secuencia en la que están resaltadas o marcadas las posiciones modificadas, tal como con un asterisco o un carácter similar o con formato en negrita, cursiva o subrayado, texto en color, o una representación de la estructura química del ácido nucleico que incluye la estructura de la base modificada.

Ejemplos

[0122] Los siguientes ejemplos específicos se deben interpretar como meramente ilustrativos, y no limitantes en modo alguno del resto de la divulgación.

Ejemplo 1: Amplificación por círculo rodante de una molécula de par cerrado circular sintética

[0123] Se proporcionaron cuatro oligodesoxirribonucleótidos como se muestra en la Tabla 1.

Tabla 1. Secuencias de oligonucleótidos

Nombre	Secuencia	SEQ ID NO
MPCc-1	CGACTTATGCATTTGGTATCTGCGCTCTGC ATATTTAAATGGAAGGAGATAGTTAAGGATA AGGGCAGAGCGCAGATAC	1
MPCc-2	CAAATGCATAAGTCGTGTCTTACCGGGTTGA TAGCGGCCGCTCGGAGAAAAGAAGTTGGAT GATGCAACCCGGTAAGACA	2
pS-T1	CCTTATCCTTAACTATCTCCTT	3
pS-T2	TAGCGGCCGCTCGGAGAAAAG	4

[0124] MPCc-1 y MPCc-2 se fosforilaron por separado en reacciones de 50 µl en las que 30 µl de oligodesoxirribonucleótido 10 µM (concentración final) se trataron con 10 µl de 10 U/µl de polinucleótido quinasa de T4 (New England Biolabs ("NEB") Cat. No. M0201S), en presencia de 5 µm de tampón de ligasa de T4 10X (NEB; el tampón madre 10X contiene ATP 10 mM). Se añadieron 14 µl de ddH_2O para dar un volumen final de 50 µl (véase Tabla 2). Las reacciones se incubaron a 37 ºC durante 30 min, seguido de inactivación enzimática a 65 ºC durante 20 min.

Tabla 2. Condiciones de la reacción de fosforilación (volúmenes en µl)

Reactivo	5'P-MPCc-1	5'P-MPCc-2
10 µM MPCc-1	30	0
10 µM MPCc-2	0	30
10 u/µl PNK de T4	1	1
Tampón de ligasa de T4 10X	5	5
ddH_2O	14	14
Volumen total	50	50

[0125] La concentración de MPCc-1 y MPCc-2 fosforiladas (5'P-MPCc-1 y 5'P-MPCc-2, respectivamente) procedente de las reacciones anteriores se ajustó a 6 µM.

[0126] MPCc-1 y MPCc-2 fosforiladas y purificadas a continuación se sometieron a desnaturalización a 95 ºC durante 5 minutos, se pusieron en hielo y se mezclaron con tampón, ddH_2O, y ligasa de T4 (NEB, Cat. No. M0202S) para producir moléculas de par cerrado circulares, como se muestra en la Tabla 3. La ligación se produjo a 25 ºC, y

se extrajeron alícuotas de 18 µl a los 10, 30 y 60 minutos. En paralelo se corrió un control negativo sin ligasa (columna L0 en la Tabla 3).

Tabla 3. Condiciones de la reacción de ligación

Reactivo	L0	L3
6 µM 5'P-MPCc-1	9	9
6 µM 5'P-MPCc-2	9	9
400 u/µl de ligasa de T4	0	3
Tampón 10X	6	6
ddH_2O	36	33
Volumen total	60	60

[0127] Los productos de ligación se combinaron con los cebadores pS-T1 y/o pS-T2, dNTPs, ADN polimerasa RepliPHI™ Phi29 (Epicentre, Cat. No. PP031010), y un tampón de reacción apropiado 10 X para la ADN polimerasa RepliPHI Phi29 como se muestra en la Tabla 4.

Tabla 4. Amplificación por círculo rodante de moléculas de par cerrado circulares

	Controles		2 cebadores		1 cebador	
Reactivo	C1	C2	L0	L3	L0	L3
10 mM dNTP	5	5	5	5	5	5
10 µM cebador pS-T1	0	0	6	6	0	0
10 µM cebador pS-T2	0	0	6	6	6	6
1 X L0_10, 30, 60 min	1	0	1	0	1	0
1 X L3_10, 30, 60 min	0	1	0	1	0	1
1000 u/µl de polimerasa phi29	1	1	1	1	1	1
Tampón 10X	5	5	5	5	5	5
ddH_2O	38	38	26	26	32	32
Volumen total	50	50	50	50	50	50

[0128] Las reacciones se prepararon sin polimerasa de Phi29, se desnaturalizaron a 95 ºC durante 5 min, y se pusieron en hielo durante 5 min. Se añadió polimerasa de Phi29 seguido de incubación a 30 ºC durante 18 horas.

[0129] 5 µl de muestras de los productos de reacción se mezclaron con 1 µl de colorante de carga 6X (0,03 % de azul de bromofenol, 0,03 % de xileno cianol FF, 60 % de glicerol, Tris-EDTA 100 mM (pH 7,6)), se calentó a 95 ºC durante 10 min, y a continuación se pusieron inmediatamente en hielo. Un segundo grupo de muestras de productos de reacción se trató de forma idéntica excepto porque también se añadió SDS al 1 %.

[0130] Las muestras se cargaron en un gel de agarosa al 0,7 % en tampón TAE 1X y electroforesis a 135 V durante 28 min. El ADN se visualizó mediante tinción de gel prefabricado GelRed™ (Biotium, Cat. No.: 41003 GelRed™ Nucleic Acid Gel Stain, diluido 10.000 veces en agua). El gel se muestra en la Fig. 9. En las muestras de las reacciones que usan los productos de reacción de ligación L3 y ambos cebadores pS-T1 y pS-T2 se observaron los productos de amplificación por círculo rodante con pesos moleculares aparentes superiores a 10 kb, pero no en las muestras que usan los controles L0 o en las muestras que carecían de cebador. Las muestras que usan los productos de la reacción de ligación L3 y ambos cebadores pS-T1 y pS-T2 que se trataron con SDS mostraban una mayor retención del producto en los pocillos, consistente con la desnaturalización de la estructura secundaria en los productos de RCA.

Ejemplo 2. Simulación de la detección de la metilación mediante la conversión de C a U por tratamiento con bisulfito con una molécula de par cerrado lineal

[0131] Se simuló la determinación de la secuencia y de las posiciones de 5-metilcitosina de un fragmento hipotético de ADN dúplex usando la conversión de restos de C a restos de U por tratamiento con bisulfito como sigue. El esquema general de este ejemplo se ilustra en Figura 12. La secuencia del ADN se muestra a continuación.

Muestra de ADN (C metilada marcada como ^{m}C)

```
5'-AGATGTGGA^mCGGGGTGGG^mCGGAGGTGGGTTGGGGC-3' (SEQ ID NO: 5)
   |||||||||/ ||||||||/ ||||||||||||||||
3'-TCTACACCTG^mCCCCACCCG^mCCTCCACCCAACCCCG-5' (SEQ ID NO: 6)
```

[0132] Las dos cadenas están conectadas mediante ligación a una secuencia enlazadora (representada como "nnnn") para dar el siguiente producto. La secuencia enlazadora es adecuada para su uso como cebador de secuenciación.

```
3'-TCTACACCTG^mCCCCACCCG^mCCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGG^mCGG
GTGGGG^mCAGGTGTAGA-5' (SEQ ID NO: 7)
```

[0133] Además, hay unida una solapa lineal de secuencia conocida (no mostrada) a cada extremo de la molécula de la SEQ ID NO: 7. La solapa en el extremo 3' es adecuada para la unión del cebador para su secuenciación o replicación. El complemento de la solapa en el extremo 5' es adecuado para la unión del cebador para su secuenciación o replicación.

[0134] El producto se trata con bisulfito de sodio, que da lugar a la conversión de los restos de citosina (pero no de la 5-metilcitosina) en uracilo, dando el siguiente producto. Los restos de uracilo recién formados están en negrita y marcados con asteriscos encima de las bases.

```
   *  * **    *** ***   * ** ***  ****     *
3'-TUTAUAUUTG^mCUUUAUUUG^mCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGG^mCGG
GTGGGG^mCAGGTGTAGA-5' (SEQ ID NO: 8)
```

[0135] Se sintetiza una cadena complementaria (marcada como SEQ ID NO: 9 a continuación) mediante la replicación del ADN que supone la hibridación de un cebador a la solapa añadida al extremo 3'.

```
3'-TUTAUAUUTGCUUUAUUUGCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGGCGGGT
   |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnnACCCCAACCCACCTCCGCCCA

GGGGCAGGTGTAGA-5' (SEQ ID NO: 8)
|||||||||||||||
CCCCGTCCACATCT-3' (SEQ ID NO: 9)
```

[0136] El dúplex anterior se secuenció en ambas direcciones; los intermedios de secuenciación se muestran a continuación. La cadena naciente, cuya secuencia se está obteniendo, es la SEQ ID NO: 10 en la reacción a y la SEQ ID NO: 11 en la reacción b.

Reacción de secuenciación a

```
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnnACCCCAACCCACCTCCGCCCA
                                        ||||||||||||||||||||||||
                                    3'-nnnnTGGGGTTGGGTGGAGGCGGGT

CCCCGTCCACATCT-3' (SEQ ID NO: 9)
|||||||||||||||
GGGGCAGGTGTAGA-5' (SEQ ID NO: 10)
```

Reacción de secuenciación b

```
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnn-3' (SEQ ID NO: 11)
   ||||||||||||||||||||||||||||||||||||||||
3'-TUTAUAUUTGCUUUAUUUGCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGGCGGGT

GGGGCAGGTGTAGA-5' (SEQ ID NO: 8)
```

[0137] Por lo tanto, las lecturas predichas a obtener a partir de estas reacciones contienen las siguientes secuencias

a: 5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGTnnnn-3' (SEQ ID NO: 10)

b: 5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnn-3' (SEQ ID NO: 11)

[0138] La secuencia de la muestra original, incluyendo el estado de metilación de la citosina, se determina mediante la aplicación de las siguientes reglas, resumidas en la Tabla 5. La cadena directa de la secuencia original es la cadena con el mismo sentido que las dos lecturas.

[0139] En las posiciones en las que la lectura a y la lectura b tienen ambas A, la cadena directa de la secuencia original también tiene A, y la cadena inversa tiene T. En las posiciones en las que la lectura a y la lectura b tienen ambas T, la cadena directa de la secuencia original también tiene T, y la cadena inversa tiene A.

[0140] Cuando la lectura a y la lectura b tienen ambas C, la cadena directa de la secuencia original tiene ^{m}C, y la cadena inversa tiene G. Cuando la lectura a y la lectura b tienen ambas G, la cadena directa de la secuencia original tiene G, y la cadena inversa tiene ^{m}C.

[0141] Cuando una lectura tiene G en una posición en la que la otra lectura tiene A, la cadena directa de la secuencia original tiene G, y la cadena inversa tiene C.

[0142] Cuando una lectura tiene T en una posición en la que la otra lectura tiene C, la cadena directa de la secuencia original tiene C, y la cadena inversa tiene G.

[0143] Las lecturas a y b están emparejadas en la columna 1 y 2 en la Tabla según la cual la lectura contiene los restos G y T en las posiciones en las que las lecturas difieren; en este ejemplo, la lectura a corresponde a la columna 1.

Tabla 5: Reglas de determinación del estado de metilación por tratamiento con bisulfito

Lecturas de secuenciación		Secuencia original	
1	2	Cadena directa (5' => 3')	Cadena inversa (3' => 5')
A	A	A	T
T	T	T	A
C	C	C metilada	G
G	G	G	C metilada
G	A	G	C
T	C	C	G

[0144] La aplicación de las reglas anteriores a las SEQ ID NOs: 10 y 11 da lugar a la recuperación (después de la secuencia enlazadora nnnn) de las secuencias originales, es decir, las SEQ ID NOs: 5 y 6. En la Figura 10A se muestra un alineamiento de las lecturas a y b con la cadena directa de la secuencia original.

Ejemplo 3. Simulación de detección de la metilación mediante la conversión de ^{m}C a T por transición fotoquímica con una molécula de par cerrado lineal

[0145] Se simuló la determinación de la secuencia y de las posiciones de 5-metilcitosina de un fragmento hipotético de ADN dúplex usando la conversión de ^{m}C a T por transición fotoquímica como sigue. El esquema general de este ejemplo se ilustra en Figura 13. La secuencia del ADN se muestra a continuación.

Muestra de ADN (C metilada marcada como ^{m}C)

```
5'-AGATGTGGA[m]CGGGGTGGG[m]CGGAGGTGGGTTGGGGC-3' (SEQ ID NO: 5)
   |||||||||/ ||||||||/ ||||||||||||||||
3'-TCTACACCTG[m]CCCCACCCG[m]CCTCCACCCAACCCCG-5' (SEQ ID NO: 6)
```

[0146] Las dos cadenas están conectadas mediante ligación a una secuencia enlazadora (representada como "nnnn") para dar el siguiente producto. La secuencia enlazadora es adecuada para su uso como cebador de secuenciación. Las solapas lineales (no mostradas) también están unidas a los extremos 3' y 5' de esta molécula.

```
3'-TCTACACCTG[m]CCCCACCCG[m]CCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGG[m]CGG
GTGGGG[m]CAGGTGTAGA-5' (SEQ ID NO: 7)
```

[0147] El producto se trata con luz para así convertir fotoquímicamente los restos de 5-metilcitosina (pero no la citosina) en timina, dando el producto siguiente. Los restos de timina recién formados están en negrita y marcados con asterisco por encima o por debajo de las bases.

```
              *         *                                        *
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGGTGGGT
GGGGTAGGTGTAGA-5' (SEQ ID NO: 12)
    *
```

[0148] Se sintetiza una cadena complementaria (marcada como SEQ ID NO: 13 a continuación) mediante la replicación del ADN usando un cebador que une la solapa conectada al extremo 3' de la molécula.

```
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGGTGGGT
   |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGCnnnnGCCCCAACCCACCTCCACCCA

GGGGTAGGTGTAGA-5' (SEQ ID NO: 12)
|||||||||||||
CCCCATCCACATCT-3' (SEQ ID NO: 13)
```

[0149] El dúplex anterior se secuenció en ambas direcciones como en el Ejemplo 2 anterior, obteniendo las siguientes lecturas.

Lectura a: 5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3' (SEQ ID NO: 14)

Lectura b: 5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3' (SEQ ID NO: 15)

[0150] La secuencia de la muestra original, incluyendo el estado de metilación de la citosina, se determina mediante la aplicación de las siguientes reglas, resumidas en la Tabla 6. La cadena directa de la secuencia original es la cadena con el mismo sentido que las dos lecturas.

[0151] En las posiciones en las que la lectura a y la lectura b tienen ambas A, la cadena directa de la secuencia original también tiene A, y la cadena inversa tiene T. En las posiciones en las que la lectura a y la lectura b tienen ambas T, la cadena directa de la secuencia original también tiene T, y la cadena inversa tiene A.

[0152] Cuando la lectura a y la lectura b tienen ambas C, la cadena directa de la secuencia original tiene C, y la cadena inversa tiene G. Cuando la lectura a y la lectura b tienen ambas G, la cadena directa de la secuencia original tiene G, y la cadena inversa tiene C.

[0153] Cuando una lectura tiene G en una posición en la que la otra lectura tiene A, la cadena directa de la secuencia original tiene G, y la cadena inversa tiene ^{m}C.

[0154] Cuando una lectura tiene T en una posición en la que la otra lectura tiene C, la cadena directa de la secuencia original tiene ^{m}C, y la cadena inversa tiene G.

[0155] Las lecturas a y b están emparejadas en la columna 1 y 2 en la Tabla 6 según la cual la lectura contiene los restos G y T en las posiciones en las que las lecturas difieren; en este ejemplo, la lectura a corresponde a la columna 1.

Tabla 6: Reglas de determinación del estado de metilación por transición fotoquímica

Lecturas de secuenciación		Secuencia original	
1	2	Cadena directa (5' => 3')	Cadena inversa (3' => 5')
A	A	A	T
T	T	T	A
C	C	C	G
G	G	G	C
G	A	G	C metilada
T	C	C metilada	G

[0156] La aplicación de las reglas anteriores a las SEQ ID NOs: 14 y 15 da lugar a la recuperación (después de la secuencia enlazadora nnnn) de las secuencias originales, es decir, las SEQ ID NOs: 5 y 6. En la Figura 10B se muestra un alineamiento de las lecturas a y b con la cadena directa de la secuencia original.

a => 5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3' (SEQ ID NO: 14)

b => 5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3' (SEQ ID NO: 15)

r => 5'-AGATGTGGA^{m}CGGGGTGGG^{m}CGGAGGTGGGTTGGGGC-3' (r_a) (SEQ ID NO: 5)

3'-TCTACACCTG^{m}CCCCACCCG^{m}CCTCCACCCAACCCCG-5' (r_b) (SEQ ID NO: 6)

Ejemplo 4: Comparación de la precisión de una secuenciación simulada de lectura única y lectura múltiple

[0157] Se descargó del GenBank la secuencia de un genoma ensamblado de *Escherichia coli*, número de acceso del GenBank U00096, longitud de 4.639.675 pb. De esta secuencia se extrajeron fragmentos seleccionados aleatoriamente con longitudes que oscilan entre 500 pb y 2000 pb. Estos fragmentos se designaron secuencias maestras.

[0158] Se generaron cinco subsecuencias a partir de las secuencias maestras introduciendo por ordenador errores de deleción y de lectura a tasas definidas, como se muestra en la Tabla 7.

[0159] Las cinco subsecuencias, que contienen errores, se sometieron a análisis de comparación de múltiples secuencias usando el algoritmo CLUSTALW (configuración por defecto). Los resultados del análisis de CLUSTALW se usaron como entrada para las "cons" del programa del paquete EMBOSS con el fin de obtener una secuencia consenso. Las "cons" del programa se describen en Rice y col., Trends Genet 16, 276-277 (2000), y Mullan y col., Brief Bioinform 3, 92-94 (2002).

[0160] Cada una de la primera subsecuencia y la secuencia consenso se compararon con la secuencia maestra, y se tabularon las frecuencias de los huecos y de los errores de lectura; véase Tabla 7. Los resultados demuestran que la formación de una secuencia consenso usando múltiples lecturas reduce la frecuencia de los errores de lectura y de los huecos en cada una de las diversas tasas de error sometidas a ensayo. Para cada grupo de tasas de errores de deleción y de lectura, se alinearon contra la secuencia maestra una única lectura simulada y una secuencia consenso determinada a partir de 5 lecturas simuladas. El número y porcentaje de posiciones con errores de lectura y huecos se determinaron como fracción de las posiciones en el alineamiento.

Tabla 7. Precisión de las secuencias consenso determinadas a partir de 5 lecturas simuladas comparadas a lecturas individuales a tasas de error variables

Tasa de errores introducidos	Longitud de la sec. maestra (nt)	Sencilla vs. Maestra		Consenso vs. Maestra	
		Errores de lectura	Huecos	Errores de lectura	Huecos
5 % de deleción 1 % de errores de lectura	816	53/816 (6,5 %)	47/816 (5,8 %)	8/817 (1,0 %)	5/817 (0,6 %)

5 % de deleción 2 % de errores de lectura	1565	90/1565 (5,8 %)	74/1565 (4,7 %)	9/1565 (1,0 %)	4/1565 (0,3 %)
1 % de deleción 30 % de errores de lectura	1589	401/1602 (25,0 %)	41/1602 (2,6 %)	90/1593 (5,6 %)	5/1593 (0,3 %)
1 % de deleción 30 % de errores de lectura	760	182/764 (23,8 %)	11/764 (1,4 %)	47/761 (6,2 %)	1/861 (0,1 %)

Ejemplo 5. Simulación de la determinación de secuencia usando una MPCc

[0161] Se proporciona una muestra de ácidos nucleicos de doble cadena como en el Ejemplo 2. Las cadenas directa e inversa de la muestra se cierran juntas por ligación de un inserto que forma una horquilla a cada extremo de la molécula como se muestra en la etapa de construcción de la MPCc de la Figura 14 para formar una molécula de par cerrado circular. Se lleva a cabo una secuenciación de una sola molécula mediante una reacción de síntesis usando un cebador que se une a uno de los insertos. Se obtienen los datos de secuencia que comprenden al menos una secuencia de la cadena directa de la muestra y al menos una secuencia de la cadena inversa de la muestra. Los datos de secuencia se analizan comparando las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular para determinar la secuencia de la muestra de ácidos nucleicos de acuerdo con la Tabla 8.

Tabla 8. Reglas de determinación de la secuencia de MPCc

Secuencia obtenida		Secuencia original	
Modelada por la cadena directa	Modelada por la cadena inversa	Cadena directa (5' => 3')	Cadena inversa (3' => 5')
A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C

[0162] Nota: en la Tabla 8 y las Tablas 9-11 a continuación, la secuencia obtenida modelada por la cadena directa corresponde a la línea superior de los datos de secuenciación (es decir, la secuencia mostrada bajo la flecha marcada "secuenciación" y encima de la flecha marcada "análisis de secuencia") en las Figuras 14-17, respectivamente. De forma similar, la secuencia obtenida modelada por la cadena inversa corresponde a la línea inferior de los datos de secuenciación en las Figuras 14-17, respectivamente.

Ejemplo 6. Simulación de la detección de metilación usando la conversión de C a U mediante el tratamiento con bisulfito con una molécula de par cerrado circular

[0163] El esquema general de este Ejemplo se muestra en la Figura 15. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende al menos una 5-metilcitosina como en el Ejemplo 2. Se forma una molécula de par cerrado circular como en el Ejemplo 5. Se realiza la conversión con bisulfito como en el Ejemplo 2. Los datos de secuencia se obtienen como en el Ejemplo 5. Los datos de secuencia se analizan comparando las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular para determinar la secuencia de la muestra de ácidos nucleicos y la posición de la al menos una 5-metilcitosina según las reglas de la Tabla 9.

Tabla 9. Reglas de determinación de la secuencia por MPCc/tratamiento con bisulfito

Secuencia obtenida		Secuencia original	
Modelada por la cadena inversa	Modelada por la cadena directa	Cadena directa (5' => 3')	Cadena inversa (3' => 5')
A	T	A	T
T	A	T	A
C	A	C	G
A	C	G	C
C	G	G	C metilada
G	C	C metilada	G

Ejemplo 7. Simulación de la detección de metilación usando la conversión de ^{m}C a T mediante transición fotoquímica con una molécula de par cerrado circular

[0164] El esquema general de este ejemplo se muestra en la Figura 16. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende al menos una 5-metilcitosina como en el Ejemplo 3. Se forma una molécula de par cerrado circular como en el Ejemplo 5. Se realiza la transición fotoquímica como en el Ejemplo 3. Los datos de secuencia se obtienen como en el Ejemplo 5. Los datos de secuencia se analizan comparando las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular para determinar la secuencia de la muestra de ácidos nucleicos y la posición de la al menos una 5-metilcitosina según las reglas de la Tabla 10.

Tabla 10. Reglas de determinación de la secuencia por MPCc/transición fotoquímica

Secuencia obtenida		Secuencia original	
Modelada por la cadena inversa	Modelada por la cadena directa	Cadena directa (5' => 3')	Cadena inversa (3' => 5')
A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C
C	A	G	C metilada
A	C	C metilada	G

Ejemplo 8. Simulación de la detección de 5-bromouracilo usando una molécula de par cerrado circular

[0165] El esquema general de este ejemplo se muestra en la Figura 17. Se proporciona una muestra de ácidos nucleicos de doble cadena que comprende al menos un 5-bromouracilo. Se forma una molécula de par cerrado circular como en el Ejemplo 5. Los datos de secuencia se obtienen como en el Ejemplo 5. Los datos de secuencia se analizan comparando las secuencias de las cadenas directa e inversa de la molécula de par cerrado circular para determinar la secuencia de la muestra de ácidos nucleicos y la posición del al menos un 5-bromouracilo según las reglas de la Tabla 11.

Tabla 11. Reglas de determinación de la secuencia por MPCc/5-bromouracilo

Secuencia obtenida		Secuencia original	
Modelada por la cadena inversa	Modelada por la cadena directa	Cadena directa (5' => 3')	Cadena inversa (3' => 5')

A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C
G	T	A	5-bromouracilo
T	G	5-bromouracilo	A

[0166] La memoria descriptiva se entiende con mayor profundidad en vista de las enseñanzas de las referencias citadas en la memoria descriptiva. Las realizaciones dentro de la memoria descriptiva proporcionan una ilustración de las realizaciones de la invención y no se deben interpretar como una limitación al alcance de la invención. La persona experta reconoce fácilmente que muchas otras realizaciones están englobadas por la invención. La mención a cualquier referencia en el presente documento no es una admisión de que dicha referencia es técnica anterior a la presente invención.

[0167] A menos que se indique lo contrario, todos los números que expresan cantidades de ingredientes, condiciones de reacción, etc. usados en la memoria descriptiva, incluyendo las reivindicaciones, se debe entender que están modificados en todos los casos por el término "aproximadamente". Por consiguiente, a menos que se indique lo contrario, los parámetros numéricos son aproximaciones y pueden variar. Como mínimo, y no como intento para limitar la aplicación de la doctrina de equivalentes al alcance de las reivindicaciones, cada parámetro se debe interpretar en vista del número de dígitos significativos y aproximaciones de redondeo ordinarias. La mención de series de números con cantidades diferentes de dígitos significativos en la memoria descriptiva no se debe interpretar que implica que los números con menos dígitos significativos dados tienen la misma precisión que los números con más dígitos significativos dados.

[0168] El uso de la palabra "un" o "una" cuando se usa junto con el término "que comprende" en las reivindicaciones y/o la memoria descriptiva puede significar "uno", pero también es consistente con el significado de "uno o más", "al menos uno", y "uno o más de uno". El uso del término "o" en las reivindicaciones se usa para indicar "y/o" a menos que se indique explícitamente para referirse exclusivamente a alternativas o que las alternativas sean mutuamente excluyentes, aunque la divulgación admite una definición que se refiere únicamente a alternativas e "y/o".

[0169] A menos que se indique lo contrario, el término "al menos" que precede a una serie de elementos se debe entender que se refiere a cada elemento de la serie.

[0170] A menos que se defina lo contrario, todos los términos técnicos y científicos usados en el presente documento tienen el mismo significado que entiende normalmente la persona experto en la materia a la que pertenece esta invención. Aunque en la práctica o puesta a prueba de la presente invención se puede usar cualquier método y material similar o equivalente a los descritos en el presente documento, los métodos y materiales preferidos se describen a continuación.

REIVINDICACIONES

1. Un método de determinación de la secuencia de una muestra de ácidos nucleicos que comprende:

a. el suministro de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que el inserto tiene una secuencia conocida;

b. la obtención de los datos de secuencia que comprende la secuencia de al menos dos unidades de inserto-muestra, en la que se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra;

c. el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (b) al comparar las secuencias con la secuencia conocida del inserto;

d. aceptar o rechazar al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (b) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos;

e. la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (d); y

f. la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas

en el que la aceptación o rechazo de al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (b) comprende la aceptación de aquellas de las al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos que están inmediatamente aguas arriba o aguas abajo de una secuencia del inserto de muestra con una puntuación superior o igual a un límite predeterminado, y el rechazo de aquellas que no lo están.

2. El método de la reivindicación 1, en el que la obtención de los datos de secuencia comprende la secuenciación de una sola molécula.

3. El método de la reivindicación 1, en el que el suministro de una molécula de ácidos nucleicos circular comprende la ligación de la muestra de ácidos nucleicos al inserto de ácidos nucleicos para formar la molécula de ácidos nucleicos circular.

4. El método de la reivindicación 1, en el que la molécula de ácidos nucleicos circular comprende al menos dos unidades de inserto-muestra.

5. El método de la reivindicación 1, en el que el inserto de ácidos nucleicos comprende un promotor y la síntesis de la molécula de ácidos nucleicos producto comprende la puesta en contacto de promotor con una ARN polimerasa que reconoce el promotor seguido de la síntesis de una molécula de ácidos nucleicos producto que comprende restos de ribonucleótidos.

6. El método de la reivindicación 1, en el que el inserto de ácidos nucleicos tiene una longitud que oscila entre 14 y 200 restos de nucleótidos.

7. El método de la reivindicación 1, en el que el grupo de secuencias aceptadas se encuentra en una forma seleccionada entre una variable multi-elementos y una cadena de datos sencilla que comprende los datos de secuencia de la etapa (b) que han sido procesados para borrar, sobrescribir, u omitir repeticiones de la secuencia de la muestra de ácidos nucleicos rechazada en la etapa (e).

8. El método de la reivindicación 1, en el que el grupo de secuencias aceptadas se encuentra en forma de variable multi-elementos de un tipo seleccionado entre una lista, conjunto, *hash*, y matriz.

9. El método de la reivindicación 1, en el que en la etapa (d) se aceptan al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos, y la determinación de la secuencia de la muestra de ácidos nucleicos comprende la determinación de una secuencia consenso basada en las al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (d).

10. El método de la reivindicación 9, en el que la secuencia consenso comprende bases representadas probabilísticamente en al menos una posición en la que difieren las al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (d).

11. El método de la reivindicación 9, en el que en la etapa (d) se aceptan al menos tres repeticiones de la

secuencia de la muestra de ácidos nucleicos, y la determinación de la secuencia consenso comprende la determinación de los votos mayoritarios de las al menos tres repeticiones de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (d).

12. El método de la reivindicación 9, en el que la secuencia consenso comprende niveles de confianza.

13. El método de la reivindicación 12, en el que los niveles de confianza se expresan en una forma seleccionada entre la frecuencia de bases, el contenido de información, y la puntuación de calidad de Phred.

14. El método de la reivindicación 12, en el que las etapas (b)-(f) de la reivindicación 1 se llevan a cabo en tiempo real, y la secuencia consenso y los niveles de confianza se actualizan en tiempo real.

15. El método de la reivindicación 14, en el que el método se realiza hasta que se alcanza un nivel de confianza mínimo establecido en un porcentaje preseleccionado de posiciones de la secuencia consenso.

16. El método de la reivindicación 15, que además comprende la generación de una alerta cuando el porcentaje de posiciones preseleccionadas alcanza el nivel de confianza mínimo establecido.

17. El método de la reivindicación 1, que además comprende la repetición de las etapas de la reivindicación 1 con al menos otra muestra de ácidos nucleicos de la misma fuente, especie, o cepa que la muestra de ácidos nucleicos de la reivindicación 1 que tiene una secuencia que se solapa parcialmente con la secuencia de la muestra de ácidos nucleicos de la reivindicación 1, determinando así al menos otra secuencia, y ensamblando la al menos otra secuencia con la secuencia de la etapa (f) para formar un cóntigo.

18. El método de la reivindicación 1, en el que se usan las puntuaciones de la etapa (c) para estimar un nivel de confianza de los datos de secuencia de la etapa (b) en su conjunto.

19. El método de la reivindicación 1, en el que el cálculo de las puntuaciones comprende la determinación del número de desemparejamientos entre los al menos dos insertos de los datos de secuencia y la secuencia conocida del inserto.

20. El método de la reivindicación 1, en el que el cálculo de las puntuaciones comprende la determinación del porcentaje de identidad de los al menos dos insertos de los datos de secuencia con la secuencia conocida del inserto.

21. El método de la reivindicación 1, en el que el cálculo de las puntuaciones comprende realizar un alineamiento entre los al menos dos insertos de los datos de secuencia y la secuencia conocida del inserto.

22. El método de la reivindicación 1, en el que las puntuaciones se generan en una base seleccionada en base a un recuento y en base a un porcentaje.

23. Un sistema que comprende un aparato de secuenciación unido de manera operable a un aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario, el almacenamiento que está codificado mediante programación que comprende un sistema operativo, un *software* de interfaz de usuario, e instrucciones que, cuando las ejecuta el procesador, opcionalmente mediante la introducción por parte del usuario, realiza un método que comprende:

a. la obtención de los datos de secuencia de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que:

(i) el inserto tiene una secuencia conocida,

(ii) los datos de secuencia comprenden la secuencia de al menos dos unidades de inserto-muestra, y

(iii) se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra;

b. el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (a) al comparar las secuencias con la secuencia conocida del inserto;

c. aceptar o rechazar al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (a) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos;

d. la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (c); y

e. la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas,

en el que la aceptación o rechazo de al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (b) comprende la aceptación de aquellas de las al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos que están inmediatamente aguas arriba o aguas abajo de una secuencia del inserto de muestra con una puntuación superior o igual a un límite predeterminado, y el rechazo de aquellas que no lo están, y

en el que se usa un resultado del sistema para producir al menos una de (i) una secuencia de una muestra de ácidos nucleicos o (ii) una indicación de que existe una base modificada en al menos una posición en una muestra de ácidos nucleicos.

24. Un almacenamiento codificado mediante programación que comprende un sistema operativo, un *software* de interfaz de usuario, e instrucciones que, cuando las ejecuta el procesador sobre un sistema que comprende un aparato de secuenciación unido de manera operable a un aparato de computación que comprende un procesador, almacenamiento, un sistema de bus, y al menos un elemento de interfaz de usuario, opcionalmente con la introducción por parte del usuario, realiza un método que comprende:

a. la obtención de los datos de secuencia de una molécula de ácidos nucleicos circular que comprende al menos una unidad de inserto-muestra que comprende un inserto de ácidos nucleicos y una muestra de ácidos nucleicos, en la que:

(i) el inserto tiene una secuencia conocida,

(ii) los datos de secuencia comprenden la secuencia de al menos dos unidades de inserto-muestra, y

(iii) se produce una molécula de ácidos nucleicos que comprende al menos dos unidades de inserto-muestra;

b. el cálculo de las puntuaciones de las secuencias de al menos dos insertos de los datos de secuencia de la etapa (a) al comparar las secuencias con la secuencia conocida del inserto;

c. aceptar o rechazar al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (a) según las puntuaciones de una o las dos secuencias de los insertos inmediatamente aguas arriba y aguas abajo de la repetición de la secuencia de la muestra de ácidos nucleicos;

d. la recopilación de un grupo de secuencias aceptadas que comprende al menos una repetición de la secuencia de la muestra de ácidos nucleicos aceptada en la etapa (c); y

e. la determinación de la secuencia de la muestra de ácidos nucleicos usando el grupo de secuencias aceptadas,

en el que la aceptación o rechazo de al menos dos de las repeticiones de la secuencia de la muestra de ácidos nucleicos de los datos de secuencia de la etapa (b) comprende la aceptación de aquellas de las al menos dos repeticiones de la secuencia de la muestra de ácidos nucleicos que están inmediatamente aguas arriba o aguas abajo de una secuencia del inserto de muestra con una puntuación superior o igual a un límite predeterminado, y el rechazo de aquellas que no lo están, y

en el que el método da lugar a un resultado usado para producir al menos una de (i) una secuencia de una muestra de ácidos nucleicos o (ii) una indicación de que existe una base modificada en al menos una posición en una muestra de ácidos nucleicos.

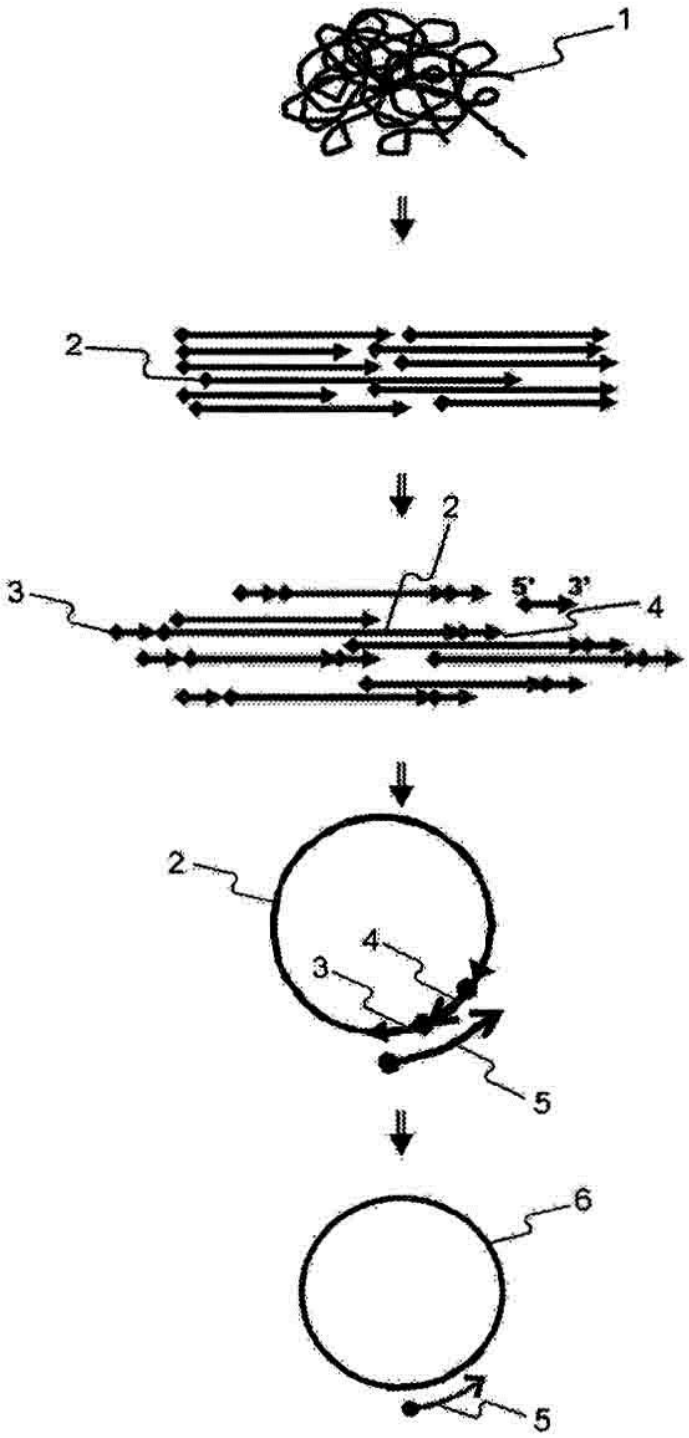


FIG. 1

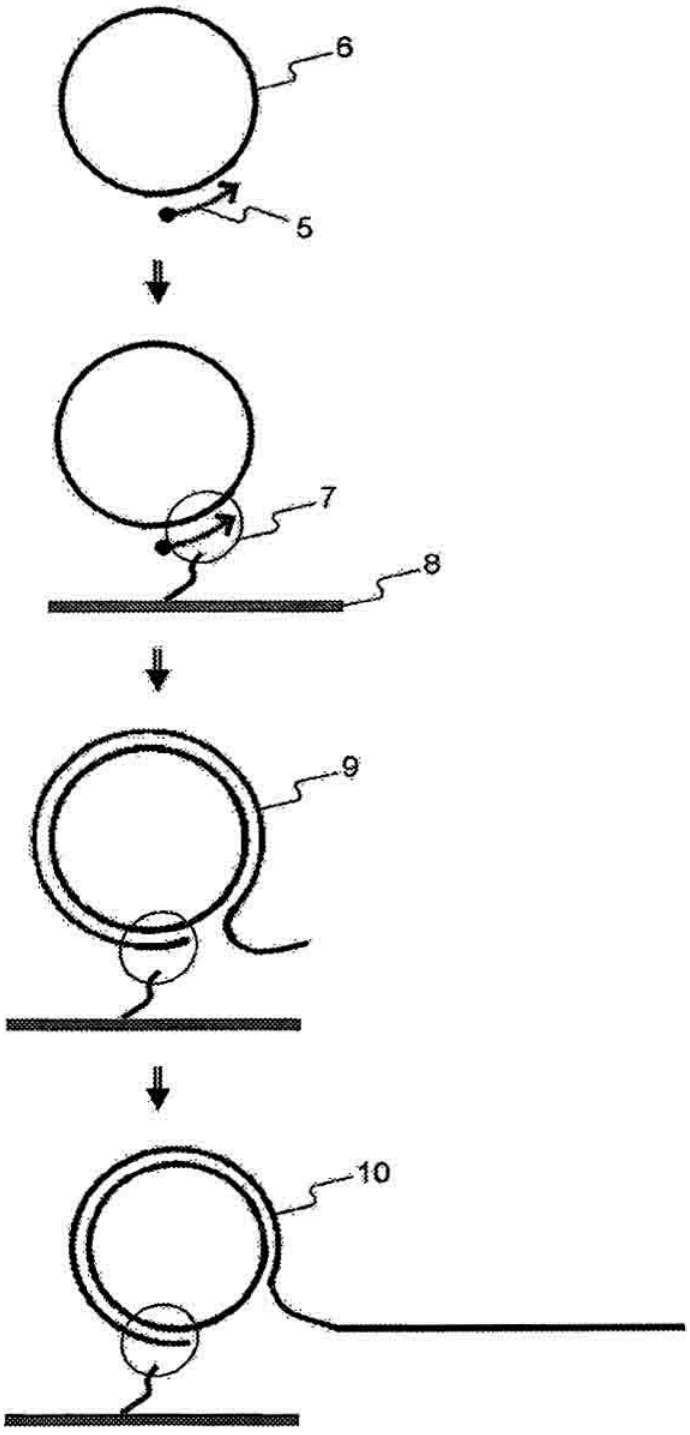


FIG. 2

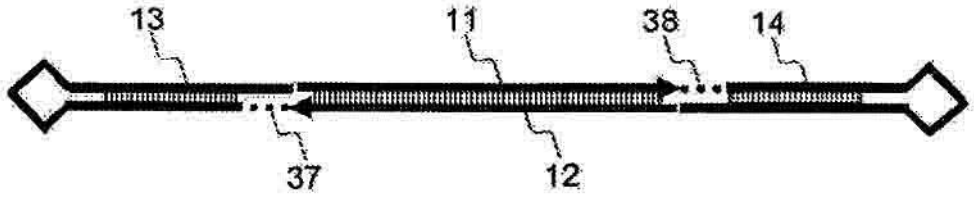

13
11
38
14
37
12

FIG. 3A

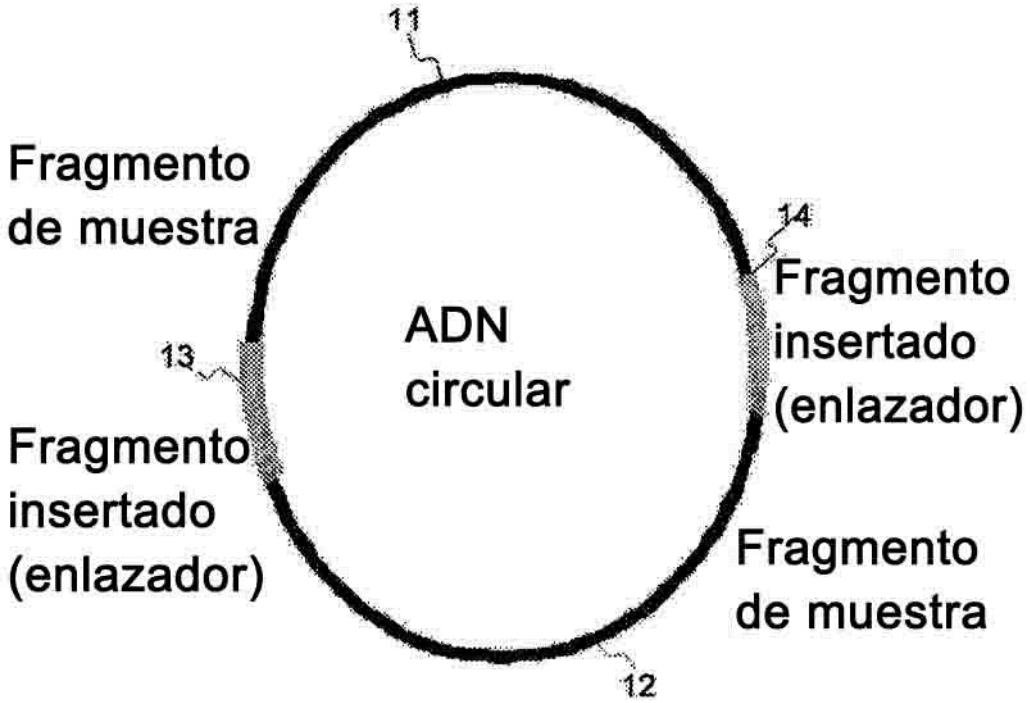

11
Fragmento de muestra
14
Fragmento insertado (enlazador)
ADN circular
13
Fragmento insertado (enlazador)
Fragmento de muestra
12

FIG. 3B

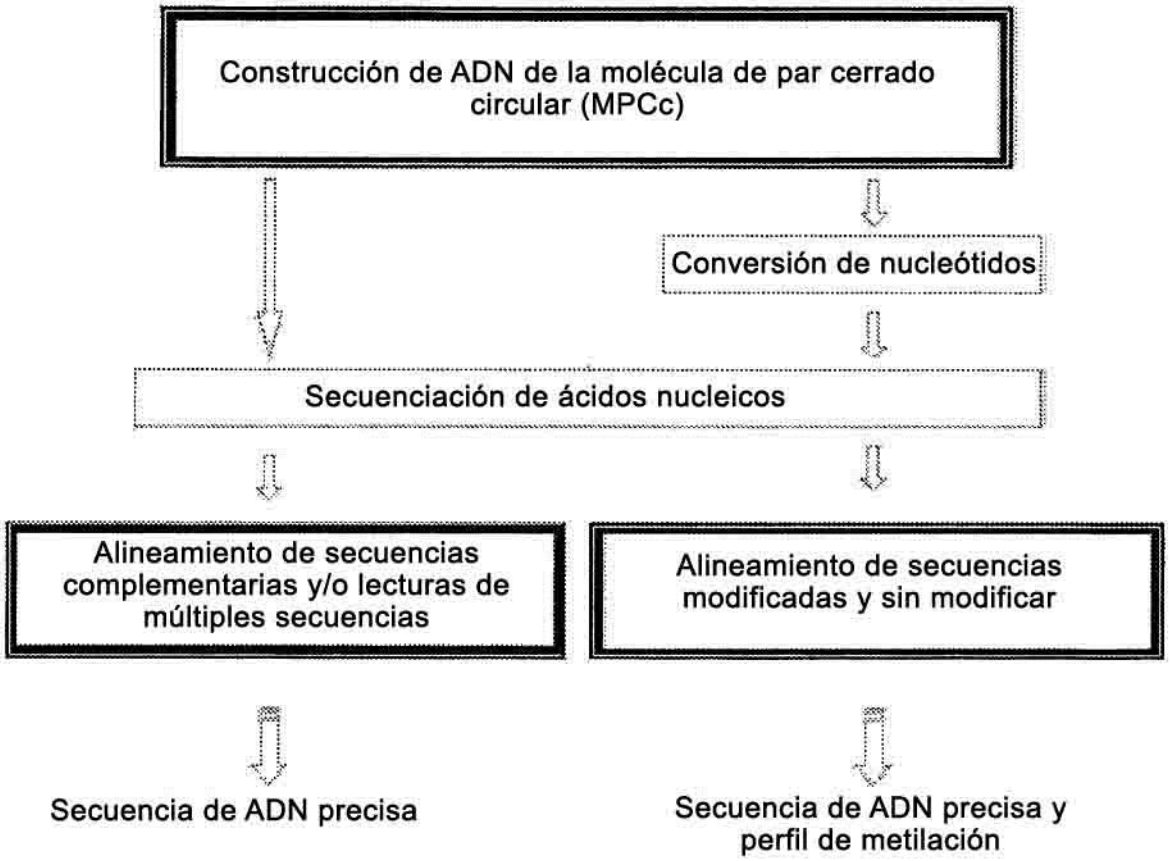

Construcción de ADN de la molécula de par cerrado circular (MPCc)
Conversión de nucleótidos
Secuenciación de ácidos nucleicos
Alineamiento de secuencias complementarias y/o lecturas de múltiples secuencias
Alineamiento de secuencias modificadas y sin modificar
Secuencia de ADN precisa
Secuencia de ADN precisa y perfil de metilación

FIG. 4

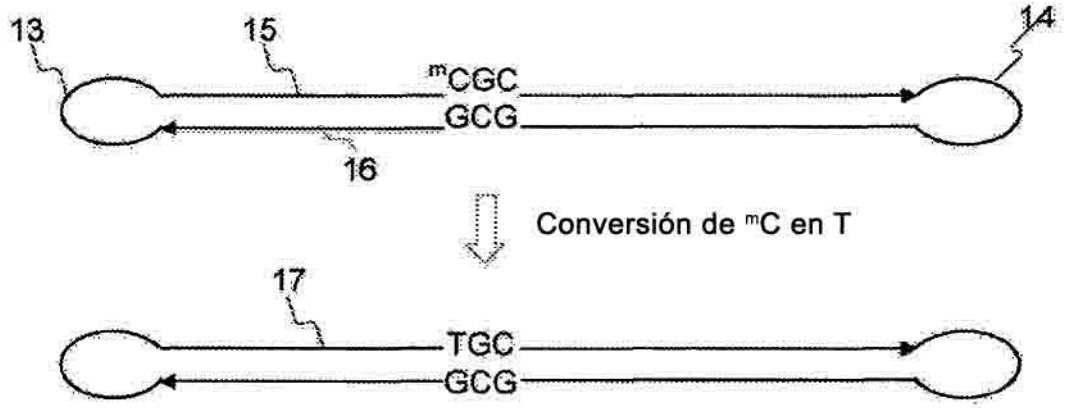

13
15
14
mCGC
GCG
16
Conversión de mC en T
17
TGC
GCG

FIG. 5A

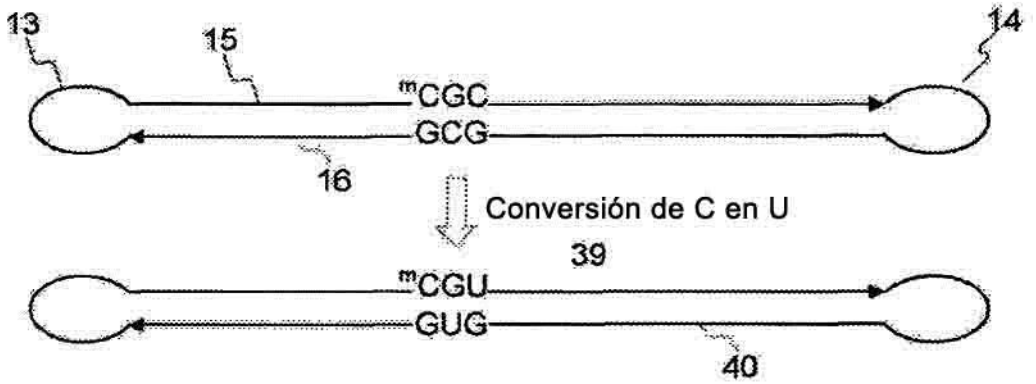

13
15
14
mCGC
GCG
16
Conversión de C en U
39
mCGU
GUG
40

FIG. 5B

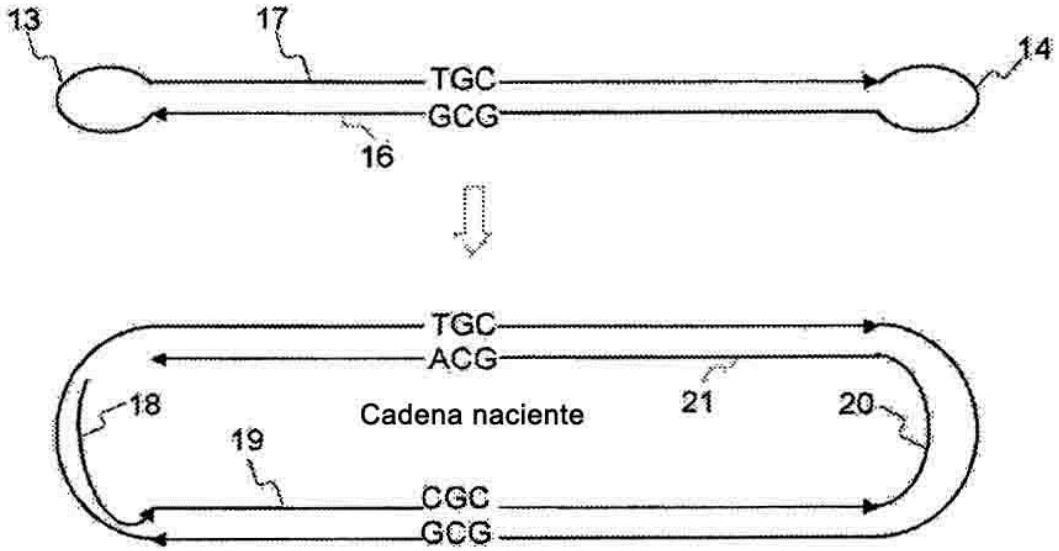

13
17
14
TGC
GCG
16
TGC
ACG
18
Cadena naciente
21
20
19
CGC
GCG

FIG. 6A

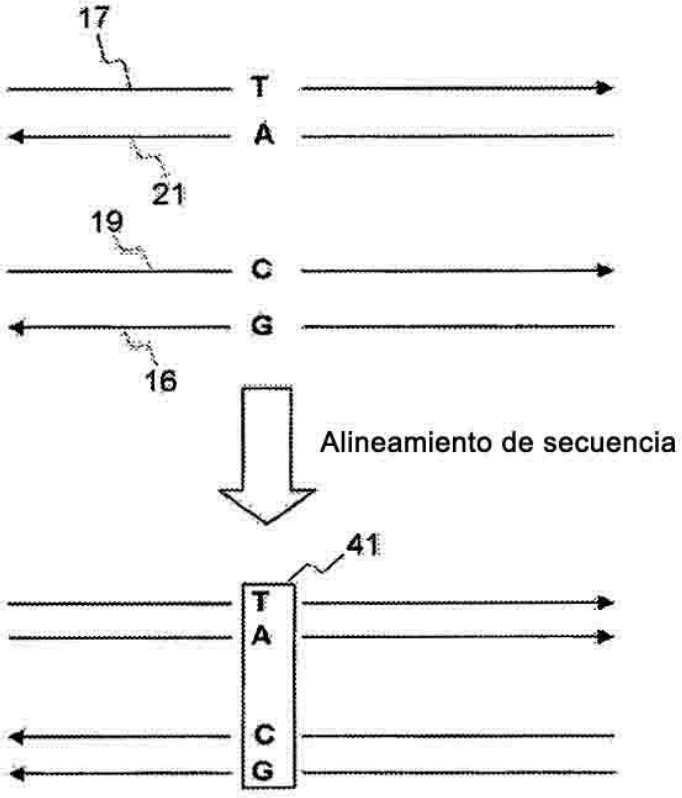
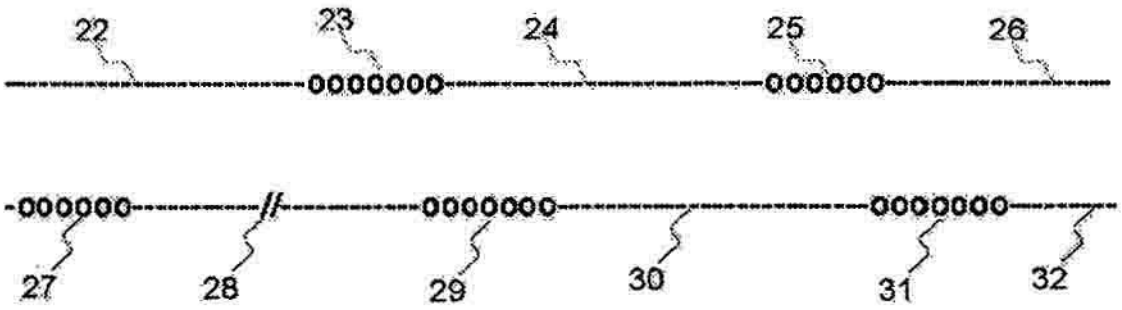

17
T
A
21
19
C
G
16
Alineamiento de secuencia
41
T
A
C
G

FIG. 6B



Resultado de la lectura de un molde circular

FIG. 7A

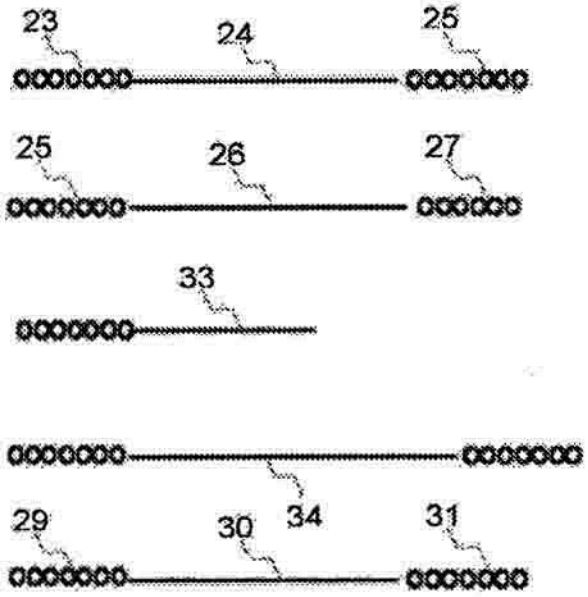


FIG. 7B

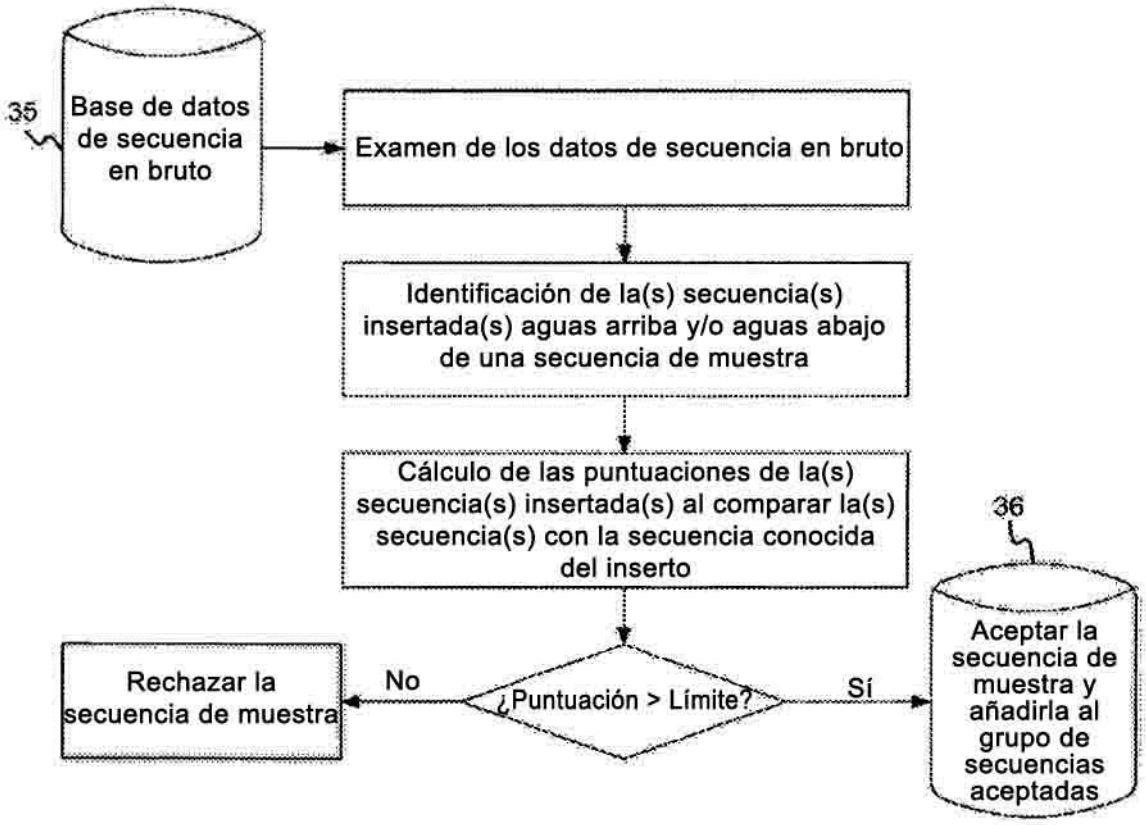

35
Base de datos de secuencia en bruto
Examen de los datos de secuencia en bruto
Identificación de la(s) secuencia(s) insertada(s) aguas arriba y/o aguas abajo de una secuencia de muestra
Cálculo de las puntuaciones de la(s) secuencia(s) insertada(s) al comparar la(s) secuencia(s) con la secuencia conocida del inserto
¿Puntuación > Límite?
No
Rechazar la secuencia de muestra
Sí
36
Aceptar la secuencia de muestra y añadirla al grupo de secuencias aceptadas

FIG. 8

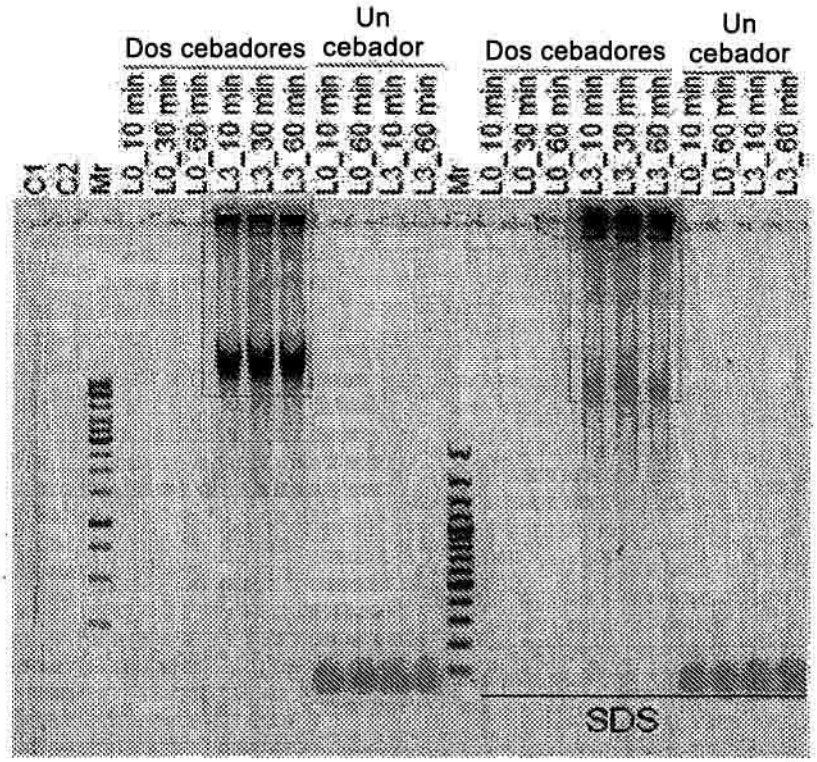

Dos cebadores
Un cebador
Dos cebadores
Un cebador
C1
C2
Mr
L0_10 min
L0_30 min
L0_60 min
L3_10 min
L3_30 min
L3_60 min
L0_10 min
L0_60 min
L3_10 min
L3_60 min
Mr
L0_10 min
L0_30 min
L0_60 min
L3_10 min
L3_30 min
L3_60 min
L0_10 min
L0_60 min
L3_10 min
L3_60 min
SDS

FIG. 9

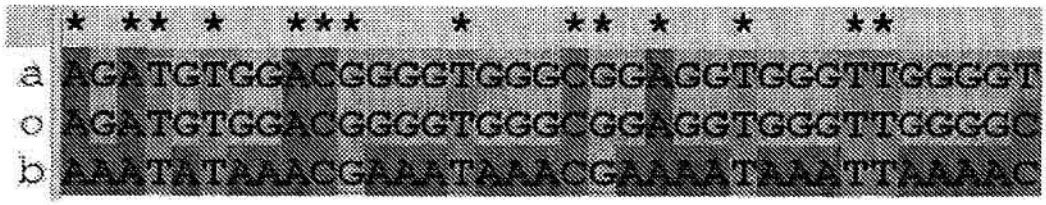


FIG. 10A

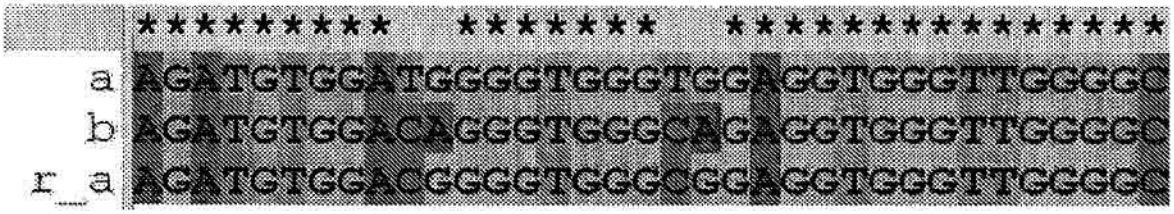


FIG. 10B

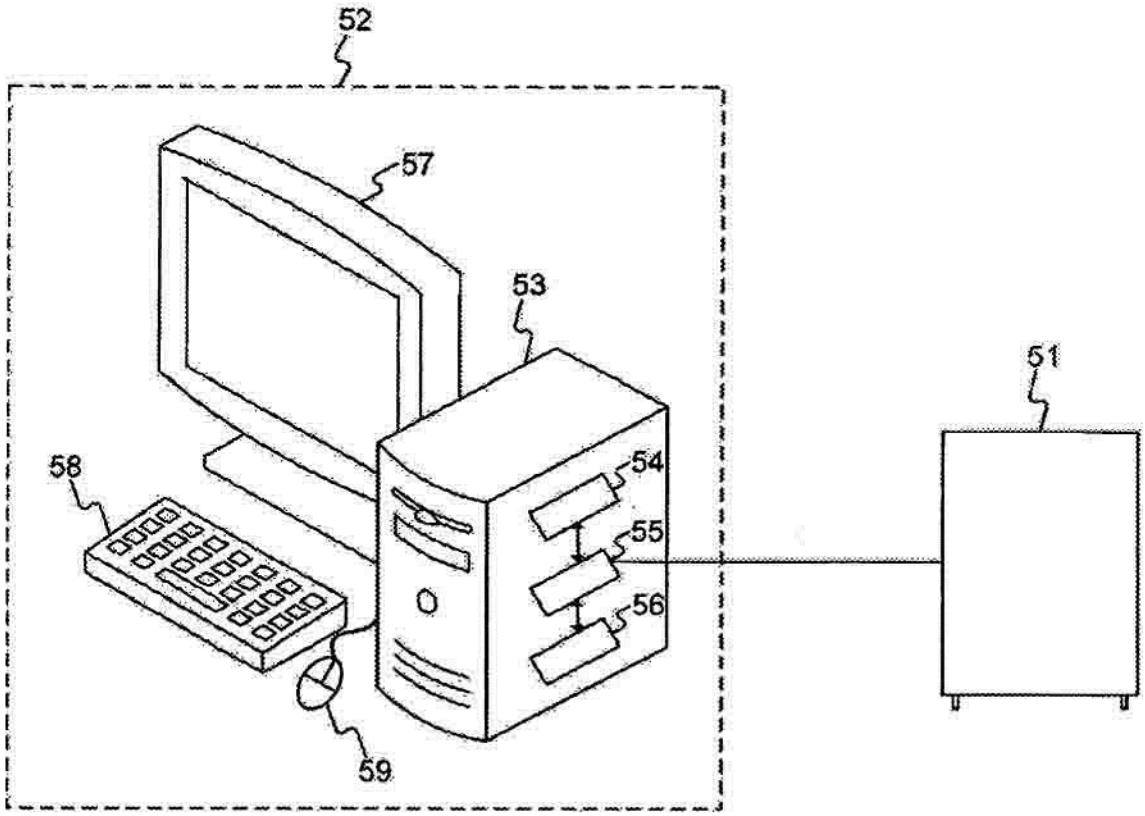


FIG. 11A

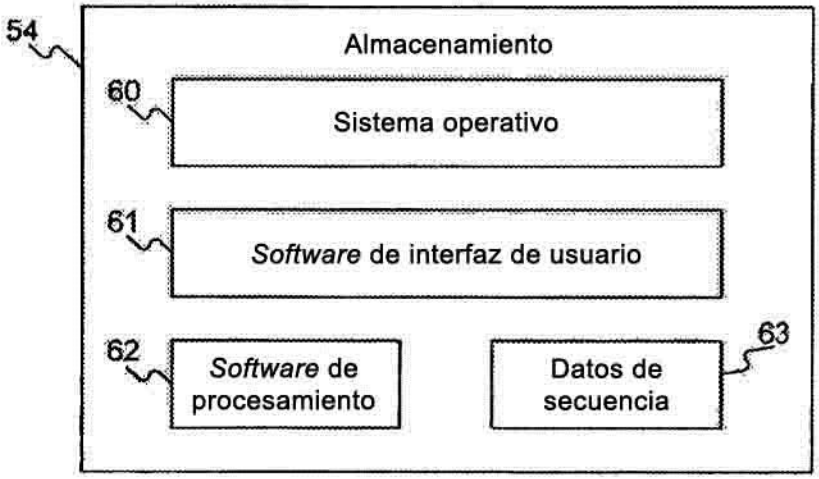
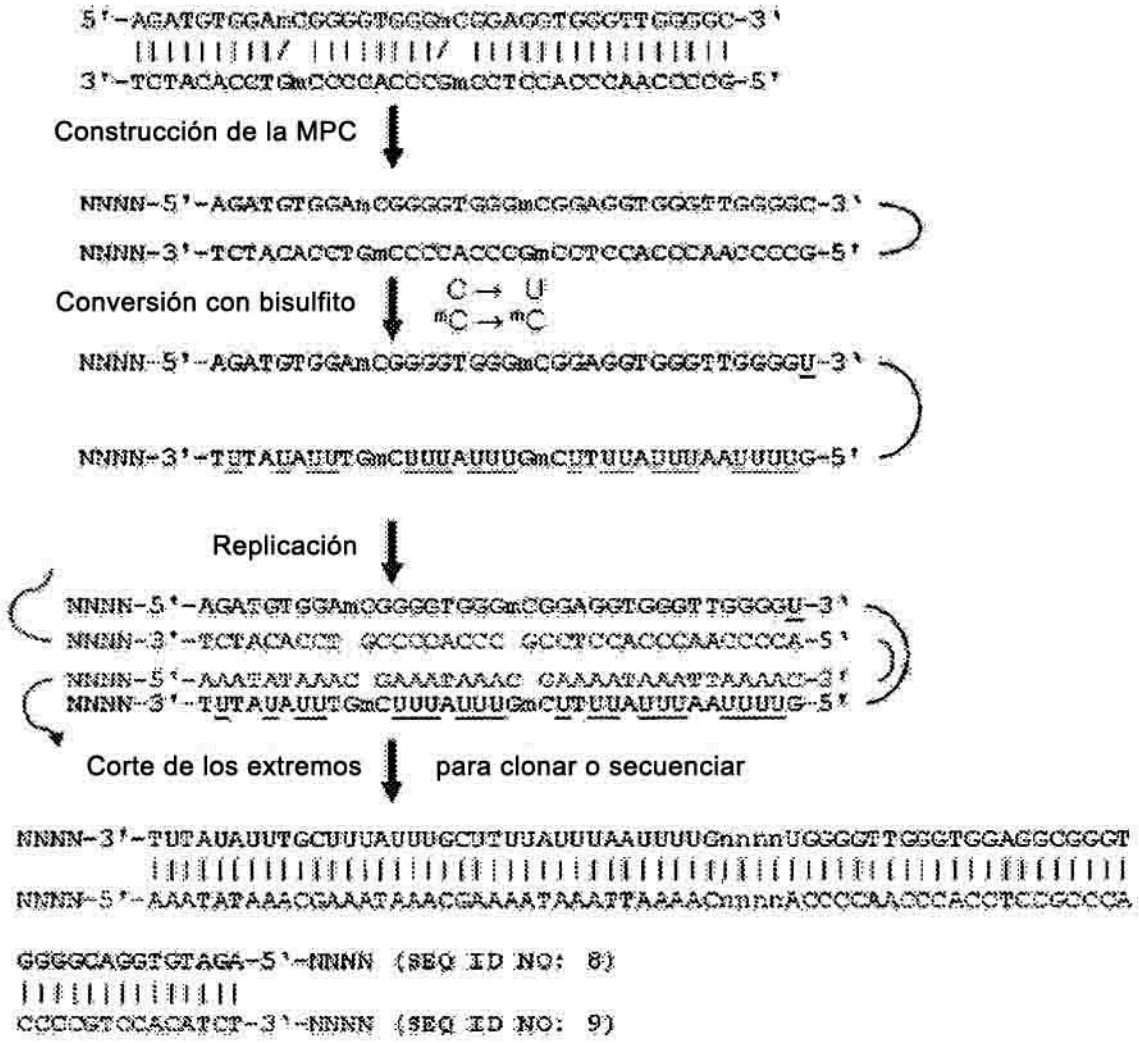


FIG. 11B



GGGGCAGGTGTAGA-5'-NNNN (SEQ ID NO: 8)

CCCCGTCCACATCT-3'-NNNN (SEQ ID NO: 9)

FIG. 12

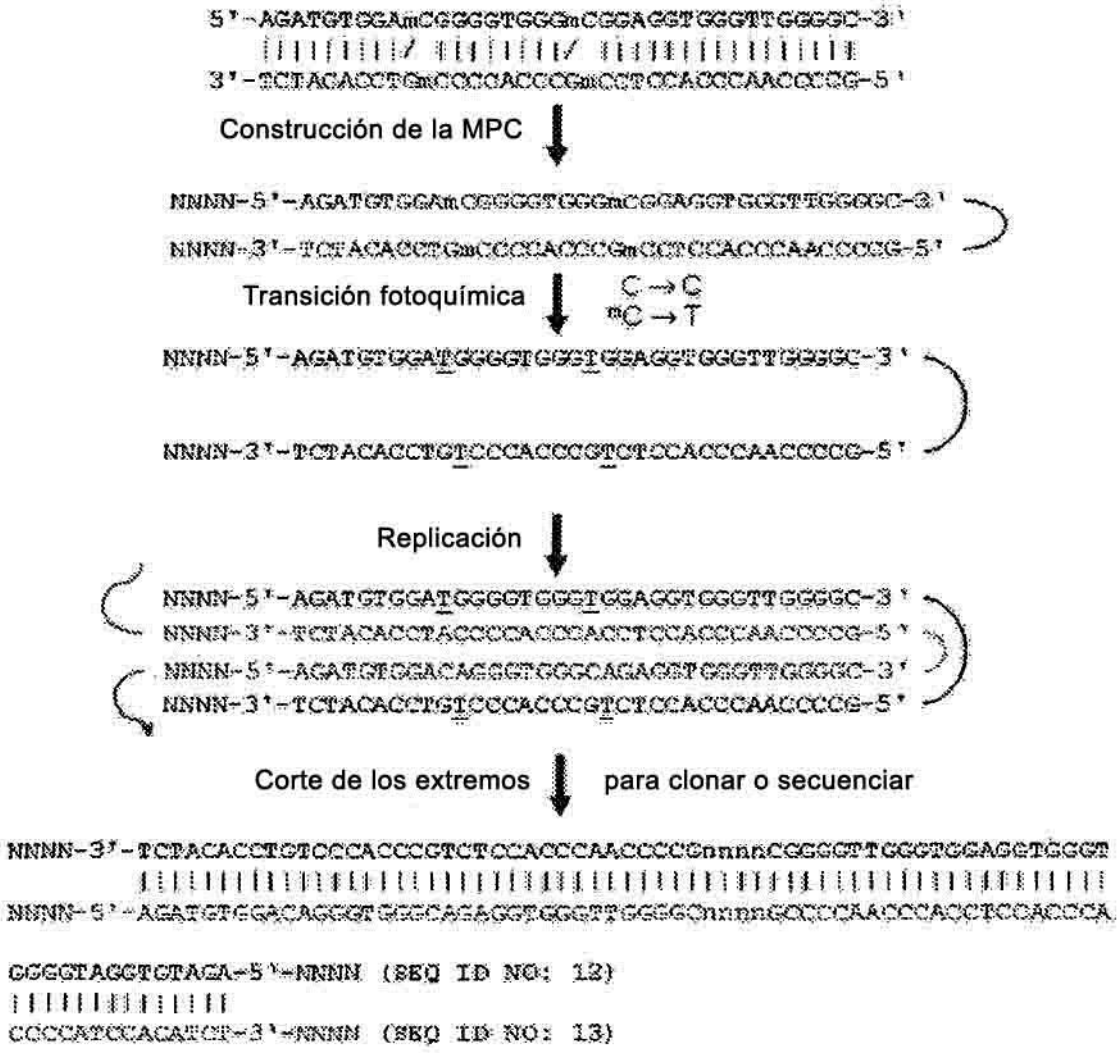

Construcción de la MPC
Transición fotoquímica
C → C
mC → T
Replicación
Corte de los extremos
para clonar o secuenciar
GGGGTAGGTGTAGA-5'-NNNN (SEQ ID NO: 12)
CCCCATCCACATCT-3'-NNNN (SEQ ID NO: 13)

FIG. 13

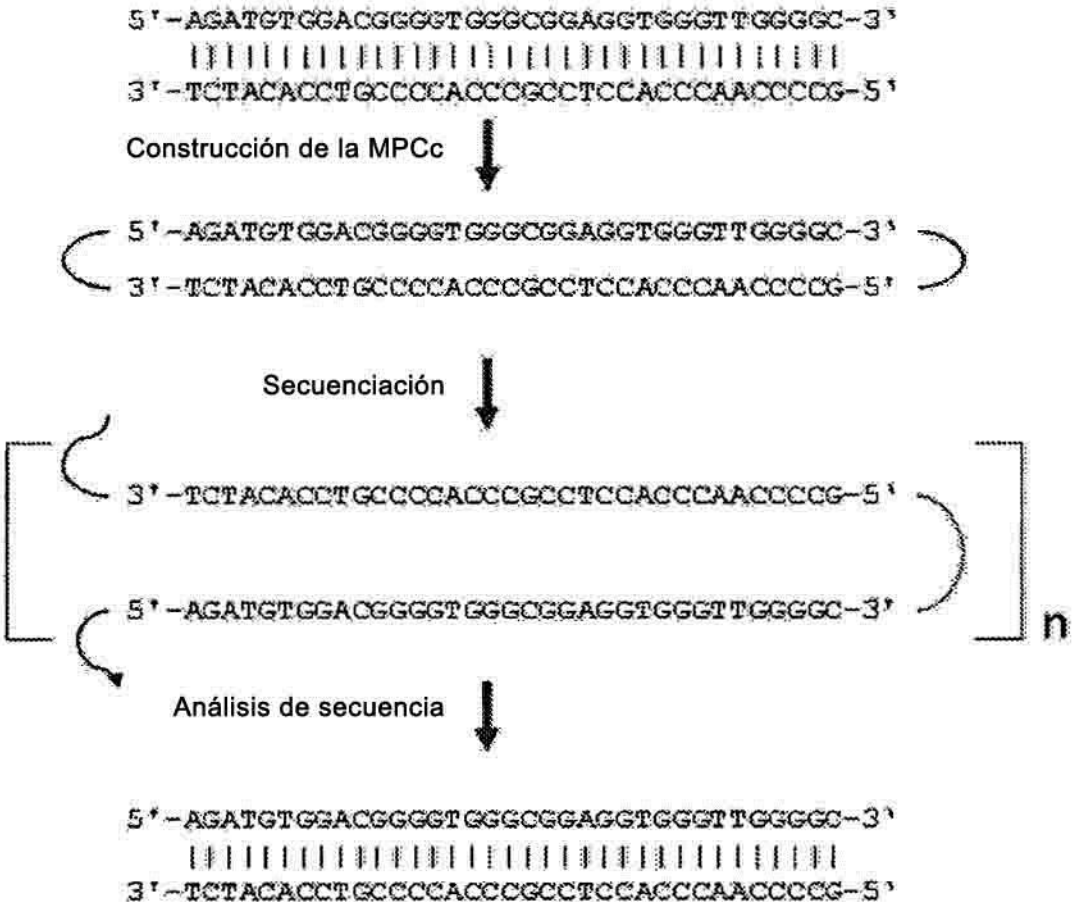

5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
Construcción de la MPCc
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
Secuenciación
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
n
Análisis de secuencia
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'

FIG. 14

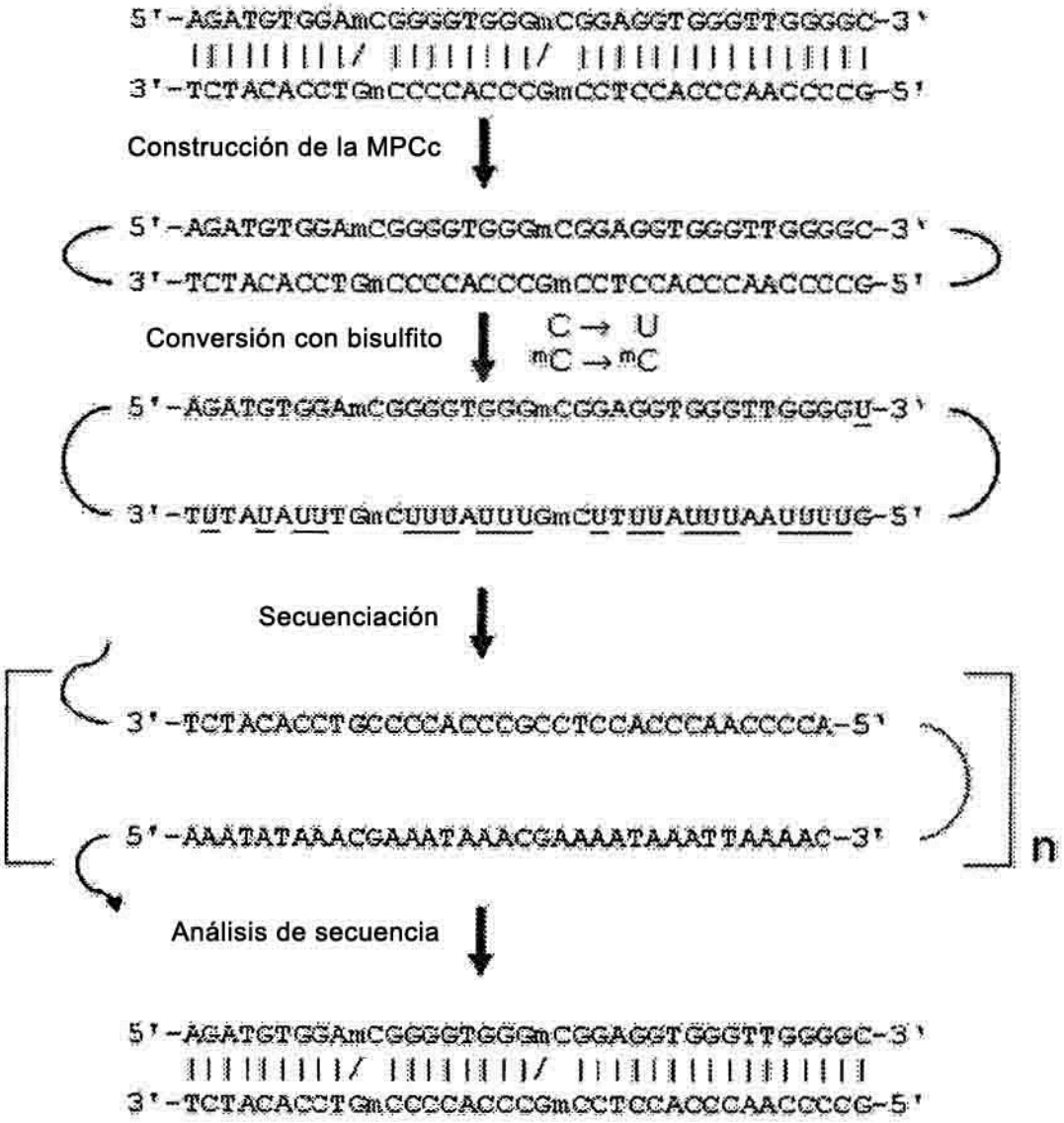

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Construcción de la MPCc
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Conversión con bisulfito
C → U
mC → mC
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGU-3'
3'-TUTAUAUUTGmCUUUAUUUGmCUTUUAUUUAAUUUG-5'
Secuenciación
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCA-5'
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAAC-3'
n
Análisis de secuencia
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'

FIG. 15

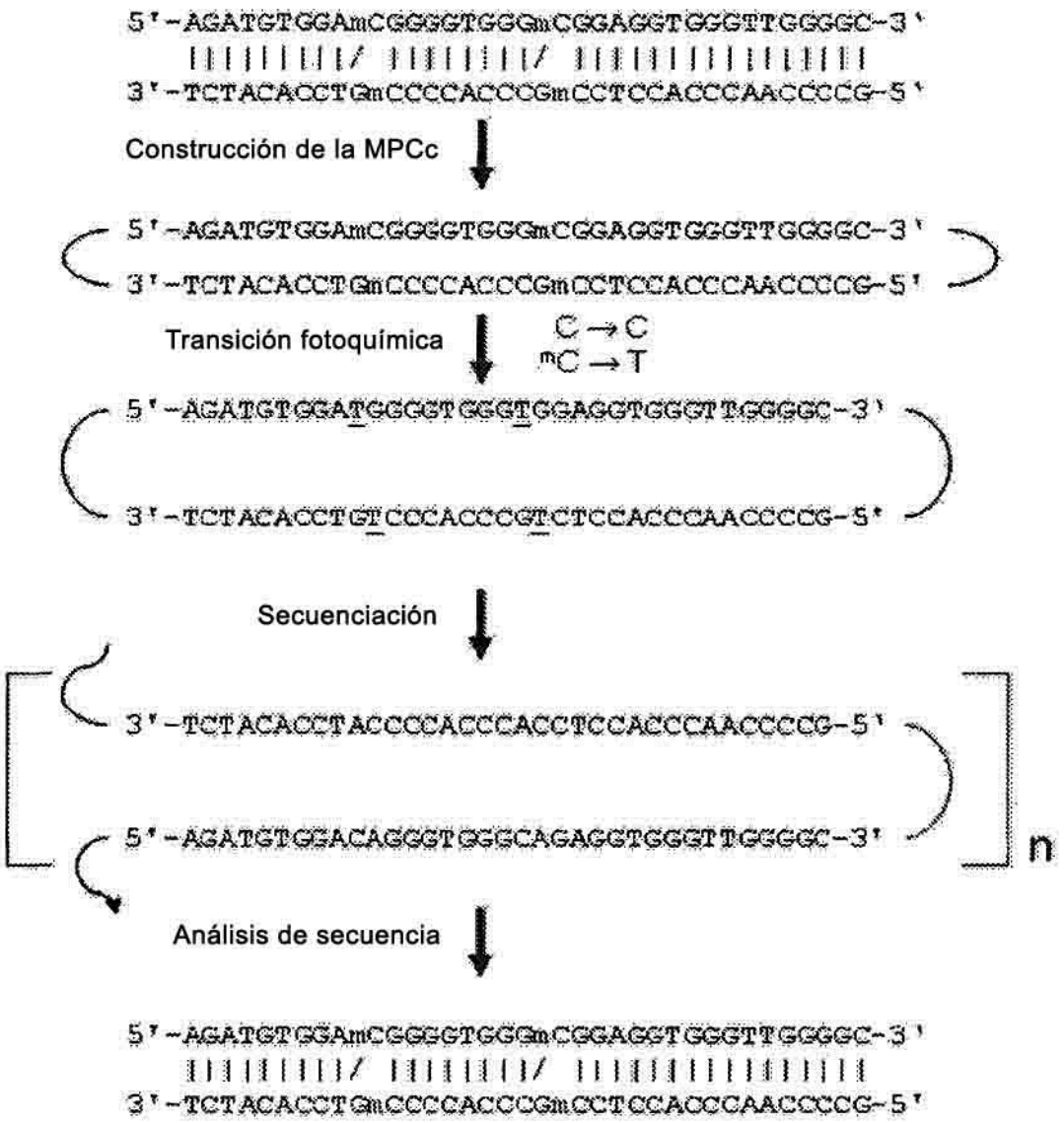

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Construcción de la MPCc
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Transición fotoquímica
C → C
mC → T
5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCG-5'
Secuenciación
3'-TCTACACCTACCCCACCCACCTCCACCCAACCCCG-5'
5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3'
n
Análisis de secuencia
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'

FIG. 16

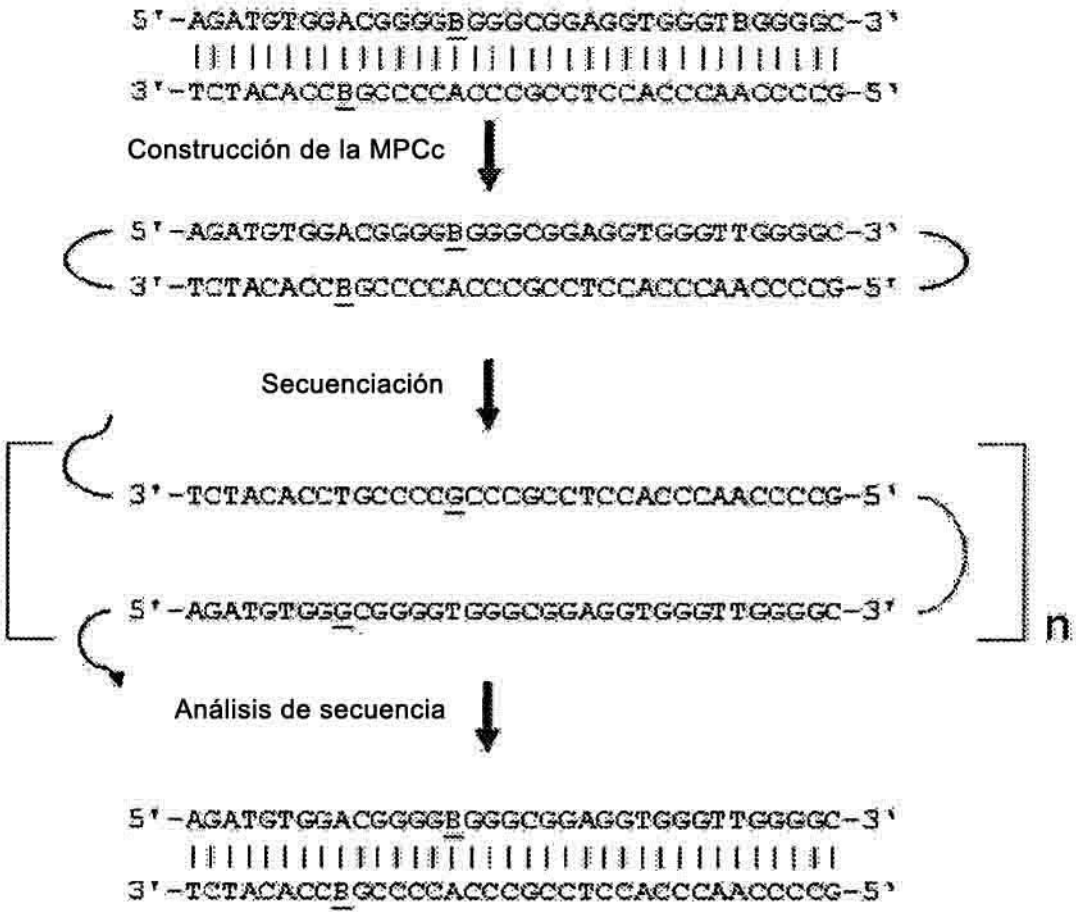

5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTBGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'
Construcción de la MPCc
5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'
Secuenciación
3'-TCTACACCTGCCCCGCCCGCCTCCACCCAACCCCG-5'
5'-AGATGTGGGCGGGGTGGGCGGAGGTGGGTTGGGGC-3'
n
Análisis de secuencia
5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'

FIG. 17